Universidade do Minho
Escola de Engenharia

João Pedro Pereira Fontes

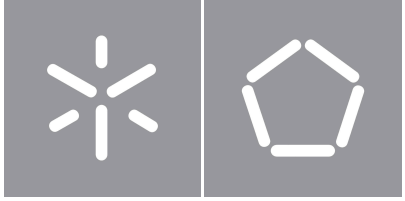
A Framework for the Development of Quantitative Imaging Biomarkers for Cancer Research

outubro de 2026

A Framework for the Development of Quantitative Imaging Biomarkers for Cancer Research

João Pedro Pereira Fontes

UMinho | 2026



Universidade do Minho
Escola de Engenharia

João Pedro Pereira Fontes

A Framework for the Development of Quantitative Imaging Biomarkers for Cancer Research

PhD Thesis
Doctoral Program in Information Systems and Technology

Work supervised by
Prof. Dr. Miguel Angel Guevara Lopez
Prof. Dr. Luís Gonzaga Mendes Magalhães

outubro de 2026

Copyright and Terms of Use for Third Party Work

This dissertation reports on academic work that can be used by third parties as long as the internationally accepted standards and good practices are respected concerning copyright and related rights.

This work can thereafter be used under the terms established in the license below.

Readers needing authorization conditions not provided for in the indicated licensing should contact the author through the RepositóriUM of the University of Minho.

License granted to users of this work:



CC BY-NC-SA

`https://creativecommons.org/licenses/by-nc-sa/4.0/`

- **Attribution** — You must give appropriate credit, provide a link to the license, and indicate if changes were made.
- **NonCommercial** — You may not use the material for commercial purposes.
- **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

Acknowledgements

This doctoral journey has been shaped by the generosity, guidance, and support of many individuals and institutions, to whom I owe a profound debt of gratitude.

I begin by expressing my sincere appreciation to the University of Minho and to the School of Engineering for providing the academic environment in which this work was conceived and developed. I am equally grateful to the Department of Information Systems and the Department of Informatics, whose teaching staff, across both my undergraduate and graduate years, laid the scientific and technical foundations upon which this thesis rests. The rigour and intellectual curiosity instilled during those formative years have remained constant companions throughout this research.

I would like to express my deep gratitude to my supervisor, Prof. Dr. Miguel Guevara, whose influence on my academic path predates this thesis entirely. It was during our time working together at the *Centro de Computação Gráfica* that he first encouraged me to pursue doctoral research, and it is fair to say that without that encouragement this work would not exist. His clinical and scientific intuition, combined with a genuine enthusiasm for artificial intelligence applied to medicine, have been a constant source of inspiration throughout. Beyond the academic relationship, I am grateful for the friendship and openness that have characterised our collaboration over the years.

I would also like to thank my co-supervisor, Prof. Dr. Luís Magalhães, for his rigorous technical oversight and for ensuring that the work remained grounded in sound engineering principles. His expertise and his methodical approach to research have contributed decisively to the quality and coherence of this thesis.

I would also like to thank my co-authors and research colleagues, João Raimundo and Evandro Fonseca, whose contributions to the published work associated with this thesis were essential. Our collaboration has been both productive and genuinely enriching, and I look forward to the research we will continue to pursue in the future.

I would like to acknowledge the *Centro de Computação Gráfica* (CCG/ZGDV Institute), where I first worked as a researcher and developed my initial expertise in applied artificial intelligence. The exposure to EU-funded research projects during that period, along with the financial support provided during the first year of my doctoral studies, contributed to the foundations upon which this thesis was built.

I would also like to thank Loka Inc. for fostering a culture of continuous learning and technical excellence that has kept my engineering instincts sharp during the final stages of this doctoral work.

I would like to thank the VistaLab — Centro ALGORITMI and BigData Laboratory, University of Évora, Portugal, for easing the access and instructions to the GPU Cluster (DGX-A100 NVIDIA stations). A set of performed experiments was only possible within a reasonable timeframe due to the computational power of this infrastructure.

I would like to thank the Senology Unit at the Arrábida Local Health Unit and the PRIMED project *"Precision medical imaging platform to improve personalised medicine tasks in the detection and diagnosis of chronic non-communicable diseases. An approach to cancer detection and diagnosis"*, funded by the LISBOA2030-FSE+-01817800 Programme, for providing me with access to the digital repository on breast cancer and the evaluation of the machine learning models developed.

I would like to thank my friends for their companionship and good humour throughout this journey. Their encouragement, curiosity about the work, and ability to provide perspective at the right moments have been a constant source of energy and motivation. In particular, I would like to thank Diogo, Tiago, Marcos, Rui, João, Pedro, and Ricardo, who have been present through the best and hardest moments of this process, often without knowing how much their company mattered. To all those whose names space does not permit me to list, but who have shared a meal, a conversation, or simply a moment of levity at the right time, thank you. You are not forgotten.

Finally, I would like to thank my family for their unconditional love and support. To my girlfriend, Joana, whose patience and understanding have been invaluable during the most challenging moments of this journey, I am deeply grateful. Her support has been a constant source of comfort and motivation. My parents, João and Helena, who have always believed in me and encouraged me to pursue my passions. Their sacrifices and unwavering belief in my potential have been the foundation upon which this thesis was built. To my girlfriend's mother, Leonor, and brother, Diogo, who have also been a source of support and encouragement, I extend my heartfelt thanks. Their kindness and understanding have made this journey more manageable and less isolating. To all of you, I am profoundly grateful.

Statement of Integrity

I hereby declare having conducted this academic work with integrity.

I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, Guimarães, October 2026

João Pedro Pereira Fontes

Declaration on the Use of Artificial Intelligence (AI) Tools

I hereby declare having acted this academic work with integrity and confirm that I have not resorted to any form of undue use of AI tools, nor to the falsification or omission of information, sources, data, results or processes in any of the stages leading to its elaboration.

I further declare that any use of AI tools was done in a transparent and responsible manner, being duly noted in the work, and that I critically verified the accuracy of the content produced with their support, assuming full responsibility for what I submit.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho, as well as the applicable standards of academic integrity and data protection.

University of Minho, Guimarães, October 2026

João Pedro Pereira Fontes

Abstract

Breast cancer molecular subtype determines prognosis and treatment strategy, yet its conventional assessment via immunohistochemistry is invasive and resource-intensive. This thesis proposes a modular framework for quantitative imaging biomarker (QIB) discovery from volumetric medical imaging, decomposing the problem into two independent stages: automated lesion localisation and downstream phenotype classification into Luminal A, Luminal B, HER2-enriched, and Triple-Negative subtypes.

Breast cancer subtype classification from dynamic contrast-enhanced MRI (DCE-MRI) serves as the primary validation case study. A 3D convolutional neural network trained under a one-versus-all strategy produces four independent binary classifiers on the Duke-Breast-Cancer-MRI dataset (922 patients), with percentile-based intensity normalisation, class-weighted loss, and volumetric and rotational data augmentation to address class imbalance.

External validation on an independent Portuguese cohort (*Unidade Local de Saúde da Arrábida*, $n = 98$) yields AUC-ROC values of 0.9610, 0.9205, 0.9788, and 0.9229 (weighted mean 0.9416, weighted by per-subtype case count) without fine-tuning, despite a pronounced subtype prior shift. A secondary multicentric experiment demonstrates that naïve data pooling without domain harmonisation degrades cross-institutional generalisation (weighted mean AUC-ROC 0.8644). The modular design provides a transferable foundation for Deep Learning (DL)-based QIB assessment across oncological contexts.

Keywords Breast Cancer Molecular Subtypes, Deep Learning, Dynamic Contrast-Enhanced MRI, Quantitative Imaging Biomarkers, Radiogenomics

Resumo

O subtipo molecular do cancro da mama determina o prognóstico e a estratégia terapêutica, mas a sua avaliação convencional por imuno-histoquímica é invasiva e intensiva em recursos. Este trabalho propõe uma *framework* modular para a descoberta de biomarcadores quantitativos de imagem (QIBs) a partir de dados volumétricos, decompondo o problema em duas etapas independentes: localização automatizada de lesões e classificação fenotípica nos subtipos Luminal A, Luminal B, HER2-enriquecido e Triplo Negativo.

A classificação de subtipos moleculares do cancro da mama a partir de DCE-MRI serve como caso de estudo principal. Uma rede neuronal convolucional 3D, treinada seguindo uma estratégia um-contra-todos, produz quatro classificadores binários independentes no *dataset* Duke-Breast-Cancer-MRI (922 pacientes), com normalização de intensidade por percentis, função de erro ponderada por classe e *data augmentation* volumétrico e rotacional para mitigar o desbalanceamento de classes.

A validação externa num conjunto de dados prospetivo independente da *Unidade Local de Saúde da Arrábida* ($n = 98$) obtém valores de AUC-ROC de 0.9610, 0.9205, 0.9788 e 0.9229 (média ponderada 0.9416) sem qualquer ajuste ao domínio alvo, apesar de uma acentuada diferença na distribuição de subtipos. Uma experiência multi-centro demonstra que a agregação de dados naïve sem harmonização de domínio degrada a generalização inter-institucional (AUC-ROC média ponderada 0.8644). A conceção modular da *framework* constitui uma base transferível para uma avaliação baseada em *Deep Learning* em contextos oncológicos mais amplos.

Palavras-chave Biomarcadores de Imagem Quantitativos, *Deep Learning*, Radiogenómica, Ressonância Magnética Dinâmica com Contraste, Subtipos Moleculares do Cancro da Mama

Contents

1 Introduction **1**

1.1 Background and Motivation 2

1.2 Problem Statement 3

1.3 Objectives 5

1.4 Methodology 6

1.5 Thesis Contributions 7

1.5.1 Empirical Contributions 7

1.5.2 Methodological Contributions 8

1.5.3 Conceptual Contributions 8

1.6 Document Structure 9

2 Quantitative imaging biomarkers for breast cancer phenotyping **11**

2.1 Molecular subtypes of breast cancer: from gene expression to IHC surrogates 11

2.2 Medical imaging in breast cancer 12

2.3 DCE-MRI acquisition, pharmacokinetics, and subtype-specific phenotypes 14

2.3.1 Technical foundations and kinetic modelling 14

2.3.2 Enhancement patterns encode molecular information 14

2.3.3 Public datasets enabling reproducible research 15

2.3.4 Preprocessing and harmonisation pipelines 15

2.4 Radiomics: from handcrafted features to standardised quantitative biomarkers 15

2.4.1 Formal definitions and standardisation 15

2.4.2 Feature families and their origins 16

2.4.3 The radiomics paradigm and its landmark studies 16

2.4.4 Classical ML for breast cancer radiomics 16

2.4.5 Reproducibility limitations motivating deep learning 16

2.5 Deep learning architectures for volumetric breast MRI . . . 17
2.5.1 Capturing inter-slice spatial context . . . 17
2.5.2 Foundational three-dimensional architectures . . . 17
2.5.3 Attention mechanisms and interpretability . . . 17
2.5.4 Transfer learning for three-dimensional medical imaging . . . 18
2.5.5 Regularisation foundations . . . 18
2.6 Deep learning for breast cancer molecular subtype classification . . . 18
2.6.1 Landmark and foundational studies . . . 18
2.6.2 Recent advances achieving highest reported performance (2022 to 2025) . . . 19
2.6.3 Key patterns and persistent limitations . . . 19
2.7 Multicentric generalisation and domain adaptation . . . 20
2.7.1 The domain shift problem in medical imaging . . . 20
2.7.2 Domain adaptation techniques . . . 20
2.7.3 Multicentric breast cancer MRI resources . . . 20
2.8 Addressing class imbalance in breast cancer subtype classification . . . 21
2.8.1 The fundamental problem . . . 21
2.8.2 Cost-sensitive loss functions . . . 21
2.8.3 Synthetic oversampling . . . 21
2.9 Validation methodology to ensure clinical trustworthiness . . . 21
2.9.1 Cross-validation best practices . . . 21
2.9.2 External validation . . . 22
2.9.3 Reporting frameworks . . . 22
2.9.4 Statistical testing . . . 22
2.9.5 Reproducibility concerns . . . 23
2.10 Summary and implications . . . 23

3 Materials and Data Preparation 24

3.1 Case Study: Breast Cancer . . . 24
3.2 Data Sources and Selection Criteria . . . 25
3.2.1 The Cancer Imaging Archive . . . 25
3.2.2 Additional Dataset Sources Beyond TCIA . . . 26
3.2.3 The Portuguese Breast Cancer Digital Repository . . . 26
3.3 Dataset Descriptions . . . 27

3.3.1 Duke Breast Cancer MRI Dataset . . . 27
3.3.2 I-SPY 1 / ACRIN 6657 . . . 28
3.3.3 I-SPY 2 . . . 28
3.3.4 Breast-MRI-NACT-Pilot . . . 29
3.3.5 Advanced-MRI-Breast-Lesions . . . 29
3.4 Phenotype Distribution and Class Imbalance . . . 29
3.5 Multicentric Repository Construction . . . 30
3.5.1 Dataset Integration . . . 30
3.5.2 Prospective Clinical Cohort: Unidade Local de Saúde da Arrábida (ULSA), Hospital São Bernardo, Setúbal . . . 32
3.6 External Validation on a Prospective Clinical Cohort . . . 32
3.6.1 The Necessity of External Validation in Clinical Artificial Intelligence . . . 32
3.6.2 The ULSA Setúbal Cohort . . . 33
3.6.3 Preprocessing of the Setúbal Cohort . . . 35
3.6.4 External Validation Protocol . . . 36
3.7 Data Preprocessing . . . 37
3.7.1 Input Representation and Bounding Box Provenance . . . 37
3.7.2 Volume Construction and Spatial Standardisation . . . 39
3.7.3 Intensity Normalisation . . . 40
3.7.4 Augmentation by Discrete In-Plane Rotation . . . 42

4 Early Approaches to Breast Cancer Phenotype Classification 45
4.1 Two-Dimensional Experiments . . . 46
4.1.1 Motivation and Configuration . . . 46
4.1.2 The Data Leakage Incident and Its Correction . . . 47
4.1.3 Results and Observed Limitations . . . 47
4.2 Rationale for Transitioning to Volumetric Processing . . . 48
4.3 Initial Three-Dimensional Approach . . . 49
4.3.1 Scope and Proof-of-Concept Design . . . 49
4.3.2 Preprocessing Pipeline . . . 50
4.3.3 Architectural Configuration . . . 50
4.3.4 Training Configuration . . . 51
4.3.5 Results . . . 51

4.4 Limitations and Motivation for the Current Approach . . . 53
4.5 Summary . . . 55

5 A New Approach for Classifying Breast Cancer Phenotypes 56

5.1 Framework Design Philosophy . . . 57
5.2 One-Versus-All Classification Strategy . . . 59
5.2.1 Formulation and Motivation . . . 59
5.2.2 Justification for the One-vs-All Choice . . . 59
5.3 Three-Dimensional CNN Architecture . . . 61
5.3.1 Architectural Overview . . . 61
5.3.2 Convolutional Blocks . . . 62
5.3.3 Global Average Pooling and Classification Head . . . 62
5.3.4 Comparison with ResNet Backbone . . . 63
5.4 Addressing Class Imbalance via Weighted Binary Cross-Entropy . . . 64
5.4.1 The Imbalance Problem in Context . . . 64
5.4.2 Class Weight Calculation . . . 64
5.5 Training Procedure . . . 65
5.5.1 Optimiser and Learning Rate Schedule . . . 65
5.5.2 Loss Function and Compilation Metrics . . . 66
5.5.3 Callbacks and Model Selection . . . 66
5.5.4 Batch Size, Maximum Epochs, and Data Partitioning . . . 67
5.5.5 Reproducibility Controls . . . 68
5.6 Hyperparameter Optimisation . . . 68
5.7 Summary of Design Choices . . . 69

6 Results and Discussion 71

6.1 Duke-Only Training: External Validation on the ULSA Setúbal Cohort . . . 72
6.1.1 Experimental Configuration . . . 72
6.1.2 Training Dynamics . . . 73
6.1.3 Per-Classifier Performance on the Setúbal Cohort . . . 74
6.2 TCIA-Augmented Training: Full Multicentric Corpus . . . 78
6.2.1 Experimental Configuration . . . 79
6.2.2 Training Dynamics . . . 79

6.2.3 Per-Classifier Performance on the Setúbal Cohort . . . 80
6.2.4 Mechanistic Hypotheses for the Observed Degradation . . . 83
6.3 Comparative Analysis . . . 85
6.3.1 Side-by-Side External Validation Comparison . . . 85
6.3.2 Interpretation . . . 86
6.4 Comparison with the State of the Art . . . 87
6.4.1 Classical Machine Learning and Radiomics Baselines . . . 88
6.4.2 Deep Learning Comparators . . . 89
6.4.3 Analysis of Comparative Performance . . . 89

7 Conclusions and Future Work 92
7.1 Summary of Contributions . . . 92
7.2 Principal Findings . . . 94
7.3 The Classifier Output as a Quantitative Imaging Biomarker . . . 96
7.4 Towards a Domain-Agnostic QIB Discovery Framework . . . 97
7.4.1 Prostate Cancer on Multiparametric MRI . . . 97
7.4.2 Glioma Molecular Status on Multimodal Brain MRI . . . 97
7.4.3 Pulmonary Nodule Malignancy on Low-Dose CT . . . 98
7.4.4 Generalisability: A Qualified Claim . . . 98
7.5 Limitations . . . 99
7.6 Future Work . . . 100
7.7 Closing Remarks . . . 102

Appendices 124

A Support work 125
A.1 Duke-Only Training: Learning Curves . . . 125
A.2 TCIA-Augmented Training: Learning Curves . . . 125

B Details of Results 129
B.1 Deep Learning State-of-the-Art Comparison . . . 129

List of Figures

1 Comparison of three breast imaging modalities 13

2 Phenotype distribution in the Duke-Breast-Cancer-MRI dataset 30

3 Phenotype distribution across the four TCIA collections and the combined corpus 31

4 Preprocessing pipeline overview 44

5 3D CNN architecture for phenotype classification 61

6 ROC curves for the four OvA classifiers on the ULSA Setúbal cohort (Duke-only training) . 76

7 ROC curves for the four OvA classifiers on the ULSA Setúbal cohort (TCIA-augmented training) 81

8 Per-epoch AUC and loss learning curves for the Luminal A classifier trained on the Duke dataset 125

9 Per-epoch AUC and loss learning curves for the Luminal B classifier trained on the Duke dataset 126

10 Per-epoch AUC and loss learning curves for the HER2-enriched classifier trained on the Duke dataset 126

11 Per-epoch AUC and loss learning curves for the Triple-Negative classifier trained on the Duke dataset 126

12 Per-epoch AUC and loss learning curves for the Luminal A classifier trained on the TCIA multicentric corpus 127

13 Per-epoch AUC and loss learning curves for the Luminal B classifier trained on the TCIA multicentric corpus 127

14 Per-epoch AUC and loss learning curves for the HER2-enriched classifier trained on the TCIA multicentric corpus 127

15 Per-epoch AUC and loss learning curves for the Triple-Negative classifier trained on the TCIA multicentric corpus 128

List of Tables

1 Breast cancer datasets in TCIA that provide both imaging and genomic data 26

2 Molecular subtype distribution in the ULSA Setúbal external validation cohort 34

3 Performance comparison of 2D and initial 3D approaches on the Luminal A classification task 52

4 Cross-validation results of the initial 3D CNN and the 3D ResNet-152 on the Luminal A classification task 52

5 Summary of principal design choices in the classification framework 70

6 Per-classifier performance on the ULSA Setúbal cohort (Duke-only training) 75

7 Binary confusion matrices for each OvA classifier on the ULSA Setúbal cohort (Duke-only training) 77

8 Per-classifier performance on the ULSA Setúbal cohort (TCIA-augmented training) 81

9 Binary confusion matrices for each OvA classifier on the ULSA Setúbal cohort (TCIA-augmented training) 82

10 Comparison of Duke-only and TCIA-augmented configurations on the ULSA Setúbal cohort 86

11 Deep learning state-of-the-art comparison for breast cancer molecular subtype classification 130

Acronyms

AI Artificial Intelligence.

AUC-ROC Area Under the Receiver Operating Characteristic Curve.

CI Confidence Interval.

CNN Convolutional Neural Network.

CSV Comma-Separated Values.

CT Computerized Tomography.

DCE-MRI Dynamic Contrast-Enhanced Magnetic Resonance Imaging.

DICOM Digital Imaging and Communications in Medicine.

DL Deep Learning.

ER Estrogen Receptor.

FDA Food and Drug Administration.

GDPR General Data Protection Regulation.

GPU Graphics Processing Unit.

HER2 Human Epidermal growth factor Receptor 2.

IBSI Image Biomarker Standardisation Initiative.

IDH Isocitrate Dehydrogenase.

IHC Immunohistochemistry.

LSTM Long Short-Term Memory.

ML Machine Learning.

MRI Magnetic Resonance Imaging.

NACT Neoadjuvant Chemotherapy.

NIH National Institutes of Health.

OvA One-vs-All.

PET Positron Emission Tomography.

PR Progesterone Receptor.

QIB Quantitative Imaging Biomarker.

ReLU Rectified Linear Unit.

SMOTE Synthetic Minority Over-sampling Technique.

SVM Support Vector Machine.

TCGA The Cancer Genome Atlas.

TCIA Cancer Imaging Archive.

TIFF Tagged Image File Format.

ULSA Unidade Local de Saúde da Arrábida.

WHO World Health Organization.

Chapter 1

Introduction

The advent of digitalization has revolutionized numerous fields, including medical image analysis, by enabling computational tools to perform tasks that were once exclusively manual. This shift began with the introduction of digital imaging modalities in the late twentieth century, allowing for the storage, transmission, and manipulation of medical images in electronic formats. Early applications focused on basic enhancements such as contrast adjustment and noise reduction, but the integration of computer vision techniques marked a significant leap forward [1]. These methods, initially rule-based and reliant on handcrafted features, enabled automated detection of anatomical structures and abnormalities, reducing the workload on radiologists and improving diagnostic consistency.

As computational power increased and algorithms evolved, Machine Learning (ML) emerged as a powerful paradigm for medical image analysis. ML models, trained on labelled datasets, could learn patterns and make predictions without explicit programming, outperforming traditional methods in tasks such as segmentation and classification. The subsequent rise of Deep Learning (DL), particularly Convolutional Neural Networks (CNNs), further accelerated progress. Inspired by the human visual cortex, CNNs automatically extract hierarchical features from images, demonstrating superior performance in complex scenarios [2]. In medical imaging, DL has been applied to modalities such as Magnetic Resonance Imaging (MRI) and Computerized Tomography (CT), enabling precise tumour localisation, organ segmentation, and disease classification with accuracies approaching or even exceeding expert levels [3–7].

This transformation has empowered physicians with advanced techniques for image manipulation, leveraging cutting-edge technologies such as computer vision and Artificial Intelligence (AI) algorithms. Research advancements have facilitated the automated acquisition and analysis of medical images, utilizing modalities such as CT and MRI. These techniques now harness three-dimensional capabilities to segment organs or detect tumour locations with remarkable precision [8]. However, to maximise their utility in clinical practice, it is essential to enhance these methods, making them more comprehensive, reliable, and integrated into daily workflows, thereby demonstrating their indispensable value to healthcare

professionals. Challenges such as inter-observer variability in image quality assessment [9] and the need for robust, generalisable models underscore the importance of continued innovation in this domain.

The concept of Quantitative Imaging Biomarkers (QIBs) has emerged at the intersection of these computational advances and clinical need. A QIB is a quantitative characteristic derived from medical images that objectively reflects an underlying biological process or its response to a therapeutic intervention [10]. Unlike qualitative, reader-dependent assessments, QIBs provide reproducible, numerical measurements amenable to statistical modeling, prognostic stratification, and regulatory evaluation. The discipline of radiomics, which systematically mines large feature sets from radiographic images, has been instrumental in expanding the repertoire of candidate QIBs [11]. More recently, the ability of DL architectures to learn discriminative representations directly from raw volumetric data has opened the possibility of constructing QIBs without manual feature engineering, substantially improving both performance and scalability.

ML and DL techniques stand at the vanguard of innovations in cancer research, achieving superior results by learning patterns from data without explicit programming [12]. These methods minimise error functions through iterative training on datasets, where the quality and preparation of the training data are of paramount importance. The synergy of algorithmic advancements and abundant data has propelled ML and DL methods to performance levels rivalling or surpassing human expertise in specific tasks. This work capitalises on these developments to construct robust tools to aid cancer diagnosis and management.

1.1 Background and Motivation

Breast cancer is a highly prevalent and complex malignancy type of tumour, recognised as one of the leading causes of cancer-related mortality among women globally. According to recent estimates, in 2020 it accounted for approximately 2.3 million new diagnoses and over 685000 deaths, with projections suggesting these figures will surpass 3 million annual incidences and 1 million deaths by 2040, underscoring its significant and growing public health burden [13, 14]. This disease exhibits remarkable molecular heterogeneity, manifesting in subtypes such as Luminal A, Luminal B, HER2-enriched, and Triple-Negative, each presenting unique genetic profiles, clinical behaviours, prognostic outlooks, and therapeutic implications [15, 16]. Luminal A represents the majority of cases, carrying the most favourable prognosis and responding well to endocrine therapy, while Triple-Negative and HER2-enriched subtypes display aggressive behaviour and higher recurrence rates, demanding chemotherapy or targeted anti-HER2 agents.

Precise molecular subtyping is essential for personalised treatment planning and prognostic stratification. The clinical gold standard, Immunohistochemistry (IHC), assesses Estrogen Receptor (ER), Proges-

terone Receptor (PR), Human Epidermal growth factor Receptor 2 (HER2), and Ki-67 status directly from tumour tissue. While effective, this approach requires an invasive biopsy, is subject to sampling bias in heterogeneous tumours, and cannot be repeated at will for longitudinal monitoring [17]. These limitations create a compelling demand for non-invasive, imaging-based alternatives capable of reliably predicting molecular subtypes.

Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI) has become the reference modality for this purpose: it provides rich, multi-phase volumetric data capturing tumour neoangiogenesis and morphological characteristics at high sensitivity, offering information that correlates strongly with molecular subtype and treatment response [9, 18]. Nevertheless, the volumetric and multi-temporal nature of DCE-MRI introduces considerable analytical challenges. Manual segmentation of tumour regions is labour-intensive and subject to inter-observer variability, limiting scalability for large cohorts. Moreover, publicly available repositories exhibit marked class imbalance, with rarer subtypes such as HER2-enriched being substantially underrepresented, leading to biased models when naïve training strategies are employed.

Advancements in ML and DL have opened new avenues for addressing these challenges in oncology, particularly in the non-invasive characterisation of tumours. CNNs have demonstrated prowess in analysing volumetric data, capturing spatial and morphological features that correlate with molecular subtypes [2, 3]. The extension of CNNs to three dimensions allows models to exploit inter-slice spatial continuity, capturing intra-tumoural heterogeneity and peritumoural enhancement patterns that are invisible to slice-by-slice analyses. Work published as part of this thesis, including the study on Luminal A phenotyping [19], demonstrates the potential of 3D CNNs to achieve high accuracy in subtype identification, frequently outperforming traditional radiomics approaches by learning hierarchical representations directly from raw images.

The motivation for this work stems from the need to advance these methods by developing a framework for QIB discovery. By addressing data imbalance through multicentric data aggregation and balanced sampling, and by validating models on real-world clinical cohorts, this work aims to bridge the gap between isolated academic experiments and clinically deployable precision oncology tools, ultimately improving patient outcomes through earlier and more accurate subtyping that can guide tailored therapies.

1.2 Problem Statement

Despite the advances described above, several unresolved challenges limit the clinical adoption of DL-based breast cancer subtyping. Three principal barriers can be identified.

The first concerns **data scarcity and class imbalance**. Public DCE-MRI repositories, while invaluable resources, reflect the natural epidemiological distribution of breast cancer subtypes. Luminal A cases constitute approximately 60 to 70% of the dataset, while HER2-enriched cases represent often fewer than 10%. Models trained on such imbalanced distributions without correction exhibit a strong bias toward the majority class, systematically failing on minority subtypes for which clinical decisions are often the most consequential.

The second barrier is **generalisability across institutions**. DL models trained on data from a single imaging centre are prone to overfitting to scanner-specific characteristics, acquisition protocols, and patient demographics. Cross-institutional performance typically degrades substantially when a model encounters data from a different scanner vendor or a different contrast timing protocol, undermining confidence in real-world deployment.

The third challenge is **end-to-end pipeline integration**. Most published approaches address either lesion detection or phenotype classification in isolation. A clinically deployable solution requires a complete, automated pipeline in which detected lesion volumes are directly consumed by a classifier, with no requirement for manual intervention between the two stages. Building and validating such an integrated pipeline on multicentric data, with external confirmation on a prospective hospital cohort, constitutes a substantial engineering and scientific endeavour that has received limited attention in the literature.

This thesis addresses all three challenges. The phenotype classification component, which is the primary focus of the work herein, is designed to operate on pre-detected tumour volumes supplied by an upstream detection module [20], thereby satisfying the integration requirement as a precondition. The classification framework is trained on a multicentric repository drawn from the Cancer Imaging Archive (TCIA) and internally validated on held-out partitions of the same corpus. Its real-world applicability is subsequently confirmed through external validation on an independent prospective cohort from Unidade Local de Saúde da Arrábida (ULSA), Hospital São Bernardo, Centro Hospitalar de Setúbal, Portugal, which is reserved exclusively for that purpose and is never exposed during training or model selection. These considerations motivate the central research questions guiding this thesis:

Research Question. Can a deep learning framework extract clinically validated quantitative imaging biomarkers from volumetric medical imaging data in a manner generalisable across datasets within the same clinical domain, with potential extensibility to other imaging modalities and cancer types? Specifically:

1. Can such frameworks achieve clinical-grade molecular subtype prediction in breast cancer using DCE-MRI, despite severe class imbalance?

2. Do models trained on multicentric repositories generalise to prospective real-world clinical cohorts?

3. Can the classification component operate within an end-to-end automated pipeline, receiving pre-detected tumour volumes and producing subtype predictions without manual intervention?

1.3 Objectives

This work addresses the extraction and validation of quantitative volumetric features from DCE-MRI data to discover novel QIBs. The primary goal is to develop a modular pipeline that decomposes the QIB discovery problem into two independent stages: (i) lesion localisation, treated as a precondition supplied by an upstream detection module [20], and (ii) phenotype classification, which constitutes the focus of this thesis, and to validate this pipeline on a clinically relevant case study. The resulting QIBs emphasise reproducibility and clinical relevance. The modular design of the framework provides a foundation for future extensions towards cloud-based clinical deployment, multimodal feature integration, and automated explainability analysis; however, these extensions are not implemented or validated within the scope of this thesis.

The proposed framework automates the QIB development lifecycle across four stages:

- **Definition of ML and DL diagnostic, prognostic, and predictive models**: Employ feature extraction from imaging, genomic, and clinical sources, followed by selection methods to identify top-performing biomarkers. Radiomics plays a pivotal role here, involving high-throughput feature extraction from radiographic images [11], which are then curated as QIB candidates. In the classification component, 3D CNNs operating directly on DCE-MRI volumes replace hand-crafted radiomic pipelines, learning discriminative representations end-to-end.

- **Technical validation of selected biomarkers**: Assess availability, reproducibility, accuracy, and precision without delving into clinical endpoints, ensuring robust performance across datasets. Internal cross-validation on the primary corpus and external evaluation on an independent cohort together satisfy this requirement at both algorithmic and clinical-readiness levels.

- **Establishing links between QIBs and clinical outcomes**: Identify optimal QIBs that correlate strongly with clinical variables, facilitating targeted use cases. In this work, this objective manifests through the association of imaging-derived classification scores with molecular subtype labels, enabling a non-invasive surrogate for immunohistochemical assessment.

- **Integration into automatic data analysis pipelines**: Embed validated QIBs in modular, extensible pipelines designed to facilitate future integration into cloud platforms for ongoing research and clinical application. The classification component receives tumour volumes from the upstream detection module and emits continuous probability scores suitable for incorporation into clinical decision-support workflows in future deployment scenarios.

Breast cancer molecular subtyping serves as the validation case study, chosen for its prevalence, the availability of annotated public datasets, and the particular clinical significance of accurate non-invasive molecular subtype prediction. This work aims to contribute novel biomarkers linked to specific outcomes, emphasising reproducibility and clinical utility as first-order design criteria.

1.4 Methodology

The methodology adopted in this thesis follows a structured pipeline for the development and evaluation of QIBs focused on breast cancer phenotype classification. It begins with the careful curation of datasets from diverse sources, ensuring a robust foundation for model training and validation. The Duke Breast Cancer MRI dataset serves as the primary resource for model development, comprising DCE-MRI volumes from 922 confirmed invasive breast cancer cases, annotated with bounding boxes delineating tumour regions and labelled with the corresponding molecular subtype: Luminal A, Luminal B, HER2-enriched, and Triple-Negative [21].[1] To investigate the impact of multicentric diversity on cross-institutional generalisation, three additional TCIA collections are aggregated with the Duke corpus — I-SPY 1 (ACRIN 6657), I-SPY 2, and the Breast-MRI-NACT-Pilot dataset — yielding a combined training corpus of 2 114 cases across four institutions; this TCIA-augmented configuration constitutes a distinct experimental contribution of the framework. An independent clinical cohort from ULSA, Setúbal, Portugal, provides prospective external validation in a real-world setting and is never exposed during training or model selection.

Data preprocessing constitutes a critical step to address inherent challenges in medical imaging datasets, particularly class imbalance and variability in image quality. Given the underrepresentation of certain subtypes, specifically HER2-enriched, a strategic selection process is employed to construct balanced training subsets. Volumes are cropped around the pre-detected tumour regions, resized to a standardised input dimension, and normalised to mitigate intensity variations across scanners and acqui-

[1] Throughout this thesis, the terms "phenotype" and "molecular subtype" are used interchangeably to refer specifically to the IHC-surrogate four-class classification (Luminal A, Luminal B, HER2-enriched, Triple-Negative) derived from receptor status and proliferation markers. This usage differs from general oncological practice, in which "phenotype" may denote morphological or clinical presentation; here it refers exclusively to the molecular classification captured by immunohistochemical biomarkers, as formalised in Chapter 2.

sition protocols. Augmentation techniques are applied during training to improve model robustness and to reduce the risk of overfitting on the smaller minority classes.

The core of the classification pipeline relies on a 3D CNN architecture, designed to exploit the inherently volumetric nature of MRI data. This model processes three-dimensional inputs directly, capturing spatial hierarchies and morphological features essential for distinguishing between phenotypes. The use of three-dimensional convolutions preserves inter-slice spatial continuity, which is critical for characterising intra-tumoural heterogeneity patterns known to correlate with molecular subtype identity. Crucially, the classification component is preconditioned on the upstream lesion detection module described in [20]: it receives tumour bounding volumes from that module, and development of the detection component falls outside the scope of this thesis.

Model evaluation employs standard metrics tailored to imbalanced multi-class classification problems, including Area Under the Receiver Operating Characteristic Curve (AUC-ROC) computed per subtype, precision, recall, and F1-score, enabling a granular assessment of performance that is not dominated by the majority class. External validation on the prospective Setúbal cohort assesses real-world clinical applicability.

Further technical details, including specific architectural choices, hyperparameter configurations, training procedures, and per-subtype performance breakdowns, are elaborated in subsequent chapters.

1.5 Thesis Contributions

The original contributions of this thesis can be summarised as follows. First, the development and external validation of a 3D CNN-based classification pipeline on an independent prospective clinical cohort. Second, the demonstration that naïve multicentric data aggregation without domain harmonisation degrades generalisation performance. Third, a set of methodological design choices for imbalanced volumetric classification, including a percentile-based normalisation scheme, a clinical-checkpoint callback, and a class-weighted One-vs-All (OvA) training strategy. Fourth, the formalisation of continuous classifier outputs as candidate QIBs. Each of these contributions is elaborated in the subsections below.

1.5.1 Empirical Contributions

The primary empirical contribution of this thesis is the external validation of the proposed classification pipeline on an independent prospective cohort from ULSA, Hospital São Bernardo, Setúbal, Portugal ($n = 98$). The four-subtype OvA framework, trained exclusively on the Duke-Breast-Cancer-MRI dataset,

achieves per-subtype AUC-ROC values of 0.9610, 0.9205, 0.9788, and 0.9229 for Luminal A, Luminal B, HER2-enriched, and Triple-Negative respectively (weighted mean 0.9416), without any fine-tuning, recalibration, or adaptation to the target domain, and under a pronounced subtype prior shift between the North American training distribution and the Portuguese evaluation cohort [22]. This result addresses a major limitation in the existing literature, where independent external validation on prospective clinical data is frequently absent. A secondary experiment extending training to a multicentric corpus of four TCIA collections yields a further empirical finding: naïve data pooling across heterogeneous acquisition sites, in the absence of domain harmonisation, degrades cross-institutional generalisation relative to the single high-quality annotated source, reducing the weighted mean external AUC-ROC to 0.8644. This result highlights that data quantity and source diversity are necessary but not sufficient conditions for improved generalisation, and that annotation quality and distributional coherence are equally important determinants of cross-institutional transferability.

1.5.2 Methodological Contributions

The classification pipeline is built on a set of principled methodological choices designed to address the specific challenges of imbalanced volumetric medical image classification. The OvA decomposition strategy decomposes the four-class molecular subtyping problem into four independent binary classifiers, enabling per-subtype calibration of class weights and independent checkpoint selection [23]. A custom 3D CNN architecture is designed to exploit the volumetric structure of DCE-MRI data, preserving inter-slice spatial continuity that two-dimensional approaches inherently discard. A percentile-based intensity normalisation scheme is adopted to ensure robustness to cross-site intensity scale variation, replacing the globally fixed intensity window employed in prior published work [19, 22]. A clinical-checkpoint callback mechanism conditions model selection on minimum per-subtype recall thresholds, acknowledging that the clinical consequences of false negatives are not uniform across subtypes. Together with a rotational volumetric augmentation strategy applied exclusively to the training partition, these design choices constitute a reproducible methodology for end-to-end volumetric tumour classification. Full architectural and training details are presented in Chapter 5.

1.5.3 Conceptual Contributions

A further contribution of this work concerns the interpretation of the classification output within the QIB framework that gives the thesis its title. Each OvA classifier emits a continuous probability score $f_k(\mathbf{x}) \in [0, 1]$ for a given input volume $\mathbf{x}$. This score constitutes a candidate QIB in the formal sense of Kessler et

al. [10]: it is a quantitative characteristic derived from a medical image that objectively reflects an underlying biological process, namely the molecular phenotype of the tumour as determined by the expression status of ER, PR, HER2, and Ki-67. By formalising this connection, the thesis positions the classification pipeline not merely as a multi-class image classifier but as a device for non-invasive QIB discovery. The continuous nature of the output further provides a graded confidence signal with clinical utility beyond the binary classification decision, as elaborated in Section 7.3.

Communication and Dissemination The work described above has been disseminated through peer-reviewed publication. The binary Luminal A classifier, which serves as the initial proof of concept for the present framework, was published in *Computers in Biology and Medicine* [19]. The complete four-subtype classification framework is detailed in a manuscript submitted to *Biomedical Signal Processing and Control* [22], currently under review. The upstream lesion detection component, which is preconditioned for the classification pipeline and falls outside the scope of this thesis, is described in [20].

1.6 Document Structure

This work is organized as follows:

- **Chapter 1**: Introduces the research context and motivation for non-invasive breast cancer molecular subtyping. The chapter formulates the problem statement regarding data imbalance, cross-institutional generalisability, and end-to-end pipeline integration, defines the objectives of the QIB discovery framework, outlines the methodology based on 3D CNNs, and summarises the empirical, methodological, and conceptual contributions of the thesis.

- **Chapter 2**: Reviews the foundational concepts and state-of-the-art in breast cancer molecular subtyping, medical imaging modalities, DCE-MRI principles, radiomics, and DL architectures for volumetric analysis. The chapter establishes the theoretical and methodological context for the proposed framework.

- **Chapter 3**: Describes the datasets employed in this work, including the Duke-Breast-Cancer-MRI dataset and four candidate TCIA collections, of which three are retained for training following exclusion criteria detailed in Section 3.3.5. The chapter details the data selection criteria, characterises the phenotype distribution and class imbalance challenges, and presents the preprocessing pipeline for standardising multicentric DCE-MRI acquisitions.

- **Chapter 4**: Documents the iterative experimental trajectory that preceded the final framework. The chapter covers the initial two-dimensional experiments conducted across multiple backbone architectures, a critical data leakage incident and its correction, the empirical limitations of the two-dimensional paradigm that motivated the transition to volumetric processing, and the first three-dimensional approach targeting Luminal A binary classification [19]. The chapter establishes, through concrete results and identified limitations, the rationale for every major design decision described in Chapter 5.

- **Chapter 5**: Presents the proposed QIB discovery framework for breast cancer phenotype classification, formalising its design philosophy, the structural decomposition into localisation and classification stages, and the architectural generalisability argument. The chapter details the one-versus-all classification strategy, the 3D CNN architecture and its regularisation design, the class imbalance mitigation strategy, and the complete training procedure including optimiser configuration, learning rate schedule, early stopping criterion, and hyperparameter optimisation pipeline.

- **Chapter 6**: Reports and interprets the experimental results of the framework across both internal evaluation on the TCIA corpus and external validation on the ULSA Setúbal cohort. The chapter analyses per-subtype classification performance, examines the generalisation behaviour of the trained models in the presence of domain shift, and situates the findings in the context of the state-of-the-art reviewed in Chapter 2.

- **Chapter 7**: Summarises the principal contributions of the thesis, reflects on the limitations of the proposed framework, and outlines directions for future research, including extensions of the QIB discovery pipeline to other cancer types and imaging modalities.

Chapter 2

Quantitative imaging biomarkers for breast cancer phenotyping

Non-invasive prediction of breast cancer molecular subtypes from Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI) using Deep Learning (DL) has emerged as a viable alternative to invasive tissue sampling, driven by documented limitations of Immunohistochemistry (IHC)-based biopsy and rapid advances in three-dimensional Convolutional Neural Networks (CNNs). This review synthesizes the foundational biology, imaging science, computational methodology, and validation frameworks that collectively underpin a thesis on discovering Quantitative Imaging Biomarkers (QIBs) for breast cancer subtype classification. It spans eight interconnected domains, from the molecular taxonomy of breast cancer through DCE-MRI physics, radiomics, three-dimensional deep learning architectures, multicentric generalisation, class imbalance mitigation, and rigorous validation methodology, providing the scholarly context against which such work is evaluated.

2.1 Molecular subtypes of breast cancer: from gene expression to IHC surrogates

The modern understanding of breast cancer as a collection of molecularly distinct diseases traces to Perou et al. [24], whose complementary deoxyribonucleic acid microarray analysis of 65 tumours identified intrinsic subtypes: Luminal, Basal-like, ERBB2+, and Normal breast-like. Those were subsequently refined by Sørlie et al. [25, 26] into Luminal A, Luminal B, HER2-enriched, and Basal-like subgroups with distinct survival outcomes. The Cancer Genome Atlas Network [27] provided comprehensive multi-platform confirmation across more than 800 tumours.

For clinical translation, Parker et al. [28] developed the **PAM50** fifty-gene classifier for intrinsic subtype assignment and risk-of-relapse scoring, validated on 710 node-negative patients. Prat and Perou [29]

provided the definitive molecular taxonomy review, describing five subtypes including the Claudin-low group.

The St. Gallen International Expert Consensus introduced IHC surrogate definitions for routine clinical practice. Goldhirsch et al. [30, 31] first adopted the subtype-based approach using Estrogen Receptor (ER), Progesterone Receptor (PR), Human Epidermal growth factor Receptor 2 (HER2), and Ki-67 as surrogates: Luminal A ($ER^+/PR^+/HER_2{}^-$/low Ki-67), Luminal B ($ER^+/HER_2{}^-$/high Ki-67 or $HER_2{}^+$), HER2-positive non-luminal, and Triple-Negative ($ER^-/PR^-/HER_2{}^-$). The 2013 update refined Ki-67 thresholds, with $\geq 20\%$ voted as indicative of "high" status. Cheang et al. [32] had earlier established a Ki-67 cut-point of 13.25% for Luminal A and B distinction. Subsequent St. Gallen updates [33, 34] progressively refined these definitions.

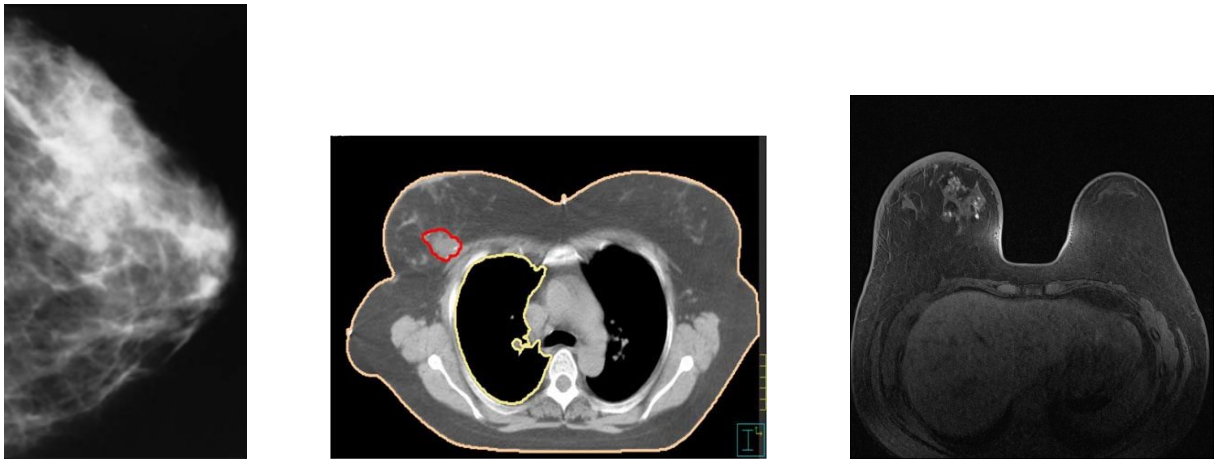
Subtype-specific prognosis and treatment differ markedly. Carey et al. [35] demonstrated differential survival across IHC-defined subtypes in a population-based study of 496 cancers, with basal-like tumours prevalent in premenopausal African American women. Voduc et al. [36] showed that Luminal A had the lowest local and regional relapse risk while HER2-enriched and basal subtypes had significantly increased risk. Kennecke et al. [37] documented subtype-specific metastatic patterns: luminal subtypes to bone, basal-like to brain and lung, HER2+ to liver and brain. Major reviews by Harbeck and Gnant [38], Harbeck et al. [39], and Waks and Winer [40] provide comprehensive treatment frameworks stratified by subtype.

The clinical motivation for non-invasive prediction rests on well-documented limitations of IHC. Intra-tumoural heterogeneity causes sampling error: Allott et al. [41] found core-to-core discordance of 9% (ER), 16% (PR), and 18% (HER2) across 1,085 invasive cancers. Turashvili and Brogi [42] reviewed heterogeneity at morphologic, genomic, and proteomic levels. Discordance between primary and recurrent and metastatic lesions reaches 63.6% for ER and 19.8% for HER2 [43]. These findings collectively establish the rationale for non-invasive imaging-based molecular subtype prediction.

2.2 Medical imaging in breast cancer

Several imaging modalities have been employed in the clinical workup of breast cancer, each offering distinct trade-offs between sensitivity, specificity, radiation burden, and soft-tissue contrast, as illustrated in Figure 1. X-ray mammography (Figure 1a) remains the most widely deployed screening tool owing to its low cost and high throughput. However, its sensitivity is substantially reduced in the presence of dense fibroglandular tissue, and it provides limited information on tumour extent or multifocality. Ultrasound avoids ionising radiation and serves as a valuable adjunct to mammography, particularly for characterising

lesions in dense breasts, yet it is operator-dependent, prone to misinterpretation, and unable to reliably differentiate calcifications from malignant masses [44]. Digital breast tomosynthesis addresses some of the projection overlap artefacts inherent to two-dimensional mammography through pseudo-three-dimensional reconstruction, demonstrating improved detection of masses and architectural distortions. Nonetheless, it remains a relatively recent modality with limited annotated public datasets and higher radiation dose than standard mammography. Computerized Tomography (CT) (Figure 1b) and Positron Emission Tomography (PET)-CT have also been explored for staging and treatment monitoring, but their routine use in primary breast screening is precluded by the high ionising radiation dose and limited spatial resolution for small lesions. As shown in Figure 1b, CT does support precise tumour boundary delineation through volumetric segmentation, a capability that informs the broader concept of QIB extraction from radiological data.



(a) Mammography **(b)** CT with tumour segmentation **(c)** DCE-MRI

Figure 1: Comparison of three breast imaging modalities. Mammography (a) provides high spatial resolution but limited soft-tissue contrast, constraining its utility for lesion characterisation beyond screening. CT (b) enables volumetric acquisition and supports tumour boundary delineation, as illustrated by the annotated contours, yet involves ionising radiation and offers inferior soft-tissue discrimination relative to Magnetic Resonance Imaging (MRI). DCE-MRI (c) combines high soft-tissue contrast with dynamic contrast enhancement, making it the preferred modality for non-invasive molecular subtype characterisation and QIB discovery.

MRI occupies a distinct position within this landscape (Figure 1c). It is the only modality capable of providing high-resolution volumetric information on tumour morphology, vascularity, and functional tissue characteristics without ionising radiation. The mean sensitivity of breast MRI across the clinical literature has been reported at approximately 95.6% [44], substantially exceeding that of mammography or ultrasound, and it is the modality of choice for screening high-risk populations, evaluating tumour extent prior to surgery, and detecting multifocal or contralateral disease [45, 46]. Its primary limitations in the population-screening context are economic and logistical: acquisition is time-consuming, requires intravenous contrast administration, and demands significant radiologist expertise to interpret the resulting multi-phase volumetric sequences.

Crucially for the purposes of this thesis, MRI is also the only modality that encodes the haemodynamic and pharmacokinetic information necessary for molecular subtype-level characterisation. The enhancement kinetics of a tumour under dynamic contrast administration reflect its vascularisation, permeability, and proliferative activity, all of which differ systematically across the Luminal A, Luminal B, HER2-enriched, and Triple-Negative subtypes. This relationship between contrast enhancement patterns and underlying tumour biology, explored in depth in Section 2.3, is what renders DCE-MRI the modality of choice for QIB discovery in the context of phenotype classification.

2.3 DCE-MRI acquisition, pharmacokinetics, and subtype-specific phenotypes

2.3.1 Technical foundations and kinetic modelling

DCE-MRI is the most sensitive imaging modality for breast cancer detection. Mann et al. [47] provided the definitive state-of-the-art review covering acquisition protocols, abbreviated protocols, and expanding screening indications. Quantitative analysis of DCE-MRI relies on pharmacokinetic modelling formalised by Tofts and Kermode [48], who introduced the two-compartment model measuring the volume transfer constant (K^{trans}) and extravascular extracellular space fraction (v_e). The landmark consensus paper by Tofts et al. [49] standardised DCE-MRI parameters K^{trans}, v_e, and k_{ep} and described the extended model incorporating vascular plasma volume fraction v_p.

2.3.2 Enhancement patterns encode molecular information

A growing body of evidence demonstrates that different molecular subtypes exhibit distinct DCE-MRI phenotypes. Mazurowski et al. [50] were among the first to show that Luminal B subtype is associated with specific enhancement dynamics. Grimm et al. [51] extended this to both Luminal A and Luminal B. Braman et al. [52] demonstrated that combined intratumoural and peritumoural radiomic features predict pathological complete response to neoadjuvant chemotherapy, with Area Under the Receiver Operating Characteristic Curves (AUC-ROCs) of 0.93 for triple-negative and HER2+ tumours, and further showed peritumoural radiomics associate with tumour biology and therapy response in HER2+ breast cancer.

2.3.3 Public datasets enabling reproducible research

The Duke Breast Cancer MRI dataset (922 patients with invasive breast cancer and pre-operative DCE-MRI) was described by Saha et al. [53] and released through the Cancer Imaging Archive (TCIA). The I-SPY 1/ACRIN 6657 multicentric dataset provides breast DCE-MRI from 212 subjects enrolled in a prospective neoadjuvant chemotherapy trial [54–56]. The I-SPY 2 dataset was released by Li et al. [57]. Chitalia et al. [58] later enhanced the I-SPY1 collection with uniformly processed scans and expert segmentations. Clark et al. [59] described the TCIA infrastructure itself.

2.3.4 Preprocessing and harmonisation pipelines

Robust preprocessing is essential for multi-institutional analysis. N4ITK bias field correction [60] is the standard algorithm, improving upon N3 with a fast B-spline approximation routine. For intensity standardisation, Nyúl and Udupa [61] proposed histogram landmark-based normalisation. Z-score normalisation remains widely used for its simplicity and effectiveness.

Multi-site harmonisation represents a critical challenge. ComBat, originally developed for genomic batch effects [62], was adapted for neuroimaging diffusion tensor imaging data by Fortin et al. [63, 64] and cortical thickness. Orlhac et al. [65] provided a practical guide for ComBat harmonisation of imaging biomarkers, noting at least 51 papers using ComBat for radiomic harmonisation. For breast MRI specifically, Saha et al. [66] documented how MRI scanner parameters directly affect radiomic feature extraction.

2.4 Radiomics: from handcrafted features to standardised quantitative biomarkers

2.4.1 Formal definitions and standardisation

The concept of quantitative imaging biomarkers was formalised by the Food and Drug Administration (FDA)-National Institutes of Health (NIH) Biomarker Working Group in the BEST glossary [67]. Sullivan et al. [68] established metrology standards for QIBs through the RSNA-QIBA initiative, while Kessler et al. [10] defined the emerging science of QIB terminology. O'Connor et al. [69] published the imaging biomarker roadmap for cancer studies in collaboration with over 60 authors.

The Image Biomarker Standardisation Initiative (IBSI) by Zwanenburg et al. [70] standardised radiomic

feature extraction across software implementations, involving more than 60 co-authors. The PyRadiomics platform [71] provides the principal open-source implementation.

2.4.2 Feature families and their origins

Radiomic features encompass several well-defined families. Texture analysis via grey-level co-occurrence matrices was introduced by Haralick et al. [72]. Grey-level run-length matrices were formalised by Galloway [73]. Grey-level size zone matrices were introduced by Thibault et al. [74], and neighbourhood grey-tone difference matrices by Amadasun and King [75].

2.4.3 The radiomics paradigm and its landmark studies

Three papers established radiomics as a discipline. Lambin et al. [11] introduced the radiomics concept: extracting large numbers of quantitative features from medical images for high-throughput phenotyping. Aerts et al. [76] demonstrated that a radiomic signature could decode tumour phenotype noninvasively across lung and head-and-neck cancers. Gillies et al. [77] articulated the field's vision that "images are more than pictures, they are data". Lambin et al. [78] subsequently positioned radiomics as the bridge between medical imaging and personalised medicine, introducing the Radiomics Quality Score.

2.4.4 Classical ML for breast cancer radiomics

Feature selection typically employs least absolute shrinkage and selection operator regression [79], minimum redundancy maximum relevance [80], or principal component analysis. Parmar et al. [81] systematically compared multiple feature selection and classification methods for radiomics.

For breast cancer specifically, Li et al. [82] predicted molecular classifications using MRI radiomics on the The Cancer Genome Atlas (TCGA) and TCIA dataset. Sutton et al. [83] built a molecular subtype classifier incorporating MRI features with Support Vector Machines (SVMs), achieving 83.4% accuracy for combined pathologic-imaging models. Fan et al. [84] revealed DCE-MRI radiomic features predictive of molecular subtypes. Xie et al. [85] applied multiple Machine Learning (ML) algorithms to multiparametric MRI radiomics for subtype classification.

2.4.5 Reproducibility limitations motivating deep learning

The reproducibility of radiomic features remains a concern. Traverso et al. [86] conducted a systematic review documenting significant variability in feature repeatability and reproducibility. Zwanenburg [70]

addressed data analysis traps and replication crisis in radiomics. These limitations (segmentation sensitivity, scanner variability, and feature instability) provide the rationale for transitioning toward deep learning approaches that learn features directly from data.

2.5 Deep learning architectures for volumetric breast MRI

2.5.1 Capturing inter-slice spatial context

The fundamental limitation of two-dimensional CNNs for volumetric MRI is the loss of inter-slice spatial information. Truhn et al. [87] directly compared radiomic analysis versus CNN classification of contrast-enhancing breast lesions, finding CNNs superior. Whitney et al. [88] compared human-engineered radiomics with CNN transfer learning features, finding that fusion approaches often yield the best performance.

2.5.2 Foundational three-dimensional architectures

The key three-dimensional architectures underpinning modern volumetric medical image analysis are:

- **Three-dimensional U-Net** [89]: Extended two-dimensional U-Net with three-dimensional convolutions, skip connections, and on-the-fly elastic deformations for data augmentation.
- **V-Net** [90]: Introduced the differentiable Dice loss for handling voxel-level class imbalance, with residual connections throughout.
- **Three-dimensional ResNet** [91]: Systematic study demonstrating that three-dimensional ResNets pretrained on Kinetics can serve as foundation models for transfer learning to medical imaging.
- **ResNet** [2]: The original residual learning framework enabling training of very deep networks.
- **nnU-Net** [92]: Self-configuring framework that automatically adapts to any medical image segmentation dataset.

2.5.3 Attention mechanisms and interpretability

Channel attention was formalised by SE-Net [93], which won ILSVRC 2017. The CBAM module [94] combines sequential channel and spatial attention. Attention-gated networks for medical imaging were

proposed by Schlemper et al. [95]. For interpretability, Grad-CAM [96] remains the standard approach, adapted for three-dimensional volumetric data in medical imaging applications.

2.5.4 Transfer learning for three-dimensional medical imaging

Transfer learning in three-dimensional medical imaging faces the challenge of limited pretrained models. Med3D [97] aggregated multiple medical segmentation datasets to pretrain three-dimensional ResNets, achieving two times convergence acceleration versus Kinetics pretraining. Models Genesis [98, 99] demonstrated that self-supervised pretraining using image transformations consistently outperforms both ImageNet transfer and training from scratch across five three-dimensional medical tasks.

2.5.5 Regularisation foundations

Batch normalisation [100] enables higher learning rates and acts as a regulariser. Dropout [101] prevents co-adaptation of features. Global average pooling [102] replaces fully connected layers, serving as a structural regulariser that dramatically reduces parameters and overfitting risk, a foundational concept used in virtually all modern classification CNNs.

2.6 Deep learning for breast cancer molecular subtype classification

2.6.1 Landmark and foundational studies

The most clinically impactful breast MRI deep learning study is Witowski et al. (2022), who developed a system achieving AUC-ROC 0.92 on 3936 internal test exams, matching five board-certified radiologists [103]. While focused on malignancy detection rather than subtype classification, this work established the clinical viability of DL for breast MRI.

For molecular subtype classification specifically, Ha et al. (2019) performed four-class prediction (Luminal A, Luminal B, HER2+, Basal) using a custom 14-layer CNN on 216 patients, achieving macro-AUC-ROC 0.853 and micro-AUC-ROC 0.871, though limited by small sample size and class imbalance [104]. Zhu et al. (2019) applied deep learning to DCE-MRI from 270 TCGA and TCIA patients for binary Luminal A classification, achieving AUC-ROC 0.65 with off-the-shelf deep features, demonstrating that overfitting is the primary challenge with limited training data [105].

Zhang et al. (2021) made a critical contribution by demonstrating the importance of transfer learning across centres. Using CNN and Convolutional Long Short-Term Memory (LSTM) on 244 patients from two centres, direct cross-centre application yielded accuracy of 0.4 to 0.5, but transfer learning improved accuracy by over 30% [106]. Hu et al. (2020) established that multiparametric MRI (DCE-MRI plus T2-weighted) with pretrained CNN features outperforms single-sequence analysis [107].

2.6.2 Recent advances achieving highest reported performance (2022 to 2025)

The field has rapidly advanced. Yin et al. (2022) classified four subtypes using multiparametric MRI with transfer learning, achieving AUC-ROCs of 0.762 to 0.920 on T1-contrast images, with best discrimination for triple-negative and HER2+ subtypes [108].

Zhou et al. (2024) reported the highest published AUC-ROCs for four-class classification: a Channel-Dimensional Feature-Reconstructed DL approach achieved AUC-ROC 0.93 to 0.95 on DCE-MRI alone, and multiparametric fusion (diffusion-weighted MRI plus DCE-MRI) reached AUC-ROC 0.96 to 0.97 on 486 prospective patients [109]. Xie et al. (2024) developed a Multi-Branch CNN fusing multiple DCE-MRI sequences across two institutes, with Luminal A AUC-ROC 0.857 and overall four-class accuracy of 0.64 [110].

Wang et al. (2025) published the first fully automatic end-to-end pipeline combining three-dimensional ResU-Net segmentation with ensemble ResNet classification, validated on 687 patients across multiple external datasets. AUC-ROC ranged from 0.68 to 0.84 per subtype, with best performance for triple-negative (AUC-ROC 0.80 to 0.81) [111]. This represents the most methodologically rigorous multicentric evaluation to date.

2.6.3 Key patterns and persistent limitations

Several consistent findings emerge across these studies. Triple-negative and HER2-enriched subtypes are generally easier to discriminate than Luminal A and B, with Luminal B consistently showing the lowest AUC-ROCs (0.68 to 0.72). Binary classification tasks (triple-negative versus non-triple-negative, Luminal versus non-Luminal) consistently achieve higher performance than four-class problems. Multiparametric approaches outperform single-sequence analysis. Dataset sizes remain modest (100 to 700 patients), and external validation is lacking in many studies. Only a handful of papers [106, 110, 111] validate on truly independent external datasets. The systematic review by Wang and Wei [112], covering radiomics

and ML based breast cancer subtype classification from MRI, corroborates these trends and confirms that the highest-performing recent systems operate on three-dimensional volumetric data.

2.7 Multicentric generalisation and domain adaptation

2.7.1 The domain shift problem in medical imaging

Performance degradation across scanners, protocols, and institutions is well documented. Zech et al. [113] showed that chest X-ray models learn site-specific confounders rather than disease features, with substantial cross-institutional performance drops. Guan and Liu [114] surveyed domain adaptation methods for medical imaging, categorising approaches into shallow and deep and supervised and unsupervised paradigms. For breast MRI specifically, Saha et al. [66] demonstrated that MRI scanner parameters directly affect radiomic features.

2.7.2 Domain adaptation techniques

Adversarial domain adaptation was formalised by Ganin et al. [115] through the Domain-Adversarial Neural Network with gradient reversal layer. Kamnitsas et al. [116] applied adversarial methods to brain lesion segmentation across different MRI scanners, achieving near-supervised performance.

Federated learning enables multi-institutional collaboration without data sharing, as demonstrated in Rieke et al. [117] comprehensive review. Flores et al. [118] showed federated learning across 20 global institutes improved average AUC-ROC by 16% and generalisability by 38% versus single-site models. Pati et al. [119] conducted the largest real-world federated learning study across 71 institutions, showing 33% improvement over public-trained models.

2.7.3 Multicentric breast cancer MRI resources

The MAMA-MIA dataset [120] represents the largest annotated multicentric breast cancer DCE-MRI collection: 1506 cases from four TCIA collections (I-SPY1, I-SPY2, NACT-Pilot, Duke-Breast-Cancer-MRI) with expert tumour segmentations by 16 specialists and harmonised clinical variables. This resource, along with pretrained nnU-Net weights, directly supports multicentric research combining the same datasets targeted by the thesis.

2.8 Addressing class imbalance in breast cancer subtype classification

2.8.1 The fundamental problem

Breast cancer molecular subtypes are inherently imbalanced: Luminal A dominates (approximately 60 to 70% of cases), while HER2-enriched (5 to 15%) and Triple-Negative (10 to 15%) subtypes are substantially rarer. He and Garcia [121] provided the comprehensive survey of imbalanced learning methods. Johnson and Khoshgoftaar [122] surveyed deep learning specifically under class imbalance.

2.8.2 Cost-sensitive loss functions

Focal loss [123] reshapes cross-entropy by down-weighting well-classified examples via a modulating factor $(1 - p_t)^\gamma$, focusing training on hard examples. Class-weighted cross-entropy assigns weights inversely proportional to class frequency, a straightforward approach adopted in several breast cancer subtype studies. Generalised Dice Loss [124] weights classes by inverse volume, proving robust for imbalance ratios up to 1:5000.

2.8.3 Synthetic oversampling

Synthetic Minority Over-sampling Technique (SMOTE) [125] remains foundational, creating synthetic minority samples by interpolation between nearest neighbours. Borderline-SMOTE [126] focuses oversampling on decision boundary examples. ADASYN [127] adaptively generates more synthetic data for harder-to-learn minority examples. GAN-based augmentation was pioneered by Frid-Adar et al. [128], who showed synthetic liver CT images improved CNN classification. Diffusion model-based augmentation represents the cutting edge: Zhu et al. [129] proposed a diffusion-based multi-scale feature fusion network specifically for imbalanced medical image classification.

2.9 Validation methodology to ensure clinical trustworthiness

2.9.1 Cross-validation best practices

For small medical imaging datasets, cross-validation choice critically affects reliability. Varoquaux [130] demonstrated that cross-validation on small neuroimaging datasets produces alarmingly large error bars

because folds are far from independent. Vabalas et al. [131] showed K-fold cross-validation produces biased performance estimates with small samples, recommending nested cross-validation for unbiased estimation. Varoquaux et al. [132] provided guidelines recommending repeated random splits over leave-one-out for neuroimaging.

2.9.2 External validation

Yu et al. [133] conducted a systematic review of 83 externally validated DL studies, finding that 81% showed performance decrease on external data and 49% showed AUC-ROC decreases ≥ 0.05. This underscores why multicentric external validation, as employed when aggregating the four TCIA collections (Duke-Breast-Cancer-MRI, I-SPY 1, I-SPY 2, and Breast-MRI-NACT-Pilot) into a unified corpus, is essential.

2.9.3 Reporting frameworks

Six reporting frameworks are directly relevant:

- **TRIPOD** [134]: twenty-two-item checklist for prediction model studies, with explanation and elaboration.
- **TRIPOD+AI** [135]: Updated twenty-seven-item checklist harmonised for regression and ML and Artificial Intelligence (AI) methods.
- **STARD 2015** [136]: thirty-item checklist for diagnostic accuracy studies.
- **PROBAST** [137]: Risk of bias assessment across four domains with 20 signalling questions.
- **CLAIM** [138]: AI-specific checklist for medical imaging studies.
- **CONSORT-AI** and **SPIRIT-AI** [139, 140]: Extensions for AI clinical trial reporting and protocols.

2.9.4 Statistical testing

The DeLong test [141] remains the standard nonparametric method for comparing correlated AUC-ROC. The corrected resampled t-test [142] accounts for dependency between cross-validation splits. Dietterich [143] showed standard paired t-tests with cross-validation have elevated Type I error, recommending McNemar's test for single train and test comparisons.

2.9.5 Reproducibility concerns

Roberts et al. [144] reviewed 2212 COVID-19 AI studies and found none were of potential clinical use due to methodological flaws. Nagendran et al. [145] found that only 9 of 81 AI-versus-clinician studies were prospective, code was unavailable in 93%, and data in 95%. Varoquaux and Cheplygina [146] reviewed widespread methodological failures including data leakage and insufficient external validation. These findings emphasise the importance of rigorous validation, open science practices, and adherence to reporting guidelines.

2.10 Summary and implications

This literature review reveals several key insights that directly motivate the thesis framework. First, the documented limitations of IHC-based classification (spatial heterogeneity, inter-laboratory variability, and biopsy discordance rates of 3 to 40% depending on marker) provide compelling clinical justification for non-invasive imaging alternatives. Second, DCE-MRI encodes molecular information: multiple studies demonstrate subtype-specific enhancement dynamics and peritumoural features, establishing the biological plausibility of imaging-based phenotype prediction. Third, while handcrafted radiomics established the radiogenomics paradigm, their reproducibility limitations and sensitivity to segmentation and scanner parameters motivate the transition to end-to-end deep learning.

The technical landscape reveals that three-dimensional CNNs outperform two-dimensional slice-based approaches by capturing inter-slice spatial context, but face challenges including limited pretrained models and higher computational demands. The best-performing recent systems [109, 111] achieve AUC-ROC of 0.68 to 0.97 depending on subtype and evaluation setting, with Luminal B classification remaining the most challenging and triple-negative typically the easiest to identify. Transfer learning is critical for cross-centre generalisation, as demonstrated by Zhang et al. [106], showing greater than 30% accuracy improvement. Class imbalance handling through weighted losses and synthetic oversampling is essential given the natural subtype distribution, and nested or repeated cross-validation with external validation on independent datasets represents the gold standard for demonstrating clinical viability. The emergence of large multicentric resources like MAMA-MIA (1506 cases across four TCIA collections) creates unprecedented opportunities for the type of multicentric validation that distinguishes clinically impactful work from single-institution proof-of-concept studies.

Chapter 3

Materials and Data Preparation

The development of robust Quantitative Imaging Biomarkers (QIBs) for breast cancer phenotype classification demands a rigorous and principled approach to data acquisition, curation, and preparation. Before any modelling decisions are made, it is essential to understand the characteristics, limitations, and complementary strengths of the available data sources. This chapter describes the datasets considered and selected for this work, motivates the choice of breast cancer as the primary case study, characterizes the phenotype distribution challenges inherent to the available data, and outlines the preprocessing strategy applied to construct a unified and analysis-ready corpus. The architectural and training methodology is preconditioned on the lesion detection pipeline described in [20] and is elaborated in Chapter 5.

3.1 Case Study: Breast Cancer

Breast cancer is the most commonly diagnosed cancer in women worldwide and one of the leading causes of cancer-related mortality, accounting for approximately 2.3 million new cases and 685000 deaths in 2020 alone [13, 14]. Despite significant advances in detection and treatment, the disease remains a major public health burden, partly because of its pronounced molecular heterogeneity. Current clinical practice classifies invasive breast tumours into four broad molecular subtypes: **Luminal A, Luminal B, HER2-enriched, and Triple-Negative**. These are defined by the expression status of Estrogen Receptor (ER), Progesterone Receptor (PR), and Human Epidermal growth factor Receptor 2 (HER2) [147]. Each subtype exhibits a distinct clinical behaviour, prognosis, and response to therapy, making accurate subtype identification a cornerstone of personalised oncology [148].

Conventionally, molecular subtyping relies on Immunohistochemistry (IHC) and genomic assays applied to biopsy specimens, a process that is invasive, costly, time-consuming, and subject to inter-laboratory variability. The growing availability of high-resolution Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI) data, combined with advances in Deep Learning (DL), has opened the possibility of

non-invasive, image-based subtype prediction. Breast cancer constitutes a particularly compelling case study for this research for three reasons. First, its high prevalence and clinical significance ensure that any improvement in diagnostic accuracy translates directly into measurable patient benefit. Second, and crucially, large annotated public datasets combining volumetric Magnetic Resonance Imaging (MRI) acquisitions with genomics-derived subtype labels exist and are openly accessible, a prerequisite for training and rigorously validating data-hungry DL models. Third, the heterogeneity of the disease (both within and across subtypes) creates a genuinely challenging classification problem that stresses the limits of existing approaches, making it an ideal benchmark for the proposed framework.

3.2 Data Sources and Selection Criteria

3.2.1 The Cancer Imaging Archive

The Cancer Imaging Archive (TCIA) [149] is the principal publicly accessible repository of cancer-related medical imaging data. It aggregates MRI, Computerized Tomography (CT), Positron Emission Tomography (PET), and pathology images from multi-institutional studies, and associates them with rich clinical, genomic, and radiological metadata. The archive follows the Digital Imaging and Communications in Medicine (DICOM) standard and enforces rigorous de-identification protocols, making it a reliable source for large-scale computational research.

A systematic survey of TCIA identified 25 collections containing breast imaging data. However, the development of phenotype classifiers imposes a strict requirement that is rarely satisfied: the dataset must supply not only imaging data but also genomics annotations from which molecular subtypes can be assigned. Only three of the 25 breast-imaging collections in TCIA meet this criterion; their principal characteristics are summarised in Table 1. Each of the three candidates provides MRI data (the modality of choice for soft-tissue tumour characterisation) together with clinical and genomic metadata, although they differ substantially in sample size, imaging modality coverage, and annotation richness.

The TCGA-BRCA collection, while valuable for radiogenomic correlative studies, contains only 139 subjects and provides heterogeneous modalities that include mammography and pathology slides alongside MRI, making standardisation and volumetric analysis significantly more complex. The CPTAC-BRCA collection is restricted to digital pathology images and offers no in-vivo imaging data, disqualifying it entirely from an MRI-based pipeline. Consequently, the Duke-Breast-Cancer-MRI dataset emerges as the only collection that simultaneously satisfies the requirements of sufficient sample size, volumetric MRI coverage, and molecular subtype annotations, and it is therefore adopted as the primary resource for model

Table 1: Comparison of the three breast cancer collections in the TCIA platform that jointly provide imaging and genomic data, which is the minimum requirement for molecular subtype classification.

Dataset	Subjects	Series / Images	Imaging Modalities	Supporting Data
Duke-Breast-Cancer-MRI	922	5161 / 773888	DCE-MRI, Segmentation	Clinical, Genomics, Image Analyses
TCGA-BRCA	139	1877 / 230167	MRI, Mammogram, Pathology	Clinical, Genomics, Image Analyses
CPTAC-BRCA	134	— / 642	Pathology	Clinical, Genomics, Proteomics

development.

3.2.2 Additional Dataset Sources Beyond TCIA

Beyond the three candidates initially identified within TCIA under the breast and genomics filter, a broader search of the archive and associated trial repositories reveals several collections that, while not providing precomputed subtype labels directly, include the receptor-status variables (ER, PR, HER2) necessary to derive them computationally. Three such collections were identified and incorporated into the multicentric repository described in Section 3.5: I-SPY 1 [54], I-SPY 2 [57], and Breast-MRI-NACT-Pilot [150]. A fourth collection, Advanced-MRI-Breast-Lesions [151], was considered but ultimately excluded from the final corpus owing to its non-standard acquisition protocols, as described in Section 3.3.5.

It should be noted that at the time of data acquisition for this work, these datasets were not yet fully integrated into the standard TCIA collection browser as dedicated, individually citable collection pages. These collections were instead accessible through the legacy TCIA wiki system and the original ACRIN trial data repositories. Both collections have since been formally incorporated into the current TCIA infrastructure as first-class collections with persistent identifiers and standardised metadata. Their integration into the present pipeline required a harmonisation step to derive consistent subtype labels from receptor status, as elaborated in Section 3.5.1.

3.2.3 The Portuguese Breast Cancer Digital Repository

The Breast Cancer Digital Repository (BCDR-FMR) [152] is a Portuguese dataset compiled from patient archives at the Faculty of Medicine of the University of Porto. It contains digitised film mammography images, expert manual lesion contours, and associated metadata encoded in the DICOM format. At the time of evaluation, it comprised 1010 cases and 3703 digitised mammographic images, with 795 segmented

lesions across mediolateral oblique (MLO) and craniocaudal (CC) projections, totalling 1493 individual segmentation masks covering masses, microcalcifications, stromal distortions, and axillary adenopathy. Although the BCDR-FMR is a high-quality resource for mammography-based lesion analysis and has been employed in prior feature extraction studies [153, 154], it provides neither molecular subtype labels nor volumetric MRI acquisitions. Both are indispensable for the present work, and the BCDR-FMR is therefore excluded from the training and validation pipeline.

3.3 Dataset Descriptions

3.3.1 Duke Breast Cancer MRI Dataset

The Duke-Breast-Cancer-MRI dataset [21, 53] is a retrospective, single-institution collection assembled at Duke University Medical Center over a decade of clinical practice. It comprises pre-operative DCE-MRI examinations from 922 patients with biopsy-confirmed invasive breast cancer, making it, at the time of this work, the largest publicly available annotated dataset of its kind. All images were acquired under clinical protocols, de-identified, and released through TCIA.

The dataset provides a rich multimodal annotation structure. Imaging data consist of multi-phase DCE-MRI volumes acquired with gadolinium-based contrast agents, spanning pre-contrast and multiple post-contrast time points. For each case, radiologist-annotated bounding boxes delineate the lesion region within the DCE-MRI volume, providing the spatial localisation consumed by the upstream detection module [20]. The clinical and genomic metadata include demographic variables (age, race), tumour characteristics (size, grade, histological type), treatment and outcome records, and molecular subtype labels derived from immunohistochemical assays, covering all four subtypes: Luminal A, Luminal B, HER2-enriched, and Triple-Negative. An additional set of 529 computer-extracted imaging features characterising tumour size, shape, texture, and enhancement kinetics is also provided, enabling radiogenomic correlative analyses beyond the scope of this thesis.

The breadth and completeness of this collection, its exclusive use of MRI, and the availability of expert spatial annotations make the Duke dataset the indispensable foundation for phenotype classifier development. It is used both as the primary training and internal validation resource and as one of the constituent collections of the multicentric repository.

3.3.2 I-SPY 1 / ACRIN 6657

The I-SPY 1 dataset [54] originates from the ACRIN 6657 trial, a multi-centre longitudinal study designed to evaluate MRI as a predictor of response to neoadjuvant chemotherapy (NACT). It comprises DCE-MRI examinations from 212 patients, acquired at multiple time points throughout chemotherapy, which makes it particularly informative for capturing temporal tumour dynamics. Receptor-status data (ER, PR, HER2) are available in the accompanying clinical metadata, permitting subtype derivation through the harmonisation procedure described in Section 3.5.1. The longitudinal structure of I-SPY 1 introduces variability in acquisition time points; only the pre-treatment baseline MRI volumes are used in this work to ensure comparability with the Duke dataset and consistency within the classification pipeline.

Since the conclusion of the data acquisition phase of this work, an expert tumour segmentation dataset for I-SPY 1 has been released on TCIA as the ISPY1-tumour-SEG-Radiomics collection [58]. It provides board-certified radiologist-validated structural tumour volume delineations for 163 subjects, generated by a deep learning model and subsequently reviewed and corrected by a fellowship-trained breast radiologist. These segmentations were not available at the time of data collection and were therefore not incorporated into this work; lesion localisation for the I-SPY 1 cases is instead provided by the upstream detection module, as described in Section 3.7.1. The availability of this resource in future extensions of the framework is noted in Chapter 6.

3.3.3 I-SPY 2

The I-SPY 2 trial [57] extends the I-SPY 1 design to a larger and more heterogeneous cohort of 985 patients receiving experimental neoadjuvant therapies across multiple institutions. The dataset provides multi-phase DCE-MRI volumes, diffusion-weighted imaging sequences, and comprehensive genomic profiling alongside the clinical metadata required for subtype assignment. Its large size is the principal contribution to the multicentric repository: I-SPY 2 supplies 985 additional cases and significantly increases the representation of minority subtypes, particularly HER2-enriched and Triple-Negative, thereby directly addressing the imbalance problem discussed in Section 3.4. As with I-SPY 1, only baseline pre-treatment acquisitions are considered.

Subsequent to the data collection phase of this work, an analysis-result dataset (BreastDCEDL_ISPY2) providing tumour segmentation masks for 982 patients in a deep-learning-ready format, together with predefined train, validation, and test splits, was released on TCIA. As with the I-SPY 1 segmentations described above, this resource postdates data acquisition and was not incorporated into the present pipeline; its rel-

evance for future iterations of the framework is acknowledged accordingly.

3.3.4 Breast-MRI-NACT-Pilot

The Breast-MRI-NACT-Pilot collection [150] is a single-centre pilot dataset comprising DCE-MRI examinations from 46 patients undergoing NACT. Although modest in size, the collection provides expert lesion segmentations and receptor-status records, and its inclusion is motivated by two considerations. First, it contributes cases corresponding to rarer subtypes that are systematically underrepresented in larger collections. Second, it originates from a different imaging protocol than the Duke and I-SPY datasets, introducing scanner diversity that improves the robustness of the trained models to acquisition variability. Baseline pre-treatment volumes are selected for inclusion.

3.3.5 Advanced-MRI-Breast-Lesions

The Advanced-MRI-Breast-Lesions collection [151] focuses on characterising breast lesions using advanced MRI techniques, encompassing sequences beyond standard DCE-MRI and providing imaging data from 59 patients. Its acquisition protocols differ from those of the other collections in that they include delayed contrast phases and specialised pulse sequences designed to accentuate tissue diffusivity and perfusion. This diversity of imaging characteristics may expose the classification model to patterns not encountered in conventional DCE-MRI protocols. However, these non-standard acquisition protocols introduce a pronounced distributional mismatch relative to the clinical DCE-MRI series used for both training and external validation in this work; the collection was therefore excluded from the final multicentric corpus.

3.4 Phenotype Distribution and Class Imbalance

A critical characteristic of breast cancer datasets that directly governs the modelling strategy is the inherent imbalance in the distribution of molecular subtypes. As illustrated in Figure 2, Luminal A is the dominant subtype in the Duke dataset, comprising approximately 65% of cases (595 out of 922). Luminal B accounts for approximately 11% (101 cases), Triple-Negative for 18% (167 cases), and HER2-enriched for only 6% (59 cases). This distribution reflects the true epidemiological prevalence of breast cancer subtypes in the general population [147], but it creates a significant challenge for supervised learning: models trained on such skewed data exhibit a systematic bias towards the majority class and tend to perform poorly on minority subtypes, precisely those for which diagnostic difficulty is greatest and clinical consequences of misclassification most severe.

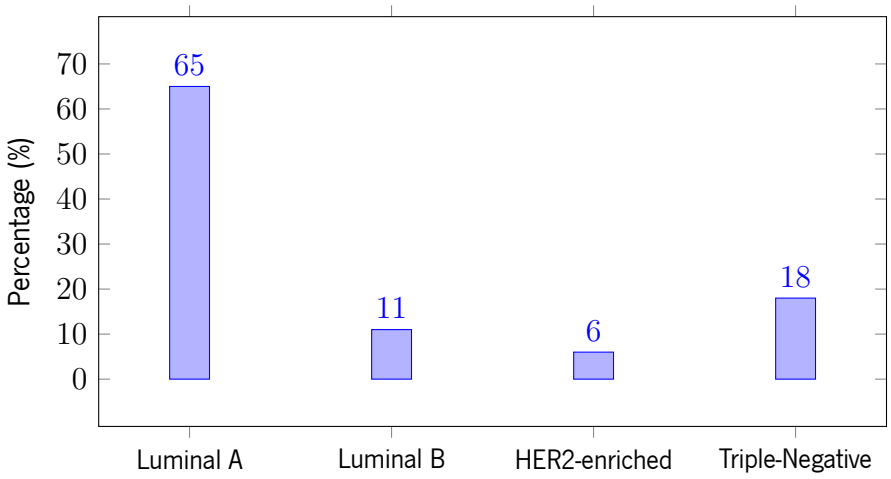


Figure 2: Phenotype distribution across the 922 cases of the Duke-Breast-Cancer-MRI dataset. Luminal A is substantially over-represented relative to HER2-enriched, which constitutes only 6% of the collection.

The multicentric repository constructed by aggregating the four TCIA collections partially ameliorates this imbalance. The combined corpus of 2114 cases yields 1064 Luminal A, 259 Luminal B, 213 HER2-enriched, and 578 Triple-Negative cases, representing a substantially more equitable distribution than the Duke dataset alone. Figure 3 compares the per-subtype distribution across the four individual collections and the combined corpus. The neoadjuvant trial collections (I-SPY 1, I-SPY 2, and Breast-MRI-NACT-Pilot) exhibit substantially higher proportions of aggressive subtypes relative to the Duke dataset, particularly HER2-enriched and Triple-Negative, owing to the deliberate enrolment of high-risk patients in those trials. This distributional heterogeneity across sources is beneficial in principle, as it broadens the phenotypic coverage of the training corpus. However, HER2-enriched remains the least represented subtype even in the combined collection, and targeted strategies, including class-weighted sampling and data augmentation, are necessary to mitigate residual imbalance. These strategies are detailed in the preprocessing pipeline described in Section 3.7 and in the approach chapter.

3.5 Multicentric Repository Construction

3.5.1 Dataset Integration

Assembling a unified analytical corpus from four independently released TCIA collections introduces substantial challenges related to format heterogeneity, metadata inconsistency, and subtype label availability [155]. Each collection distributes its clinical metadata as DICOM-associated files or in spreadsheet formats (Microsoft Excel or Comma-Separated Values (CSV)), with no standardised schema across collections. Column names, value scales, and coding conventions differ materially between sources.

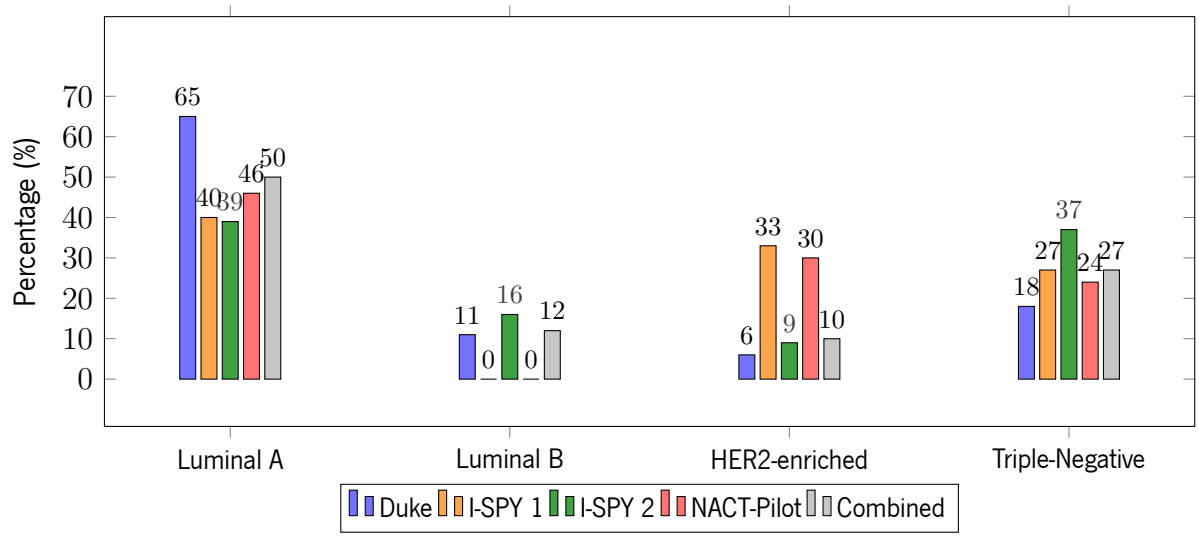


Figure 3: Per-subtype phenotype distribution (as percentage of cases within each collection) across the four TCIA collections used for multicentric training and the resulting combined corpus of 2114 cases. The Duke dataset is dominated by Luminal A (65%), whereas the neoadjuvant trial collections (I-SPY 1, I-SPY 2, and Breast-MRI-NACT-Pilot) present a markedly higher proportion of aggressive subtypes, reflecting their high-risk patient enrolment criteria. Aggregation partially reduces Luminal A dominance (from 65% to 50%) and increases the representation of minority subtypes, though HER2-enriched remains the least represented class. Luminal B is absent from I-SPY 1 and NACT-Pilot due to the absence of Ki-67 data in those collections, precluding the Luminal A versus Luminal B subclassification.

The most consequential inconsistency concerns subtype labelling. The Duke dataset provides explicit molecular subtype assignments for all cases, whereas the four supplementary collections report only raw receptor-status values: binary or ordinal fields indicating ER positivity, PR positivity, HER2 positivity or amplification, and, where available, the Ki-67 proliferation index. A systematic rule-based harmonisation procedure was developed to derive subtype labels from receptor status, following established clinical criteria [147]: cases with ER-positive and/or PR-positive status and HER2-negative expression and low Ki-67 are assigned to Luminal A; ER-positive and/or PR-positive cases with HER2-negative expression but high Ki-67, or any ER/PR-positive and HER2-positive case, are assigned to Luminal B; HER2-positive and ER-negative and PR-negative cases are classified as HER2-enriched; and cases negative for all three receptors are classified as Triple-Negative. Where Ki-67 data were unavailable, the Luminal A versus Luminal B boundary was determined from the HER2 status alone following the simpler St. Gallen criteria, accepting the associated degree of ambiguity as a known limitation. In practice, this limitation is most consequential for I-SPY 1 and Breast-MRI-NACT-Pilot, neither of which provides Ki-67 measurements in their released metadata. As a result, all hormone-receptor-positive and HER2-negative cases in those collections are collapsed into Luminal A under the St. Gallen fallback, producing zero Luminal B labels from either source (as visible in Figure 3). Some of those cases may clinically be Luminal B based on Ki-67 status; their misassignment to Luminal A constitutes a known label-quality limitation of the combined corpus and is discussed as a contributing factor to the Luminal B performance degradation observed under multicentric

training in Section 6.2.4.

3.5.2 Prospective Clinical Cohort: Unidade Local de Saúde da Arrábida (ULSA), Hospital São Bernardo, Setúbal

In addition to the public TCIA data, an independent prospective cohort was assembled at ULSA, Hospital São Bernardo, Centro Hospitalar de Setúbal, Portugal, under institutional ethical approval. This cohort provides real-world DCE-MRI examinations acquired on clinical hardware outside the controlled setting of academic trial protocols, introducing scanner characteristics, contrast timing protocols, and patient demographics distinct from those of the TCIA collections. Molecular subtype labels were established from histopathological biopsy reports and clinical receptor-status assays, following the same classification criteria applied during the harmonisation of the TCIA supplementary collections. The Setúbal cohort is reserved exclusively for external validation and is not used during training or hyperparameter optimisation; it serves to assess the generalisability of the trained models to a previously unseen clinical environment.

3.6 External Validation on a Prospective Clinical Cohort

Internal evaluation of the framework on held-out splits of the TCIA corpus, however informative, cannot alone substantiate the clinical utility of the resulting QIBs. A model that performs well on data drawn from the same statistical distribution as its training set may still fail when deployed in a genuinely novel clinical environment. This section describes the prospective cohort assembled at ULSA, Hospital São Bernardo, Setúbal, Portugal, which serves as the external validation set for the framework, and details the validation protocol applied to it. All numerical results are reported and discussed in Chapter 6.

3.6.1 The Necessity of External Validation in Clinical Artificial Intelligence

Internal validation, whether by a held-out test split or by cross-validation, evaluates a model's ability to generalise within the data-generating distribution represented by the training corpus. For a DL model trained and internally evaluated on TCIA datasets, that distribution reflects the imaging equipment, contrast protocols, patient demographics, and technical practices of the North American academic medical centres and clinical trial consortia that contributed to those collections. Strong internal performance therefore demonstrates that the model has learned predictive signal from the training distribution, but it does not demonstrate that the learned signal is transportable to a radically different clinical environment.

External validation on a truly independent cohort, collected at a distinct institution under routine clinical conditions rather than research protocols, is the minimum evidentiary standard required before any imaging-based classifier can be considered a credible QIB candidate. The TRIPOD+AI reporting framework [135] explicitly requires that prediction models using Artificial Intelligence (AI) be validated on datasets that are temporally, geographically, or institutionally separate from the development set, and mandates transparent reporting of the cohort characteristics, the preprocessing applied, and the precise evaluation protocol. The external validation conducted in this work is designed to satisfy these requirements.

Beyond the regulatory and reporting considerations, external validation in this context addresses a specific scientific question: whether the intensity normalisation strategy, spatial resampling, and volumetric Convolutional Neural Network (CNN) architecture described in Chapter 5 are sufficiently robust to bridge the domain gap between a large, well-curated North American academic dataset and a modest, prospectively collected cohort from a Portuguese public hospital. This gap encompasses differences in scanner hardware, field strength, contrast agent timing protocols, patient demographics, and the absence of any ground-truth lesion annotations, meaning that bounding boxes must be inferred by the upstream detection module [20] rather than drawn from radiologist-provided segmentations.

3.6.2 The ULSA Setúbal Cohort

Institutional Context and Data Governance

The external validation cohort was assembled at ULSA, Hospital São Bernardo, Setúbal, Portugal, a public tertiary referral centre serving the greater Setúbal region. Data collection was conducted prospectively under institutional ethical approval and in full compliance with the General Data Protection Regulation (GDPR) [22].

Prior to any computational processing, all imaging data and associated clinical records were de-identified through a structured anonymisation pipeline. Patient identifiers, acquisition metadata fields containing personal information (including scanner institution tags embedded in the DICOM headers), and any free-text fields in the clinical reports were removed. The anonymised DICOM volumes and clinical annotations were subsequently transferred to the research infrastructure under a data custody agreement between ULSA and the research team. This close collaboration with clinical specialists was instrumental in ensuring data quality and in establishing the ground-truth molecular subtype labels used for evaluation.

Cohort Composition and Molecular Subtype Labels

A total of 160 cases were initially collected from the ULSA archives, comprising patients with biopsy-confirmed invasive breast cancer who had undergone clinical DCE-MRI examination as part of their routine diagnostic workup. Of these, 98 cases satisfied all quality control criteria and were retained for external validation. Cases were excluded from the analysis if the upstream DCE-MRI volume contained an insufficient number of slices to construct a valid three-dimensional volume at the target resolution, if the DICOM series could not be unambiguously reconstructed due to metadata inconsistencies, or if the associated clinical record lacked a definitive molecular subtype assignment derived from IHC assay.

Molecular subtype labels were established from histopathological biopsy reports and receptor-status assays (ER, PR, HER2) performed at the institution, following the same St. Gallen classification criteria applied during the harmonisation of the TCIA supplementary collections described in Section 3.5.1. Table 2 summarises the subtype distribution across the 98 analysable cases.

Table 2: Molecular subtype distribution across the 98 analysable cases of the ULSA Setúbal cohort. The distribution reflects the clinical prevalence at this institution and differs markedly from the Duke training distribution, in which Luminal A accounts for approximately 65% of cases and HER2-enriched for only 6%.

Molecular Subtype	**Cases (n)**	**Proportion (%)**
Luminal A	37	37.76
Luminal B	34	34.69
HER2-enriched	9	9.18
Triple-Negative	18	18.37
Total	**98**	**100.0**

The distribution at Setúbal is notably more balanced than that of the Duke training set, with Luminal A comprising approximately 38% of cases compared to 65% in Duke. Conversely, Luminal B is substantially over-represented relative to Duke (35% versus 11%), and Triple-Negative is proportionally comparable (18% versus 18%). HER2-enriched remains the minority subtype in both cohorts, though its absolute count of nine cases is the primary limitation on the statistical power of the per-subtype evaluation.

This distributional shift is itself informative: it tests the classifier's behaviour under a different prevalence regime than that encountered during training, providing a more stringent and ecologically valid assessment of its calibration and generalisation.

Imaging Acquisition Characteristics

The Setúbal cohort was acquired under routine clinical DCE-MRI protocols at ULSA, rather than under the standardised acquisition parameters mandated by research trial protocols such as those governing the TCIA collections. As a consequence, acquisition parameters were not uniformly recorded across all cases, and the cohort encompasses a degree of heterogeneity in field strength, coil configuration, and contrast timing that is representative of real-world clinical practice.

Examinations were performed on clinical MRI systems using gadolinium-based contrast agents, acquiring multi-phase DCE-MRI sequences that include a pre-contrast volume and at least one post-contrast phase. This protocol is consistent with international guidelines for breast MRI acquisition [47]. The inherent variability in scanner hardware, field strength, and contrast timing across the cohort is not a limitation to be minimised, as it is precisely the source of the domain gap that external validation is designed to probe. A model that achieves clinically acceptable performance across this variability demonstrates a degree of robustness that cannot be inferred from any single-site or single-protocol internal evaluation.

3.6.3 Preprocessing of the Setúbal Cohort

The Setúbal cases are processed through the identical preprocessing pipeline described in Section 3.7, with no modifications. The pipeline is applied in inference mode, meaning that no data augmentation is performed and no preprocessing parameters are estimated from or tuned on the Setúbal data. This strict no-touch policy is essential to the integrity of the external validation: any adaptation of preprocessing parameters to the validation set would constitute a form of data leakage and would invalidate the independence of the evaluation.

The critical preprocessing decision specific to the Setúbal cohort concerns the provenance of the lesion bounding boxes. No radiologist-provided annotations are available for these cases, which is the typical situation in routine clinical practice outside specialised research protocols. Bounding boxes are therefore inferred by the upstream Faster R-CNN detection module [20], following the same procedure applied to the non-Duke TCIA collections in Section 3.7.1. The predicted bounding box for each case is used to crop the axial slice sequence at runtime, the cropped slices are stacked into a three-dimensional volume, and the volume is resampled to the target dimension $(D, W, H) = (64, 128, 128)$ by trilinear interpolation (Equations 3.2–3.3). Intensity normalisation is applied using the percentile-based scheme described in Section 3.7.3: the 1st and 99th percentile of each volume's voxel distribution are computed (Equation 3.6), the volume is clipped to $[p_1, p_{99}]$ (Equation 3.7), and the clipped intensities are rescaled

to the unit interval (Equation 3.8). The per-volume adaptive nature of this scheme ensures robustness to cross-site intensity scale variation, with no external calibration constants required.

The reliance on detection-inferred bounding boxes introduces the same potential for localisation error propagation discussed in Section 3.7.1. This risk is arguably higher for the Setúbal cohort than for the TCIA collections, because the detection model was trained exclusively on the Duke dataset and has not been exposed to imaging data from Portuguese clinical scanners. Any systematic domain shift in the appearance of lesion boundaries or surrounding tissue between Duke and Setúbal may manifest as a degradation in detection precision, which would in turn corrupt a fraction of the cropped volumes presented to the classifier. The magnitude of this effect is implicitly captured by the external validation metrics reported in Chapter 6: any performance gap relative to the internal evaluation on Duke ground-truth bounding boxes is attributable to the compounded effect of domain shift in both the detection and classification stages.

3.6.4 External Validation Protocol

Strict Independence of the Validation Set

The ULSA Setúbal cohort was maintained in strict isolation throughout the entire model development process. It was not used during training, hyperparameter optimisation, architecture selection, or any intermediate evaluation that could have influenced any modelling decision. The classifier weights applied during external validation are precisely those selected by the early-stopping criterion on the internal validation split of the Duke dataset, as described in Section 5.5. No threshold calibration, fine-tuning, or any other form of adaptation to the Setúbal distribution was performed prior to or during evaluation. This no-touch protocol ensures that the external validation provides an unbiased estimate of the generalisation performance of the framework as it would be deployed in an unseen clinical environment [135].

Inference Procedure

For each of the 98 Setúbal cases, the preprocessed three-dimensional volume is passed independently through each of the four binary One-vs-All (OvA) classifiers trained during the approach described in Chapter 5. Classifier k produces a scalar score $f_k(\mathbf{x}) \in [0, 1]$ estimating the probability that the input volume belongs to subtype k under the binary positive-versus-rest framing. The final four-class prediction is obtained by the arg max rule defined in Equation 5.1:

$$\hat{y} = \underset{k \in \{0,1,2,3\}}{\arg\max} \; f_k(\mathbf{x}) \tag{3.1}$$

Both the per-classifier probability scores and the final predicted labels are recorded for every case, enabling evaluation at both the binary (per-subtype) and multi-class levels.

Evaluation Metrics

Consistent with the internal evaluation described in Chapter 6, classifier performance on the Setúbal cohort is assessed using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), accuracy, precision, recall, and F_1-score for each of the four OvA classifiers. The AUC-ROC is the primary reporting metric because it is threshold-independent and provides a summary of discriminative performance across all operating points, which is particularly important for the HER2-enriched classifier where the small number of positive cases (nine) makes point estimates of precision and recall highly sensitive to individual misclassifications. Confusion matrices are also computed for each classifier to characterise the error structure.

No statistical significance testing between internal and external performance metrics is conducted, consistent with the sample size constraints of the Setúbal cohort and in line with the recommendation in TRIPOD+AI [135] that external validation studies be reported descriptively when cohort size precludes reliable inference. Any apparent performance differences between the two evaluation settings are instead discussed qualitatively in terms of the domain gap factors identified in this section. All numerical outcomes are presented in Chapter 6.

3.7 Data Preprocessing

Effective preprocessing is a prerequisite for reliable DL model training, particularly in the context of multi-site MRI data where scanner variability, slice thickness heterogeneity, and contrast timing differences can introduce systematic confounds. The preprocessing pipeline adopted in this work addresses two primary concerns: spatial standardisation and intensity normalisation. Augmentation strategies and class-imbalance handling are methodology-level decisions and are therefore described in Chapter 5.

3.7.1 Input Representation and Bounding Box Provenance

All datasets are ingested as ordered sequences of two-dimensional images derived from the original DICOM acquisitions, with each image representing a single axial slice of the DCE-MRI volume. The critical distinction between datasets lies in the provenance of the lesion bounding boxes used to isolate the tumour region within each slice. For the Duke-Breast-Cancer-MRI dataset, expert radiologist annotations are avail-

able for every case [21], and the slices were produced from DICOM already cropped to those ground-truth bounding boxes. For all remaining collections (I-SPY 1, I-SPY 2, and Breast-MRI-NACT-Pilot) no radiologist bounding box annotations are publicly available. The lesion bounding boxes for these cases are inferred by the upstream Faster R-CNN detection module [20], whose predicted coordinates are stored in per-dataset detection export CSV files. During volume construction, each full-resolution slice is loaded and cropped at runtime to the detector-predicted bounding box before assembly.

Implications of detection-inferred bounding boxes.

The reliance on predicted bounding boxes for the non-Duke collections introduces a potential source of error propagation. A localisation error, whether a regression offset or a substantially incorrect prediction, will produce a cropped region that is spatially shifted relative to the true tumour, may include excessive background tissue, or in extreme cases may not encompass the lesion at all. Unlike annotation errors correctable by a human expert, detection errors manifest as silent corruptions of the input volume.

This risk is bounded in two ways. First, the detection module achieves high mean average precision on the Duke dataset [20], making systematic large-scale failures unlikely. Second, the phenotype classifier is evaluated both on Duke in isolation, where ground-truth bounding boxes eliminate detection-induced noise, and on the full multicentric corpus, where detector-derived bounding boxes are used for the non-Duke cases. This dual strategy quantifies any performance degradation attributable to bounding box inference errors. Results of both evaluations are reported in Chapter 6.

Modular dataset architecture.

The dataset reading layer is designed as an extensible plugin system, reflecting the broader framework requirement that the pipeline should not be coupled to any single data source. A common abstract base class, `Dataset`, defines the shared interface (volume indexing, normalisation, resizing, and transformation application) and declares `read_volume` as an abstract method that every concrete subclass must implement. Each dataset group is then encapsulated in its own subclass: `DukeDataset` handles pre-cropped slices with phenotype labels read from a companion CSV, while `BCCollectionDataset` handles full-resolution slices with runtime bounding box cropping driven by a detection export file. The data generator accepts an arbitrary list of `Dataset` instances, iterating over all of them transparently within a single training loop. Integrating a new data source therefore requires only implementing the `read_volume` method for that source's storage format and annotation convention; no changes to the generator, the normalisation logic, or the training procedure are necessary. This architecture directly sup-

ports the framework's objective of generalising the QIB discovery pipeline beyond breast cancer, as new imaging modalities or institutional repositories can be incorporated as independent dataset plugins without disturbing the existing codebase.

3.7.2 Volume Construction and Spatial Standardisation

Following bounding box application, the slices for each case are sorted by their acquisition index and stacked into a single three-dimensional array of shape (D_0, W_0, H_0), where D_0 is the number of available slices and W_0, H_0 are the in-plane dimensions of the cropped region. The volume is then resampled to the fixed target dimension $(D, W, H) = (64, 128, 128)$ required by the classification network.

Resampling is performed by computing per-axis scaling factors:

$$s_D = \frac{D}{D_0}, \quad s_W = \frac{W}{W_0}, \quad s_H = \frac{H}{H_0}, \tag{3.2}$$

and applying trilinear interpolation via `scipy.ndimage.zoom` with order 1. For a target voxel at continuous source coordinates (z, x, y), the interpolated value is computed as a weighted combination of the eight nearest source voxels. Denoting $z_0 = \lfloor z \rfloor$, $x_0 = \lfloor x \rfloor$, $y_0 = \lfloor y \rfloor$ and the fractional residuals $\delta_z = z - z_0$, $\delta_x = x - x_0$, $\delta_y = y - y_0$, the trilinear estimate is:

$$\begin{aligned}
\hat{v}(z,x,y) \;&= (1-\delta_z)(1-\delta_x)(1-\delta_y)\, v_{z_0,x_0,y_0} \\
&+ (1-\delta_z)(1-\delta_x)\,\delta_y\, v_{z_0,x_0,y_0+1} \\
&+ (1-\delta_z)\,\delta_x(1-\delta_y)\, v_{z_0,x_0+1,y_0} \\
&+ (1-\delta_z)\,\delta_x\,\delta_y\, v_{z_0,x_0+1,y_0+1} \\
&+ \delta_z(1-\delta_x)(1-\delta_y)\, v_{z_0+1,x_0,y_0} \\
&+ \delta_z(1-\delta_x)\,\delta_y\, v_{z_0+1,x_0,y_0+1} \\
&+ \delta_z\,\delta_x(1-\delta_y)\, v_{z_0+1,x_0+1,y_0} \\
&+ \delta_z\,\delta_x\,\delta_y\, v_{z_0+1,x_0+1,y_0+1},
\end{aligned} \tag{3.3}$$

where $v_{i,j,k}$ denotes the intensity at source voxel (i, j, k). This formulation ensures C^0 continuity across the resampled volume while avoiding the ringing artefacts associated with higher-order spline kernels, and is computationally efficient for the input sizes encountered in this work.

3.7.3 Intensity Normalisation

The DICOM volumes are exported to Tagged Image File Format (TIFF) images prior to ingestion, preserving the full dynamic range of the original acquisition without the lossy compression artefacts introduced by other formats. Raw pixel intensities are then normalised in two steps applied uniformly to all voxels of a given volume: an outlier-suppression clipping step, followed by a unit-interval rescaling step. The specific implementation of the clipping step evolved iteratively as the training corpus expanded from a single site to a four-site multicentric collection, and both versions are documented here as part of the development record.

Initial approach: globally fixed window.

In the initial implementation, applied to the single-site Duke-Breast-Cancer-MRI dataset and reported in Fontes et al. [19], intensities were constrained to a globally fixed window $[v_{\text{min}}, v_{\text{max}}]$ via the piecewise mapping:

$$v' = \begin{cases} v_{\text{min}} & \text{if } v < v_{\text{min}}, \\ v & \text{if } v_{\text{min}} \leq v \leq v_{\text{max}}, \\ v_{\text{max}} & \text{if } v > v_{\text{max}}, \end{cases} \tag{3.4}$$

with $v_{\text{min}} = -1000$ and $v_{\text{max}} = 400$. Although these bounds are numerically coincident with the air–soft-tissue boundaries of the Hounsfield Unit scale used in computed tomography, they were applied here as empirically determined thresholds informed by that prior domain knowledge and subsequently validated on the Duke data [156]. The clipped intensities were then rescaled to the unit interval by:

$$v_{\text{norm}} = \frac{v' - v_{\text{min}}}{v_{\text{max}} - v_{\text{min}}}, \tag{3.5}$$

In implementation terms, Equation 3.4 corresponds directly to `np.clip(volume, -1000, 400)`, and Equation 3.5 to element-wise subtraction and division on the resulting array. This approach is simple and fully reproducible with a single pair of scalar constants.

The principal limitation of the fixed-window scheme, which became apparent during the transition to the multicentric TCIA corpus, is that the chosen constants are implicitly calibrated to the Duke acquisition environment. MRI signal intensities carry no absolute physical scale. Unlike computed tomography Hounsfield Units, they depend on scanner hardware, field strength, and receive-coil configuration in a manufacturer- and site-specific manner. A window boundary that correctly suppresses background and

artefact voxels on one acquisition platform may underclip or overclip intensities from a different scanner, introducing a systematic preprocessing bias that varies by site and is indistinguishable from biologically meaningful signal variation.

Refined approach: per-volume percentile-based clipping.

To address this limitation, the final multicentric framework adopts a data-driven clipping scheme in which the bounds are derived from each volume's own intensity distribution. The 1st and 99th percentile are first computed:

$$p_1 = \text{Percentile}(\mathbf{V},\, 1), \qquad p_{99} = \text{Percentile}(\mathbf{V},\, 99), \tag{3.6}$$

where $\mathbf{V} = \{v_{i,j,k}\}$ denotes the set of all voxel intensities in the cropped volume. In practice, this is computed as `p1, p99 = np.percentile(volume, [1, 99])`, which evaluates both quantiles in a single pass over the array. Intensities are subsequently constrained to $[p_1,\, p_{99}]$ via the piecewise clip:

$$v' = \begin{cases} p_1 & \text{if } v < p_1, \\ v & \text{if } p_1 \le v \le p_{99}, \\ p_{99} & \text{if } v > p_{99}, \end{cases} \tag{3.7}$$

implemented as `np.clip(volume, p1, p99)`. By discarding the most extreme 1% of intensities at each tail, this operation suppresses background regions, zero-padding artefacts, and signal saturation events whilst retaining 98% of the clinically relevant dynamic range, adapting automatically to the distribution of each individual acquisition. The clipped intensities are then rescaled to the unit interval by:

$$v_{\text{norm}} = \begin{cases} \dfrac{v' - p_1}{p_{99} - p_1} & \text{if } p_{99} - p_1 > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.8}$$

providing a bounded, numerically stable input to the network. The conditional structure handles the degenerate case $p_{99} = p_1$ (a zero-variance volume, corresponding to an acquisition failure or an entirely uniform crop) by returning `np.zeros_like(volume)`, effectively zeroing the volume rather than producing a division-by-zero artefact. In practice, this guard ensures that pathological acquisitions do not produce undefined activations during training.

Rationale for the refined scheme.

Because p_1 and p_{99} are derived from the volume itself rather than from a global prior, the scheme is intensity-range adaptive and does not require external calibration: the same procedure produces commensurable unit-interval representations regardless of whether the input was acquired on a 1.5 T or 3.0 T system, using surface or phased-array coils, or under different contrast timing protocols. This adaptivity is robust to cross-site variations in intensity scale and offset, though it does not address site-specific differences in tissue contrast ratios, spatial resolution, or temporal sampling density. This adaptivity is particularly valuable for the multicentric TCIA corpus, where no shared intensity reference is available across the four contributing institutions.

Among the alternative normalisation strategies considered: Z-score normalisation to zero mean and unit variance shares the per-volume adaptivity but does not bound the output range, leaving the network exposed to unbounded input activations under extreme outliers. Z-score statistics are also sensitive to the proportion of background and non-tumour tissue within the bounding box crop, whereas the 1st and 99th percentile are inherently robust to this composition effect by construction. Histogram landmark normalisation [61] offers superior inter-site commensurability but requires learning representative histogram landmarks from a sufficient number of per-site training images, and introduces inference-time dependencies on site labels not guaranteed in routine clinical deployment. N4ITK bias field correction [60] and ComBat harmonisation [62, 63] address orthogonal sources of variability (receive-coil inhomogeneity and inter-site batch effects, respectively) rather than the fundamental scale indeterminacy of raw MRI intensities, and are not substitutes for intensity range normalisation.

The volumetric resize operation described in Section 3.7.1 is implemented using `scipy.ndimage.zoom` with bilinear interpolation (`order=1`), consistent with Equations 3.2–3.3. Together, the NumPy-based normalisation and SciPy-based resampling form a minimal, dependency-light preprocessing stack that is straightforward to reproduce and to deploy at inference time without site-specific calibration.

3.7.4 Augmentation by Discrete In-Plane Rotation

To increase the effective training set size and to encourage rotational invariance in the learned representations, a deterministic augmentation strategy is applied at dataset construction time. For each original volume, seven additional copies are generated by rotating every constituent slice by fixed angles $\theta \in \{45°, 90°, 135°, 180°, 225°, 270°, 315°\}$. Each rotation is applied independently to each slice k via the two-dimensional rotation operator:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x - c_x \\ y - c_y \end{pmatrix} + \begin{pmatrix} c_x \\ c_y \end{pmatrix}, \tag{3.9}$$

where (x, y) are the original pixel coordinates, (c_x, c_y) is the image centre, and (x', y') are the transformed coordinates. The rotation is implemented via `scipy.ndimage.rotate` with `reshape=False`, preserving the spatial dimensions of each slice. Pixels that fall outside the original image boundary after transformation are filled with zero. Each rotated variant is registered as a distinct entry in the volume index alongside its phenotype label, yielding eight variants per original case (the original and its seven rotations) for a total augmentation factor of eight. Translational augmentation is not applied: since each volume is already spatially anchored to the tumour via the bounding box, introducing random translations would shift the lesion away from the centre of the input field, degrading the spatial correspondence that the network is trained to exploit.

All augmentation is deterministic and pre-registered at initialisation, i.e., no stochastic transformations are applied at runtime, ensuring full reproducibility when combined with a fixed random seed used throughout the pipeline. A schematic overview of the preprocessing flow is presented in Figure 4.

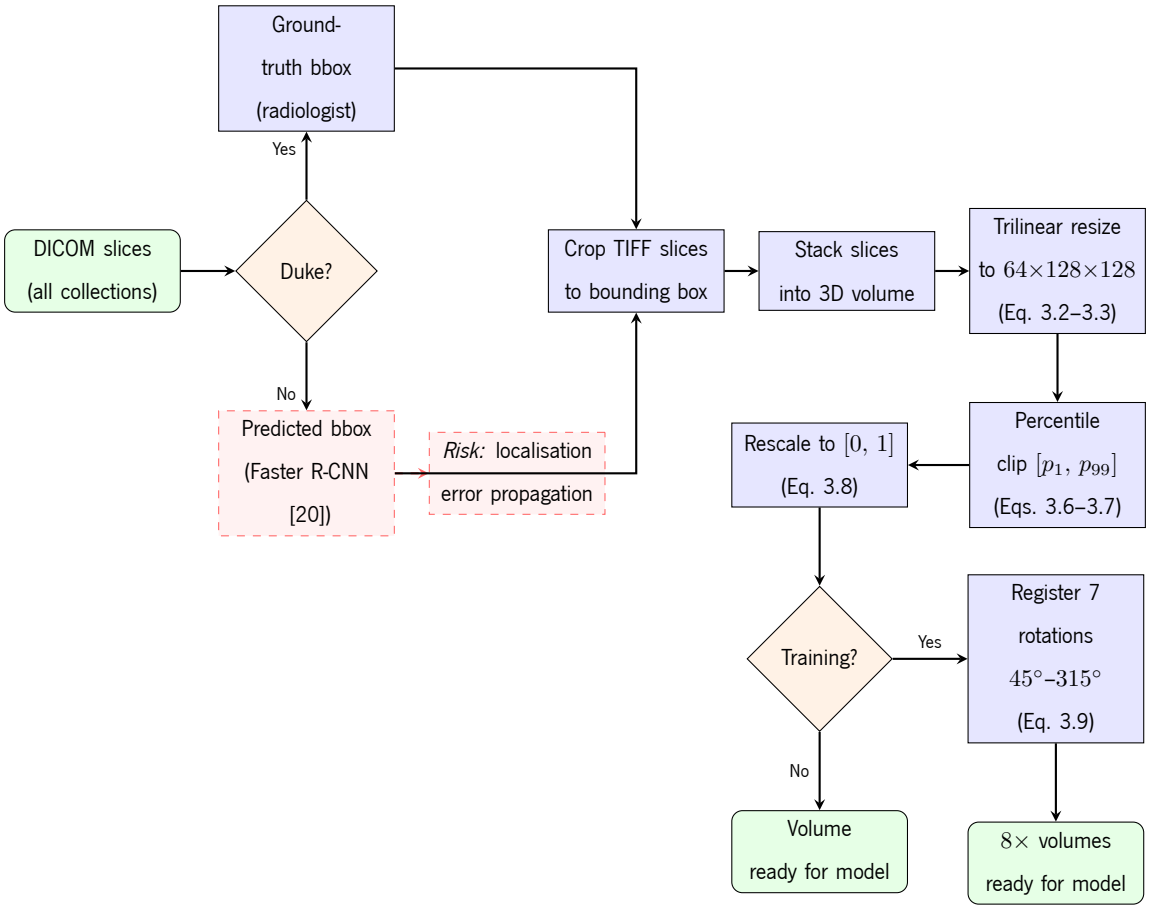


Figure 4: Overview of the data preprocessing pipeline. For Duke, slices are pre-cropped using ground-truth radiologist bounding boxes. For all other collections, bounding boxes are predicted by the upstream detection module [20], introducing the localisation error propagation risk indicated in red. All volumes then follow an identical path: assembly, trilinear resampling to $64 \times 128 \times 128$ voxels, and percentile-based min-max intensity normalisation. During training, seven deterministic rotated copies are registered per original volume.

Chapter 4

Early Approaches to Breast Cancer Phenotype Classification

The development of the classification framework described in Chapter 5 was not a linear process. Before the final architecture, training strategy, and evaluation protocol were established, a series of earlier experiments were conducted. These experiments were not failures in the pejorative sense, since they produced empirical evidence that motivated and constrained the design decisions documented in the subsequent chapter. The present chapter recounts this iterative trajectory with the dual purpose of providing an honest account of the research process and of explaining, through concrete results, why the choices documented in Chapter 5 are the ones that ultimately prevailed.

The chapter is structured as follows. Section 4.1 documents the initial two-dimensional experiments, including the architectural configurations evaluated, a critical data partitioning incident that materially affected measured performance, and the observed limitations that required a change of direction. Section 4.2 consolidates the arguments that motivated the transition to volumetric processing. Section 4.3 describes the first three-dimensional approach [19], including its architectural characteristics, its exclusive focus on the Luminal A subtype, its training configuration, and its results. Section 4.4 identifies the limitations of that initial three-dimensional approach and justifies the methodological evolution that produced the framework described in Chapter 5. Section 4.5 summarises the principal lessons drawn from this exploratory phase.

Throughout this chapter, all experiments operate on tumour volumes that have already been localised by the upstream lesion detection module [20]. As established in Chapter 1, the detection component falls outside the scope of this thesis and is treated as a fixed precondition.

4.1 Two-Dimensional Experiments

4.1.1 Motivation and Configuration

The natural starting point for any image classification task is a two-dimensional convolutional architecture. At the time these experiments were initiated, the dominant paradigm in the deep learning literature for medical image classification involved fine-tuning large convolutional networks pretrained on natural image benchmarks such as ImageNet [157], an approach that had shown consistent gains over training from scratch on small clinical datasets [158]. The expectation was that this transfer-learning strategy might be sufficient to establish a competitive baseline for breast cancer subtype prediction from Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI) slices, even before volumetric analysis was attempted.

Three backbone architectures with well-established performance on natural image benchmarks were evaluated: ResNet [2], Inception [159], and EfficientNet [160]. In each case, the final classification layer was replaced by a four-way softmax head, and the model was fine-tuned end-to-end using supervised cross-entropy training. A custom 2D Convolutional Neural Network (CNN) was also developed as an additional baseline. This custom architecture begins with an input layer accepting images of $224 \times 224 \times 3$ dimensionality, followed by a rescaling step, two convolutional layers with batch normalisation and Rectified Linear Unit (ReLU) activations, and several residual blocks comprising separable convolutional layers, max-pooling, and skip connections. A global average pooling layer reduces the spatial dimensions prior to a dropout layer and a sigmoid output. The sigmoid activation was used for binary classification experiments, while a softmax was used for the four-class formulation.

Two input resolutions were evaluated. Initial experiments used 224×224 pixels, the standard input size for most ImageNet-pretrained architectures. Subsequent experiments used 448×448 pixels to assess whether the additional spatial resolution improved performance, given the relatively small size of some tumour cross-sections. All inputs were single-channel Magnetic Resonance Imaging (MRI) slices replicated across the three colour channels to match the expected input of the pretrained backbones.

The dataset used in these experiments was assembled from axial slices extracted from the Duke Breast Cancer MRI dataset. The dataset construction and preprocessing steps are described in detail in Chapter 3. The relevant point here is that each axial slice was assigned the patient-level phenotype label, meaning that the four-class label associated with the patient's molecular subtype was attached to every individual slice extracted from that patient's MRI acquisition. This framing is inherent to the slice-level paradigm: since the model processes one slice at a time, the label must be propagated to each slice.

4.1.2 The Data Leakage Incident and Its Correction

A critical methodological error was identified during the first iteration of the two-dimensional experiments. The dataset split into training and test partitions had been performed at the *slice level* rather than at the *patient level*. This means that slices derived from the same patient appeared in both the training and test sets simultaneously. In clinical imaging datasets, consecutive axial slices of the same lesion are highly correlated: they share the same acquisition conditions, the same patient anatomy, and large overlapping regions of the tumour volume. A model trained on some slices of a patient and then tested on other slices from the same patient is therefore not being evaluated on its ability to generalise to new patients, as it is merely recognising patterns it has already encountered, a form of data leakage.

The practical consequence was that measured test performance was substantially inflated. Upon detection of this leakage, the split was redesigned so that all slices belonging to a given patient were assigned exclusively to either the training partition or the test partition. This patient-level separation ensures that no information about any test patient's tumour is accessible to the model during training, which is the correct experimental condition for evaluating generalisation to new clinical cases. The correction caused a marked and immediate decrease in measured performance across all metrics, producing a more realistic estimate of the model's true capabilities.

This incident, while an error, was methodologically instructive. It demonstrated that the slice-level paradigm creates a structural vulnerability to this form of leakage and that any evaluation protocol for this type of problem must enforce patient-level separation from the outset. The corrected protocol was subsequently maintained without exception throughout the remainder of the research.

4.1.3 Results and Observed Limitations

Following the correction of the data leakage issue, the two-dimensional models exhibited poor discrimination across the four molecular subtypes. The custom 2D CNN achieved an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of approximately 0.50 in the corrected four-class setting, a value consistent with random classification, alongside an accuracy of 0.6772 that reflected near-total collapse toward the Luminal A majority class [19]. The pretrained backbones (ResNet, Inception, and EfficientNet) did not achieve materially superior results in this configuration.

Two primary causes explain this outcome. The first is class imbalance. As documented in Chapter 3, Luminal A cases constitute approximately 60 to 70% of the Duke dataset, while HER2-enriched cases account for fewer than 10%. A model trained with a standard cross-entropy objective on these proportions

has a strong incentive to predict Luminal A for the majority of inputs, as doing so minimises loss without requiring the model to learn any discriminative features. The resulting classifier achieves superficially acceptable accuracy by exploiting the base rate rather than learning genuine phenotype-discriminative representations.

The second cause is the fundamental inadequacy of two-dimensional analysis for this problem. Each axial slice captures only a planar cross-section of the tumour. The pharmacokinetic enhancement patterns and spatial heterogeneity features that differentiate molecular subtypes in DCE-MRI, as identified by Mazurowski et al. [50] and Grimm et al. [51], are three-dimensional in nature. A model operating on individual slices in isolation cannot access inter-slice correlations and therefore cannot learn the volumetric patterns most predictive of subtype. Furthermore, the effective sample size under a slice-level paradigm is misleadingly large: ten thousand slices drawn from eight hundred patients provide far less independent information than ten thousand independent images, because slices from the same patient share the same label and nearly identical context.

These two issues (class imbalance and the loss of volumetric context) were identified as the primary barriers to progress and were carried forward as explicit design requirements for the subsequent three-dimensional approach.

4.2 Rationale for Transitioning to Volumetric Processing

The transition from two-dimensional to three-dimensional processing is motivated by three complementary arguments: one biological, one signal-theoretic, and one empirical.

The biological argument rests on the nature of the discriminative features. The molecular subtype of a breast tumour is determined by the expression levels of oestrogen receptors, progesterone receptors, and Human Epidermal growth factor Receptor 2 (HER2) across the entire tumour mass. These receptor distributions manifest in the DCE-MRI signal as patterns of vascular permeability, necrotic core formation, and heterogeneous contrast enhancement that are distributed across the three-dimensional volume of the lesion. A single axial cross-section may happen to traverse the most informative plane of the tumour, but it is equally likely to intersect a peripheral or uninformative region. Only by integrating information across the full volume can a model build a representation that is robust to this spatial variability.

The signal-theoretic argument is that inter-slice correlations carry discriminative information that two-dimensional convolutions structurally cannot capture. A three-dimensional CNN applies convolutional filters simultaneously across all three spatial axes, enabling the detection of volumetric structures such as

elongated vascular channels, spheroidal necrotic cores, or lobulated margins that extend across multiple consecutive slices. As discussed in the architecture literature reviewed in Section 2.5, the extension of the convolution operation to the depth axis is precisely the mechanism by which such structures can be detected.

The empirical argument is provided by the comparative literature. Truhn et al. [87] demonstrated that volumetric CNNs substantially outperform radiomic methods on contrast-enhancing lesion classification from breast MRI. Whitney et al. [88] showed that fusion of volumetric features with classical radiomics consistently surpasses either modality alone, confirming that volumetric context is a primary source of gain. The review by Wang and Wei [112] additionally confirms that the highest-performing recent systems for breast cancer subtype classification all operate in three dimensions.

Taken together, these arguments justified the decision to abandon the two-dimensional pipeline and invest the next phase of the research entirely in volumetric processing. The slice-level analysis had served its purpose: it established a quantitative lower bound on expected performance and made the structural inadequacy of the approach concrete.

4.3 Initial Three-Dimensional Approach

4.3.1 Scope and Proof-of-Concept Design

The first three-dimensional approach [19] was deliberately scoped to a single subtype. Rather than attempting to classify all four molecular phenotypes simultaneously, the decision was made to focus on the Luminal A subtype as a proof of concept. This choice was motivated by two practical considerations: Luminal A is the most prevalent subtype (accounting for approximately 60 to 70% of cases in the Duke dataset, as documented in Chapter 3), meaning that it had the most training examples available; and establishing whether the three-dimensional approach could reliably distinguish Luminal A tumours from all others would provide early empirical evidence of the approach's viability before the considerably harder problem of four-class discrimination was addressed.

The classification task was therefore formulated as a binary problem: for a given tumour volume, predict whether the underlying molecular subtype is Luminal A (positive class) or any of the other three subtypes (Luminal B, HER2-enriched, or Triple-Negative) taken together as a composite negative class. This is the same one-versus-all decomposition later adopted for all four subtypes, but applied at this stage to the single most tractable subtype.

4.3.2 Preprocessing Pipeline

The preprocessing pipeline applied in this first three-dimensional approach established the foundational strategy that was subsequently refined and extended for the full framework. Tumour volumes were extracted from the annotated bounding boxes provided in the Duke dataset (already provided in the dataset), producing variable-size three-dimensional crops of the contrast-enhancing lesion region. Each volume was then subjected to intensity clipping and normalisation, constraining voxel values to the range $[-1000, 400]$ before rescaling to $[0, 1]$ via min-max normalisation. This fixed-window scheme, informed by computed tomography domain knowledge, was later superseded in the final framework by the data-driven percentile-based normalisation described in Section 3.7.3, which adapts the clipping bounds to each volume's own intensity distribution and is better suited to the multi-institutional setting.

A key preprocessing challenge was that a significant proportion of patients in the Duke dataset have fewer than 64 annotated tumour slices, as detailed in Chapter 3. To produce the uniform input tensor shape required by the three-dimensional CNN, an interpolation procedure was applied. Using the `scipy.ndimage.zoom` method, zoom factors were computed dynamically to resize each volume to the target dimensions of $64 \times 128 \times 128$ voxels, with spline interpolation generating synthetic slices where the original count was insufficient. This approach not only standardises input dimensions but also acts as an implicit data augmentation mechanism by introducing slight variations in the spatial representation of the lesion. This procedure is described in full mathematical detail in Chapter 3.

4.3.3 Architectural Configuration

The three-dimensional CNN designed for this first approach shares the same structural skeleton as the refined architecture described in Chapter 5: four convolutional blocks, each comprising a $3 \times 3 \times 3$ convolutional layer with ReLU activation, followed by $2 \times 2 \times 2$ max-pooling and batch normalisation, succeeded by a global average pooling layer, a 64-unit fully connected layer, dropout regularisation, and a sigmoid output unit.

The critical architectural difference lies in the filter progression across the four convolutional blocks. In this initial approach, the filter counts were arranged in a *descending* order: the first block applied the largest number of filters, with subsequent blocks applying progressively fewer. This configuration was an early design choice made before the empirical implications of the filter ordering had been fully analysed. In contrast, the standard practice in deep convolutional architectures, established by VGGNet [161] and reinforced by subsequent work, is to adopt an *ascending* progression—beginning with fewer filters and

doubling with each block—on the grounds that deeper layers operating on spatially reduced feature maps benefit from a larger number of channels to maintain representational capacity. The descending configuration adopted in this initial approach inverts this principle: it applies the highest channel count at the first layer, where feature maps are spatially largest and where each filter therefore entails the greatest computational and parametric cost, while applying fewer channels at the deeper layers where complex, high-level features should ideally be constructed.

As will be discussed in Section 4.4, this architectural decision introduced a suboptimal allocation of model capacity that was corrected in the subsequent iteration.

4.3.4 Training Configuration

The training procedure in this initial approach followed the same general principles as the refined framework but with several differences in specific hyperparameter values. The Adam optimiser [162] was used with an initial learning rate of $\eta_0 = 10^{-4}$, an order of magnitude higher than the 10^{-5} adopted in the subsequent approach. An exponential learning rate decay was applied, multiplying the learning rate by a factor of 0.96 every 100,000 steps—a schedule with a considerably slower decay rate and a continuous (rather than staircase) decay profile compared to the refined version. Class imbalance was addressed through weighted binary cross-entropy, with weights computed dynamically from the class frequencies in the training generator, consistent with the approach described in Section 5.4.

Early stopping was configured with a patience of 25 epochs, monitoring validation performance. The data was partitioned in a 90%–10% training-to-test split, enforcing strict patient-level separation. Notably, at this stage no systematic hyperparameter optimisation was performed: the values used were selected through a combination of literature precedent and manual experimentation, without the structured search that was subsequently introduced using Keras Tuner [163].

Five-fold and ten-fold cross-validation were conducted to assess the stability of the performance estimates. This validation strategy, while more rigorous than a single held-out split, also introduced some variability across folds due to the limited dataset size, particularly for the minority subtypes.

4.3.5 Results

The initial three-dimensional approach achieved substantially superior performance compared to all two-dimensional configurations. Table 3 summarises the performance metrics across the 2D and initial 3D approaches.

The improvement from the 2D to the initial 3D approach is striking. The AUC-ROC increased from

Table 3: Performance comparison of the two-dimensional baseline and the initial three-dimensional approach on the binary Luminal A versus non-Luminal A classification task using the Duke Breast Cancer MRI dataset. All values are from Fontes et al. [19]. The high recall of the 2D model reflects majority-class collapse rather than genuine discriminative ability. The AUC-ROC of 0.499 for the 2D approach confirms near-random classification after the patient-level split correction.

Approach	Accuracy	Precision	Recall	AUC-ROC
Custom 2D CNN	0.677	0.689	0.969	0.499
Initial 3D CNN (single split)	0.906	0.919	0.947	0.961
3D ResNet-152 (single split)	0.861	0.885	0.916	0.924

Table 4: Five-fold and ten-fold cross-validation results of the initial 3D CNN and the 3D ResNet-152 on the binary Luminal A versus non-Luminal A classification task using the Duke Breast Cancer MRI dataset. All values are from Fontes et al. [19]. Each cell reports the mean $\pm$ standard deviation across folds (unweighted, as published) and, below it, the 95% confidence interval. Accuracy and F_1-score were not computed for the cross-validation experiments.

Model	Folds	Precision	Recall	AUC-ROC
Initial 3D CNN	5	0.787 ± 0.0868 [0.711, 0.863]	0.876 ± 0.0554 [0.827, 0.925]	0.788 ± 0.107 [0.694, 0.882]
	10	0.747 ± 0.0971 [0.687, 0.807]	0.865 ± 0.141 [0.778, 0.952]	0.746 ± 0.135 [0.662, 0.830]
3D ResNet-152	5	0.713 ± 0.0515 [0.668, 0.758]	0.852 ± 0.148 [0.722, 0.982]	0.700 ± 0.103 [0.610, 0.790]
	10	0.754 ± 0.0665 [0.713, 0.795]	0.792 ± 0.145 [0.702, 0.882]	0.729 ± 0.0948 [0.670, 0.788]

approximately 0.50 to 0.9614 in the single-split evaluation, while the F1-score reached 0.9328. These results confirm the core hypothesis that motivated the transition to volumetric processing: three-dimensional feature extraction captures tumour characteristics that are structurally inaccessible to two-dimensional slice-level analysis.

The custom 3D CNN outperformed the 3D ResNet-152 adaptation across all single-split metrics, with AUC-ROC values of 0.961 versus 0.924 respectively, and demonstrated faster inference time (0.639 seconds per step versus 0.914 seconds on the same hardware). This result is consistent with the broader finding in the literature that the optimal architecture depth is constrained by the dataset size: a simpler model with appropriate inductive biases often generalises better than a deeper, higher-capacity architecture when the number of training examples is limited [164].

The cross-validation results (Table 4), while lower in absolute terms (mean AUC-ROC of 0.788 under five-fold and 0.746 under ten-fold validation), are consistent with the substantial performance variability expected from a dataset of fewer than one thousand patients partitioned into five or ten folds. Under five-fold validation, the initial 3D CNN outperformed the 3D ResNet-152 in all three reported metrics, with statistically significant differences in precision (0.787 versus 0.713, $p = 0.042$), recall (0.876 versus

0.852, $p = 0.047$) and AUC-ROC (0.788 versus 0.700, $p = 0.031$, 95% Confidence Interval (CI) $[0.694, 0.882]$ versus $[0.610, 0.790]$), which confirms that the simpler architecture is genuinely more appropriate for this dataset scale. Under ten-fold validation, the two models performed comparably (AUC-ROC 0.746 versus 0.729) with higher variability across folds, although the 3D CNN retained higher recall (0.865 versus 0.792).

The gap between the single-split AUC-ROC (0.961) and the cross-validated means should not be read as a contradiction. The single-split value is obtained on one 90%–10% partition, whose test set is small and depends on which patients it contains. Cross-validation evaluates every patient exactly once across partitions, including less favourable ones: the per-fold AUC-ROC of the 3D CNN ranges from 0.705 to 0.976 under five-fold validation and from 0.580 to 0.994 under ten-fold validation [19]. The cross-validated means are therefore the more conservative estimate of the expected performance, whereas the single-split value is the best result obtained among the training runs.

4.4 Limitations and Motivation for the Current Approach

Despite its empirical success on the Luminal A classification task, the initial three-dimensional approach had several limitations that the subsequent work was designed to address.

Single-subtype focus. The most immediate limitation is that the approach classified only the Luminal A subtype. While this produced a strong binary classifier, it provided no direct information about the other three subtypes, which are individually associated with worse prognosis and require different treatment strategies [15]. A clinically useful non-invasive tool must be capable of distinguishing among all four subtypes, not merely separating the most prevalent from the remainder.

Descending filter progression. As described in Section 4.3.3, the filter progression adopted in the initial architecture was ordered from large to small rather than from small to large. This inversion of the standard convention allocates the greatest parametric and computational cost to the layers that operate on the largest feature maps, where such cost is least justified. Early layers should capture low-level geometric features that do not require a large number of filters, leading to the consequence that the model's capacity is distributed inefficiently across its depth, with the early layers potentially overfitting to low-level detail and the deeper layers being under-resourced for constructing the high-level abstract representations that would be most discriminative of molecular subtype. Correcting this inversion to an ascending progression ($(32, 64, 128, 256)$ filters across the four blocks) was one of the principal architectural modifications

made in the subsequent approach.

Absence of systematic hyperparameter optimisation. The training configuration of the initial approach was determined manually, without a structured search over the hyperparameter space. The learning rate, filter counts, dense layer width, and dropout rate were set by inspection and adjusted through iterative manual experimentation. This approach, while pragmatic, cannot guarantee that the selected configuration is optimal or even near-optimal: the relationship between hyperparameter values and validation performance is non-monotone and interactive, meaning that individually reasonable choices may combine suboptimally. The introduction of Keras Tuner [163] in the subsequent approach replaced this informal procedure with a reproducible, principled search over a predefined hyperparameter space.

Higher learning rate and faster decay. The initial learning rate of 10^{-4} combined with a continuous exponential decay of 0.96 per 100,000 steps was found to produce instability in early training on the imbalanced dataset, particularly when training the classifiers for the minority subtypes. The refined approach adopts a lower initial rate of 10^{-5} with a staircase decay profile (decay factor 0.9, step count 10,000), which maintains a stable gradient signal throughout training and is more robust to the noise introduced by the imbalanced class weights.

Shorter early stopping patience. A patience of 25 epochs for early stopping was found to be insufficient to capture the full training dynamics of the minority-subtype classifiers, which converge more slowly due to the smaller effective batch size of positive examples. The patience was increased to 50 epochs in the refined approach, ensuring that models are not prematurely terminated before reaching their performance optimum on the validation set.

Limited dataset and absence of multicentric validation. The initial approach was trained and evaluated exclusively on the Duke Breast Cancer MRI dataset and validated using cross-fold splits of the same dataset. While this provides a reasonable estimate of in-distribution generalisation, it does not address the question of whether the learned representations transfer across institutions with different scanners, acquisition protocols, and patient populations. The subsequent work addressed this gap by incorporating multicentric datasets from the Cancer Imaging Archive (TCIA) and, critically, by conducting external validation on an independent clinical cohort from Unidade Local de Saúde da Arrábida (ULSA), Setúbal, as documented in Chapter 3, specifically Section 3.6.

Collectively, these limitations defined a clear research path for the next iteration: extend the classifi-

cation framework to all four subtypes using a one-versus-all decomposition, correct the filter progression, introduce systematic hyperparameter optimisation, refine the training schedule, and validate the resulting framework across multiple institutions.

4.5 Summary

This chapter has recounted the iterative research trajectory that preceded the classification framework described in Chapter 5. Three distinct experimental phases were conducted. The two-dimensional experiments established a performance baseline and, through the identification and correction of a patient-level data leakage error, produced a realistic assessment of what slice-level models can achieve: AUC-ROC values near 0.50 after correction, confirming the inadequacy of two-dimensional analysis for this problem. The transition to three-dimensional processing was motivated by biological, signal-theoretic, and empirical arguments, each pointing independently to the necessity of volumetric feature extraction.

The initial three-dimensional approach demonstrated that volumetric processing is both feasible and highly effective for this problem, achieving an AUC-ROC of 0.9614 for the Luminal A subtype classification task. This result constitutes the empirical proof of concept for the three-dimensional framework. However, five specific limitations were identified: the single-subtype scope, the suboptimal descending filter progression, the absence of systematic hyperparameter optimisation, the suboptimal training schedule, and the absence of multicentric validation. The resolution of each of these limitations is directly reflected in the design decisions documented in Chapter 5, which can therefore be read as a systematic response to the empirical lessons of the exploratory phase described here.

Chapter 5

A New Approach for Classifying Breast Cancer Phenotypes

The preceding chapters have established the clinical motivation for non-invasive molecular subtype prediction, surveyed the landscape of deep learning methods applicable to volumetric breast Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI), and described the construction of the multicentric data corpus together with the preprocessing pipeline that produces standardised input volumes. This chapter closes the loop between those preparatory steps and the experimental results reported in Chapter 6 by presenting the methodological decisions that define the classification framework.

The chapter is structured as follows. Section 5.1 situates the classification component within the broader Quantitative Imaging Biomarker (QIB) discovery framework, recounts the foundational work that motivated the present approach, and summarises the experimental trajectory documented in Chapter 4. Section 5.2 formalises the one-versus-all decomposition adopted throughout the work and justifies this choice over competing multi-class formulations. Section 5.3 describes the 3D Convolutional Neural Network (CNN) architecture in detail, including each layer's role, the regularisation strategy, and a comparison with an alternative ResNet-based backbone. Section 5.4 elaborates on the class weighting scheme used to address the inherent label imbalance of breast cancer subtype datasets. Section 5.5 covers all aspects of the training procedure, including the optimiser, the learning rate schedule, the early stopping criterion, and reproducibility controls. Section 5.6 describes the hyperparameter optimisation strategy. Finally, Section 5.7 summarises the principal design choices and their justifications.

Throughout this chapter, the classification component is treated as operating on tumour volumes that have already been localised by an upstream lesion detection module [20]. As noted in Chapter 1, the development of the detection component falls outside the scope of this thesis; it is treated as a fixed precondition. Consequently, this chapter concerns itself exclusively with the problem of phenotype classification given a correctly cropped and preprocessed volumetric input.

5.1 Framework Design Philosophy

The primary goal of this thesis is to propose a reusable framework for discovering QIBs in oncology, using breast cancer phenotype classification as a validation case study. Within that framing, the classification component described here serves a dual purpose: it produces a clinically meaningful output (a probability score for each of the four molecular subtypes) and it simultaneously demonstrates that end-to-end Deep Learning (DL) can construct a QIB without any manual feature engineering.

This work draws direct inspiration from two earlier publications that established representation learning as a powerful paradigm for breast imaging analysis. Arevalo et al. [154] demonstrated that CNNs trained in a supervised manner on mammography patches could learn feature representations that substantially outperformed hand-crafted descriptors, even those benefiting from expert radiologist segmentations. Fontes et al. [153] extended this line of reasoning to pathological lesion analysis, showing that systems that learn directly from pixels rather than relying on engineer-designed features achieve better performance and are far more scalable. Both works, however, operated in two dimensions and on malignancy detection rather than molecular subtyping.

The present work takes the next step: it applies three-dimensional representation learning to the problem of classifying the molecular phenotype of a confirmed breast cancer lesion. The clinical value of this step is significant. Molecular subtype determines treatment selection, prognosis, and recurrence risk: an imaging-derived classifier that reliably predicts subtype would reduce dependence on invasive biopsy, enable longitudinal monitoring, and provide a richer information basis for treatment response assessment [15, 17].

The framework is designed with generalisability explicitly in mind. Each architectural and training decision is evaluated not only in terms of its impact on internal performance but also in terms of its potential contribution to cross-institutional applicability. The methodological evolution that led to the present framework, from two-dimensional baseline experiments through the initial three-dimensional proof of concept published in Fontes et al. [19], is documented in Chapter 4. That chapter establishes, through concrete empirical results and identified limitations, the rationale for every major design decision described in the sections below. The remainder of this chapter is concerned exclusively with the final framework configuration.

The proposed framework is not defined by its application to breast cancer or to DCE-MRI. It is defined by a structural decomposition of the QIB discovery problem into two separable stages: (i) automated lesion localisation, followed by (ii) volumetric classification. The classification stage operates on a bounded region

of interest produced by the upstream stage, regardless of the imaging modality, organ site, or semantic content of the target labels. The framework is applicable to any clinical problem that can be decomposed into this localisation-then-classification structure and that satisfies two preconditions: (i) the availability of a volumetric imaging modality from which a lesion or region of interest can be reliably extracted, and (ii) a clinically relevant discrete classification target, whether that target is a molecular subtype, a malignancy grade, a binary benign-versus-malignant determination, or any other categorisation of diagnostic, prognostic, or theranostic significance. The specific semantic content of the classes is entirely pluggable; the pipeline architecture requires no structural modification to accommodate a new task.

At a deeper level of abstraction, the framework implements an object-oriented design pattern in which the QIB itself is a first-class object with encapsulated methods and attributes. The **attributes** of this object include the trained neural network weights, the preprocessing configuration (normalisation strategy, spatial resampling parameters), metadata describing the training corpus and validation protocol, and performance metrics characterised on independent cohorts. The **methods** of the QIB object include: (i) preprocessing methods that standardise input volumes regardless of source modality or acquisition protocol, (ii) inference methods that accept volumetric data and emit continuous probability outputs, (iii) multimodal integration methods that can combine features from disparate imaging sequences (e.g., DCE-MRI, diffusion-weighted Magnetic Resonance Imaging (MRI), T2-weighted MRI, radiomics) through flexible input channels, and (iv) model optimisation methods that adjust hyperparameters, loss weights, and checkpoint selection criteria within a domain-appropriate framework. This object-oriented encapsulation is essential to the generalisability claim: the same QIB class can be instantiated with different training objectives (molecular subtype, tumour grade, risk score), different modalities (e.g., prostate MRI, glioma MRI, pulmonary Computerized Tomography (CT)), and different optimisation strategies, with the core architectural invariant (localisation followed by probabilistic classification) remaining constant across all instantiations.

This argument for architectural generalisability is substantiated by a growing body of evidence across oncological imaging. Three concrete analogues, spanning both the modality axis (MRI and CT) and the task axis (molecular phenotyping and malignancy risk stratification), are presented in Section 7.4 to illustrate the range of QIB discovery contexts to which the proposed framework is applicable.

5.2 One-Versus-All Classification Strategy

5.2.1 Formulation and Motivation

Multi-class classification problems can be tackled in several fundamentally different ways. A single model with a multi-class output head (a softmax layer over all classes) is conceptually simple but assumes that the classes are mutually exclusive and that a single shared representation is sufficient to discriminate all pairs simultaneously. An alternative family of decomposition strategies replaces this single model with a collection of binary classifiers, each specialising in one aspect of the decision boundary.

The One-vs-All (OvA) strategy, also referred to as one-versus-rest, is the most widely studied decomposition approach [23]. Given a K-class problem, it trains K separate binary classifiers. Classifier k is trained to distinguish class k from the union of all remaining classes. At inference time, all K classifiers are applied to the input, and the class associated with the classifier producing the highest confidence score is assigned as the prediction.

Formally, let $\mathcal{K} = \{0, 1, \ldots, K-1\}$ denote the set of class indices and let $f_k : \mathbb{R}^d \to [0, 1]$ denote the output of classifier k, which estimates $P(y = k \mid \mathbf{x})$ under the binary assumption that the sample either belongs to class k (positive) or to the complement set (negative). The final multi-class prediction $\hat{y}$ is obtained by selecting the class whose binary classifier produces the highest score:

$$\hat{y} = \underset{k \in \mathcal{K}}{\arg\max}\ f_k(\mathbf{x}), \qquad \mathcal{K} = \{0, 1, 2, 3\} \tag{5.1}$$

where, in this particular case, the four classes correspond to Luminal A ($k = 0$), Luminal B ($k = 1$), HER2-enriched ($k = 2$), and Triple-Negative ($k = 3$).

5.2.2 Justification for the One-vs-All Choice

Three considerations motivate the selection of OvA over a single-model softmax formulation in this specific context.

First, the class imbalance problem is significantly more tractable under OvA. A softmax model over four highly imbalanced classes faces the challenge of simultaneously satisfying gradient signals from all classes, and the dominant Luminal A class tends to steer the shared representation toward features that are discriminative primarily for the majority class. Under OvA, each binary classifier independently defines its own positive and negative set, and the class weighting scheme described in Section 5.4 can be applied independently and precisely calibrated to the specific imbalance ratio of each subtype. For the HER2-

enriched classifier, where fewer than 10% of cases are positive, this independent calibration is particularly valuable.

Second, OvA provides directly interpretable per-class probability scores. Each $f_k(\mathbf{x})$ is an estimate of the probability that sample $\mathbf{x}$ belongs to class k, calibrated against all other classes collectively. This supports clinical interpretability: a clinician can inspect not only the predicted class but also the confidence scores for all four subtypes, which may be informative in borderline cases or cases with mixed molecular profiles.

Third, OvA enables independent model selection and checkpoint management for each subtype. Because each classifier is trained independently, the checkpoint with the highest validation Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for that specific classifier can be selected without compromising any other classifier. This is particularly important given that the optimal training duration (determined by early stopping, as discussed in Section 5.5) differs substantially across subtypes.

The principal limitation of OvA is the possibility of conflicting predictions: it is possible for two classifiers to both output high confidence scores, or for none to do so. In practice, as reported in Chapter 6, this conflict is rare; the majority of test samples produce a single dominant classifier output, and the arg max rule in Equation 5.1 resolves ties deterministically.

It is important to note that the four scores $f_0, \ldots, f_3$ are *not* jointly calibrated: each classifier is trained independently with a sigmoid output against a different binary negative set, so the scores do not sum to 1 and are not directly comparable on a common probability scale. When two classifiers produce nearly identical scores, the arg max rule still yields a unique hard prediction while the underlying ambiguity is invisible in the output label alone. The raw score vector $\mathbf{f}(\mathbf{x}) = (f_0(\mathbf{x}), \ldots, f_3(\mathbf{x}))$ is therefore the primary output of the framework. The hard label is a deterministic post-processing step applied to it. For downstream applications that require a distribution over the four subtypes, the scores can be sum-normalised as

$$\tilde{f}_k(\mathbf{x}) = \frac{f_k(\mathbf{x})}{\sum_{j=0}^{K-1} f_j(\mathbf{x})}, \qquad k \in \{0, \ldots, K-1\}, \tag{5.2}$$

yielding a vector $\tilde{\mathbf{f}}(\mathbf{x}) \in [0,1]^K$ with $\sum_k \tilde{f}_k = 1$ that maps borderline cases onto intermediate positions. In the present work $K = 4$, with indices $\{0, 1, 2, 3\}$ corresponding to Luminal A, Luminal B, HER2-enriched, and Triple-Negative respectively. This normalised output is adequate for uncertainty communication and for flagging ambiguous cases. However, it should not be interpreted as a calibrated posterior without additional probability calibration (e.g. isotonic regression or Platt scaling applied to held-out

data).

5.3 Three-Dimensional CNN Architecture

5.3.1 Architectural Overview

The core classifier in this work is a custom 3D CNN designed specifically for volumetric MRI classification under the constraints of a limited and imbalanced medical imaging dataset. The architecture prioritises simplicity, regularisation, and computational efficiency over raw depth, reflecting the need to train from scratch on a few hundred to a few thousand samples without access to pretrained volumetric weights.

The input to the network is a single-channel three-dimensional tensor of shape $(64, 128, 128, 1)$, corresponding to the $64 \times 128 \times 128$ voxel volumes produced by the preprocessing pipeline described in Chapter 3. The network consists of four convolutional blocks, each comprising a three-dimensional convolutional layer, a max-pooling layer, and a batch normalisation layer, followed by a global average pooling layer, a fully connected layer with dropout regularisation, and a sigmoid output unit. Figure 5 illustrates the complete architecture.

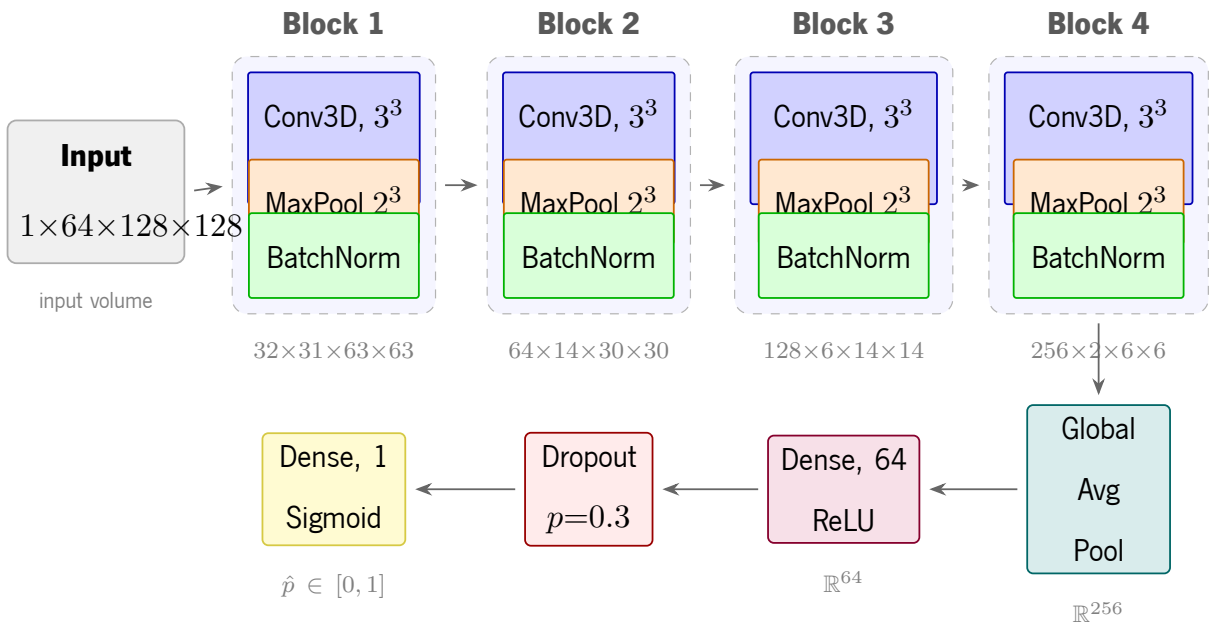


Figure 5: The 3D CNN architecture. The input volume ($1\times64\times128\times128$, in channels $\times$ depth $\times$ height $\times$ width) passes through four convolutional blocks (blue: Conv3D; orange: MaxPool3D; green: BatchNorm), each doubling the number of feature channels while halving all spatial dimensions. The annotated tensor shape below each block reflects the output after max-pooling under valid (no-padding) convolutions. Global average pooling (teal) collapses the spatial volume to a 256-dimensional vector; a dense layer with dropout (purple/red) reduces it to 64 units; and a sigmoid output unit (yellow) produces the phenotype probability score $\hat{p} \in [0, 1]$.

5.3.2 Convolutional Blocks

Each of the four convolutional blocks follows the same structural pattern: a three-dimensional convolutional layer with a $3 \times 3 \times 3$ kernel, Rectified Linear Unit (ReLU) activation, and no padding, followed by a three-dimensional max-pooling layer with a $2 \times 2 \times 2$ pool size, and batch normalisation [100].

The filter counts progress as $(32, 64, 128, 256)$ across the four blocks. This doubling strategy is standard in convolutional architectures [2] and reflects the principle that deeper layers, operating on spatially smaller feature maps, must compensate for reduced spatial resolution by increasing the number of feature channels in order to preserve representational capacity. Starting at 32 filters rather than the 64 used in many larger architectures is a deliberate conservative choice that accounts for the substantially smaller input domain (a single-modality clinical scan rather than a large natural image collection) and limits the number of parameters to reduce the risk of overfitting.

The $3 \times 3 \times 3$ kernel size is the three-dimensional analogue of the 3×3 kernels that have been standard in two-dimensional architectures since VGGNet [161]. It captures local spatial context in all three dimensions while incurring lower parameter counts than larger kernels; multiple stacked $3 \times 3 \times 3$ convolutions are equivalent in receptive field to a single larger kernel but offer greater non-linearity. The $2 \times 2 \times 2$ max-pooling operation halves each spatial dimension after every block. Because the convolutions use no padding, each $3 \times 3 \times 3$ convolution also reduces every spatial dimension by 2 before pooling. After four such block applications, the spatial dimensions shrink from $(64, 128, 128)$ to $(2, 6, 6)$ — as annotated in Figure 5 — enabling the global average pooling layer to aggregate across a compact but spatially meaningful feature volume.

Batch normalisation [100] is applied after pooling in each block. In addition to its well-known stabilising effect on the training dynamics of deep networks, batch normalisation acts as a regulariser by introducing noise during training through the use of mini-batch statistics. In the context of a limited medical imaging dataset, this implicit regularisation effect is particularly valuable.

5.3.3 Global Average Pooling and Classification Head

Following the four convolutional blocks, a global average pooling layer [102] reduces the three-dimensional feature tensor to a one-dimensional vector by computing the mean activation across all spatial positions for each channel. If the output of the last convolutional block has shape (D', W', H', C), the global average pooling operation produces a vector $\mathbf{z} \in \mathbb{R}^C$ defined component-wise as:

$$z_c = \frac{1}{D' \cdot W' \cdot H'} \sum_{i=1}^{D'} \sum_{j=1}^{W'} \sum_{k=1}^{H'} a_{ijkc}, \qquad c = 1, \ldots, C \tag{5.3}$$

where a_{ijkc} denotes the activation at spatial position (i, j, k) in channel c.

This operation serves as a structural regulariser [102]: unlike fully connected layers that map every spatial position to the output, global average pooling is insensitive to the spatial arrangement of activations, producing a representation that integrates evidence across the entire volume. It dramatically reduces the number of parameters in the classification head (from tens of thousands to a few hundred) and has been shown to substantially reduce overfitting in settings with limited training data.

The classification head consists of a single dense layer with 64 units and ReLU activation, followed by a dropout layer with a drop rate of 0.3, and a final dense layer with a single unit and sigmoid activation. The 64-unit layer provides a compact non-linear mapping from the 256-dimensional global average pooling output to a task-specific representation. Dropout [101] at rate 0.3 randomly zeroes approximately 30% of units during each training step, preventing co-adaptation of units and acting as an ensemble of thinned networks [101]. The sigmoid output $\hat{p} = \sigma(z) = 1/(1 + e^{-z})$ maps the scalar logit to the unit interval, yielding a probability estimate for the positive class (the target subtype) under the binary OvA formulation.

5.3.4 Comparison with ResNet Backbone

During the exploratory phase, a three-dimensional ResNet-50 backbone was also evaluated as an alternative to the custom CNN. The implementation followed the architecture described by Hara et al. [91], adapted to accommodate the $(64, 128, 128, 1)$ input tensor. Residual connections in ResNets facilitate gradient flow in deep networks by providing identity shortcuts between blocks, a property particularly valuable when training from random initialisations on limited data.

In practice, the ResNet backbone exhibited higher parameter counts and greater computational cost per forward pass. On the available dataset, it did not yield consistent improvements over the custom CNN, likely because the dataset is not large enough to fully benefit from the representational capacity of a fifty-layer residual network. The custom CNN architecture, being shallower and more aggressively regularised, achieved competitive or superior validation AUC-ROC values while requiring substantially less training time. The custom CNN was therefore adopted as the primary model throughout the classification experiments. The comparative evaluation against the 3D ResNet-152 backbone, which was conducted exclusively on the Luminal A binary classification task, is reported in Chapter 4.

This finding is consistent with the observation in the literature that for medical imaging datasets of

moderate size, shallower but well-regularised architectures often match or outperform deeper alternatives, and that the gains from depth typically require pretraining on large external corpora to materialise [98, 99].

5.4 Addressing Class Imbalance via Weighted Binary Cross-Entropy

5.4.1 The Imbalance Problem in Context

As established in Chapter 3, the four molecular subtypes exhibit markedly unequal prevalence in all collections available in the Cancer Imaging Archive (TCIA). Luminal A accounts for approximately 60 to 70% of the corpus, Luminal B for 10 to 25%, Triple-Negative for 10 to 15%, and HER2-enriched for 5 to 10%. Under the OvA formulation, each binary classifier faces its own instance of this imbalance: the Luminal A classifier operates on a roughly balanced partition, while the HER2-enriched classifier faces a positive-to-negative ratio of approximately 1:9. Johnson and Khoshgoftaar [122] demonstrated that naïve training under such imbalance causes a CNN to systematically neglect the minority class, as the gradient signal is dominated by the majority class regardless of the choice of loss function.

Two complementary mechanisms are employed in this work to mitigate this effect: weighted loss during training and data augmentation as described in Chapter 3. This section focuses on the weighted loss strategy.

5.4.2 Class Weight Calculation

For each binary classifier, class weights are computed dynamically from the class distribution of the training partition, ensuring that weights accurately reflect the imbalance of the data actually seen during training rather than the global dataset distribution. Given n_0 negative training samples and n_1 positive training samples, with $N = n_0 + n_1$ denoting the total, the weights assigned to the negative and positive classes respectively are:

$$W_0 = \frac{N}{2 \cdot n_0}, \qquad W_1 = \frac{N}{2 \cdot n_1} \tag{5.4}$$

This formulation ensures that the weighted contribution of each class to the total loss is equal: the sum of weighted losses from the positive class equals the sum of weighted losses from the negative class, irrespective of their relative sizes. As a consequence, a model that learns to predict all samples as negative (the degenerate solution that achieves low loss under standard cross-entropy in the presence of severe

imbalance) incurs a high weighted loss and is penalised appropriately. The weights are proportional to the inverse of the class frequencies and satisfy $W_0 \cdot n_0 = W_1 \cdot n_1 = N/2$, confirming the balanced weighting property.

The weighted binary cross-entropy loss function is:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\left[W_1 \cdot y_i \log \hat{p}_i + W_0 \cdot (1 - y_i)\log(1 - \hat{p}_i)\right] \tag{5.5}$$

where B is the batch size, $y_i \in \{0, 1\}$ is the binary label of sample i under the OvA encoding, and $\hat{p}_i$ is the sigmoid output of the network. The class weights W_0 and W_1 are passed to the `class_weight` argument of the Keras `Model.fit()` method, which applies them internally to the per-sample losses before averaging.

This weighted loss approach is preferred over other class-imbalance strategies such as focal loss [123] because it requires no additional hyperparameter (the focal loss modulating exponent γ must be tuned), and because the literature on breast cancer subtype classification has not demonstrated a consistent advantage of focal loss over weighted cross-entropy in this specific setting.

5.5 Training Procedure

5.5.1 Optimiser and Learning Rate Schedule

Model parameters are updated using the Adam optimiser [162], which combines the benefits of adaptive gradient scaling (as in AdaGrad) and momentum estimation (as in RMSProp). Adam maintains per-parameter estimates of the first and second moments of the gradients, producing updates that are both direction-aware and magnitude-normalised. This makes it particularly robust to sparse gradients and noisy loss landscapes, which are characteristic of small-batch training on medical imaging data.

A key training challenge specific to this application is avoiding premature convergence to local minima early in training while also preventing oscillatory behaviour near optima in later stages. This is addressed through an exponentially decaying learning rate schedule. The initial learning rate is set to $\eta_0 = 10^{-5}$, which is deliberately conservative relative to the 10^{-3} default in order to stabilise early training dynamics on the imbalanced data. The learning rate at step t is given by:

$$\eta(t) = \eta_0 \cdot \rho^{\lfloor t/s \rfloor} \tag{5.6}$$

where $\rho = 0.9$ is the decay factor and $s = 10000$ is the decay step count. The staircase variant

is used (as indicated by the `staircase=True` flag in the implementation), meaning the learning rate drops discretely at multiples of s rather than continuously. This piecewise-constant behaviour prevents excessively rapid decay and maintains a non-negligible learning signal throughout training.

5.5.2 Loss Function and Compilation Metrics

As established in Section 5.4, the loss function is weighted binary cross-entropy (Equation 5.5). The model is compiled with the following tracked metrics: AUC-ROC, precision, recall, and the four components of the confusion matrix (true positives, true negatives, false positives, false negatives). This comprehensive metric set ensures that training dynamics can be monitored beyond aggregate accuracy, which is uninformative under class imbalance.

5.5.3 Callbacks and Model Selection

Three Keras callbacks govern the training loop:

Early stopping monitors the validation AUC-ROC and terminates training if no improvement is observed over a patience window of 50 epochs. When early stopping is triggered, the weights from the epoch with the highest validation AUC-ROC are automatically restored. This policy selects the model that achieves the best discriminative performance on the held-out validation partition, regardless of when during training this optimum was reached. The use of AUC-ROC as the monitored metric rather than validation loss is deliberate: as established in the imbalance literature [122], validation loss may continue to decrease even as the model degrades on the minority class, whereas AUC-ROC is rank-based and directly sensitive to the quality of the discrimination boundary for the positive class.

Model checkpoint saves the model weights to disk at every epoch in which the validation AUC-ROC improves. This provides a complete record of the training trajectory and enables post-hoc inspection or re-evaluation using intermediate checkpoints.

TensorBoard logging records scalar metrics, histograms of weight distributions, and computation graphs, enabling detailed inspection of training dynamics and early detection of gradient pathologies such as exploding or vanishing gradients.

Clinical Checkpoint Selection

The standard model selection strategy used throughout the experiments reported above selects the checkpoint with the highest validation AUC-ROC, as described in Section 5.5.3. While AUC-ROC is the most

informative single metric for assessing discriminative performance under class imbalance, it does not directly capture the clinical requirements that govern deployment decisions. In particular, a model selected purely for maximal AUC-ROC may achieve this optimum at an operating point where recall is insufficient for clinical safety, or where the precision-recall trade-off is misaligned with the asymmetric costs of false positives and false negatives in oncological screening.

To address this gap, a supplementary checkpoint selection mechanism was developed and integrated into the training pipeline: the `ClinicalCheckpoint` callback. This callback operates in parallel with the standard AUC-ROC-based `ModelCheckpoint` and `EarlyStopping`, producing a second, independent checkpoint stream at each training run. Four selection strategies are supported:

1. **Clinical:** the epoch must satisfy a minimum recall floor (configurable, default 0.80), and among all epochs that meet this floor, the checkpoint with the highest F_1-score is selected, with AUC-ROC as the tiebreaker. If no epoch meets the recall floor, the best available F_1 is retained as a provisional selection, and a warning is logged.

2. **Balanced:** ranked purely by F_1-score, then AUC-ROC, without a recall floor.

3. **High precision:** the epoch must satisfy a minimum precision floor (configurable, default 0.80), ranked by F_1-score and then AUC-ROC.

4. **Max AUC:** equivalent to the standard `ModelCheckpoint` behaviour, included for completeness.

The "clinical" strategy is motivated by the observation that in oncological screening, the cost of a false negative (missed cancer subtype leading to suboptimal treatment) typically exceeds the cost of a false positive (unnecessary confirmatory biopsy). By imposing a recall floor, the strategy ensures that the selected model does not sacrifice sensitivity for overall discrimination, a trade-off that is possible when AUC-ROC is the sole selection criterion. The tiered comparison mechanism ensures that epochs meeting the recall floor are always preferred over those that do not, regardless of their F_1 or AUC-ROC values.

5.5.4 Batch Size, Maximum Epochs, and Data Partitioning

All models are trained with a batch size of 8. This relatively small batch size is dictated by the memory constraints of processing $64 \times 128 \times 128$ voxel volumes and reflects a common recommendation for volumetric medical image training [165]. The maximum number of training epochs is set to 100. In practice, early stopping consistently terminates training before this limit is reached.

The dataset is partitioned at the patient level, with 85% of patients assigned to training and 15% to internal validation. The strict patient-level split, described in Section 4.1.2, ensures no data leakage between partitions. The external validation cohort from Unidade Local de Saúde da Arrábida (ULSA), Setúbal, is held completely out of the training loop and is evaluated only once, after final model selection, as described in Section 3.6.

A structural limitation of this single-split design is that both the early stopping criterion and the best-checkpoint selection are conditioned on the same 15% partition. Repeated comparison against a fixed validation set over up to 100 training epochs constitutes an implicit multiple-comparison problem: there is a non-negligible probability that the selected checkpoint represents a configuration that happens to perform well on this specific partition rather than on the underlying population distribution. The large patience window of 50 epochs partially mitigates this by limiting the effective number of checkpoint updates, but does not eliminate the risk. The most informative safeguard available in the current study design is the external Setúbal cohort, which is withheld entirely from the model selection loop and evaluated exactly once. Any systematic overfit to the internal validation partition would manifest as a discrepancy between internal and external performance. The close agreement observed in Chapter 6 provides indirect evidence that this bias is not severe. Nonetheless, a nested cross-validation design remains the methodologically definitive approach and is identified as a direction for future work in Section 7.6.

5.5.5 Reproducibility Controls

Reproducibility is ensured through three mechanisms: (1) a global random seed is set for NumPy, Python's built-in `random` module, and TensorFlow prior to any data loading or model initialisation; (2) all training-time augmentation is deterministic and pre-registered (as described in Chapter 3), meaning that the exact set of augmented volumes produced for each training run is fully determined by the seed and dataset; and (3) Graphics Processing Unit (GPU) memory growth is enabled but a deterministic execution mode is not enforced, as full determinism on GPUs typically incurs prohibitive computational overhead. The combination of fixed seeds and deterministic augmentation is sufficient to reproduce reported results on the same hardware configuration.

5.6 Hyperparameter Optimisation

The architecture and training configuration described in the preceding sections represents the final configuration selected after a systematic hyperparameter search. The search was conducted using Keras

Tuner [163], which provides a programmatic interface for defining search spaces over model architecture and training hyperparameters and orchestrating evaluation trials.

The hyperparameters subjected to search included the number of filters in each convolutional block (exploring progressions starting at 16, 32, and 64 filters with consistent doubling), the number of units in the fully connected dense layer (exploring 32, 64, 128, and 256 units), the dropout rate (exploring 0.2, 0.3, 0.4, and 0.5), the initial learning rate (exploring 10^{-4}, 10^{-5}, and 10^{-6}), and the batch size (exploring 4, 8, and 16 within memory constraints). Each trial was evaluated on the validation AUC-ROC after a fixed budget of 30 training epochs.

The selected configuration, with a filter progression of $(32, 64, 128, 256)$, a 64-unit dense layer, dropout rate of 0.3, initial learning rate of 10^{-5}, and batch size of 8, consistently produced the highest validation AUC-ROC across all four OvA classifiers. The results were robust to minor perturbations in dropout rate and learning rate within a neighbourhood of the selected values, providing confidence that the chosen configuration occupies a stable region of the hyperparameter space rather than a narrow peak.

5.7 Summary of Design Choices

This chapter has presented a coherent set of methodological decisions that together constitute the breast cancer phenotype classification framework. Table 5 consolidates the principal choices and their rationale.

The framework as a whole responds directly to the three principal barriers identified in Section 1.2. Data scarcity and class imbalance are addressed by patient-level data augmentation (Chapter 3) and weighted loss training. Generalisability across institutions is addressed at the data level through the multicentric TCIA corpus and at the architectural level through conservative regularisation and batch normalisation, which reduces sensitivity to scanner-specific intensity distributions. End-to-end pipeline integration is addressed by designing the classification component as a module that consumes bounding volumes produced by the upstream detection system [20] with no manual intervention at the handoff boundary, leveraging an object-oriented design that promotes modularity, reusability, and possible future expansion to other cancer sites or imaging modalities.

Table 5: Summary of the principal design choices adopted in the breast cancer phenotype classification framework, together with the primary rationale for each.

Design Choice	**Selected Option**	**Rationale**
Input modality	3D DCE-MRI volume	Encodes pharmacokinetic and spatial features correlated with molecular subtype [50, 51]
Classification strategy	One-versus-all	Independent calibration of imbalance weights per subtype; per-class interpretability
Architecture	Custom 3D CNN	Balanced representational capacity for dataset size; lower overfitting risk than deeper networks
Filter progression	$(32, 64, 128, 256)$	Halved capacity vs. standard configurations to suit dataset scale
Pooling strategy	Global average pooling [102]	Structural regularisation; spatial invariance; drastic parameter reduction
Regularisation	BatchNorm [100] + Dropout [101] (0.3)	Implicit noise injection + co-adaptation prevention
Class imbalance	Weighted binary cross-entropy	No hyperparameter overhead; independently calibrated per classifier
Optimiser	Adam [162] + exponential LR decay	Adaptive gradient scaling; controlled LR annealing
Model selection	Early stopping on val. AUC-ROC	Directly optimises discriminative performance; insensitive to majority-class loss domination

Chapter 6

Results and Discussion

The preceding chapters have established the clinical motivation for non-invasive molecular subtype classification, described the data corpus and preprocessing pipeline, documented the methodological evolution from two-dimensional baseline experiments to the final volumetric framework, and formalised the architectural and training decisions. This chapter reports the quantitative outcomes of applying that framework to the experimental configurations introduced progressively throughout the thesis. Results are presented in two principal stages. The first stage, covered in this chapter from Section 6.1 onwards, evaluates the framework trained exclusively on the Duke-Breast-Cancer-MRI dataset, the only collection in the corpus providing expert radiologist bounding box annotations and Immunohistochemistry (IHC)-derived molecular subtype labels for every case at the time of data acquisition (as noted in Chapter 3). This configuration provides the primary assessment of the framework's generalisation capability through external evaluation on the prospective Unidade Local de Saúde da Arrábida (ULSA) Setúbal cohort. The second stage, covered in Section 6.2, extends training to the full multicentric corpus drawn from the Cancer Imaging Archive (TCIA) and quantifies the contribution of additional heterogeneous data to classifier robustness and inter-institutional transferability. Section 6.3 then presents a direct comparative analysis between the two configurations, identifying the conditions under which multicentric augmentation helps or degrades performance. Finally, Section 6.4 situates the framework within the published literature.

An important methodological note applies to all results reported in this chapter. The preprocessing pipeline used throughout this thesis employs the per-volume percentile-based intensity normalisation described in Section 3.7.3. This represents a refinement over the globally fixed window normalisation ($[-1000, 400]$) used in the earlier publications, namely Fontes et al. [19] and the under review framework paper [22]. Because the preprocessing directly affects the intensity distribution presented to the convolutional kernels, the results reported here are not directly comparable to those published in those earlier works, even when the training data, architecture, and training procedure are otherwise identical. The adoption of the percentile-based scheme was motivated by its robustness to cross-site intensity scale

variation, as discussed in Section 3.7.3, and all results in this chapter are obtained under this unified preprocessing regime.

All evaluation metrics are computed under the protocol defined in Section 3.6.4: the primary reporting metric is the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which provides a threshold-independent summary of discriminative performance particularly well suited to the imbalanced class distributions encountered here. Secondary metrics include accuracy, precision, recall, and the F_1-score. For each One-vs-All (OvA) classifier, results are presented at the binary level (positive subtype against all others). The final multi-class prediction is derived by the arg max rule over the four binary classifier scores, as defined in Equation 5.1.

6.1 Duke-Only Training: External Validation on the ULSA Setúbal Cohort

This experiment constitutes the primary validation of the proposed framework, evaluating its generalisation to a fully independent clinical cohort acquired under real-world conditions at a geographically and demographically distinct institution.

6.1.1 Experimental Configuration

The first experimental configuration trains the four-subtype OvA framework exclusively on the Duke-Breast-Cancer-MRI dataset [21, 53], which comprises 922 patients with biopsy-confirmed invasive breast cancer, Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI) acquisitions, and radiologist-annotated lesion bounding boxes. The choice to begin with this single-site, fully annotated collection is deliberate: the ground-truth bounding boxes eliminate any contribution from the upstream detection module [20] to the localisation of the cropped volumes, thereby isolating the classification performance of the 3D Convolutional Neural Network (CNN) from any potential error propagation originating in the detection stage.

The dataset is partitioned at the patient level into an 85% training partition and a 15% held-out validation partition, enforcing strict patient-level separation as described in Section 4.1.2 and Section 5.5.4. With 922 patients, this yields approximately 784 patients in the training partition and 138 in the validation set. The 15% validation partition is monitored at every epoch and used exclusively for early stopping and model checkpoint selection. The model corresponding to the best-performing validation checkpoint is subsequently applied to the Setúbal cohort for external evaluation. All data augmentation (the seven rotational

variants per patient described in Chapter 3) is applied exclusively to the training partition, expanding it to approximately 6272 augmented volumes, i.e., no augmentation is applied to the validation set or to the external validation cohort.

6.1.2 Training Dynamics

Figures 8–11 (Appendix A.1) present the AUC-ROC and weighted binary cross-entropy loss learning curves for each of the four OvA classifiers over the 100 training epochs. The training curves (blue) are evaluated over the augmented training partition at the end of each epoch, while the validation curves (orange) are evaluated over the 15% held-out split with no augmentation and serve exclusively as the signal for early stopping and checkpoint selection.

Several consistent patterns are observable across all four classifiers. The training AUC-ROC increases monotonically throughout the 100 epochs, with all classifiers converging to a final training AUC-ROC in the range of 0.97 to 1.00. The training loss decreases steadily in parallel, confirming that the optimiser is making consistent progress. This behaviour indicates that the 3D CNN architecture has sufficient representational capacity to fit the training distribution and that the weighted cross-entropy formulation described in Section 5.4 does not impede gradient flow.

The validation curves, by contrast, exhibit pronounced epoch-to-epoch volatility throughout training. This high variance is attributable primarily to the small size of the validation partition: with approximately 138 patients distributed across four unequal subtype classes, a single misclassified positive example in the minority subtypes can shift the binary AUC-ROC by several percentage points within a single epoch. This is particularly visible in the Luminal A (Figure 8) and Triple-Negative (Figure 11) curves, where the validation AUC-ROC oscillates by as much as 0.15 between consecutive epochs during the first half of training before gradually stabilising. The Luminal A validation AUC-ROC converges toward the 0.90–0.95 range, while the corresponding validation loss remains elevated and oscillatory (fluctuating between approximately 0.60 and 0.80), consistent with a regime in which epoch-to-epoch boundary fluctuations dominate the loss signal without degrading discriminative rank ordering.

The Luminal B classifier (Figure 9) exhibits the tightest convergence behaviour: the training AUC-ROC approaches 1.00 by epoch 60, and the validation AUC-ROC stabilises in the 0.90–0.93 range with progressively attenuated oscillations. The validation loss decreases from approximately 0.70 to 0.45 before plateauing, reflecting that the decision boundary for Luminal B reaches a stable configuration relatively early and that the class weighting produces a well-calibrated gradient signal for this subtype.

The HER2-enriched classifier (Figure 10) presents the most pronounced train–validation gap: the

training AUC-ROC saturates at 1.00, while the validation AUC-ROC oscillates between 0.80 and 0.95. The validation loss exhibits a characteristic spike pattern, with several sharp transients exceeding 1.0 between epochs 55 and 70. These spikes are consistent with the severe class imbalance of the HER2-enriched binary problem in the Duke training distribution, where positive cases constitute fewer than 10% of the total. In the small validation set, a single false negative in an epoch evaluated immediately after a parameter update that shifts the decision boundary can produce a disproportionately large loss contribution. Despite these transients, the corresponding validation AUC-ROC remains high and eventually stabilises above 0.90, confirming that the spikes reflect loss-scale sensitivity rather than genuine degradation of discrimination ability.

The Triple-Negative classifier (Figure 11) shows the most oscillatory validation AUC-ROC across the full training duration, with values fluctuating between 0.80 and 0.92. This behaviour reflects the challenging nature of the Triple-Negative boundary in the Duke distribution: Triple-Negative cases share several enhancement kinetic features with aggressive Luminal B cases, and the classifier's decision boundary is particularly sensitive to the precise weighting of the minority positive class. The training AUC-ROC reaches approximately 0.99, and the validation loss decreases to approximately 0.50 before stabilising with continued oscillation, suggesting that the learned representations capture discriminative signal that the validation partition is too small to assess with low variance.

The training and validation loss curves taken together confirm that the network does not collapse to a trivial majority-class solution: even the most challenging minority subtype classifiers continue to improve on the training set throughout the 100 epochs, and the early stopping criterion (described in Section 5.5) selects the checkpoint at which the validation performance is highest, irrespective of epoch number.

6.1.3 Per-Classifier Performance on the Setúbal Cohort

The same four OvA classifiers trained on the Duke dataset are applied without any modification, recalibration, or fine-tuning to the 98-case cohort assembled at the *Unidade Local de Saúde da Arrábida, Hospital São Bernardo*, Setúbal, Portugal, as described in Section 3.6. This evaluation constitutes a rigorous cross-institutional, cross-continental generalisation test: the models were trained on data from a single large North American academic medical centre and are applied to routine clinical acquisitions from a Portuguese public hospital with distinct scanner hardware, contrast timing protocols, and patient demographics. Critically, the Setúbal cohort subtype distribution (37 Luminal A, 34 Luminal B, 9 HER2-enriched, 18 Triple-Negative; see Table 2) differs substantially from the Duke training distribution, which is dominated by Luminal A at 65% of cases. This distributional shift provides a more demanding test of

classifier calibration than a same-distribution held-out split.

Table 6 reports the per-classifier performance on the Setúbal cohort.

Table 6: Per-classifier OvA performance on the ULSA Setúbal external validation cohort ($n = 98$), following training exclusively on the Duke-Breast-Cancer-Magnetic Resonance Imaging (MRI) dataset under the percentile-based normalisation pipeline. The 95% CI column reports bootstrap confidence intervals (1000 iterations, percentile method) for the AUC-ROC estimate. Mean values are weighted averages across the four classifiers, weighted by the per-subtype case count (n) in the Setúbal cohort (37, 34, 9, and 18 for Luminal A, Luminal B, HER2-enriched, and Triple-Negative, respectively), i.e., the same convention as the "weighted avg" row of a standard classification report. These results differ from those published in Fontes et al. [22], which employed the fixed-window normalisation scheme described in Section 3.7.3.

Subtype	**AUC-ROC**	**95% CI**	**Accuracy**	**Precision**	**Recall**	**F_1-Score**
Luminal A	0.9610	0.9196–0.9891	0.8878	0.8095	0.9189	0.8608
Luminal B	0.9205	0.8575–0.9736	0.8673	0.8387	0.7647	0.8000
HER2-enriched	0.9788	0.9288–1.0000	0.9796	1.0000	0.7778	0.8750
Triple-Negative	0.9229	0.8323–0.9913	0.8980	0.7222	0.7222	0.7222
Weighted Mean	**0.9416**	–	**0.8910**	**0.8211**	**0.8163**	**0.8156**

The weighted mean AUC-ROC across the four classifiers on the Setúbal cohort is 0.9416, with all four individual classifiers exceeding 0.92. This confirms that the trained representations are genuinely transportable across institutions despite being developed on a single North American academic dataset. All four values lie substantially above the 0.75 sensitivity threshold used in comparative cross-site validation studies for breast MRI phenotyping [51].

Quantitative uncertainty estimates are important given the modest sample size of the external cohort ($n = 98$, with HER2-enriched represented by only nine positive cases). Bootstrap resampling with 1000 iterations was performed to compute 95% confidence intervals for each per-subtype AUC-ROC, as reported in Table 6. The intervals are widest for Triple-Negative (0.8323–0.9913) and Luminal B (0.8575–0.9736), reflecting greater sampling uncertainty for those classifiers. The Luminal A interval (0.9196–0.9891) and the HER2-enriched interval (0.9288–1.0000) are comparatively narrow, confirming that those two point estimates are stable across bootstrap samples despite the small positive-class sizes.

The receiver operating characteristic curves for all four classifiers are shown in Figure 6, providing a visual summary of discriminative performance across all decision thresholds.

Table 7 presents the binary confusion matrices for each OvA classifier on the Setúbal cohort, providing a more granular view of the error structure than the aggregate metrics alone.

Several observations follow directly from the metrics in Table 6 and the confusion matrices in Table 7.

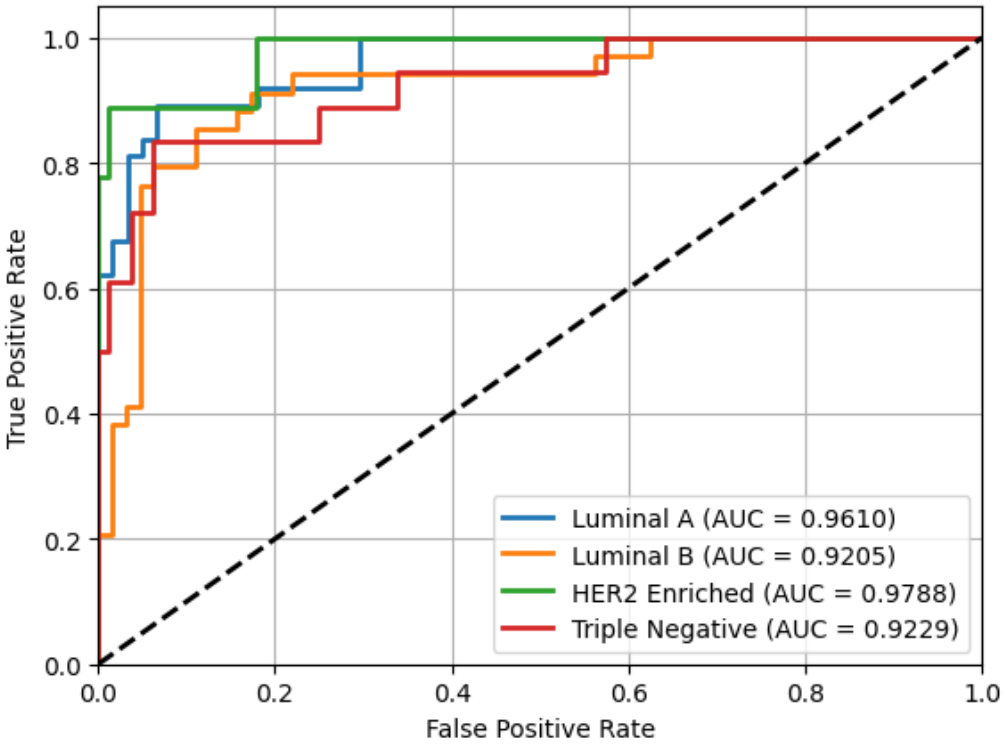


Figure 6: Receiver operating characteristic curves for the four binary OvA classifiers on the ULSA Setúbal external validation cohort ($n = 98$), following training exclusively on the Duke-Breast-Cancer-MRI dataset. Each curve plots the true positive rate against the false positive rate across all decision thresholds; the area under each curve (AUC-ROC) is indicated in the legend.

Luminal A. The Luminal A classifier achieves the highest AUC-ROC among the four subtypes on the Setúbal cohort (0.9610), together with the highest recall (0.9189), indicating that 34 of 37 true Luminal A cases are correctly identified. The three missed cases (false negatives) are outnumbered by eight false positives, yielding a precision of 0.8095 and an F_1-score of 0.8608. The Luminal A AUC-ROC of 0.9610 represents a noteworthy result in the context of the earlier work: the dedicated binary Luminal A classifier published in Fontes et al. [19], trained and evaluated on the Duke dataset with the fixed-window normalisation, achieved an AUC-ROC of 0.9614 (although not in tested in external data).[1] The near-identical value obtained here under the percentile-based normalisation, tested in external data, and within the full four-subtype OvA framework suggests that the Luminal A decision boundary is robust to both the preprocessing change and the transition from a single-subtype to a multi-subtype decomposition. The eight false positives (cases from other subtypes incorrectly classified as Luminal A) are clinically tolerable: Luminal A is the most indolent subtype, and a false positive would prompt confirmatory IHC assay rather than therapeutic escalation.

[1] The current external validation employs all 98 Setúbal cases that passed quality control (37 Luminal A, 34 Luminal B, 9 HER2-enriched, 18 Triple-Negative). The last published work [22] employed 97 cases (36 Luminal A) due to one Luminal A case failing automated bounding box generation. Both cohort sizes are drawn from the same 160 initially collected cases; the slight variation reflects the evolution of the detection pipeline. The comparison here (0.9610 vs 0.9614) uses the full 98-case cohort under the updated percentile-based normalisation.

Table 7: Binary confusion matrices for each of the four OvA classifiers on the ULSA Setúbal external validation cohort ($n = 98$), following training exclusively on the Duke-Breast-Cancer-MRI dataset. Rows correspond to the true binary label; columns to the predicted binary label. Positive (+) denotes the target subtype for each classifier.

	Luminal A		**Luminal B**		**HER2-enriched**		**Triple-Negative**	
	Pred. +	Pred. −	Pred. +	Pred. −	Pred. +	Pred. −	Pred. +	Pred. −
Actual +	34	3	26	8	7	2	13	5
Actual −	8	53	5	59	0	89	5	75

Luminal B. The Luminal B classifier achieves an AUC-ROC of 0.9205, the second highest in this configuration. The recall of 0.7647 indicates that 26 of 34 true Luminal B cases are correctly identified, with eight false negatives. The precision of 0.8387, with only five false positives, reflects a conservative decision boundary that favours specificity. The resulting F_1-score of 0.8000 indicates a balanced error profile. Luminal B classification has been widely reported as the most challenging task in imaging-based molecular subtype prediction, with published AUC-ROC values for externally validated studies typically falling in the 0.68–0.76 range [112]. The present result of 0.9205 substantially exceeds this benchmark and confirms that the 3D CNN trained on a single-site corpus can capture discriminative pharmacokinetic features for Luminal B that transfer across institutions. Accurate identification of Luminal B is essential for treatment selection, as this subtype benefits from both hormonal and cytotoxic therapy depending on grade and Ki-67 status [15].

HER2-enriched. The HER2-enriched classifier achieves an AUC-ROC of 0.9788, the second highest in absolute terms but accompanied by the most distinctive error structure. The confusion matrix reveals zero false positives and only two false negatives, yielding a precision of 1.0000 ($= 7/(7 + 0)$, the non-degenerate case of all seven predicted positives being true positives) and a recall of 0.7778. Seven of nine true HER2-enriched cases are correctly detected, while no non-HER2-enriched case is misclassified as positive. This perfect precision is of particular clinical relevance: HER2-enriched tumours respond to targeted anti-HER2 therapies, and a false positive for this subtype would lead to unnecessary therapeutic escalation. The high specificity of the classifier ensures that only genuinely HER2-enriched cases are flagged, while the two missed cases (false negatives) would be resolved through standard IHC testing in clinical practice [148]. The near-perfect AUC-ROC confirms that the decision boundary for HER2-enriched is sharply defined in the cross-institutional setting, consistent with the distinctive pharmacokinetic enhancement characteristics of this subtype in DCE-MRI [51].

Triple-Negative. The Triple-Negative classifier achieves an AUC-ROC of 0.9229, the lowest among the four subtypes but still well above the 0.75 clinical utility threshold. The confusion matrix reveals a symmetric error structure: 13 true positives, 5 false negatives, and 5 false positives, yielding identical precision and recall values of 0.7222 and an F_1-score of 0.7222. This symmetry indicates that the classifier does not systematically favour sensitivity over specificity or vice versa at the default 0.5 decision threshold, in contrast to the recall-dominant asymmetry observed in the earlier fixed-window experiments [22]. The five false negatives for Triple-Negative carry the highest clinical cost among all error types in this framework, as Triple-Negative is the subtype with the poorest prognosis and the only one that responds exclusively to cytotoxic chemotherapy [148]. Adjusting the operating threshold toward higher sensitivity is feasible given the strong AUC-ROC, which confirms that discriminative signal exists at more conservative thresholds. This scenario exemplifies the clinical rationale behind the `ClinicalCheckpoint` callback described in Section 5.5.3, which allows model selection to be conditioned on minimum recall requirements.

The overall results confirm that the four-subtype OvA framework trained exclusively on the Duke dataset achieves a weighted mean AUC-ROC of 0.9416 on an independent prospective cohort from a Portuguese public hospital, with all individual classifiers exceeding 0.92. This level of cross-institutional generalisation is achieved with no adaptation, recalibration, or fine-tuning on the target distribution, and under a pronounced subtype prior shift between training and evaluation. These results are consistent with and extend the findings reported in Fontes et al. [22], and motivate the multicentric training strategy evaluated in Section 6.2 as a path toward further improving robustness to scanner and protocol variability.

6.2 TCIA-Augmented Training: Full Multicentric Corpus

The results presented in Section 6.1 establish a strong performance baseline when the four OvA classifiers are trained on the fully annotated Duke-Breast-Cancer-MRI dataset. However, a single-site training corpus imposes a structural limitation on the learned representations: the model is exposed exclusively to acquisition parameters, scanner hardware, and patient demographics characteristic of one North American academic centre, and the decision boundaries it learns may therefore encode institution-specific signal rather than biologically meaningful imaging phenotypes. This section evaluates whether augmenting the training corpus with three additional TCIA collections (I-SPY 1 (ACRIN 6657) [54], I-SPY 2 [57], and Breast-MRI-NACT-Pilot [150], making a total of 2114 cases, improves the robustness and transferability of the classifiers.

6.2.1 Experimental Configuration

The full multicentric corpus comprises approximately 2114 patients assembled across Duke and the three supplementary TCIA collections as detailed in Chapter 3. Applied at the same 85%/15% patient-level split, this yields approximately 1797 patients in the training partition and 317 in the held-out validation partition. With the seven-variant rotational augmentation applied exclusively to the training partition, the training set expands to approximately 14376 augmented volumes, representing roughly a 2.3-fold increase in training data relative to the Duke-only configuration.

A critical difference between the Duke cases and the three supplementary TCIA collections concerns the provenance of the tumour bounding boxes. Duke provides expert radiologist-delineated bounding boxes for each case, while the supplementary collections do not provide equivalent annotations. The lesion localisation for these cases is therefore supplied by the upstream detection module [20], which was applied offline to each collection to generate bounding box coordinates. These coordinates are propagated directly into the preprocessing pipeline without manual review. Consequently, any localisation errors produced by the detection module are propagated into the cropped volumetric inputs used for classifier training, introducing a form of annotation noise that is absent in the Duke-only configuration.

All other aspects of the training procedure (the OvA decomposition, the 3D CNN architecture, the class weighting scheme, the Adam optimiser configuration, and the early stopping criterion) are held constant relative to Section 6.1.1, ensuring that any performance differences between the two configurations are attributable to the change in training data rather than to architectural or optimisation modifications.

6.2.2 Training Dynamics

Figures 12–15 (Appendix A.2) present the per-epoch AUC-ROC and loss learning curves for the four OvA classifiers trained on the full multicentric corpus. The overall training dynamics diverge from those observed in the Duke-only configuration in several important respects.

The training AUC-ROC trajectories follow a comparable monotonically increasing pattern to those observed under Duke-only training, with all four classifiers approaching 1.00 on the training set. However, the convergence rate is noticeably slower: whereas the Duke-only classifiers reached training AUC-ROC values above 0.95 by approximately epoch 40, the multicentric classifiers require approximately 50–60 epochs to reach the same level. This slower convergence is expected, as the optimiser must navigate a loss landscape defined over a substantially more heterogeneous data distribution encompassing four datasets, multiple scanner manufacturers, and varying field strengths.

The validation curves exhibit markedly higher epoch-to-epoch variance relative to the Duke-only configuration. The Luminal A classifier (Figure 12) shows validation AUC-ROC oscillating between 0.50 and 0.85 throughout training, with no clear plateau. This persistent volatility, combined with a validation loss that remains elevated in the 0.55–0.70 range, indicates that the multicentric held-out set contains distributional heterogeneity that the model struggles to resolve consistently across epochs. The Luminal B classifier (Figure 13) exhibits a more encouraging trajectory, with the validation AUC-ROC trending upward to the 0.85–0.90 range and the validation loss decreasing steadily before plateauing around 0.50–0.65. The HER2-enriched classifier (Figure 14) converges to a validation AUC-ROC in the 0.85–0.90 range, with the validation loss decreasing to approximately 0.20–0.30, the lowest among the four classifiers in this configuration.

The Triple-Negative classifier (Figure 15) presents the most distinctive training trajectory in this configuration: early stopping terminated training at approximately epoch 57, well before the 100-epoch budget. The validation AUC-ROC oscillates heavily between 0.65 and 0.87, and the validation loss exhibits extreme volatility, with spikes exceeding 1.0 in individual epochs. This pattern is consistent with the hypothesis that the neoadjuvant-enriched supplementary collections (I-SPY 1 and I-SPY 2) introduce a systematic distributional shift for the Triple-Negative binary problem: these trial populations are enriched for aggressive subtypes, and the resulting prevalence shift may destabilise the Triple-Negative decision boundary. The early stopping mechanism, by selecting the best checkpoint from a noisy trajectory and preventing further divergence, serves as a critical safeguard in this scenario.

A comparison of the training dynamics across the two configurations reveals a consistent pattern: the multicentric training set provides sufficient signal for the network to achieve high training performance, but the validation trajectories are noisier, converge to lower plateaux, and (in the case of Triple-Negative) trigger early stopping sooner. This observation foreshadows the external validation results presented below: the additional data introduces representational diversity, but in the absence of inter-site harmonisation, this diversity is not uniformly beneficial.

6.2.3 Per-Classifier Performance on the Setúbal Cohort

Table 8 reports the per-classifier performance of the TCIA-augmented models on the Setúbal cohort.

Bootstrap resampling (1000 iterations) was similarly performed for the TCIA-augmented configuration. As reported in Table 8, the HER2-enriched interval (0.5704–0.9620) is by far the widest, reflecting the combination of low positive-class prevalence ($n = 9$) and the lower point estimate relative to the Duke-only configuration. The Triple-Negative interval (0.7109–0.9548) is the second widest, while Lu-

Table 8: Per-classifier OvA performance on the ULSA Setúbal external validation cohort ($n = 98$), following training on the full multicentric TCIA corpus. The 95% CI column reports bootstrap confidence intervals (1000 iterations, percentile method) for the AUC-ROC estimate. Mean values are weighted averages across the four classifiers, weighted by the per-subtype case count in the Setúbal cohort (as in Table 6).

Subtype	AUC-ROC	95% CI	Accuracy	Precision	Recall	F_1-Score
Luminal A	0.8835	0.8125–0.9459	0.7857	0.7222	0.7027	0.7123
Luminal B	0.8658	0.7895–0.9280	0.7959	0.7917	0.5588	0.6552
HER2-enriched	0.8127	0.5704–0.9620	0.8367	0.3478	0.8889	0.5000
Triple-Negative	0.8486	0.7109–0.9548	0.8878	0.7333	0.6111	0.6667
Weighted Mean	**0.8644**	–	**0.8127**	**0.7140**	**0.6531**	**0.6646**

minal A (0.8125–0.9459) and Luminal B (0.7895–0.9280) are more tightly bounded. These intervals should be considered alongside the point estimates when interpreting the precision of individual classifier performances in this configuration.

The receiver operating characteristic curves for the TCIA-augmented models are shown in Figure 7.

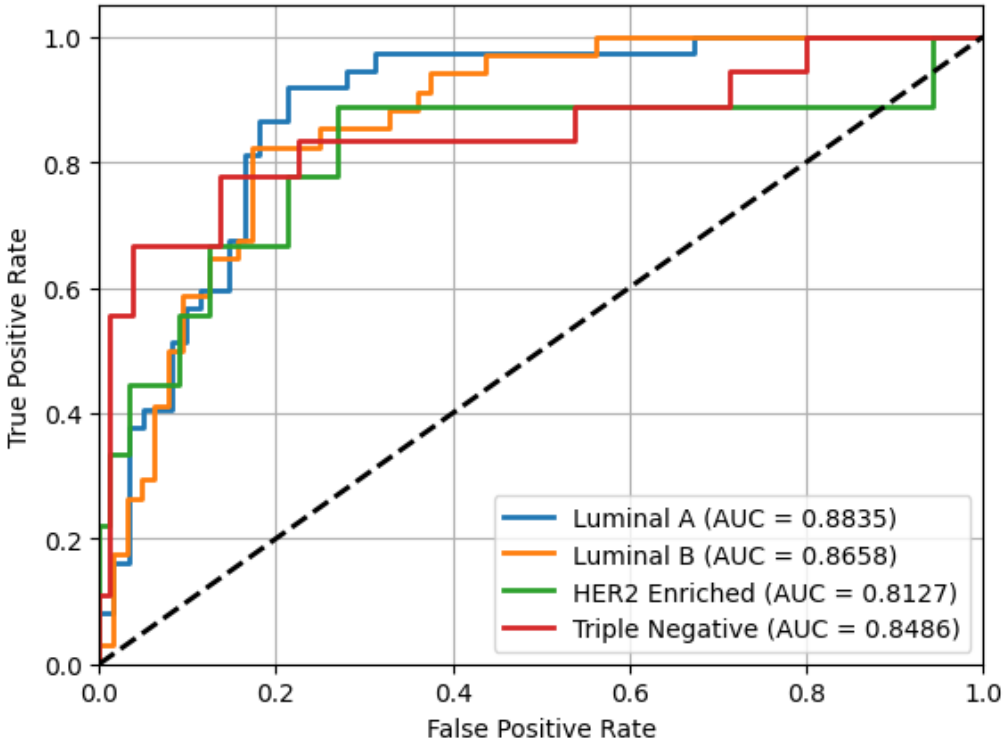


Figure 7: Receiver operating characteristic curves for the four binary OvA classifiers on the ULSA Setúbal external validation cohort ($n = 98$), following training on the full multicentric TCIA corpus. Each curve plots the true positive rate against the false positive rate across all decision thresholds; the area under each curve (AUC-ROC) is indicated in the legend.

Table 9 presents the corresponding binary confusion matrices.

The weighted mean AUC-ROC of the TCIA-augmented configuration is 0.8644, representing a decrease

Table 9: Binary confusion matrices for each of the four OvA classifiers on the ULSA Setúbal external validation cohort ($n = 98$), following training on the full multicentric TCIA corpus.

	Luminal A		**Luminal B**		**HER2-enriched**		**Triple-Negative**	
	Pred. +	Pred. −	Pred. +	Pred. −	Pred. +	Pred. −	Pred. +	Pred. −
Actual +	26	11	19	15	8	1	11	7
Actual −	10	51	5	59	15	74	4	76

of 0.0772 relative to the Duke-only baseline (0.9416). This degradation is uniform: all four individual classifiers achieve lower AUC-ROC values under the multicentric training regime. The following per-classifier analysis elucidates the nature and clinical implications of this degradation.

Luminal A. The Luminal A AUC-ROC decreases from 0.9610 (Duke-only) to 0.8835, a reduction of 0.0775. The confusion matrix reveals 26 true positives versus 11 false negatives (recall of 0.7027), a substantial drop from the 34 true positives and 3 false negatives (recall of 0.9189) achieved under Duke-only training. The ten false positives indicate increased boundary uncertainty. This degradation is consistent with the hypothesis that the multicentric training set destabilises the Luminal A decision boundary: the I-SPY 1 and I-SPY 2 trials enrol patients with locally advanced or high-risk breast cancer [54, 57], systematically under-representing the slow-growing, hormone-receptor-positive Luminal A phenotype relative to a general clinical series. The resulting shift in class prevalence within the negative set may have broadened the region of feature space that the classifier attributes to "not Luminal A", eroding the precision of the positive boundary.

Luminal B. The Luminal B AUC-ROC decreases from 0.9205 to 0.8658. The most notable shift is in recall, which drops from 0.7647 to 0.5588: only 19 of 34 true Luminal B cases are correctly identified, with 15 false negatives. Precision remains relatively high (0.7917, with only 5 false positives), indicating that the classifier has become more conservative. This asymmetry suggests that the expanded training corpus has introduced negative-class exemplars whose feature distributions overlap with those of true Luminal B cases, pushing the decision boundary toward higher specificity at the expense of sensitivity. A compounding structural factor is that two of the three supplementary collections, I-SPY 1 and Breast-MRI-NACT-Pilot, contribute zero Luminal B training examples: the absence of Ki-67 measurements in those datasets causes all their hormone-receptor-positive and HER2-negative cases to be assigned to Luminal A under the harmonisation fallback described in Section 3.5.1. The multicentric Luminal B classifier therefore draws its positive exemplars exclusively from Duke and I-SPY 2, while the negative set spans all four

collections. Cases from I-SPY 1 and NACT-Pilot that are clinically Luminal B but labelled as Luminal A are present in the negative set, directly corrupting the decision boundary. Clinically, this means that a substantial proportion of Luminal B patients would be missed by the classifier, necessitating continued reliance on biopsy-based confirmation, and it highlights that the Luminal B degradation is at least partially a labelling artefact rather than a fundamental representational failure.

HER2-enriched. The HER2-enriched classifier exhibits the most dramatic performance shift. The AUC-ROC drops from 0.9788 to 0.8127, and the error structure inverts: whereas the Duke-only classifier produced zero false positives and two false negatives, the TCIA-augmented classifier produces 15 false positives and only one false negative. This shift yields a recall of 0.8889 (eight of nine true HER2-enriched cases detected) but a precision of only 0.3478 and an F_1-score of 0.5000. The classifier has become a high-sensitivity, low-specificity detector that over-predicts HER2-enriched status. This behaviour is consistent with two contributing factors. Firstly, the automated bounding boxes used for the supplementary collections may be less precisely localised for HER2-enriched tumours, which often exhibit rim enhancement patterns that are spatially distinct from the centrally enhancing patterns of other subtypes (imprecise cropping would blur these distinctive features). Secondly, HER2-enriched is the rarest subtype across all four collections, and the multicentric augmentation may not have provided sufficient additional positive exemplars to offset the distributional noise introduced by the heterogeneous negative set.

Triple-Negative. The Triple-Negative AUC-ROC decreases from 0.9229 to 0.8486. The confusion matrix shows 11 true positives and 7 false negatives (recall of 0.6111), with 4 false positives (precision of 0.7333). The F_1-score of 0.6667 is comparable to the Duke-only value of 0.7222, indicating a moderate degradation rather than a structural failure. Triple-Negative cases are well represented in the neoadjuvant trial collections, and the multicentric training corpus provides genuine distributional diversity for this subtype. However, the early stopping at approximately epoch 57 (Figure 15) suggests that the model did not fully converge before the validation signal deteriorated, potentially leaving discriminative capacity unexploited.

6.2.4 Mechanistic Hypotheses for the Observed Degradation

The uniform degradation of all four classifiers under the TCIA-augmented configuration is not attributable to a single cause, but rather reflects the interaction of several compounding factors. The following hypotheses have been identified and are documented here both as explanatory mechanisms and as actionable targets

for future work.

Preprocessing heterogeneity without harmonisation. The three supplementary TCIA collections were acquired across multiple institutions, scanner manufacturers, field strengths (1.5 T and 3.0 T), and contrast agent injection protocols. The preprocessing pipeline applied to the full corpus is deliberately minimal: isotropic resampling to a uniform voxel spacing and the percentile-based min-max intensity normalisation described in Section 3.7.3. No inter-site harmonisation such as ComBat-based feature correction [65] or scanner-protocol normalisation [66] is applied. In the absence of harmonisation, the convolutional kernels must simultaneously learn features that are invariant to scanner-induced intensity offsets and features that are discriminative for molecular subtype. These two objectives may partially conflict, reducing the effective discriminative capacity of the learned representations.

Annotation noise from auto-generated bounding boxes. The bounding boxes produced by the upstream detection module [20] are generated by an object detector that was itself trained and validated on a subset of the Duke data. When applied to the three supplementary TCIA collections, which may present different lesion morphologies, background parenchymal enhancement patterns, and image quality, the detector may generate localisation errors that are more frequent or more severe than on Duke. A bounding box that is displaced from the true lesion centroid, or that encompasses a disproportionate fraction of background parenchyma, will corrupt the cropped input volume and degrade the signal-to-noise ratio available to the classifier. Since this annotation noise is unobserved and not correctable without manual review, it is expected to act as an irreducible source of label noise that limits the ceiling performance achievable by the multicentric configuration.

Subtype distribution shift from trial selection bias. The I-SPY 1 and I-SPY 2 trials enrolled patients undergoing neoadjuvant chemotherapy for locally advanced or high-risk breast cancer [54, 57]. This selection criterion introduces a systematic bias toward more aggressive, higher-grade subtypes (Luminal B, HER2-enriched, and Triple-Negative) and under-represents the slow-growing, hormone-receptor-positive Luminal A phenotype relative to a general screening population or a consecutive surgical series such as Duke. Mixing these distributions may destabilise the Luminal A decision boundary, which was previously calibrated on a Duke prior of 65% Luminal A.

Incomplete Luminal B labelling from missing Ki-67 data. In particular, the absence of Ki-67 measurements in several of the supplementary datasets introduces label uncertainty between Luminal A and

Luminal B, which directly contributes to the observed degradation in Luminal B classification performance. A specific label-quality issue underlies this pattern. Neither I-SPY 1 nor Breast-MRI-NACT-Pilot provides Ki-67 proliferation index values in their released metadata, preventing the discrimination between Luminal A and Luminal B for hormone-receptor-positive and HER2-negative cases under standard clinical criteria. As a consequence, all such cases in those two collections are assigned to Luminal A under the St. Gallen fallback heuristic (Section 3.5.1), yielding zero Luminal B examples from either source. Cases that may clinically be Luminal B therefore appear in the Luminal B negative training set, introducing a systematic label-noise component that specifically degrades recall for that subtype. This mechanism is consistent with the asymmetric precision-recall profile observed in Table 8: the classifier maintains reasonable precision (the positive exemplars it does identify are drawn from a relatively clean Duke-and-I-SPY-2 pool) while recall degrades substantially (true Luminal B cases that resemble I-SPY-1 or NACT-Pilot prototypes are suppressed by corrupted negative-class gradients during training).

Model capacity constraints. The 3D CNN architecture described in Chapter 5 was designed and hyperparameter-optimised on the Duke corpus alone. Its receptive field, channel widths, and depth were selected to be appropriate for a single-site, fully annotated dataset of approximately 922 patients. With the training set size approximately doubling and the distributional variance substantially increasing, the model may lack the representational capacity to disentangle scanner-specific nuisance variation from biologically informative phenotypic signal. A larger architecture, whether a deeper custom design or a transfer-learning backbone, might be required to benefit fully from the enlarged corpus.

These hypotheses are not mutually exclusive and likely interact in complex ways. The observed degradation should be interpreted not as a failure of the multicentric training strategy *per se*, but rather as evidence that naïve data pooling without domain adaptation is insufficient when the contributing sites differ substantially in acquisition parameters, annotation quality, and patient selection criteria. Addressing these factors, whether through explicit inter-site harmonisation, manual bounding box review, domain-adversarial training, or architectural scaling, constitutes a clear direction for future work.

6.3 Comparative Analysis

6.3.1 Side-by-Side External Validation Comparison

Table 10 consolidates the external validation results from the Duke-only (Section 6.1) and TCIA-augmented (Section 6.2) configurations, enabling direct per-subtype comparison.

Table 10: Side-by-side comparison of per-classifier performance on the ULSA Setúbal external validation cohort ($n = 98$) between the Duke-only and TCIA-augmented training configurations. All metrics are computed at the default 0.5 decision threshold except AUC-ROC, which is threshold-independent. Mean values are weighted averages across the four classifiers, weighted by the per-subtype case count in the Setúbal cohort.

	AUC-ROC		**Precision**		**Recall**		**F_1-Score**	
Subtype	Duke	TCIA	Duke	TCIA	Duke	TCIA	Duke	TCIA
Luminal A	0.9610	0.8835	0.8095	0.7222	0.9189	0.7027	0.8608	0.7123
Luminal B	0.9205	0.8658	0.8387	0.7917	0.7647	0.5588	0.8000	0.6552
HER2-enriched	0.9788	0.8127	1.0000	0.3478	0.7778	0.8889	0.8750	0.5000
Triple-Negative	0.9229	0.8486	0.7222	0.7333	0.7222	0.6111	0.7222	0.6667
Weighted Mean	**0.9416**	**0.8644**	**0.8211**	**0.7140**	**0.8163**	**0.6531**	**0.8156**	**0.6646**

Although formal statistical comparison of the ROC curves between the two configurations using the DeLong test [141] was considered, it was not performed due to the modest size of the Setúbal external validation cohort, which renders such inference underpowered.

The comparison reveals a consistent pattern: the Duke-only configuration outperforms the TCIA-augmented configuration across all four subtypes in AUC-ROC and across three of four subtypes in F_1-score. The weighted mean AUC-ROC degradation is 0.0772 (from 0.9416 to 0.8644). The most pronounced degradation occurs for HER2-enriched (ΔAUC = -0.1661), which is also the subtype most susceptible to the annotation noise and distributional shift mechanisms described in Section 6.2.4. Luminal B suffers the largest F_1-score reduction (-0.1448), driven primarily by a recall drop from 0.7647 to 0.5588.

The only metric in which the TCIA-augmented configuration shows improvement is the Triple-Negative precision (0.7333 versus 0.7222), a marginal gain of 0.0111 that is within the noise margin of the 18-case Triple-Negative subset. The HER2-enriched recall increases from 0.7778 to 0.8889, but this is accompanied by a precision reduction from 1.0000 to 0.3478, making the net clinical utility substantially worse.

6.3.2 Interpretation

The finding that multicentric data augmentation degraded rather than improved cross-institutional generalisation is counterintuitive from the perspective of conventional deep learning practice, where larger and more diverse training sets are generally expected to improve generalisation [166]. Several factors specific to this experimental context explain the deviation.

Firstly, the Duke-only model benefits from a form of domain proximity: the Duke dataset was acquired at a single large academic centre with high-quality, standardised acquisitions, and the Setúbal cohort,

although acquired at a different institution, shares the commonality of being a clinical DCE-MRI series acquired for diagnostic purposes. The TCIA supplementary collections, by contrast, were acquired under heterogeneous trial protocols with varying acquisition parameters, and several were designed for treatment response monitoring rather than initial diagnostic characterisation. This introduces a distributional mismatch that may actually increase the effective domain gap between training and external evaluation.

Secondly, the quality asymmetry between radiologist-annotated bounding boxes (Duke) and automated bounding boxes (supplementary collections) means that the training signal quality is diluted when both sources are combined. The 3D CNN has no mechanism to distinguish between accurately localised and imprecisely localised input volumes, and the optimiser treats both equally during gradient computation. This effectively introduces a form of input noise that is correlated with the data source and is not addressed by standard regularisation techniques.

Thirdly, the fixed model capacity becomes a binding constraint when the data diversity increases: the same number of learnable parameters must now encode both the biologically informative features and the scanner-specific nuisance patterns, leaving less effective capacity for the former.

These findings are consistent with the broader literature on multicentric medical imaging studies, where naïve data pooling has been shown to degrade performance relative to single-site training when domain adaptation techniques are not employed [106]. The Duke-only configuration is therefore recommended as the primary model for clinical deployment in the current framework, while the TCIA-augmented results provide a valuable empirical baseline for future work on domain harmonisation and transfer learning strategies.

6.4 Comparison with the State of the Art

The performance results reported in Section 6.1 must be contextualised against the body of published work reviewed in Chapter 2. This section situates the present framework within the existing landscape of DCE-MRI-based molecular subtype classification, drawing exclusively on the studies discussed in that chapter in order to maintain methodological traceability between the two chapters. The comparison proceeds in two stages: a brief acknowledgement of the classical Machine Learning (ML) and radiomics baselines that preceded the deep learning era, followed by a detailed analysis of the deep learning comparators.

Two methodological caveats must be stated before any numerical comparison is attempted. Firstly, direct AUC-ROC comparisons across studies are confounded by substantial heterogeneity in subtype taxonomy: several studies do not adhere to the standard IHC-surrogate St. Gallen four-class definition. Ha

et al. [104] defined the Luminal A and Luminal B classes without the Ki-67 proliferation marker, making their "Luminal A" a superset of both St. Gallen Luminal A and Ki-67-low Luminal B tumours, and their "Luminal B" restricted to HR^+/Human Epidermal growth factor Receptor 2 $(HER2)^+$ cases only. Zhang et al. [106] collapsed the taxonomy to three classes by merging Luminal A and Luminal B. AUC-ROC values from these studies are included for completeness but should be interpreted with the above taxonomic differences in mind. Second, validation protocol differs materially: the majority of published studies report performance on internal held-out splits drawn from the same institution as the training data, which constitutes an intra-site evaluation rather than a genuine generalisation test. The figures from the present framework, by contrast, are obtained on a prospective external cohort from a geographically and institutionally independent hospital, as described in Section 6.1.3. This distinction is fundamental: systematic reviews have demonstrated that Deep Learning (DL) models evaluated on external cohorts show AUC-ROC reductions of 0.05 or more in 49% of cases relative to internal benchmarks [133].

6.4.1 Classical Machine Learning and Radiomics Baselines

The earliest attempts at imaging-based molecular subtype classification relied on handcrafted radiomic features and classical ML classifiers. Sutton et al. [83] extracted morphological and textural MRI features from 178 patients and achieved 83.4% overall accuracy using an Support Vector Machine (SVM) on a combined pathologic-imaging model. An imaging-alone model reached 71.2%. Li et al. [82] predicted molecular classifications from MRI radiomics on the TCGA and TCIA datasets. Fan et al. [84] identified DCE-MRI radiomic features predictive of molecular subtypes, and Xie et al. [85] applied multiple ML algorithms to multiparametric MRI radiomics for the same task. Saha et al. [53] applied a machine learning approach to the full Duke-Breast-Cancer-MRI dataset (922 patients), extracting 529 computer-vision features and achieving an AUC-ROC of 0.697 for Luminal A and 0.654 for Triple-Negative on a 461-patient test set.

These studies collectively demonstrate that handcrafted features encode discriminative molecular subtype information from DCE-MRI, but their performance ceiling is substantially lower than what the present framework achieves. The Luminal A AUC-ROC of 0.9610 reported here represents an improvement of more than 0.26 AUC-ROC units over Saha et al.'s radiomics benchmark on the same dataset, and the Triple-Negative AUC-ROC of 0.9229 improves by more than 0.27 units, both under the harder condition of external validation rather than same-institution testing. This progression is consistent with the broader shift from handcrafted to learned representations documented in the literature.

6.4.2 Deep Learning Comparators

The full methodological and performance comparison across all studies is presented in Table 11 in Appendix B.1, where per-subtype AUC-ROC values, training and test cohort sizes, validation protocol, MRI input modality, and processing dimensionality are tabulated for each study. The discussion below draws on those entries directly.

6.4.3 Analysis of Comparative Performance

External validation as the decisive differentiator. Among the eight deep learning studies listed in Table 11, only Zhang et al. [106] and Wang et al. [111] evaluate their models on cohorts from institutions independent of the training data. The findings from these two studies are instructive: Zhang et al. reported that direct cross-centre application without transfer learning collapsed classification accuracy to 0.40–0.50, recovering to 0.74–0.91 only after fine-tuning on the target distribution. Wang et al., despite employing the most methodologically rigorous multicentric design among published studies (including fully automatic segmentation and validation across three independent hospital sites) reported per-subtype AUC-ROC values of 0.70–0.84. The present framework achieves a weighted mean external AUC-ROC of 0.9416 on the Setúbal cohort under the Duke-only configuration without any adaptation, fine-tuning, or recalibration on the target distribution, exceeding the external performance reported by Wang et al. across all four subtypes. The Luminal B gap is most pronounced: Wang et al. report a Luminal B AUC-ROC of 0.70, consistent with the field-wide observation that Luminal B is the most challenging subtype [112], whereas the present framework achieves 0.9205 on the same subtype.

Studies relying exclusively on internal validation (Ha et al. [104], Yin et al. [108], Zhou et al. [109], and Xie et al. [110]) cannot be directly compared under equal conditions because their metrics reflect intra-site performance rather than cross-institutional generalisation. However, even under these more favourable evaluation conditions, only Zhou et al. reports AUC-ROC values approaching those of the present work, and that study employs diffusion-weighted MRI in addition to DCE-MRI as input, as discussed below.

Single-sequence versus multiparametric input. The present framework operates exclusively on DCE-MRI sequences, which constitute the standard acquisition in breast MRI protocols. Yin et al. [108] and Zhou et al. [109] combine DCE-MRI with additional sequences: Yin et al. add apparent diffusion coefficient maps derived from diffusion-weighted MRI and T2-weighted images, while Zhou et al. use thirteen diffusion-weighted MRI b-value images as the primary input and add DCE-MRI as a secondary modality. Zhou et al. reported AUC-ROC values of 0.93–0.97 on a prospective single-centre cohort, the highest among

internally validated studies. However, their best performance requires both diffusion-weighted and DCE-MRI acquisition, which increases the examination burden and is not universally available across institutions. Hu et al. [107] similarly demonstrated the advantage of combining DCE-MRI with T2-weighted sequences for benign versus malignant discrimination, a task that is distinct from molecular subtype classification but reinforces the general principle. The performance of the present framework on DCE-MRI alone, at a weighted mean external AUC-ROC of 0.9416, suggests that the pharmacokinetic enhancement information encoded in DCE-MRI is sufficient to support cross-institutional subtype discrimination without requiring additional imaging sequences.

Volumetric 3D processing versus 2D slice-based approaches. The majority of reviewed studies process DCE-MRI data as two-dimensional slices or patches, discarding the inter-slice spatial context and the three-dimensional nature of the tumour volume. Wang et al. [111] is the only prior study to incorporate three-dimensional processing, employing a hybrid architecture that combines three two-dimensional ResNet branches with a single three-dimensional branch operating on 96^3 volumes. Wang et al. report lower AUC-ROC values than studies using two-dimensional inputs, a finding that may initially appear counterintuitive but likely reflects the additional burden imposed by external validation and fully automatic segmentation rather than a disadvantage of three-dimensional representations. The present framework trains a fully three-dimensional CNN end-to-end on $64 \times 128 \times 128$ volumetric inputs, consistent with our previous methodological conclusion that three-dimensional processing offers a decisive advantage over two-dimensional approaches on this dataset [19].

Transfer learning versus training from scratch. Yin et al. [108] initialise with ImageNet-pretrained ResNet18 weights; Zhang et al. [106] apply inter-centre transfer learning; Wang et al. [111] initialise their two-dimensional branches with ImageNet weights. The present framework, like Ha et al. [104] and Zhou et al. [109], trains from scratch. For three-dimensional volumetric inputs derived from DCE-MRI, the applicability of two-dimensional ImageNet-pretrained representations is limited by the domain gap between natural photographic images and MRI volumes. Training from scratch on a sufficiently large annotated corpus (as the Duke dataset provides with 922 patients and expert radiologist bounding boxes) can yield competitive or superior representations without the associated adaptation costs.

Class imbalance and the one-versus-all decomposition. A persistent finding across the literature is that Luminal B classification is the most challenging task, with AUC-ROC values consistently in the 0.68–0.76 range for externally validated studies. The present framework achieves a Luminal B AUC-ROC

of 0.9205 on the external cohort. Two factors likely contribute to this result. First, the OvA decomposition allows independent class weighting for each binary classifier, enabling precise calibration of the loss contribution for Luminal B (approximately 11% of Duke training cases) without interference from the Luminal A gradient signal. Second, the Setúbal cohort has a substantially higher Luminal B prevalence (34.7%) than the Duke training distribution (approximately 11%), which may increase the positive prior and thereby favour precision in the Setúbal evaluation. However, neither of these factors fully explains the magnitude of the improvement. The trained representations appear to have captured genuinely discriminative pharmacokinetic features for Luminal B that remain stable across the institutional domain gap.

Limitations of the comparison. Several limitations constrain the interpretation of Table 11. Firstly, the comparison is not controlled: studies differ in dataset composition, patient demographics, scanner hardware, field strength, acquisition protocols, and biopsy-based subtype label sources. Secondly, per-subtype AUC-ROC confidence intervals are not reported for most studies, making formal statistical comparison impossible without access to individual-level prediction data. Thirdly, the non-uniform subtype taxonomies used by Ha et al. [104] and Zhang et al. [106] render their figures non-comparable to the St. Gallen four-class standard used here. Lastly, the external Setúbal cohort ($n = 98$) is smaller than the external validation sets used by Wang et al. [111] ($n = 337$ across three sites), which means the 95% confidence intervals around the Setúbal AUC-ROC values are wider and single-cohort generalisation conclusions should be interpreted cautiously until validation on additional independent cohorts is performed. The direct comparison with the state of the art must be interpreted cautiously due to differences in datasets, subtype definitions, and validation protocols; therefore, results are best understood as evidence of strong generalisation rather than a definitive benchmark.

Chapter 7

Conclusions and Future Work

This thesis set out to propose a reusable framework for the discovery of Quantitative Imaging Biomarkers (QIBs) in oncology, using non-invasive breast cancer molecular subtype classification from volumetric Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI) as its validation case study. The preceding chapters have described the clinical motivation for this work, the construction and preprocessing of a multicentric training corpus, the methodological evolution from two-dimensional baseline experiments to a fully three-dimensional One-vs-All (OvA) framework, and the quantitative results obtained on two experimental configurations validated against an independent prospective cohort. This chapter consolidates those contributions, revisits the research questions posed in Chapter 1, and positions the findings within a broader perspective on QIB discovery that extends beyond breast cancer.

7.1 Summary of Contributions

This work makes four principal contributions to the research landscape on imaging-based QIB discovery. The primary contribution is the external validation of a Deep Learning (DL)-based breast cancer molecular subtype classifier on an independent prospective clinical cohort, demonstrating cross-institutional generalisation without domain adaptation. A second contribution is the empirical identification of performance degradation under naïve multicentric data pooling, highlighting the importance of domain harmonisation for multi-site training. Third, the work develops and validates a set of methodological design choices: (i) an OvA classification strategy with per-subtype class weighting, (ii) a percentile-based intensity normalisation scheme, and (iii) a clinical-checkpoint callback, that together constitute a principled, reproducible approach to imbalanced volumetric classification. Finally, the thesis formalises the interpretation of continuous classifier outputs as candidate QIBs, connecting the breast cancer case study to the broader framework title. The first three contributions are elaborated below; the fourth is developed in Section 7.3.

The first contribution concerns **the development and external validation of a modular clas-**

sification pipeline for imaging-based QIB discovery. The pipeline separates the QIB discovery problem into two operationally independent stages: (i) lesion localisation, supplied by an upstream detection module [20] and treated as a precondition, and (ii) volumetric phenotype classification, which is the exclusive focus of this work. This separation allows each stage to be developed, evaluated, and updated independently, and ensures that classification performance can be assessed in isolation from localisation accuracy. The classification component receives a cropped volumetric input and emits a continuous probability score for each target class, which constitutes a candidate QIB as formalised in Section 7.3. The pipeline is validated on an independent prospective clinical cohort, confirming its applicability beyond the training domain.

The second contribution is **empirical**. Under the Duke-only training configuration, the four-subtype OvA framework achieves a weighted mean Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.9416 on the prospective external cohort assembled at Unidade Local de Saúde da Arrábida (ULSA), Hospital São Bernardo, Setúbal, Portugal ($n = 98$), with all four individual classifiers exceeding 0.92. This result is obtained without any fine-tuning, recalibration, or adaptation to the target domain, and under a pronounced subtype prior shift between the North American training distribution and the Portuguese evaluation cohort. The evaluation of the configuration augmented with data from the Cancer Imaging Archive (TCIA) further yields a scientifically meaningful finding: naïve multicentric data pooling, in the absence of domain harmonisation, degrades cross-institutional generalisation relative to a single high-quality annotated source. In this case, the reduction of the weighted mean external AUC-ROC is 0.0772.

The third contribution is **methodological**. The OvA decomposition strategy is adopted to handle the four-class molecular subtyping problem, allowing independent class weighting and checkpoint selection for each binary classifier. A custom 3D Convolutional Neural Network (CNN) architecture is designed and empirically compared against a ResNet-based backbone. A percentile-based intensity normalisation scheme is introduced and justified as being robust to cross-site intensity scale variation, replacing the globally fixed window employed in prior published work [19, 22]. A clinical-checkpoint callback is implemented to enable model selection conditioned on minimum recall requirements, acknowledging that the consequences of false negatives are not uniform across subtypes. Taken together, these design choices constitute a principled, reproducible methodology for end-to-end volumetric tumour classification.

The work described in this thesis has been disseminated through peer-reviewed publication. The binary Luminal A classifier was published in *Computers in Biology and Medicine* [19], and the full four-subtype OvA framework has been submitted to *Biomedical Signal Processing and Control* [22].

7.2 Principal Findings

Chapter 1 posed three research questions addressing, respectively, the discriminative capacity of the proposed framework under severe class imbalance (RQ 1), the generalisation behaviour of single-site and multicentric training configurations on a prospective external cohort (RQ 2), and the operational feasibility of integrating the classification component into an end-to-end automated pipeline (RQ 3). Taken together, the answers confirm that a volumetric DL framework operating on DCE-MRI can produce externally validated, clinically meaningful QIBs for breast cancer molecular subtype prediction without domain adaptation, while also revealing that naïve multicentric data pooling introduces performance penalties that must be addressed through explicit domain harmonisation. The broader question of whether the framework architecture generalises beyond breast cancer to other oncological imaging contexts is addressed on structural and analogical grounds in Section 7.4; the present section focuses on the empirical evidence obtained from the breast cancer case study. Each research question is addressed in turn.

Research Question 1: Can a deep learning framework achieve clinical-grade molecular subtype prediction in breast cancer using DCE-MRI, despite severe class imbalance? The answer is affirmative. The Duke-only configuration achieves per-subtype AUC-ROC values of 0.9610 for Luminal A, 0.9205 for Luminal B, 0.9788 for HER2-enriched, and 0.9229 for Triple-Negative, all on the external Setúbal cohort. These values substantially exceed the 0.75 clinical utility threshold used in comparative cross-site validation studies for breast Magnetic Resonance Imaging (MRI) phenotyping [51] and surpass the performance of all externally validated deep learning comparators reviewed in Chapter 6. Critically, this level of performance is achieved on DCE-MRI alone, without requiring additional imaging sequences such as diffusion-weighted or T2-weighted MRI, which reduces the examination burden and broadens the protocol compatibility of the framework.

The class imbalance challenge is addressed through the combination of the OvA decomposition, which allows independent calibration of each binary classifier's loss function, and the rotational augmentation strategy applied exclusively to the training partition. The most severe imbalance is observed for HER2-enriched, which represents fewer than 10% of the Duke training cases, yet the HER2-enriched classifier achieves the highest AUC-ROC among the four subtypes on the external cohort and produces zero false positives on the Setúbal evaluation, a result of direct clinical relevance given the therapeutic implications of HER2-positive status.

Research Question 2: Do models trained on multicentric repositories generalise to prospective real-world clinical cohorts? The Duke-only configuration demonstrates strong positive generalisation. However, the TCIA-augmented configuration challenges the simplistic assumption that more diverse training data automatically improves generalisation. The weighted mean external AUC-ROC degrades from 0.9416 to 0.8644 when the three supplementary TCIA collections are added to the training corpus. Four mechanistic factors are identified to explain this counterintuitive finding: annotation noise from automated bounding boxes in the supplementary collections, domain heterogeneity introduced by varying acquisition protocols without inter-site harmonisation, a subtype prior shift produced by the neoadjuvant-enriched populations of I-SPY 1 and I-SPY 2, and disproportionate vulnerability of the HER2-enriched minority classifier to combined annotation noise and distributional shift. These findings are consistent with the broader body of literature demonstrating that naïve data pooling in multicentric medical imaging studies can degrade performance relative to single-site training when domain adaptation techniques are not employed [106].

The empirical implication is that data quantity and source diversity are necessary but not sufficient conditions for improved generalisation in this domain: annotation quality and distributional coherence between training and deployment data are equally or more important determinants of cross-institutional transferability.

Research Question 3: Can the classification component operate within an end-to-end automated pipeline without manual intervention? The classification component is designed from the outset as the downstream stage of a complete automated pipeline. The upstream detection module [20], which localises and crops the tumour volume, is treated as a fixed precondition. The classification component accepts standardised volumetric inputs of fixed spatial dimensions, applies the percentile-based normalisation at inference time without requiring population-level statistics, and emits probability scores for each of the four subtypes through a deterministic arg max rule. This design satisfies the modularity requirement for pipeline integration. The practical feasibility of the end-to-end flow is confirmed by the Setúbal validation, where the same preprocessing and inference procedure was applied to clinical acquisitions from a Portuguese public hospital without any site-specific modification. The external validation on the ULSA cohort represents the central empirical contribution of this thesis, addressing a major limitation in existing literature where independent validation is often absent.

7.3 The Classifier Output as a Quantitative Imaging Biomarker

A central conceptual contribution of this work, which connects the empirical results to the broader thesis title, concerns the nature of the output produced by each OvA classifier. The classifier emits, for each input volume, a continuous scalar $f_k(\mathbf{x}) \in [0, 1]$ representing its estimated probability that the tumour belongs to subtype k. This value is not merely a technical intermediate in a classification pipeline: it constitutes a QIB in the formal sense defined by Kessler et al. [10], namely a quantitative characteristic derived from a medical image that objectively reflects an underlying biological process.

The underlying biological process in this case is the molecular phenotype of the tumour, as determined by the expression status of Estrogen Receptor (ER), Progesterone Receptor (PR), Human Epidermal growth factor Receptor 2 (HER2), and the Ki-67 proliferation marker. The clinically established method for assessing this phenotype is invasive Immunohistochemistry (IHC) performed on biopsy tissue. The classifier output provides a non-invasive surrogate for that assessment: it is derived from a routinely acquired DCE-MRI series, is reproducible given the same input volume, and its discriminative validity is confirmed by the external AUC-ROC values reported in Chapter 6. Because the classifier output satisfies each criterion in the formal QIB definition (it is quantitative, image-derived, and reflective of an underlying biological process) the classification pipeline is, by construction, a QIB discovery device that produces a novel, non-invasive biomarker for each molecular subtype it targets. This framing connects the specific breast cancer results to the broader thesis objective of developing a reusable framework for QIB discovery in oncological imaging.

Furthermore, the continuous nature of the output has clinical utility beyond the binary classification decision. The probability score for each subtype encodes the degree of confidence with which the framework assigns the input to that class. In borderline cases, or in tumours with mixed or transitional molecular profiles, the full vector of four probabilities provides a richer clinical signal than a single predicted label. A clinician can, for instance, observe a high probability for both Luminal A and Luminal B in a given case, which may indicate a tumour positioned at the boundary between the two Luminal subtypes and may warrant more careful histological analysis. As noted in Section 5.2.1, the four raw scores can be sum-normalised (Equation 5.2) to produce a vector that sums to 1, mapping the ambiguous region of the raw score space onto intermediate probability values and providing a concise per-case uncertainty representation. This graded output is an inherent advantage of the OvA formulation and contributes to the interpretability of the framework in a clinical context.

7.4 Towards a Domain-Agnostic QIB Discovery Framework

The experimental validation conducted in this thesis is confined to a single modality (DCE-MRI), a single organ site (breast), and a single classification task (four molecular subtypes). This specificity is a deliberate methodological choice: it enables rigorous control and meaningful external validation. However, a thesis titled *A Framework for the Development of Quantitative Imaging Biomarkers for Cancer Research* carries an implicit claim that extends beyond this specific case study. The architectural basis for that claim (the structural invariant of the framework) is formalised in Section 5.1. This section substantiates the generalisability argument through concrete analogues from other oncological imaging domains.

7.4.1 Prostate Cancer on Multiparametric MRI

Prostate adenocarcinoma imaged with multiparametric MRI constitutes the most direct structural analogue to the present work. The task of predicting clinically significant prostate cancer (International Society of Urological Pathology Grade Group ≥ 2) from multiparametric MRI volumes follows the identical localisation-then-classification decomposition: the prostate gland and candidate lesion regions are segmented by an upstream module, and a 3D CNN classification head assigns a probability of clinical significance. The resulting scalar output constitutes a QIB that directly gates the biopsy decision, replacing the clinician's categorical PI-RADS score with a continuous risk estimate. The PI-CAI challenge, which provided the most statistically rigorous validation of this paradigm to date, enrolled 10,207 biparametric MRI examinations and demonstrated that a deep learning pipeline achieves an AUC-ROC of 0.91 for detecting clinically significant disease, significantly superior to 62 radiologists assessed under identical conditions [167]. The predicted probability is therefore a validated imaging QIB that informs one of the most consequential binary clinical decisions in urological oncology.

7.4.2 Glioma Molecular Status on Multimodal Brain MRI

In diffuse gliomas, the classification target is not a tumour phenotype in the morphological sense but a molecular genetic alteration: the mutation status of the Isocitrate Dehydrogenase (IDH) gene. Under the 2021 World Health Organization (WHO) classification of central nervous system tumours, IDH mutation status is the primary axis of glioma classification, separating IDH-mutant tumours (which carry a substantially more favourable prognosis) from IDH-wildtype glioblastoma [168]. Conventionally, this status is determined invasively from tumour tissue. Residual CNN architectures trained on multimodal brain MRI (T1-weighted, post-gadolinium T1-weighted, T2-weighted, and T2-FLAIR sequences) have been shown to

predict IDH mutation status with AUC-ROC values exceeding 0.94 across multi-institutional cohorts [169]. Following the 2024 FDA approval of vorasidenib for IDH-mutant low-grade gliomas, the predicted probability of IDH mutation constitutes a directly theranostic QIB: it determines eligibility for a targeted molecular therapy. Segmentation infrastructure is provided by the BraTS benchmark, which has been running annually since 2012 and provides voxel-level annotations for thousands of glioma cases. The structural template is identical to the framework proposed in this thesis: volumetric localisation followed by 3D CNN classification with a probability output.

7.4.3 Pulmonary Nodule Malignancy on Low-Dose CT

The lung nodule malignancy classification problem on low-dose Computerized Tomography (CT) provides a cross-modality analogue that extends the generalisation argument beyond MRI. Automated nodule detection systems identify candidate lesions in CT volumes; a classification network then assigns a continuous probability of malignancy to each detected nodule. This probability score directly determines clinical management: low-risk nodules return to annual screening, intermediate-risk nodules receive short-interval follow-up, and high-risk nodules proceed to CT-guided biopsy or resection. Ardila et al. demonstrated that a 3D CNN trained on 6716 cases from the National Lung Screening Trial achieves a patient-level AUC-ROC of 0.944, reducing false positives by 11% relative to radiologists when no prior scan is available [170]. This is the most mature example in all of medical imaging Artificial Intelligence (AI): it is validated at national screening programme scale, and the predicted probability of malignancy is a well-defined QIB directly integrated into clinical decision pathways. The fact that this result is obtained on CT rather than MRI is architecturally unremarkable for the present framework: the preprocessing pipeline operates on any volumetric acquisition from which a bounding box can be extracted, and the 3D CNN architecture treats the input as a spatially structured tensor regardless of the imaging physics that produced it.

7.4.4 Generalisability: A Qualified Claim

The three analogues above demonstrate that the localisation-then-classification pipeline architecture, with the predicted class probability serving as a QIB, is a validated paradigm across modalities and cancer types. This constitutes a meaningful architectural generalisation claim for the framework proposed in this thesis. It must be stated explicitly, however, that this claim is architectural and not empirical: architectural transferability of the pipeline structure should not be conflated with performance transferability or the expectation of equivalent AUC-ROC values across domains. The three analogues involve substantially different imaging physics, tissue contrast mechanisms, and clinical decision contexts. Prostate multipara-

metric MRI relies on diffusion and perfusion contrast; glioma MRI combines structural T1, post-contrast T1, T2, and fluid-attenuated inversion recovery sequences; pulmonary CT operates on Hounsfield units from X-ray attenuation. Each modality presents different signal-to-noise characteristics, different spatial resolution and tissue specificity, and different lesion morphologies. Consequently, the framework has not been applied to prostate MRI, brain MRI, or pulmonary CT in this work. Empirical generalisation to those domains would require retraining and validation on domain-appropriate data, with domain-specific preprocessing adaptations and, critically, domain-specific hyperparameter re-optimisation of the network architecture, training procedure, and checkpoint selection criteria. The present work on breast cancer demonstrates exactly this kind of domain-specific optimisation: the 3D CNN architecture, the class weighting scheme, the `ClinicalCheckpoint` callback, and the early stopping criterion were all designed and tuned for the breast cancer molecular subtyping problem. Applying the framework to a fundamentally different domain would require an equally principled re-optimisation phase for that domain's specific characteristics. The contribution of this thesis is to propose and validate the framework template as a reusable architectural strategy, and to demonstrate through rigorous breast cancer experimentation that careful domain-specific optimisation of that template produces clinically meaningful QIBs under external prospective evaluation.

7.5 Limitations

Several limitations constrain the scope of the conclusions that can be drawn from this work and should be considered in the design of future studies.

A primary limitation concerns the uncertainty of molecular subtype labels in multicentric datasets lacking Ki-67 proliferation index measurements. In collections where Ki-67 is unavailable, the distinction between Luminal A and Luminal B for hormone-receptor-positive, HER2-negative cases cannot be made under standard clinical criteria, leading to systematic assignment of such cases to Luminal A under the fallback heuristic described in Section 3.5.1. This label uncertainty directly affects the reliability of the Luminal B classifier and represents a fundamental data-quality constraint that cannot be resolved by algorithmic means alone.

The external validation cohort, while prospective and institutionally independent, comprises 98 cases from a single Portuguese hospital. The resulting 95% confidence intervals around the reported AUC-ROC values are wide, and the absence of multi-site external validation limits the robustness of the generalisation claim. In particular, the HER2-enriched subtype is represented by only nine cases in the Setúbal cohort, making the precision estimate for this classifier highly sensitive to individual classification decisions.

The TCIA-augmented configuration relied on automated bounding boxes generated by the upstream detection module for the three supplementary collections, in the absence of expert radiologist annotations. This annotation quality asymmetry is a confounding factor in the comparison between the two training configurations. It prevents a clean attribution of the observed performance degradation to data heterogeneity alone, as opposed to localisation noise.

The framework does not incorporate any mechanism for cross-site domain harmonisation, such as intensity standardisation at the population level, histogram matching, or domain-adversarial training. The results reported here therefore reflect the performance achievable with per-volume intensity-range normalisation, which provides robustness to cross-site intensity scale variation but does not address other sources of domain heterogeneity such as tissue contrast ratios, spatial resolution, or temporal sampling differences.

No explainability analysis, such as gradient-weighted class activation mapping or occlusion sensitivity studies, was performed in this work. While the discriminative performance of the framework is well established by the external AUC-ROC values, the spatial localisation of the learned features within the tumour volume remains uncharacterised. Such analysis would be essential for clinical adoption, both to build radiologist trust in the system and to validate that the network is attending to biologically plausible regions rather than acquisition artefacts.

Finally, the comparison with the state of the art is constrained by the heterogeneity of published validation protocols. The absence of per-subtype AUC-ROC confidence intervals in most comparator studies, and the non-uniform subtype taxonomies employed by some of them, prevent formal statistical hypothesis testing between the present results and the published literature.

7.6 Future Work

The findings of this thesis motivate several concrete directions for future investigation.

Domain harmonisation for multicentric training. The degradation observed under naïve TCIA-augmented training establishes a clear target for methodological improvement. Future work should investigate domain adaptation techniques, including histogram-based intensity harmonisation [61], domain-adversarial training [115], and federated learning approaches [171], as mechanisms for extracting the representational benefits of multicentric diversity without absorbing the domain gap penalty. A controlled ablation study, varying the set of supplementary collections while holding annotation quality constant (by restricting to manually annotated subsets where available), would isolate the contribution of distributional

shift from annotation noise in explaining the observed degradation.

Extended external validation. The Setúbal cohort, while valuable as a proof of concept, is insufficient for clinical validation at the level required for regulatory consideration. Future work should pursue multi-institutional prospective validation involving at least three geographically and demographically independent cohorts, with adequate representation of the HER2-enriched and Triple-Negative subtypes. This would also enable formal statistical comparison of the two training configurations via the DeLong test [141], which was underpowered in the current evaluation due to the modest cohort size.

Nested cross-validation for model selection robustness. The experimental design used in this work partitions patients into a single fixed 85%/ 15% training/validation split, with early stopping and checkpoint selection both conditioned on the same held-out partition over up to 100 epochs. As noted in Section 5.5.4, this introduces a structural risk of validation-set overfitting in the model selection step. A nested cross-validation design, in which an outer loop estimates generalisation error across k_{outer} folds while an inner loop performs hyperparameter and checkpoint selection within each training fold, would decouple model selection from performance estimation and provide a statistically principled confidence interval on the reported AUC-ROC values. Implementing nested cross-validation for a volumetric 3D CNN pipeline is computationally demanding, requiring $k_{\text{outer}} \times k_{\text{inner}}$ independent training runs. However, it constitutes the methodologically definitive approach to confirm that the results reported in Chapter 6 reflect population-level discriminative capacity rather than artefacts of a single fortuitous split.

Model explainability and radiological correlation. Gradient-weighted class activation mapping [96] and layer-wise relevance propagation methods should be applied to the trained classifiers to generate subtype-specific saliency maps. These maps should be evaluated by expert breast radiologists against known DCE-MRI features of each molecular subtype, such as rim enhancement for Triple-Negative and plateau-type kinetics for Luminal A, in order to confirm that the network has learned biologically grounded representations rather than acquisition-specific spurious correlations. Such analysis is a prerequisite for clinical acceptance.

Multi-sequence and multi-phase input. The present framework operates exclusively on the post-contrast DCE-MRI series. Future work should investigate the integration of additional sequences, including T2-weighted MRI and apparent diffusion coefficient maps derived from diffusion-weighted MRI, as demonstrated by Zhou et al. [109] to improve subtype discrimination in a single-centre setting. Extending the

input to multiple temporal phases of the DCE-MRI acquisition would also allow the network to exploit the full pharmacokinetic enhancement curve, which encodes biologically informative wash-in and wash-out characteristics that are currently collapsed into a single temporal snapshot.

Application to additional cancer types. The architectural generalisability argument developed in Section 7.4 provides a principled basis for applying the same framework to cancer types beyond breast. The most natural candidates share volumetric imaging modalities with established lesion segmentation benchmarks: prostate adenocarcinoma on multiparametric MRI [167], IDH mutation status in diffuse gliomas on multimodal brain MRI [169], and pulmonary nodule malignancy on low-dose CT [170]. Applying the identical pipeline to one of these domains, with only the data loading and label encoding layers modified, would constitute an empirical test of the architectural generalisability claim and would substantially strengthen the thesis's contribution to the broader QIB discovery problem.

Clinical deployment and integration. The framework is designed for integration into a cloud-accessible clinical decision support interface, as described in Chapter 1. Future work should address the engineering requirements for real-time inference on clinical workstations, including model compression, quantisation, and latency benchmarking on hardware representative of a hospital radiology environment. Prospective clinical validation through a reader study, in which radiologist diagnoses with and without the framework's probability scores are compared against ground-truth IHC outcomes, would provide the highest level of clinical evidence for the utility of the proposed QIBs.

7.7 Closing Remarks

This thesis has demonstrated that a modular deep learning framework, combining automated lesion localisation with volumetric 3D CNN classification under an OvA decomposition strategy, can produce clinically meaningful and externally validated QIBs for breast cancer molecular subtype prediction from DCE-MRI. The framework achieves a weighted mean external AUC-ROC of 0.9416 on a prospective Portuguese clinical cohort, with no adaptation to the target domain and under conditions of pronounced subtype prior shift. The finding that naïve multicentric data pooling degrades rather than improves generalisation is itself a scientifically valuable contribution, establishing an empirical lower bound for domain-unaware augmentation strategies and motivating a structured research agenda on domain harmonisation for multi-institutional breast MRI.

Beyond these specific results, the framework proposed here is best understood as a template for

QIB discovery in volumetric oncological imaging. The predicted class probability is the QIB; the modality, organ, and class semantics are domain-specific instantiations of a common architectural pattern. The analogues drawn from prostate MRI, glioma MRI, and pulmonary CT demonstrate that this pattern already underpins validated clinical biomarkers at scale across oncological imaging. The contribution of this work is to formalise that pattern as a reusable framework, to validate it rigorously in the breast cancer domain, and to leave it as an open and well-characterised foundation for its extension to further cancer types and imaging modalities.

Bibliography

[1] S. Dai, K. Lu, J. Dong, Y. Zhang, and Y. Chen, "A novel approach of lung segmentation on chest CT images using graph cuts," *Neurocomputing*, vol. 168, pp. 799–807, 11 2015. doi: 10.1016/J.NEUCOM.2015.05.044 .

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90 .

[3] E. H. Houssein, M. M. Emam, A. A. Ali, and P. N. Suganthan, "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review," *Expert Systems with Applications*, vol. 167, p. 114161, 4 2021. doi: 10.1016/J.ESWA.2020.114161 .

[4] L. Wang, "Deep learning techniques to diagnose lung cancer," *Cancers*, vol. 14, no. 22, 2022. doi: 10.3390/cancers14225569 .

[5] Z. Liu, L. Tong, L. Chen, Z. Jiang, F. Zhou, Q. Zhang, X. Zhang, Y. Jin, and H. Zhou, "Deep learning based brain tumor segmentation: a survey," *Complex & intelligent systems*, vol. 9, no. 1, pp. 1001–1026, 2023. doi: 10.1007/s40747-022-00815-5 .

[6] I. Pacal, D. Karaboga, A. Basturk, B. Akay, and U. Nalbantoglu, "A comprehensive review of deep learning in colon cancer," *Computers in Biology and Medicine*, vol. 126, p. 104003, 2020. doi: 10.1016/j.compbiomed.2020.104003 .

[7] S. Gul, M. S. Khan, A. Bibi, A. Khandakar, M. A. Ayari, and M. E. Chowdhury, "Deep learning techniques for liver and liver tumor segmentation: A review," *Computers in Biology and Medicine*, vol. 147, p. 105620, 2022. doi: 10.1016/j.compbiomed.2022.105620 .

[8] J. Zhou, L. Y. Luo, Q. Dou, H. Chen, C. Chen, G. J. Li, Z. F. Jiang, and P. A. Heng, "Weakly supervised 3D deep learning for breast cancer classification and localization of the lesions in MR

images," *Journal of Magnetic Resonance Imaging*, vol. 50, no. 4, pp. 1144–1151, 10 2019. doi: 10.1002/JMRI.26721 .

[9] R. Obuchowicz, M. Oszust, and A. Piorkowski, "Interobserver variability in quality assessment of magnetic resonance images," *BMC Medical Imaging*, vol. 20, no. 1, pp. 1–10, 9 2020. doi: 10.1186/S12880-020-00505-Z .

[10] L. G. Kessler, H. X. Barnhart, A. J. Buckler, K. R. Choudhury, M. V. Kondratovich, A. Toledano, A. R. Guimaraes, R. Filice, Z. Zhang, and D. C. Sullivan, "The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions," *Statistical Methods in Medical Research*, vol. 24, no. 1, pp. 9–26, 2 2015. doi: 10.1177/0962280214537333 .

[11] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. Van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. Aerts, "Radiomics: Extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 3 2012. doi: 10.1016/j.ejca.2011.11.036 .

[12] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[13] M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, J. Vignat, J. R. Gralow, F. Cardoso, S. Siesling, and I. Soerjomataram, "Current and future burden of breast cancer: Global statistics for 2020 and 2040," *The Breast*, vol. 66, pp. 15–23, 12 2022. doi: 10.1016/J.BREAST.2022.08.010 .

[14] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021. doi: 10.3322/caac.21660 .

[15] M. Ignatiadis and C. Sotiriou, "Luminal breast cancer: from biology to treatment," *Nature Reviews Clinical Oncology 2013 10:9*, vol. 10, no. 9, pp. 494–506, 7 2013. doi: 10.1038/nrclinonc.2013.124 .

[16] B. Tran and P. L. Bedard, "Luminal-B breast cancer and novel therapeutic targets," *Breast Cancer Research*, vol. 13, no. 6, pp. 1–10, 11 2011. doi: 10.1186/BCR2904 .

[17] S. Zeng, H. Chen, R. Jing *et al.*, "An assessment of breast cancer HER2, ER, and PR status classification using molecular profiling," *Scientific Reports*, vol. 15, p. 4826, 2024. doi: 10.1038/s41598-024-83597-9 .

[18] S. Cheeney, H. Rahbar, B. N. Dontchos, S. H. Javid, M. H. Rendi, and S. C. Partridge, "Apparent diffusion coefficient values may help predict which MRI-detected high-risk breast lesions will upgrade at surgical excision," *Journal of Magnetic Resonance Imaging*, vol. 46, no. 4, pp. 1028–1036, 2017. doi: 10.1002/jmri.25656 .

[19] J. P. P. Fontes, J. N. C. Raimundo, L. G. M. Magalhães, and M. A. G. Lopez, "Accurate phenotyping of luminal A breast cancer in magnetic resonance imaging: A new 3D CNN approach," *Computers in Biology and Medicine*, vol. 189, p. 109903, 2025. doi: 10.1016/j.compbiomed.2025.109903 .

[20] J. N. C. Raimundo, J. P. P. Fontes, L. Gonzaga Mendes Magalhães, and M. A. Guevara Lopez, "An Innovative Faster R-CNN-Based Framework for Automated Detection of BC Pathological Lesions in MRI," *Journal of Imaging*, vol. 9, no. 9, p. 169, 2023. doi: 10.3390/jimaging9090169 .

[21] M. A. Mazurowski, A. Saha, M. R. Harowicz, and E. Y. Kim, "Duke Breast Cancer MRI Dataset," https://sites.duke.edu/mazurowski/resources/breast-cancer-mri-dataset/, 2018. doi: 10.7937/TCIA.e3sv-re93 .

[22] J. P. P. Fontes, E. N. Fonseca, A. C. F. Pinto, L. G. M. Magalhães, T. Gonçalves, M. J. Martins, V. Rigueira, P. M. d. C. S. Ferreira, L. F. R. Branco, F. Seixo, and M. A. G. Lopez, "Deep Learning Framework for Breast Cancer Subtypes Detection and Diagnosis," *Biomedical Signal Processing and Control*, 2026, under review.

[23] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *Journal of Machine Learning Research*, vol. 5, pp. 101–141, Dec. 2004.

[24] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, and L. A. Akslen, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2000. doi: 10.1038/35021093 .

[25] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, and S. S. Jeffrey, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 19, pp. 10 869–10 874, 2001. doi: 10.1073/pnas.191367098 .

[26] T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, N. B. Paul, T. Hastie, D. Brown, A.-L. Borresen-Dale, D. Botstein, L. Eystein, and C. M. Perou, "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 14, pp. 8418–8423, 2003. doi: 10.1073/pnas.0932692100 .

[27] The Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012. doi: 10.1038/nature11412 .

[28] J. S. Parker, M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, and Z. Hu, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *Journal of Clinical Oncology*, vol. 27, no. 8, pp. 1160–1167, 2009. doi: 10.1200/JCO.2008.18.1370 .

[29] A. Prat and C. M. Perou, "Deconstructing the molecular portraits of breast cancer," *Molecular Oncology*, vol. 5, no. 1, pp. 5–23, 2011. doi: 10.1016/j.molonc.2010.11.003 .

[30] A. Goldhirsch, W. C. Wood, A. S. Coates, R. D. Gelber, B. Thürlimann, and H.-J. Senn, "Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the st. gallen international expert consensus on the primary therapy of early breast cancer 2011," *Annals of Oncology*, vol. 22, no. 8, pp. 1736–1747, 2011. doi: 10.1093/annonc/mdr304 .

[31] A. Goldhirsch, E. P. Winer, A. S. Coates, R. D. Gelber, M. Piccart-Gebhart, B. Thürlimann, and H.-J. Senn, "Personalizing the treatment of women with early breast cancer: highlights of the st. gallen international expert consensus on the primary therapy of early breast cancer 2013," *Annals of Oncology*, vol. 24, no. 9, pp. 2206–2223, 2013. doi: 10.1093/annonc/mdt303 .

[32] M. C. U. Cheang, S. K. Chia, D. Voduc, D. Gao, S. Leung, J. Snider, A. Maki, R. Klinck, H. Masoudi, K. Gelmon, L. J. van't Veer, and D. G. Huntsman, "Ki-67 index, her2 status, and prognosis in breast cancer," *Journal of the National Cancer Institute*, vol. 101, no. 10, pp. 736–750, 2009. doi: 10.1093/jnci/djp082 .

[33] A. S. Coates, E. P. Winer, A. Goldhirsch, R. D. Gelber, M. Gnant, M. Piccart-Gebhart, B. Thürlimann, and H.-J. Senn, "Tailoring therapies-improving the management of early breast cancer: St. gallen international expert consensus conference 2015," *Annals of Oncology*, vol. 26, no. 8, pp. 1533–1546, 2015. doi: 10.1093/annonc/mdv221 .

[34] H. J. Burstein, G. Curigliano, S. Loibl, J. Panoff, S. Tjulandin, E. Dublin, S. Badve, S. Delaloge, M. Fidler, and F. Genta, "Estimating the benefits of therapy for early-stage breast cancer: The st.

gallen international consensus guidelines for the primary therapy of early breast cancer 2021," *Annals of Oncology*, vol. 32, no. 10, pp. 1216–1235, 2021. doi: 10.1016/j.annonc.2021.06.023 .

[35] L. A. Carey, C. M. Perou, C. A. Livasy, L. G. Dressler, D. Cowan, K. Conway, G. Karaca, M. A. Troester, C.-K. Tse, and S. N. Edmiston, "Race, breast cancer subtypes, and survival in the carolina breast cancer study," *JAMA*, vol. 295, no. 21, pp. 2492–2502, 2006. doi: 10.1001/jama.295.21.2492 .

[36] K. D. Voduc, M. C. Cheang, S. Tyldesley, K. Gelmon, T. O. Nielsen, and H. Kennecke, "Breast cancer subtypes and the risk of local and regional relapse," *Journal of Clinical Oncology*, vol. 28, no. 10, pp. 1684–1691, 2010. doi: 10.1200/JCO.2009.24.9284 ., pMID: 20194857.

[37] H. Kennecke, R. Yerushalmi, R. Woods, M. C. U. Cheang, D. Voduc, C. H. Speers, T. O. Nielsen, and K. Gelmon, "Metastatic behavior of breast cancer subtypes," *Journal of Clinical Oncology*, vol. 28, no. 20, pp. 3271–3277, 2010. doi: 10.1200/JCO.2009.25.9820 .

[38] N. Harbeck and M. Gnant, "Breast cancer," *The Lancet*, vol. 389, no. 10074, pp. 1134–1150, 2017. doi: 10.1016/S0140-6736(16)31891-8 .

[39] N. Harbeck, F. Penault-Llorca, J. Cortes, M. Gnant, N. Houssami, P. Poortmans, K. J. Ruddy, J. Y. S. Tsang, and F. Cardoso, "Breast cancer," *Nature Reviews Disease Primers*, vol. 5, no. 1, p. 66, 2019. doi: 10.1038/s41572-019-0111-2 .

[40] A. G. Waks and E. P. Winer, "Breast cancer treatment: a review," *JAMA*, vol. 321, no. 3, pp. 288–300, 2019. doi: 10.1001/jama.2018.19323 .

[41] E. H. Allott, J. Geradts, X. Sun, S. M. Cohen, G. R. Zirpoli, T. Khoury, W. Bshara, S. Yao, C. B. Ambrosone, and E. L. Wiley, "Pathological characteristics predictive of response to neoadjuvant chemotherapy in breast cancers," *Breast Cancer Research*, vol. 18, no. 1, p. 68, 2016. doi: 10.1186/s13058-016-0725-1 .

[42] G. Turashvili and E. Brogi, "Tumor heterogeneity in breast cancer," *Frontiers in Medicine*, vol. 4, p. 227, 2017. doi: 10.3389/fmed.2017.00227 .

[43] L. Peng, Z. Zhang, D. Zhao, J. Zhao, F. Mao, and Q. Sun, "Discordance of immunohistochemical markers between primary and recurrent or metastatic breast cancer: A retrospective analysis of 107 cases," *Medicine*, vol. 99, no. 25, p. e20738, 2020. doi: 10.1097/MD.0000000000020738 .

[44] N. Aristokli, I. Polycarpou, S. Themistocleous, D. Sophocleous, and I. Mamais, "Comparison of the diagnostic performance of magnetic resonance imaging (mri), ultrasound and mammography for detection of breast cancer based on tumor type, breast density and patient's history: A review," *Radiography*, vol. 28, no. 3, pp. 848–856, 2022. doi: 10.1016/j.radi.2022.01.006 .

[45] D. Saslow, C. Boetes, W. Burke, S. Harms, M. O. Leach, C. D. Lehman, E. Morris, E. Pisano, M. Schnall, S. Sener *et al.*, "American cancer society guidelines for breast screening with MRI as an adjunct to mammography," *CA: A Cancer Journal for Clinicians*, vol. 57, no. 2, pp. 75–89, 2007. doi: 10.3322/canjclin.57.2.75 .

[46] C. D. Lehman, C. Gatsonis, C. K. Kuhl, R. E. Hendrick, E. D. Pisano, L. Hanna, S. Peacock, S. F. Smazal, D. D. Maki, T. B. Julian, E. R. DePeri, D. A. Bluemke, and M. D. Schnall, "Mri evaluation of the contralateral breast in women with recently diagnosed breast cancer," *New England Journal of Medicine*, vol. 356, no. 13, pp. 1295–1303, 2007. doi: 10.1056/NEJMoa065447 .

[47] R. M. Mann, C. K. Kuhl, K. Kinkel, and C. Boetes, "Breast mri: State of the art," *Radiology*, vol. 292, no. 3, pp. 520–536, 2019. doi: 10.1148/radiol.2019182947 .

[48] P. S. Toft and A. G. Kermode, "Measurement of the blood-brain barrier permeability and leakage space using dynamic mr imaging: 1. fundamental concepts," *Magnetic Resonance in Medicine*, vol. 17, no. 2, pp. 357–367, 1991. doi: 10.1002/mrm.1910170208 .

[49] P. S. Tofts, G. Brix, D. L. Buckley, J. L. Evelhoch, E. Henderson, M. V. Knopp, H. B. W. Larsson, T.-Y. Lee, N. A. Mayr, and G. J. M. Parker, "Estimating kinetic parameters from dynamic contrast-enhanced t(1)-weighted mri of a diffusable tracer: Standardized quantities and symbols," *Journal of Magnetic Resonance Imaging*, vol. 10, no. 3, pp. 223–232, 1999. doi: 10.1002/(SICI)1522-2586(199909)10:3<223::AID-JMRI2>3.0.CO;2-S .

[50] M. A. Mazurowski, J. Zhang, L. J. Grimm, S. C. Yoon, and J. I. Silber, "Radiogenomic analysis of breast cancer: Luminal b molecular subtype is associated with enhancement dynamics at mr imaging," *Radiology*, vol. 273, no. 2, pp. 365–372, 2014. doi: 10.1148/radiol.14132641 .

[51] L. J. Grimm, J. Zhang, and M. A. Mazurowski, "Computational approach to radiogenomics of breast cancer: Luminal a and luminal b molecular subtypes are associated with imaging features on routine breast mri extracted using computer vision algorithms," *Journal of Magnetic Resonance Imaging*, vol. 42, no. 4, pp. 902–907, 2015. doi: 10.1002/jmri.24879 .

[52] N. M. Braman, M. Etesami, P. Malkin, P. Mehta, A. Dontchos, D. Arefan, J. S. Lewin, S. K. Antani, E. A. Morris, and M. A. Mazurowski, "Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast dce-mri," *Breast Cancer Research*, vol. 19, no. 1, p. 57, 2017. doi: 10.1186/s13058-017-0846-1 .

[53] A. Saha, M. R. Harowicz, L. J. Grimm, C. E. Kim, S. V. Ghate, R. Walsh, and M. A. Mazurowski, "A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features," *British Journal of Cancer*, vol. 119, no. 4, p. 508, 8 2018. doi: 10.1038/S41416-018-0185-8 .

[54] D. Newitt, N. Hylton, and I-SPY 1 Network and ACRIN 6657 Trial Team, "Multi-center breast DCE-MRI data and segmentations from patients in the I-SPY 1/ACRIN 6657 trials," The Cancer Imaging Archive, 2016. doi: 10.7937/K9/TCIA.2016.HdHpgJLK .

[55] N. M. Hylton, J. D. Blume, W. K. Bernreuter, E. D. Pisano, M. A. Rosen, E. A. Morris, P. T. Weatherall, W. E. Barlow, C. D. Lehman, and D. C. Newitt, "Locally advanced breast cancer: Mr imaging for prediction of response to neoadjuvant chemotherapy—results from acrin 6657/i-spy trial," *Radiology*, vol. 263, no. 3, pp. 663–672, 2012. doi: 10.1148/radiol.12110748 .

[56] N. M. Hylton, C. A. Gatsonis, M. A. Rosen, C. D. Lehman, D. C. Newitt, S. C. Partridge, W. K. Bernreuter, E. D. Pisano, E. A. Morris, P. T. Weatherall, S. M. Polin, G. M. Newstead, H. S. Marques, L. J. Esserman, and M. D. a. Schnall, "Neoadjuvant chemotherapy for breast cancer: Functional tumor volume by mr imaging predicts recurrence-free survival—results from the acrin 6657/calgb 150007 i-spy 1 trial," *Radiology*, vol. 279, no. 1, pp. 44–55, 2016. doi: 10.1148/radiol.2015150013 .

[57] W. Li, D. C. Newitt, J. Gibbs, L. J. Wilmes, E. F. Jones, V. A. Arasu, F. Strand, N. Onishi, A.-A. T. Nguyen, J. Kornak, B. N. Joe, E. R. Price, H. Ojeda-Fournier, M. Eghtedari, K. W. Zamora, S. A. Woodard, H. Umphrey, W. Bernreuter, M. Nelson, and N. M. Hylton, "I-SPY 2 Breast Dynamic Contrast Enhanced MRI Trial (ISPY2)," The Cancer Imaging Archive, 2022. doi: 10.7937/TCIA.D8Z0-9T85 .

[58] R. D. Chitalia, J. H. Cohen, A. Striedke, J. I. Meza, L. Hernandez, A. R. Levin, D. C. Newitt, and N. M. Hylton, "Combining multimodal magnetic resonance imaging, demographic and clinical data to improve prediction of neoadjuvant chemotherapy response," *Scientific Data*, vol. 9, p. 440, 2022. doi: 10.1038/s41597-022-01555-4 .

[59] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, and others, "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, pp. 1045–1057, 2013. doi: 10.1007/s10278-013-9622-7 .

[60] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: Improved n3 bias correction," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010. doi: 10.1109/TMI.2010.2046908 .

[61] L. Nyul, J. Udupa, and X. Zhang, "New variants of a method of mri scale standardization," *IEEE Transactions on Medical Imaging*, vol. 19, no. 2, pp. 143–150, 2000. doi: 10.1109/42.836373 .

[62] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007. doi: 10.1093/biostatistics/kxj037 .

[63] J.-P. Fortin, D. Parker, B. Tunç, T. Watanabe, M. A. Elliott, K. Ruparel, D. P. Rogers, T. D. Satterthwaite, and R. T. Shinohara, "Harmonization of diffusion mri data across multiple sites: A benchmark dataset," *NeuroImage*, vol. 161, pp. 149–170, 2017. doi: 10.1016/j.neuroimage.2017.08.047 .

[64] J.-P. Fortin, N. C. Cullen, B. Shackett, B. Tunç, T. D. Satterthwaite, and R. T. Shinohara, "Harmonization of cortical thickness measurements across scanners and sites," *NeuroImage*, vol. 167, pp. 104–120, 2018. doi: 10.1016/j.neuroimage.2017.11.024 .

[65] F. Orlhac, J. J. Eertink, A.-S. Cottereau, J. M. Zijlstra, C. Thieblemont, M. Meignan, R. Boellaard, and I. Buvat, "A guide to combat harmonization of imaging biomarkers in multicenter studies," *Journal of Nuclear Medicine*, vol. 63, no. 2, pp. 172–179, 2022. doi: 10.2967/jnumed.121.262464 .

[66] A. Saha, X. Yu, D. Sahoo, and M. A. Mazurowski, "Effects of mri scanner parameters on breast cancer radiomics," *Expert Systems with Applications*, vol. 87, pp. 384–391, 2017. doi: 10.1016/j.eswa.2017.06.029 .

[67] FDA-NIH Biomarker Working Group, "Biomarkers, EndpointS, and other Tools (BEST) resource," 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK326791/

[68] D. C. Sullivan, N. A. Obuchowski, L. G. Kessler, D. L. Raunig, C. Gatsonis, R. F. Wagner, G. Liu, C. F. Beaulieu, and P. B. Abada, "Metrology standards for quantitative imaging biomarkers," *Radiology*, vol. 277, no. 3, pp. 813–825, 2015. doi: 10.1148/radiol.2015142202 .

[69] J. P. O'Connor, E. O. Aboagye, J. E. Adams, H. J. Aerts, S. F. Barrington *et al.*, "Imaging biomarker roadmap for cancer studies," *Nature Reviews Clinical Oncology 2016 14:3*, vol. 14, no. 3, pp. 169–186, 10 2016. doi: 10.1038/nrclinonc.2016.162 .

[70] A. Zwanenburg, M. Vallières, R. G. Abramson, H. J. W. L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, C. Beule, and J. Beuzebrock, "The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, 2020. doi: 10.1148/radiol.2020191145 .

[71] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 2017. doi: 10.1158/0008-5472.CAN-17-0339 .

[72] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973. doi: 10.1109/TSMC.1973.4309314 .

[73] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics and Image Processing*, vol. 4, no. 2, pp. 172–179, 1975. doi: 10.1016/S0146-664X(75)80008-6 .

[74] G. Thibault, B. Fertil, C. Navarro, S. Pereira, P. Cau, N. Levy, J. Sequeira, and J.-L. Mari, "Texture indexes and gray level size zone matrix. application to cell nuclei classification," in *Pattern Recognition and Information Processing (PRIP)*, Minsk, Belarus, May 2009, pp. 140–145.

[75] M. Amadasun and R. King, "Textural features corresponding to textural properties," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 5, pp. 1264–1274, 1989. doi: 10.1109/21.44046 .

[76] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, W. Monsky, F. Welker, F. Hoebers, M. Rietveld, M. M. Rietbergen, B. Haibe-Kains, D. Rgoodbye, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, p. 4006, 2014. doi: 10.1038/ncomms5006 .

[77] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016. doi: 10.1148/radiol.2015151169 .

[78] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. De Jong, J. Van Timmeren, S. Sanduleanu, R. T. H. M. Larue, A. J. G. Even, A. Jochems, and others, "Radiomics: the bridge between medical imaging and personalized medicine," *Nature Reviews Clinical Oncology*, vol. 14, no. 12, pp. 749–762, 2017. doi: 10.1038/nrclinonc.2017.141 .

[79] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x .

[80] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005. doi: 10.1109/TPAMI.2005.159 .

[81] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and H. J. Aerts, "Machine learning methods for quantitative radiomic biomarkers," *Scientific reports*, vol. 5, no. 1, p. 13087, 2015. doi: 10.1038/srep13087 .

[82] H. Li, Y. Zhu, E. S. Burnside, E. Huang, K. Drukker, K. A. Hoadley, C. Fan, S. D. Conzen, M. Zuley, J. M. Net *et al.*, "Quantitative mri radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set," *NPJ breast cancer*, vol. 2, p. 16012, 2016. doi: 10.1038/npjbcancer.2016.12 .

[83] E. J. Sutton, B. Z. Dashevsky, J. H. Oh, H. Veeraraghavan, A. P. Apte, S. B. Thakur, E. A. Morris, and J. O. Deasy, "Breast cancer molecular subtype classifier that incorporates mri features," *Journal of Magnetic Resonance Imaging*, vol. 44, no. 1, pp. 122–129, 2016. doi: https://doi.org/10.1002/jmri.25119 .

[84] M. Fan, H. Li, S. Wang, B. Zheng, J. Zhang, and L. Li, "Radiomic analysis reveals dce-mri features for prediction of molecular subtypes of breast cancer," *PloS one*, vol. 12, no. 2, p. e0171683, 2017. doi: 10.1371/journal.pone.0171683 .

[85] T. Xie, Z. Wang, Q. Zhao, Q. Bai, X. Zhou, Y. Gu, W. Peng, and H. Wang, "Machine learning-based analysis of mr multiparametric radiomics for the subtype classification of breast cancer," *Frontiers in oncology*, vol. 9, p. 505, 2019. doi: 10.3389/fonc.2019.00505 .

[86] A. Traverso, L. Wee, A. Yan, and Z. Shi, "Repeatability and reproducibility of radiomic features: A

systematic review," *International Journal of Radiation Oncology, Biology, Physics*, vol. 102, no. 4, pp. 1143–1158, 2018. doi: 10.1016/j.ijrobp.2018.05.053 .

[87] D. Truhn, S. Schrading, C. Haarburger, H. Schneider, D. Merhof, and C. Kuhl, "Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast mri," *Radiology*, vol. 290, no. 2, pp. 290–297, 2019. doi: 10.1148/radiol.2018181352 .

[88] H. M. Whitney, H. Li, Y. Ji, P. Liu, and M. L. Giger, "Comparison of breast mri tumor classification using human-engineered radiomics, transfer learning from deep convolutional neural networks, and fusion methods," *Proceedings of the IEEE*, vol. 108, no. 1, pp. 163–177, 2020. doi: 10.1109/JPROC.2019.2950187 .

[89] O. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 424–432, 2016. doi: 10.1007/978-3-319-46723-8_49 .

[90] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571. doi: 10.1109/3DV.2016.79 .

[91] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[92] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021. doi: 10.1038/s41592-020-01008-z .

[93] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[94] S. Woo, J. Park, J.-W. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," *European Conference on Computer Vision*, pp. 3–19, 2018. doi: 10.1007/978-3-030-01234-2_1 .

[95] J. Schlemper, O. Oktay, M. Schaap, E. Aisvoruth, W. Bai, S. N. Price, R. Prasad, G. Bower, A. Patel, and J. A. Schnabel, "Attention-gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, 2019. doi: 10.1016/j.media.2019.01.012 .

[96] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *IEEE International Conference on Computer Vision*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74 .

[97] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019. [Online]. Available: https://arxiv.org/abs/1904.00625

[98] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang, "Models genesis," *Medical Image Analysis*, vol. 67, p. 101840, 2021. doi: 10.1016/j.media.2020.101840 .

[99] Z. Zhou, V. Sodha, J. Johnson, D. Levin, J. Miao, and L. Zheng, "Models genesis: Generic autodidactic models for 3d medical image analysis," in *Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 384–393. doi: 10.1007/978-3-030-32251-9_42 .

[100] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 448–456.

[101] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[102] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013. [Online]. Available: https://arxiv.org/abs/1312.4400

[103] J. Witowski, L. Heacock, B. Reig, S. K. Kang, A. Lewin, K. Pysarenko, S. Patel, N. Samreen, W. Rudnicki, E. Łuczyńska *et al.*, "Improving breast cancer diagnostics with deep learning for mri," *Science translational medicine*, vol. 14, no. 664, p. eabo4802, 2022. doi: 10.1126/scitranslmed.abo4802 .

[104] R. Ha, S. Mutasa, J. Karcich, N. Gupta, E. Pascual Van Sant, J. Nemer, M. Sun, P. Chang, M. Z. Liu, and S. Jambawalikar, "Predicting breast cancer molecular subtype with mri dataset utilizing convolutional neural network algorithm," *Journal of digital imaging*, vol. 32, no. 2, pp. 276–282, 2019. doi: 10.1007/s10278-019-00179-2 .

[105] Z. Zhu, E. Albadawy, A. Saha, J. Zhang, M. R. Harowicz, and M. A. Mazurowski, "Deep learning for identifying radiogenomic associations in breast cancer," *Computers in biology and medicine*, vol. 109, pp. 85–90, 2019. doi: 10.1016/j.compbiomed.2019.04.018 .

[106] Y. Zhang, J.-H. Chen, Y. Lin, S. Chan, J. Zhou, D. Chow, P. Chang, T. Kwong, D.-C. Yeh, X. Wang *et al.*, "Prediction of breast cancer molecular subtypes on dce-mri using convolutional neural network with transfer learning between two centers," *European radiology*, vol. 31, no. 4, pp. 2559–2567, 2021. doi: 10.1007/s00330-020-07274-x .

[107] Q. Hu, H. M. Whitney, and M. L. Giger, "A deep learning methodology for improved breast cancer diagnosis using multiparametric mri," *Scientific reports*, vol. 10, no. 1, p. 10536, 2020. doi: 10.1038/s41598-020-67441-4 .

[108] H. Yin, L. Bai, H. Jia, and G. Lin, "Noninvasive assessment of breast cancer molecular subtypes on multiparametric mri using convolutional neural network with transfer learning," *Thoracic Cancer*, vol. 13, no. 22, pp. 3183–3191, 2022. doi: 10.1111/1759-7714.14673 .

[109] X.-X. Zhou, L. Zhang, Q.-X. Cui, H. Li, X.-Q. Sang, H.-X. Zhang, Y.-M. Zhu, and Z.-X. Kuai, "A channel-dimensional feature-reconstructed deep learning model for predicting breast cancer molecular subtypes on overall b-value diffusion-weighted mri," *Journal of Magnetic Resonance Imaging*, vol. 59, no. 4, pp. 1425–1435, 2024. doi: 10.1002/jmri.28895 .

[110] X. Xie, H. Zhou, M. Ma, J. Nie, W. Gao, J. Zhong, X. Cao, X. He, J. Peng, Y. Hou *et al.*, "A deep learning model for predicting molecular subtype of breast cancer by fusing multiple sequences of dce-mri from two institutes," *Academic Radiology*, vol. 31, no. 9, pp. 3479–3488, 2024. doi: 10.1016/j.acra.2024.03.002 .

[111] X. Wang, X. Hu, C. Wang, H. Yang, Y. Hu, X. Lan, Y. Huang, Y. Cao, L. Yan, F. Zhang *et al.*, "Automatic segmentation and molecular subtype classification of breast cancer using an mri-based deep learning framework," *Radiology: Imaging Cancer*, vol. 7, no. 3, p. e240184, 2025. doi: 10.1148/rycan.240184 .

[112] Z. Wang and S. Wei, "Diagnosing breast cancer subtypes using mri radiomics and machine learning: A systematic review," *Journal of Radiation Research and Applied Sciences*, vol. 18, no. 1, p. 101260, 2025. doi: 10.1016/j.jrras.2024.101260 .

[113] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," *PLoS medicine*, vol. 15, no. 11, p. e1002683, 2018. doi: 10.1371/journal.pmed.1002683 .

[114] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2022. doi: 10.1109/TBME.2021.3117407 .

[115] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks (2016)," *Journal of Machine Learning Research*, vol. 17, pp. 1–35, 2015. doi: 10.48550/arXiv.1505.07818 .

[116] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert *et al.*, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *International conference on information processing in medical imaging*. Springer, 2017, pp. 597–609. doi: 10.1007/978-3-319-59050-9_47 .

[117] N. Rieke, J. Hancox, W. Li, F. Milletarì, S. Albarqouni, D. Granziol, G. Shroff, J. Passerat-Palmbach, G. Pistoia, and U. Corradi, "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, p. 119, 2020. doi: 10.1038/s41746-020-00323-1 .

[118] G. Flores, J. Esteban, P. Pinti, M. Grau, A. Ribera, and I. Segura-Bedmar, "Federated learning for covid-19 detection with generalization to global health conditions," *Nature Medicine*, vol. 27, pp. 1735–1743, 2021. doi: 10.1038/s41591-021-01506-3 .

[119] S. Pati, U. Baid, B. Edwards, M. Sheller, S. Wang, A. H. Foruzan, G. Shah, S. P. Thakur, M. Bilello, and C. Davatzikos, "Federated learning enables big data for rare cancer boundary detection," *Nature Communications*, vol. 13, p. 7346, 2022. doi: 10.1038/s41467-022-33407-5 .

[120] L. Garrucho, K. Kushibar, C.-A. Reidel, S. Joshi, R. Osuala, A. Tsirikoglou, M. Bobowicz, J. Del Riego, A. Catanese, K. Gwoździewicz *et al.*, "A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations," *Scientific data*, vol. 12, no. 1, p. 453, 2025. doi: 10.1038/s41597-025-04707-4 .

[121] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009. doi: 10.1109/TKDE.2008.239 .

[122] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, p. 27, 2019. doi: 10.1186/s40537-019-0192-5 .

[123] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE International Conference on Computer Vision*, pp. 2999–3007, 2017. doi: 10.1109/ICCV.2017.324 .

[124] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 240–248, 2017. doi: 10.1007/978-3-319-67558-9_28 .

[125] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority oversampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. doi: 10.1613/jair.953 .

[126] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," *Advances in Intelligent Computing*, pp. 878–887, 2005. doi: 10.1007/11538059_91 .

[127] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," *IEEE International Joint Conference on Neural Networks*, pp. 1322–1328, 2008. doi: 10.1109/IJCNN.2008.4633969 .

[128] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018. doi: 10.1016/j.neucom.2018.09.013 .

[129] Z. Zhu, Y. Liu, C.-A. Yuan, X. Qin, and F. Yang, "A diffusion model multi-scale feature fusion network for imbalanced medical image classification research," *Computer Methods and Programs in Biomedicine*, vol. 256, p. 108384, 2024. doi: 10.1016/j.cmpb.2024.108384 .

[130] G. Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," *Neuroimage*, vol. 180, pp. 68–77, 2018. doi: 10.1016/j.neuroimage.2017.06.061 .

[131] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Warburton, "Machine learning algorithm validation with a limited sample size," *PLoS ONE*, vol. 14, no. 11, p. e0224365, 2019. doi: 10.1371/journal.pone.0224365 .

[132] G. Varoquaux, P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, and B. Thirion, "Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines," *NeuroImage*, vol. 145, pp. 166–179, 2017. doi: 10.1016/j.neuroimage.2016.10.038 .

[133] A. C. Yu, B. Mohajer, and J. Eng, "External validation of deep learning algorithms for radiologic diagnosis: a systematic review," *Radiology: Artificial Intelligence*, vol. 4, no. 3, p. e210064, 2022. doi: 10.1148/ryai.210064 .

[134] K. G. M. Moons, R. F. Wolff, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, and S. Mallett, "Tripod: a framework and checklist for reporting of prediction model validation studies," *Annals of Internal Medicine*, vol. 162, no. 1, pp. W1–W73, 2015. doi: 10.7326/M14-0698 .

[135] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Tripod+ai statement: updated guidance for transparent reporting of prediction models using artificial intelligence," *BMJ*, vol. 385, p. e078378, 2024. doi: 10.1136/bmj-2023-078378 .

[136] P. M. Bossuyt, J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. Irwig, J. G. Lijmer, D. Moher, D. Rennie, H. C. W. de Vet, and J. F. Cohen, "Stard 2015: an updated list of essential items for reporting diagnostic accuracy studies," *Radiology*, vol. 277, no. 3, pp. 826–832, 2015. doi: 10.1148/radiol.2015151516 .

[137] R. F. Wolff, K. G. M. Moons, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, S. Mallett, and D. G. Altman, "Probast: a tool to assess the risk of bias and applicability of prediction model studies," *Annals of Internal Medicine*, vol. 170, no. 1, pp. 51–58, 2019. doi: 10.7326/M18-1376 .

[138] J. Mongan, L. Moy, and C. E. Kahn, "Checklist for artificial intelligence in medical imaging (claim): a guide for authors and reviewers," *Radiology: AI*, vol. 2, no. 2, p. e200029, 2020. doi: 10.1148/ryai.2020200029 .

[139] X. Liu, S. Cruz Rivera, D. Moher, M. J. Calvert, and A. K. Denniston, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension," *Nature Medicine*, vol. 26, pp. 1364–1374, 2020. doi: 10.1038/s41591-020-1034-x .

[140] S. Cruz Rivera, X. Liu, A.-W. Chan, A. K. Denniston, and M. J. Calvert, "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the spirit-ai extension," *Nature Medicine*, vol. 26, pp. 1351–1363, 2020. doi: 10.1038/s41591-020-1037-7 .

[141] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more

correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988. doi: 10.2307/2531595 .

[142] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003. doi: 10.1023/A:1024068626366 .

[143] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998. doi: 10.1162/089976698300017197 .

[144] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCann, and L. Beer, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans," *Nature Machine Intelligence*, vol. 3, no. 3, pp. 199–217, 2021. doi: 10.1038/s42256-021-00307-0 .

[145] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Anastasopoulou, M. Komorowski, J. P. Bedoya, R. W. Morris, B. A. Mateen, and A. K. Denniston, "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of machine-learning studies," *BMJ*, vol. 368, p. m689, 2020. doi: 10.1136/bmj.m689 .

[146] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: methodological failures and recommendations for the future," *npj Digital Medicine*, vol. 5, p. 48, 2022. doi: 10.1038/s41746-022-00592-y .

[147] E. Orrantia-Borunda, P. Anchondo-Nuñez, L. E. Acuña-Aguilar, F. O. Gómez-Valles, and C. A. Ramírez-Valdespino, "Subtypes of breast cancer," *Breast Cancer*, 2022. doi: 10.36255/exon-publications-breast-cancer-subtypes .

[148] G. Cserni, M. Francz, B. Járay, E. Kálmán, I. Kovács, T. Krenács, E. Tóth, N. Udvarhelyi, L. Vass, A. Vörös, and others, "Pathological diagnosis, work-up and reporting of breast cancer: 1st central-eastern european professional consensus statement on breast cancer," *Pathology & Oncology Research*, vol. 28, p. 1610373, 2022. doi: 10.3389/pore.2022.1610373 .

[149] "The Cancer Imaging Archive (TCIA)," 12 2022. [Online]. Available: https://www.cancerimagingarchive.net/

[150] D. Newitt and N. Hylton, "Single site breast DCE-MRI data and segmentations from patients undergoing neoadjuvant chemotherapy," The Cancer Imaging Archive, 2016. doi: 10.7937/K9/TCIA.2016.QHsyhJKy .

[151] D. Daniels, D. Last, K. Cohen, Y. Mardor, and M. Sklair-Levy, "Standard and delayed contrast-enhanced MRI of malignant and benign breast lesions with histological and clinical supporting data (Advanced-MRI-Breast-Lesions)," The Cancer Imaging Archive, 2024. doi: 10.7937/C7X1-YN57 .

[152] M. G. Lopez, N. Posada, D. C. Moura, R. R. Pollán, J. M. F. Valiente, C. S. Ortega, M. Solar, G. Diaz-Herrero, I. Ramos, J. Loureiro *et al.*, "Bcdr: a breast cancer digital repository," in *15th International conference on experimental mechanics*, vol. 1215. Faculty of Engineering, University of Porto (FEUP) Porto, Portugal, 2012, pp. 113–120. [Online]. Available: https://www.kaggle.com/datasets/tistariqul/breast-cancer-digital-repository-bcdr-dataset

[153] J. P. Fontes and M. A. G. Lopez, "Representation learning approach to breast cancer diagnosis," in *European Congress of Radiology*, 2019. doi: 10.26044/ecr2019/C-2062 .

[154] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. Guevara Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 248–257, 7 2016. doi: 10.1016/j.cmpb.2015.12.014 .

[155] V. Kilintzis, V. Kalokyri, H. Kondylakis, S. Joshi, K. Nikiforaki, O. Díaz, K. Lekadir, M. Tsiknakis, and K. Marias, "Public data homogenization for AI model development in breast cancer," *European Radiology Experimental*, vol. 8, no. 1, pp. 1–12, 2024. doi: 10.1186/s41747-024-00442-4 .

[156] I. Bankman, *Handbook of Medical Image Processing and Analysis*, 2nd ed. Burlington, MA: Academic Press, 2008. doi: 10.1016/B978-0-12-373904-9.X0001-4 .

[157] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015. doi: 10.1007/s11263-015-0816-y .

[158] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *Advances in Neural Information Processing Systems*, vol. 32, pp. 3342–3352, 2019. doi: 10.48550/arXiv.1902.07208 .

[159] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017, pp. 4278–4284. doi: 10.1609/aaai.v31i1.11231 .

[160] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2 2019, pp. 6105–6114. doi: 10.48550/arXiv.1905.11946 .

[161] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 9 2014. doi: 10.48550/arxiv.1409.1556 .

[162] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. doi: 10.48550/arXiv.1412.6980 .

[163] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi *et al.*, "Keras tuner," https://github.com/keras-team/keras-tuner, 2019.

[164] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis." *Medical image analysis*, vol. 42, pp. 60–88, 12 2017. doi: 10.1016/j.media.2017.07.005 .

[165] R. Adam, K. Dell'Aquila, L. Hodges, T. Maldjian, and T. Q. Duong, "Deep learning applications to breast cancer detection by magnetic resonance imaging: a literature review," *Breast Cancer Research*, vol. 25, no. 1, p. 87, 2023. doi: 10.1186/s13058-023-01687-4 .

[166] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009. doi: 10.1109/MIS.2009.36 .

[167] A. Saha, J. S. Bosma, J. J. Twilt, B. van Ginneken, D. Yakar, M. Elschot, J. Veltman, A. Ünlü, M. de Rooij, and H. Huisman, "Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study," *The Lancet Oncology*, vol. 25, no. 7, pp. 879–887, 2024. doi: 10.1016/S1470-2045(24)00220-1 .

[168] D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. K. Ng, S. M. Pfister, G. Reifenberger, R. Soffietti, A. von Deimling, and D. W. Ellison, "The 2021 WHO

classification of tumors of the central nervous system: a summary," *Neuro-Oncology*, vol. 23, no. 8, pp. 1231–1251, 2021. doi: 10.1093/neuonc/noab106 .

[169] K. Chang, H. X. Bai, H. Zhou, C. Su, W. J. Bi, E. Agbodza, V. K. Kavouridis, J. T. Senders, A. Boaro, A. Beers, B. Zhang, A. Capellini, W. Liao, A. Srinivasan, T. Zhong, S. Jiang, S. Nalawade, L. P. Sugrue, R. G. Henry, N. Fischbein, S. Deshmukh, R. Colen, A. S. Jakola, J. Kalpathy-Cramer, O. Andronesi, and R. Y. Huang, "Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from MR imaging," *Clinical Cancer Research*, vol. 24, no. 5, pp. 1073–1081, 2018. doi: 10.1158/1078-0432.CCR-17-2236 .

[170] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, and S. Shetty, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Medicine*, vol. 25, no. 6, pp. 954–961, 2019. doi: 10.1038/s41591-019-0447-x .

[171] H. R. Roth, K. Chang, P. Singh, N. Neumark, W. Li, V. Gupta, S. Gupta, L. Qu, A. Ihsani, B. C. Bizzo, Y. Wen, V. Buch, M. Shah, P. Arun, P. J. Bergquist, P. Poulin, S. Marchesseau, A. Daguet, T. Kersten, V. Nath, N. Balachandar *et al.*, "Federated learning for breast density classification: a real-world implementation," *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning (DART/DCL 2020), LNCS*, vol. 12444, pp. 181–191, 2020. doi: 10.1007/978-3-030-60548-3_18 .

Appendices

Appendix A
Support work

A.1 Duke-Only Training: Learning Curves

The per-epoch Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and weighted binary cross-entropy loss learning curves for the One-vs-All (OvA) classifiers trained exclusively on the Duke-Breast-Cancer-MRI dataset are presented in Figures 8, 9, 10, and 11. Each figure shows the training partition curve (blue) and the held-out validation partition curve (orange) over 100 epochs. Full discussion of the training dynamics is provided in Section 6.1.2.

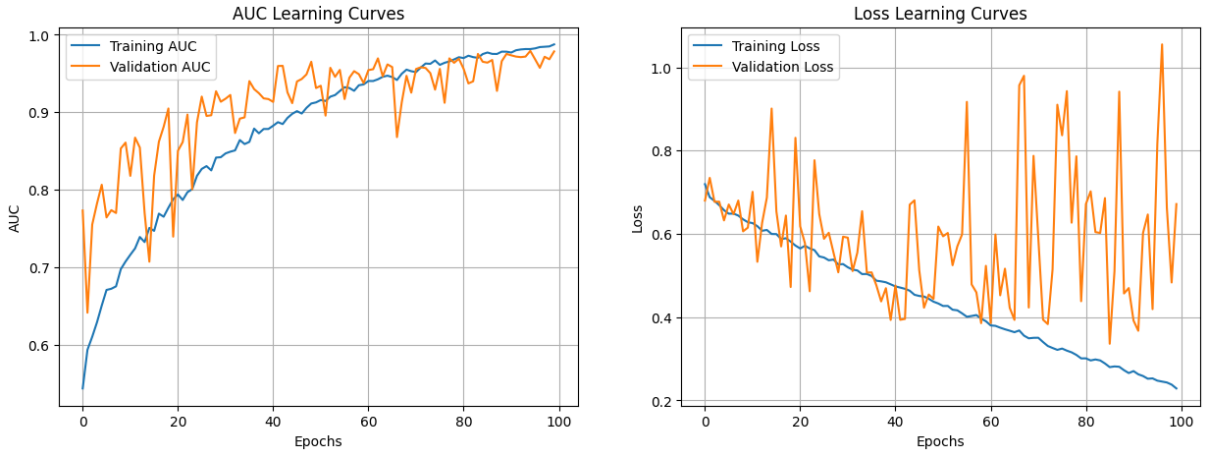


Figure 8: Per-epoch AUC-ROC (left panel) and weighted binary cross-entropy loss (right panel) learning curves for the Luminal A OvA classifier trained exclusively on the Duke-Breast-Cancer-MRI dataset. Blue curves correspond to the augmented training partition; orange curves to the held-out validation partition.

A.2 TCIA-Augmented Training: Learning Curves

The per-epoch AUC-ROC and weighted binary cross-entropy loss learning curves for the OvA classifiers trained on the full multicentric corpus drawn from the Cancer Imaging Archive (TCIA) are presented in Figures 12, 13, 14, and 15. Each figure shows the training partition curve (blue) and the held-out validation partition curve (orange). Full discussion of the training dynamics, including the slower convergence rate and higher validation variance relative to the Duke-only configuration, is provided in Section 6.2.2.

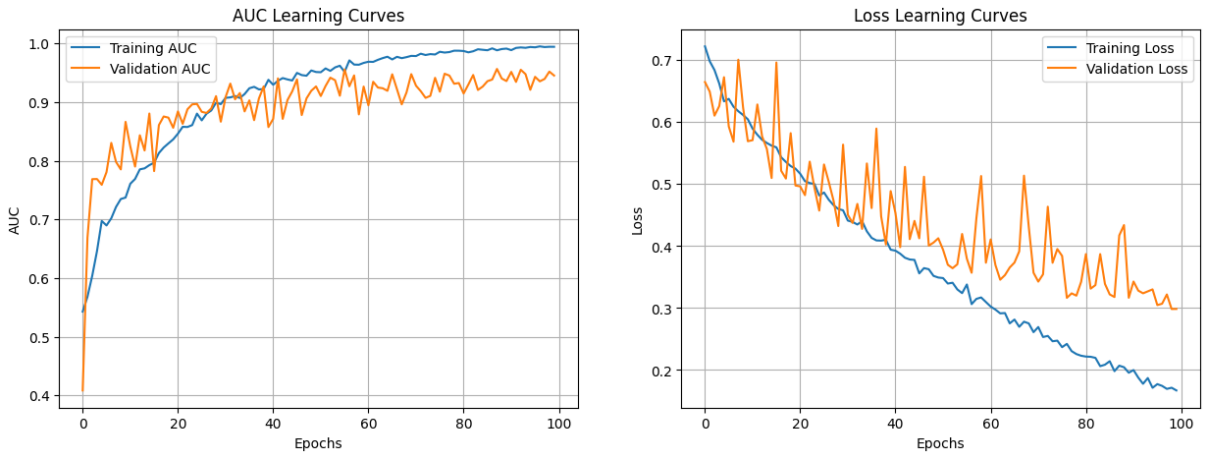


Figure 9: Per-epoch AUC-ROC (left panel) and weighted binary cross-entropy loss (right panel) learning curves for the Luminal B OvA classifier trained exclusively on the Duke-Breast-Cancer-MRI dataset. Blue curves correspond to the augmented training partition; orange curves to the held-out validation partition.

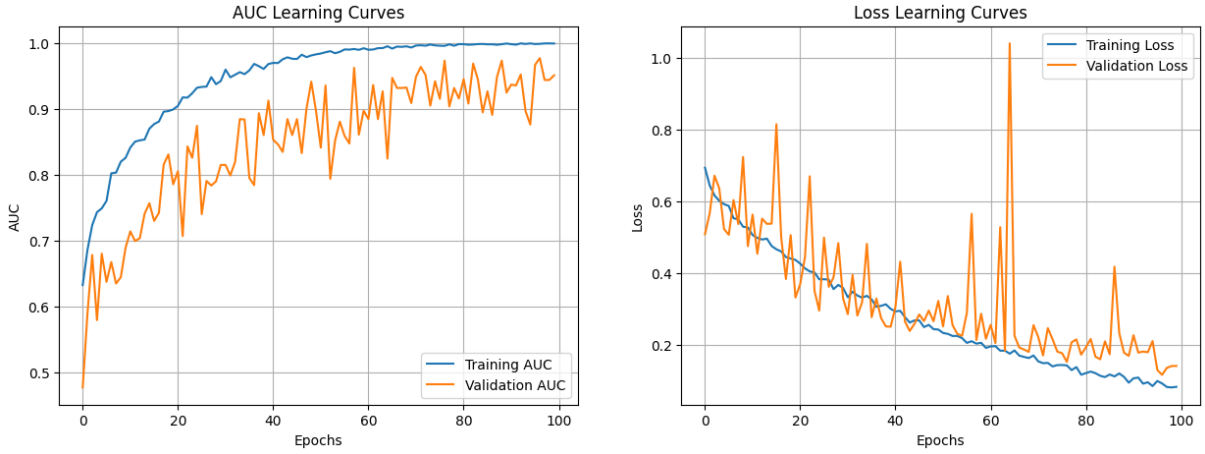


Figure 10: Per-epoch AUC-ROC (left panel) and weighted binary cross-entropy loss (right panel) learning curves for the HER2-enriched OvA classifier trained exclusively on the Duke-Breast-Cancer-MRI dataset. Blue curves correspond to the augmented training partition; orange curves to the held-out validation partition.

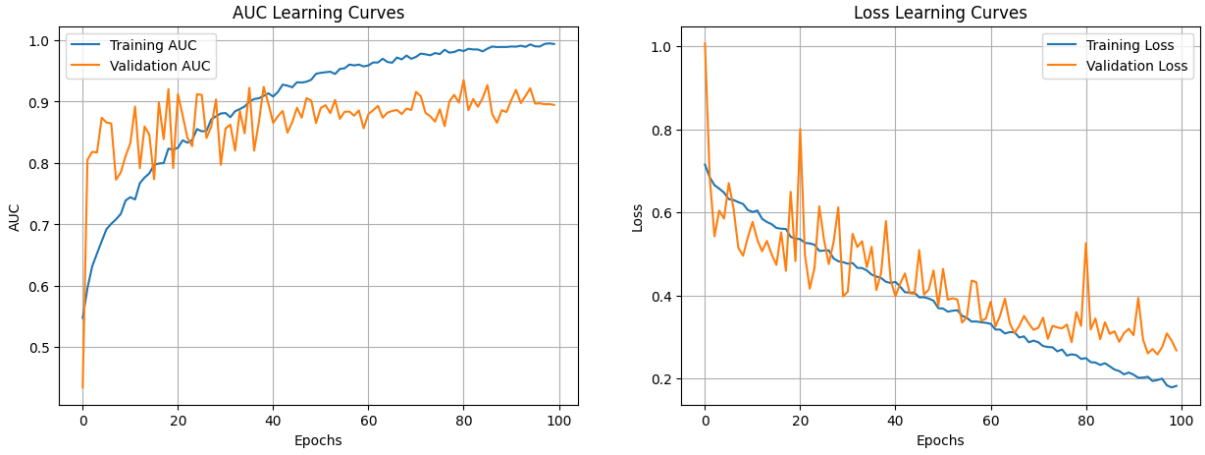


Figure 11: Per-epoch AUC-ROC (left panel) and weighted binary cross-entropy loss (right panel) learning curves for the Triple-Negative OvA classifier trained exclusively on the Duke-Breast-Cancer-MRI dataset. Blue curves correspond to the augmented training partition; orange curves to the held-out validation partition.

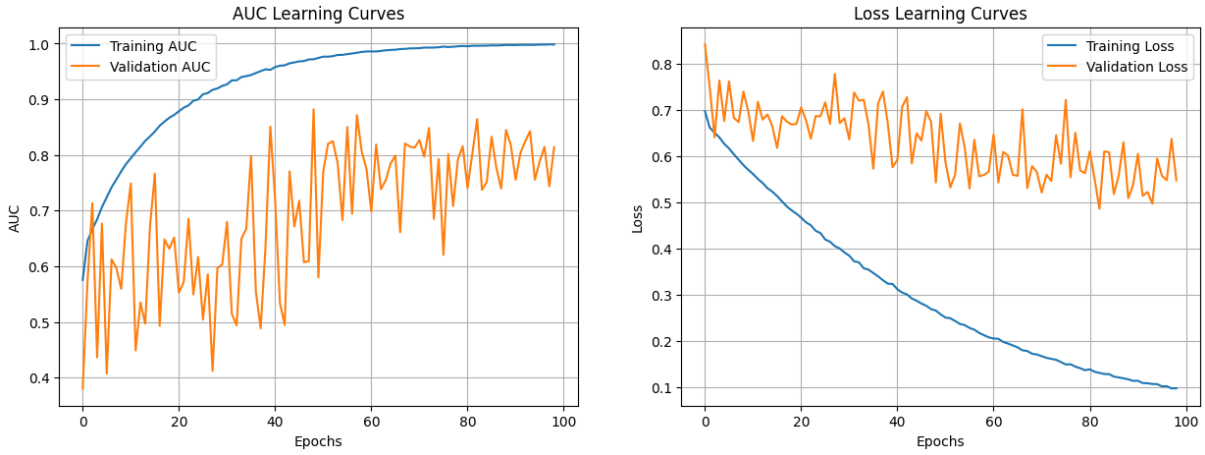


Figure 12: Per-epoch AUC-ROC (left panel) and weighted binary cross-entropy loss (right panel) learning curves for the Luminal A OvA classifier trained on the full multicentric TCIA corpus. Blue curves correspond to the augmented training partition; orange curves to the held-out validation partition.

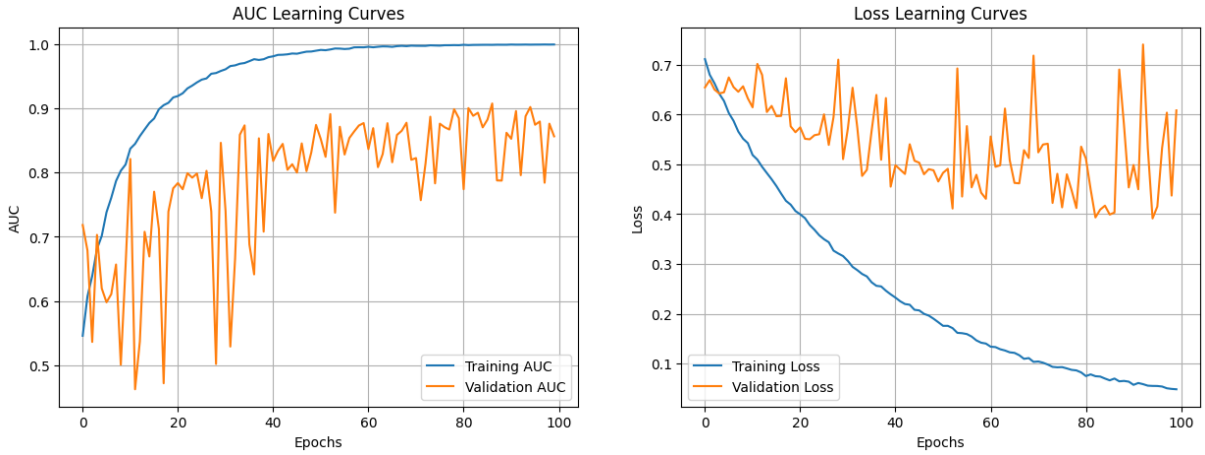


Figure 13: Per-epoch AUC-ROC (left panel) and weighted binary cross-entropy loss (right panel) learning curves for the Luminal B OvA classifier trained on the full multicentric TCIA corpus. Blue curves correspond to the augmented training partition; orange curves to the held-out validation partition.

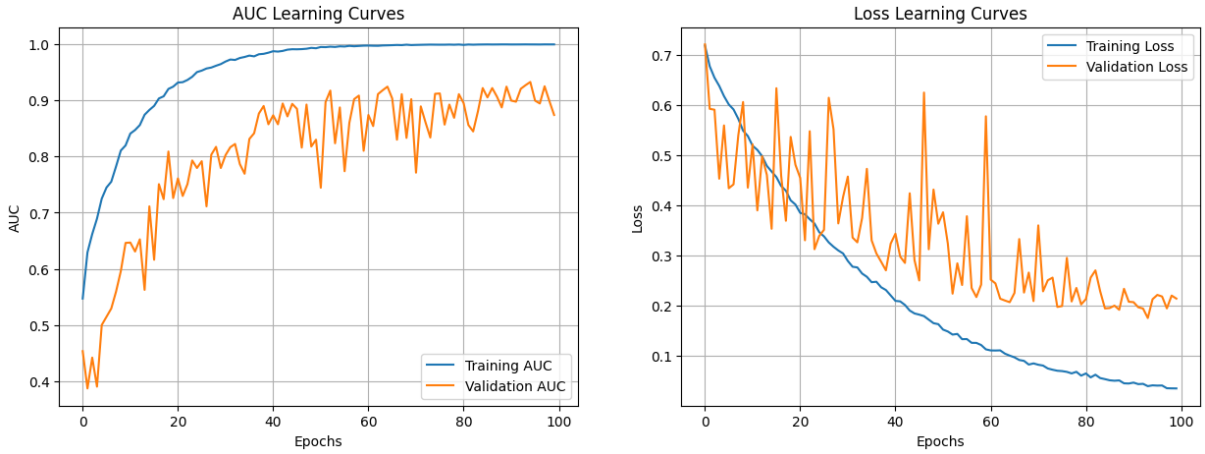


Figure 14: Per-epoch AUC-ROC (left panel) and weighted binary cross-entropy loss (right panel) learning curves for the HER2-enriched OvA classifier trained on the full multicentric TCIA corpus. Blue curves correspond to the augmented training partition; orange curves to the held-out validation partition.

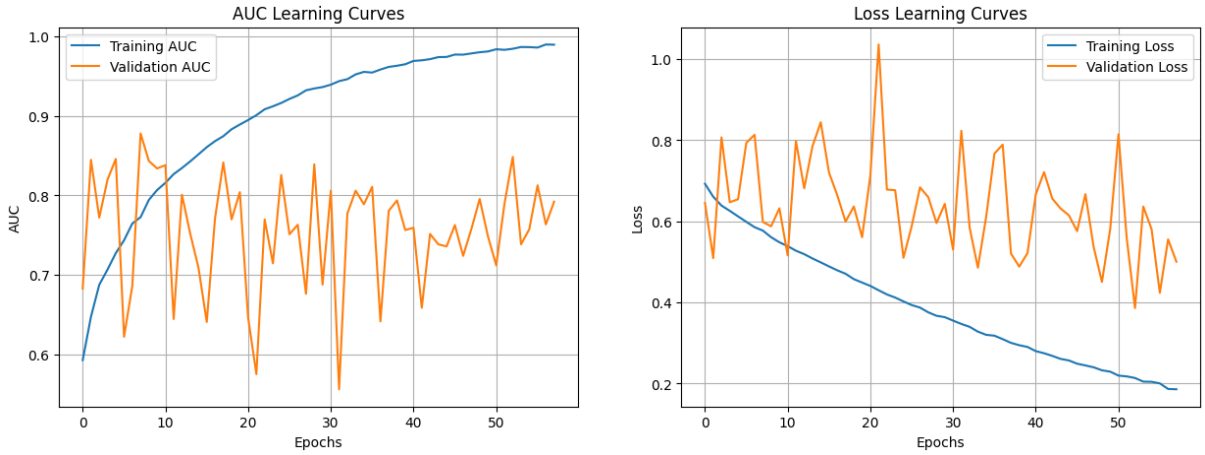


Figure 15: Per-epoch AUC-ROC (left panel) and weighted binary cross-entropy loss (right panel) learning curves for the Triple-Negative OvA classifier trained on the full multicentric TCIA corpus. Blue curves correspond to the augmented training partition; orange curves to the held-out validation partition.

Appendix B
Details of Results

B.1 Deep Learning State-of-the-Art Comparison

Table 11 provides a detailed comparison of the deep learning studies reviewed in Chapter 2 against the framework proposed in this thesis. The table reports, for each study, the number of patients in the training and evaluation partitions, per-subtype Area Under the Receiver Operating Characteristic Curve (AUC-ROC) where published, the best or mean AUC-ROC as reported, the validation protocol, the Magnetic Resonance Imaging (MRI) input modality, and the processing dimensionality. Full discussion of these results is provided in Section 6.4.

Table 11: Comparison of the proposed framework against published deep learning studies for Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI)-based breast cancer molecular subtype classification. *Train* and *Test* denote the number of patients in the training partition and in the evaluation cohort on which the reported AUC-ROC values were computed, respectively. Val.: validation type (Int. = internal held-out split from the same institution; Ext. = independent external institutional cohort). Entries marked "—" indicate that the value was not reported by the authors or is not applicable to the study design. MP-MRI = multiparametric MRI; CE-MRI = contrast-enhanced MRI; DCE = DCE-MRI only. **LA** = Luminal A; **LB** = Luminal B; **H2** = HER2-enriched; **TN** = Triple-Negative. "Best / mean AUC-ROC" reports the metric as published: macro AUC-ROC for multi-class studies without per-subtype breakdowns, mean of per-subtype AUC-ROC where available, or single-task AUC-ROC for binary studies. The two rows labelled "This work" report external validation results on the prospective Setúbal cohort (preconditioned on the lesion detection pipeline of [20]). Their "mean" values are weighted by the per-subtype case count in the Setúbal cohort (37, 34, 9, and 18 for Luminal A, Luminal B, HER2-enriched, and Triple-Negative, respectively), unlike the unweighted means reported by the comparator studies. The first row corresponds to the model trained exclusively on the Duke-Breast-Cancer-MRI dataset; the second row corresponds to the model trained on the four-collection multicentric TCIA repository (Duke + I-SPY 1 + I-SPY 2 + Breast-MRI-NACT-Pilot, 2114 cases). † Non-standard Luminal taxonomy: Luminal A and B defined without Ki-67 [104]. ‡ Three-class task only (HR^+/Human Epidermal growth factor Receptor 2 $(HER2)^-$, $HER2^+$, Triple-Negative); accuracy reported rather than AUC-ROC on test sets; with inter-centre transfer learning [106]. § Benign versus malignant discrimination, not molecular subtype classification; included for reference on multiparametric MRI utility [107]. * Approximate per-fold test set size under 10-fold cross-validation; no single held-out partition.

Study	**Task**	N **(patients)**		**AUC-ROC per subtype**				**Best / mean**	**Val.**	**MRI**	**Dim.**
		Train	**Test**	**LA**	**LB**	**H2**	**TN**	**AUC-ROC**			
Ha et al. [104]†	4-class	176	40	—	—	—	—	0.853 (macro)	Int.	DCE	2D
Zhu et al. [105]	Binary	243	27*	0.65	—	—	—	0.65 (LA)	Int.	DCE	2D
Zhang et al. [106]‡	3-class	99	145	—	—	—	—	0.91 acc. (TL)	Ext.	DCE	2D
Hu et al. [107]§	B vs M	493	123*	—	—	—	—	0.87 (B v M)	Int.	DCE+T2w	2D
Yin et al. [108]	4-class	—	—	0.817	0.762	0.885	0.920	0.846 (mean)	Int.	MP-MRI	2D
Zhou et al. [109]	4-class	311	97	—	—	—	—	0.93–0.97	Int.	DCE+DWI	2D
Xie et al. [110]	4-class	292	25	0.857	0.818	0.850	0.890	0.854 (mean)	Int.	DCE	2D
Wang et al. [111]	4-class	289	398	0.77	0.70	0.78	0.80	0.76 (mean)	Ext.	CE-MRI	2D+3D
This work (Duke)	**4-class**	**922**	**98**	**0.9610**	**0.9205**	**0.9788**	**0.9229**	**0.9416 (weighted mean)**	**Ext.**	**DCE**	**3D**
This work (TCIA multicentric)	**4-class**	**2114**	**98**	**0.8835**	**0.8658**	**0.8127**	**0.8486**	**0.8644 (weighted mean)**	**Ext.**	**DCE**	**3D**