

UNIVERSITY OF PADUA

DEPARTMENT OF INFORMATION ENGINEERING

Master's Degree in Bioengineering

“Cross-Site Transportability in Biomedical
Machine Learning: Evaluation and Adaptation Strategies”

Supervisor:

Prof. Morten Gram Pedersen

Candidate:

William Bertaiola

Student ID: 2163485

Co-supervisor:

Prof. Carlos Sáez Silvestre
Dr. David Fernández Narro

ACADEMIC YEAR 2025/2026

Facta, non verba

Abstract

In recent years, machine learning has assumed a relevant role in the biomedical field for clinical, diagnostic, and prognostic applications. However, the effective use of these models in real world clinical settings is often limited by their reduced transferability, defined as the ability to maintain reliable predictive performance when applied to data, populations, or environments different from those used during training. This limitation is particularly critical in multi site biomedical studies, where differences in acquisition protocols, instrumentation, patient populations, and data characteristics may generate domain shift and compromise model performance outside the original training context.

The objective of this thesis is to systematically investigate cross-site transportability in biomedical machine learning by comparing different adaptation and harmonization strategies across heterogeneous domains. Beyond comparing predictive performance, the thesis also aims to analyze how dataset and domain characteristics influence model behavior across different settings.

The experimental evaluation focused on supervised classification tasks designed to simulate realistic cross-site scenarios by separating source domains from target domain. Five approaches were compared: a baseline, a statistical harmonization method based on ComBat, a domain weighting strategy, a transductive learning approach, and a hybrid strategy combining ComBat harmonization and transductive adaptation. Each approach was evaluated using Logistic Regression, Random Forest, and Histogram Gradient Boosting classifiers under multiple cross-site configurations. To support the interpretation of the results, domain shift characterization and pattern discovery were performed using the dashi framework, while meta learning and SHAP analyses were used to investigate the relationship between domain characteristics and model performance. Performance was assessed using complementary evaluation metrics considering both in-distribution (ID) and out-of-distribution (OOD) behavior.

The study was conducted on three biomedical datasets representing different clinical domains and data modalities. UCI Heart Disease included 920 samples from 4 domains and represented a structured clinical setting. ABIDE II included 1114 subjects from 19 sites, with 17 sites eligible as target domains, and represented a multisite neuro clinical setting based on phenotypic variables related to autism spectrum disorder. TCGA-BRCA included 829 samples from 13 sites and represented a high dimensional genomic setting related to breast cancer. Together, these datasets provided heterogeneous scenarios with different levels of

dimensionality, domain variability, and cross-site complexity.

The results showed that no single adaptation strategy consistently outperformed all others across datasets, models, and evaluation criteria. UCI Heart Disease represented the most favourable transportability scenario, where target aware methods, especially transductive adaptation, provided the clearest benefits. ABIDE II showed a more heterogeneous and target site dependent behaviour, with moderate advantages from adaptation depending on the metric considered. TCGA-BRCA was the most challenging dataset, where the baseline often remained the strongest option for OOD discrimination, although transductive methods improved calibration. Overall, these findings suggest that cross-site transportability is a context dependent property rather than an intrinsic characteristic of a model. Adaptation should therefore be selected according to the type of domain shift, the available target domain information, and the practical objective of the biomedical application.

Abstract

Negli ultimi anni, il machine learning ha assunto un ruolo rilevante in ambito biomedico per applicazioni cliniche, diagnostiche e prognostiche. Tuttavia, l'impiego efficace di questi modelli in contesti clinici reali è spesso limitato dalla loro ridotta trasferibilità, definita come la capacità di mantenere prestazioni predittive affidabili quando vengono applicati a dati, popolazioni o ambienti differenti da quelli utilizzati durante l'addestramento. Questa limitazione risulta particolarmente critica negli studi biomedici multisito, nei quali differenze nei protocolli di acquisizione, nella strumentazione, nelle popolazioni di pazienti e nelle caratteristiche dei dati possono generare fenomeni di domain shift e compromettere le prestazioni del modello al di fuori del contesto di addestramento originale.

L'obiettivo di questa tesi è analizzare sistematicamente la trasferibilità cross-site nel machine learning biomedico, confrontando differenti strategie di adattamento e armonizzazione tra domini eterogenei. Oltre al confronto delle prestazioni predittive, la tesi si propone di analizzare come le caratteristiche dei dataset e dei domini influenzino il comportamento dei modelli in contesti differenti.

La valutazione sperimentale si è concentrata su task di classificazione supervisionata progettati per simulare scenari cross-site realistici, separando i domini sorgente dal dominio target. Sono stati confrontati cinque approcci: un metodo baseline, un metodo di armonizzazione statistica basato su ComBat, una strategia di pesatura dei domini, un approccio di apprendimento trasduttivo e una strategia ibrida che combina l'armonizzazione mediante ComBat con l'adattamento trasduttivo. Ciascun approccio è stato valutato utilizzando i classificatori Logistic Regression, Random Forest e Histogram Gradient Boosting in molteplici configurazioni cross-site. Per supportare l'interpretazione dei risultati, la caratterizzazione del domain shift e l'identificazione di pattern sono state effettuate mediante il framework dashi, mentre analisi di meta-learning e SHAP sono state utilizzate per studiare la relazione tra le caratteristiche dei domini e le prestazioni dei modelli. Le prestazioni sono state valutate mediante metriche complementari, considerando sia il comportamento in-distribution (ID) sia quello out-of-distribution (OOD).

Lo studio è stato condotto su tre dataset biomedici rappresentativi di differenti domini clinici e modalità di dati. UCI Heart Disease comprendeva 920 campioni provenienti da 4 domini e rappresentava un contesto clinico basato su dati strutturati. ABIDE II comprendeva 1114 soggetti provenienti da 19 siti, di cui 17 idonei a essere utilizzati come domini target, e

rappresentava un contesto neuroclinico multisito basato su variabili fenotipiche relative al disturbo dello spettro autistico. TCGA-BRCA comprendeva 829 campioni provenienti da 13 siti e rappresentava un contesto genomico ad alta dimensionalità relativo al carcinoma mammario. Nel complesso, questi dataset hanno permesso di analizzare scenari eterogenei caratterizzati da differenti livelli di dimensionalità, variabilità tra domini e complessità cross-site.

I risultati hanno mostrato che nessuna singola strategia di adattamento ha ottenuto prestazioni sistematicamente superiori rispetto alle altre in tutti i dataset, modelli e criteri di valutazione. UCI Heart Disease ha rappresentato lo scenario più favorevole in termini di trasferibilità, nel quale i metodi che utilizzano informazioni sul dominio target, in particolare l'adattamento trasduttivo, hanno fornito i benefici più evidenti. ABIDE II ha mostrato un comportamento più eterogeneo e dipendente dal sito target, con vantaggi moderati derivanti dall'adattamento a seconda della metrica considerata. TCGA-BRCA si è rivelato il dataset più complesso, nel quale il metodo baseline è spesso rimasto l'opzione più efficace per la discriminazione OOD, sebbene i metodi trasduttivi abbiano migliorato la calibrazione. Nel complesso, questi risultati suggeriscono che la trasferibilità cross-site sia una proprietà dipendente dal contesto, piuttosto che una caratteristica intrinseca di un modello. La strategia di adattamento dovrebbe pertanto essere selezionata in funzione del tipo di domain shift, delle informazioni disponibili sul dominio target e dell'obiettivo pratico dell'applicazione biomedica.

List of Acronyms

ML	Machine Learning
LR	Logistic Regression
RF	Random Forest
HGB	Histogram Gradient Boosting
UCI	University of California Irvine
ABIDE	Autism Brain Imaging Data Exchange
TCGA-BRCA	The Cancer Genome Atlas Breast Invasive Carcinoma
ID	In-Distribution
OOD	Out-of-Distribution
ID–OOD	In-Distribution–Out-of-Distribution
ROC-AUC	Receiver Operating Characteristic Area Under the Curve
PR-AUC	Precision–Recall Area Under the Curve
F1	F1-score
ECE	Expected Calibration Error
DASHI	Dataset Shift Identification
MSV	Multi Source Variability
GPD	Global Probabilistic Deviation
SPO	Source Probabilistic Outlyingness
SHAP	SHapley Additive exPlanations

Contents

Abstract	4
List of Acronyms	8
1 Introduction	17
1.1 Machine Learning in Biomedical Applications	17
1.2 Domain Shift and Cross site Transportability	18
1.3 Objectives	21
2 Background and Related Work	23
2.1 Cross-Site Transportability and Out-of-Distribution Generalization	23
2.2 Cross-Site Adaptation and Harmonization Strategies	25
2.2.1 ComBat Harmonization	25
2.2.2 Domain Weighting	26
2.2.3 Transductive Learning	27
2.2.4 Hybrid Harmonization and Transductive Learning	28
2.3 Calibration, Robustness and Model Reliability	29
3 Materials and Methods	31
3.1 Overview of the Selected Datasets	31
3.1.1 Dataset Preprocessing	32
3.1.2 UCI Heart Disease	32
3.1.3 ABIDE II	34
3.1.4 TCGA-BRCA	36
3.2 Machine Learning Models	39
3.2.1 Logistic Regression	39
3.2.2 Random Forest	40
3.2.3 Histogram Gradient Boosting	41
3.3 dashi Benchmark	42
3.4 Cross-Site Experimental Protocol	44
3.4.1 Source and Target Domains	44
3.4.2 Target Split Strategy	45
3.4.3 Adaptation Strategies (M0–M4)	46
3.4.4 ID and OOD Evaluation	47

3.5	Evaluation Metrics	48
3.6	Transportability Characterization Analyses	50
3.6.1	Bootstrap Stability Analysis	50
3.6.2	Entropy, Sample Size and Site Distance	51
3.6.3	Metadata Shift Analysis	52
3.6.4	Clustered Heatmaps	53
3.6.5	SHAP Meta Learning Analysis	54
4	Results	56
4.1	Descriptive analysis of the databases	56
4.2	dashi Analysis	57
4.2.1	UCI Heart Disease	57
4.2.2	ABIDE II	59
4.2.3	TCGA-BRCA	61
4.3	Performance of Adaptation Strategies	64
4.3.1	UCI Heart Disease	64
4.3.2	ABIDE II	68
4.3.3	TCGA-BRCA	72
4.4	Cross Dataset Comparison	76
4.5	Robustness and Reliability Analyses	77
4.5.1	Calibration	77
4.5.2	Entropy, Sample Size and Site Distance	78
4.5.2.1	UCI Heart Disease	79
4.5.2.2	ABIDE II	81
4.5.2.3	TCGA-BRCA	83
4.5.3	Source Target Covariate Shift Analysis	86
4.5.3.1	UCI Heart Disease	86
4.5.3.2	ABIDE II	87
4.5.3.3	TCGA-BRCA	88
4.6	Pattern Discovery and Meta learning Analysis	91
4.6.1	Correlation Heatmaps	91
4.6.1.1	UCI Heart Disease	92
4.6.1.2	ABIDE II	93
4.6.1.3	TCGA-BRCA	93
4.6.2	SHAP Meta Learning Analysis	94
4.6.2.1	UCI Heart Disease	95
4.6.2.2	ABIDE II	97
4.6.2.3	TCGA-BRCA	98
5	Discussion	101
5.1	Overall Interpretation	101

5.2	Dataset Specific Interpretation	102
5.2.1	UCI Heart Disease	102
5.2.2	ABIDE II	102
5.2.3	TCGA-BRCA	103
5.3	Adaptation Strategies and Evaluation Criteria	103
5.4	Calibration and Reliability	105
5.5	Transportability	105
5.6	Implications for Cross-Site Biomedical Machine Learning	106
5.7	Limitations and Future Directions	107
6	Conclusions	109
6.1	Main Findings	109
6.2	Final Remarks	110
	Bibliography	111

List of Figures

1.1	Overall workflow of the proposed framework for evaluating cross-site trans- portability in biomedical machine learning. The benchmark compares five methodological settings (M0–M4): baseline (M0), ComBat harmonization (M1), domain weighting (M2), transductive learning (M3), and hybrid har- monization with transductive learning (M4). Dataset characterization is per- formed using <i>dashi</i> , while model performance is assessed using the area under the receiver operating characteristic curve ROC-AUC, F1-score, bal- anced accuracy, expected calibration error ECE, and the in-distribution ID out-of-distribution OOD performance gap. Model interpretation is performed using Shapley additive explanations SHAP.	22
4.1	Site wise sample size and class distribution across the three study datasets. For UCI Heart Disease, Class 0 indicates absence of heart disease and Class 1 indicates presence of heart disease. For ABIDE II, Class 0 corresponds to Control subjects and Class 1 to Autism Spectrum Disorder (ASD). For TCGA-BRCA, Class 0 corresponds to Stage I–II tumours and Class 1 to Stage III–IV tumours.	56
4.2	Two dimensional Multi Source Variability representation for the UCI Heart Disease dataset. Point size reflects domain sample size, while color indicates the source overlap score (SPO).	58
4.3	Source to target accuracy heatmap generated by <i>dashi</i> for the UCI Heart Disease dataset. Rows represent the training domain and columns represent the target domain.	58
4.4	Two dimensional Multi Source Variability representation for the ABIDE II dataset. Point size reflects domain sample size, while color indicates the source overlap score (SPO).	60
4.5	Source to target accuracy heatmap generated by <i>dashi</i> for the ABIDE II dataset. Rows represent the training domain and columns represent the target domain.	60
4.6	Two dimensional Multi Source Variability representation for the TCGA- BRCA dataset. Point size reflects domain sample size, while color indicates the source overlap score (SPO).	62

4.7	Source to target accuracy heatmap generated by dashi for the TCGA-BRCA dataset. Rows represent the training domain and columns represent the target domain.	63
4.8	Bootstrap mean OOD balanced accuracy across UCI target sites for Logistic Regression. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations.	65
4.9	Bootstrap mean OOD balanced accuracy across UCI target sites for Random Forest. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations.	65
4.10	Bootstrap mean OOD balanced accuracy across UCI target sites for Histogram Gradient Boosting. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations. . .	66
4.11	ID–OOD balanced accuracy gap across UCI target sites for Logistic Regression. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.	67
4.12	ID–OOD balanced accuracy gap across UCI target sites for Random Forest. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.	67
4.13	ID–OOD balanced accuracy gap across UCI target sites for Histogram Gradient Boosting. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.	68
4.14	Bootstrap mean OOD balanced accuracy across ABIDE II target sites for Logistic Regression. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations. . .	69
4.15	Bootstrap mean OOD balanced accuracy across ABIDE II target sites for Random Forest. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations. . .	70
4.16	Bootstrap mean OOD balanced accuracy across ABIDE II target sites for Histogram Gradient Boosting. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations.	71
4.17	ID–OOD balanced accuracy gap across ABIDE II target sites for Logistic Regression. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.	71
4.18	ID–OOD balanced accuracy gap across ABIDE II target sites for Random Forest. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.	72
4.19	ID–OOD balanced accuracy gap across ABIDE II target sites for Histogram Gradient Boosting. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.	72

4.20	Bootstrap mean OOD balanced accuracy across TCGA-BRCA target sites for Logistic Regression. Bars represent the five adaptation strategies M_0 – M_4 , and error bars indicate the standard deviation across 100 bootstrap iterations.	73
4.21	Bootstrap mean OOD balanced accuracy across TCGA-BRCA target sites for Random Forest. Bars represent the five adaptation strategies M_0 – M_4 , and error bars indicate the standard deviation across 100 bootstrap iterations. . .	73
4.22	Bootstrap mean OOD balanced accuracy across TCGA-BRCA target sites for Histogram Gradient Boosting. Bars represent the five adaptation strategies M_0 – M_4 , and error bars indicate the standard deviation across 100 bootstrap iterations.	74
4.23	ID–OOD ROC-AUC gap across TCGA-BRCA target sites for Logistic Regression. Rows represent the adaptation strategies M_0 – M_4 , while columns represent the target sites.	74
4.24	ID–OOD ROC-AUC gap across TCGA-BRCA target sites for Random Forest. Rows represent the adaptation strategies M_0 – M_4 , while columns represent the target sites.	75
4.25	ID–OOD ROC-AUC gap across TCGA-BRCA target sites for Histogram Gradient Boosting. Rows represent the adaptation strategies M_0 – M_4 , while columns represent the target sites.	75
4.26	Mean OOD ROC-AUC across datasets, adaptation strategies, and machine learning models. Error bars represent variability across target sites.	76
4.27	Delta OOD ROC-AUC relative to M_0 as a function of class entropy in UCI Heart Disease.	80
4.28	Delta OOD ROC-AUC relative to M_0 as a function of target sample size in UCI Heart Disease.	80
4.29	Delta OOD ROC-AUC relative to M_0 as a function of source–target distance in UCI Heart Disease.	81
4.30	Delta OOD ROC-AUC relative to M_0 as a function of class entropy in ABIDE II.	82
4.31	Delta OOD ROC-AUC relative to M_0 as a function of target sample size in ABIDE II.	82
4.32	Delta OOD ROC-AUC relative to M_0 as a function of source–target distance in ABIDE II.	83
4.33	Delta OOD ROC-AUC relative to M_0 as a function of class entropy in TCGA-BRCA.	84
4.34	Delta OOD ROC-AUC relative to M_0 as a function of target sample size in TCGA-BRCA.	84
4.35	Delta OOD ROC-AUC relative to M_0 as a function of source–target distance in TCGA-BRCA.	85
4.36	Variable importance for site prediction in UCI Heart Disease	86

4.37	Metadata shift clustermap across UCI Heart Disease target sites. Values are reported as standardized Z-scores.	87
4.38	Metadata importance for site prediction in ABIDE II.	88
4.39	Metadata shift clustermap across ABIDE II target sites. Values are reported as standardized Z-scores.	89
4.40	Metadata importance for site prediction in TCGA-BRCA.	90
4.41	Metadata shift clustermap across TCGA-BRCA target sites. Values are reported as standardized Z-scores.	90
4.42	Pearson correlation heatmap between site distance descriptors and transportability indicators in UCI Heart Disease.	92
4.43	Pearson correlation heatmap between site distance descriptors and transportability indicators in ABIDE II.	93
4.44	Pearson correlation heatmap between site distance descriptors and transportability indicators in TCGA-BRCA.	94
4.45	Mean absolute SHAP importance for UCI Heart Disease across methods M0–M4.	95
4.46	SHAP feature importance rankings for OOD ROC-AUC in UCI Heart Disease across methods M0–M4.	96
4.47	SHAP beeswarm plot for OOD ROC-AUC in UCI Heart Disease across methods M0–M4.	96
4.48	Mean absolute SHAP importance for ABIDE II across methods M0–M4. . .	97
4.49	SHAP feature importance rankings for OOD ROC-AUC in ABIDE II across methods M0–M4.	98
4.50	SHAP beeswarm plot for OOD ROC-AUC in ABIDE II across methods M0–M4.	98
4.51	Mean absolute SHAP importance for TCGA-BRCA across methods M0–M4. . .	99
4.52	SHAP feature importance rankings for OOD ROC-AUC in TCGA-BRCA across methods M0–M4.	100
4.53	SHAP beeswarm plot for OOD ROC-AUC in TCGA-BRCA across methods M0–M4.	100

List of Tables

3.1	Overview of the biomedical datasets used in this thesis.	31
3.2	Predictive variables included in the UCI Heart Disease dataset.	33
3.3	Variables included in the ABIDE II strict dataset.	35
3.4	Clinical, barcode derived, and quality control variables retained in the TCGA- BRCA dataset.	37
3.5	Adaptation strategies evaluated in the cross-site benchmark.	46
4.1	Variables included in the site distance correlation heatmaps.	91

Chapter 1

Introduction

1.1 Machine Learning in Biomedical Applications

Machine learning (ML) represents a branch of artificial intelligence focused on the development of algorithms capable of automatically learning from data, identifying patterns and relationships among them. In recent years, the increasing availability and complexity of biomedical data have made machine learning an increasingly relevant tool in the field of research and healthcare. [1]

In the biomedical field, a dataset can be defined as a collection of data related to patients or biological samples in which each instance of the dataset is described by a series of characteristics or features, namely numerical or categorical variables used by the model to learn patterns and relationships in the data. The machine learning model is therefore trained on these datasets through a specific algorithm with the objective of performing predictions, classifications, or supporting decision making processes on new data. [1]

The approaches to machine learning are divided in supervised and unsupervised learning. In supervised learning, the model is trained on labeled data, where each input sample is associated with a predefined output. This approach is used in the biomedical field for tasks of diagnosis, prognosis, and clinical decision support. Supervised learning are divided into classification and regression tasks. Classification has the objective of assigning data to discrete categories, while regression is used to predict continuous values. [1]

Unsupervised learning analyzes unlabeled data with the objective of autonomously identifying patterns, relationships, or hidden structures. In the biomedical field, these techniques are used for patient clustering, dimensionality reduction identifying most influential variables, and the identification of anomalies or disease subtypes. Such approaches are particularly useful for the analysis of complex, high dimensional datasets. [1]

The use of machine learning has contributed to progress in the fields of diagnosis, prognosis, treatment personalization and drug discovery. However, the application of these models requires a critical evaluation not only of their predictive performance, but also of their

robustness, reliability, interpretability, and clinical applicability. [1, 2]

A central challenge concerns the trade off between accuracy and interpretability, since highly complex models may achieve strong predictive performance but result difficult to interpret. This limited interpretability can represent a relevant issue in medicine, where understanding the clinical and biological basis of model decisions is essential to support trust, guide treatment development, and enable the practical adoption of clinical decision support systems.

A central challenge concerns the ability of machine learning models to maintain their performance outside the context in which they were developed. In many biomedical studies, models are trained and validated on data collected under specific conditions, whereas their real clinical value depends on their ability to generalize to different institutions, populations, acquisition protocols, and technological platforms. This issue introduces the broader problem of domain shift and cross site transportability, which represents the central focus of this thesis.

1.2 Domain Shift and Cross site Transportability

A common assumption of most machine learning models is that the data used during training and testing originate from the same probability distribution. Under this assumption, the statistical properties observed during model development remain stable when the model is applied to new data, allowing the performance obtained during validation to be interpreted as a reliable estimate of future performance. In real world biomedical applications this assumption is often violated. Data collected across different institutions, populations, acquisition settings, technological platforms, or time periods can exhibit systematic differences, leading to a mismatch between training data and data encountered during deployment. [3, 4, 2]

This phenomenon is referred to as dataset shift or distribution shift and occurs whenever the probability distribution of the training data differs from that of the testing data. Traditional machine learning methods implicitly assume that any such differences are negligible, but this assumption rarely holds in clinical practice. The presence of dataset shift can compromise the generalization ability of predictive models, causing a deterioration in performance when they are applied to previously unseen environments [3].

From a theoretical perspective, a model trained on samples generated from a source distribution $P_{\text{train}}(X, Y)$ is optimized to operate under those specific statistical conditions. When the same model is applied to data generated from a different distribution $P_{\text{test}}(X, Y)$, the assumptions underlying the learning process may no longer be valid. As a consequence, the performance observed during model development may no longer reflect the expected behaviour in the target domain, potentially reducing accuracy, robustness, and reliability [3].

More generally, given a dataset and a modelling strategy, it is possible to learn a predictive function that estimates a target variable y from observations x , namely $P(y|x)$. However, if the conditions generating the data change between training and deployment, it becomes necessary to understand whether the learned relationships remain valid. Characterizing and managing these changes is therefore a fundamental requirement for the successful application of machine learning in real world biomedical settings [3].

The causes of dataset shift are numerous. In healthcare applications, shifts may arise from demographic differences between populations, changes in disease prevalence, modifications of diagnostic procedures, the introduction of new technologies, or differences in data acquisition and management processes [5]. These factors can alter both the distribution of individual variables and the relationships among them, creating discrepancies between the data used for model development and those encountered during practical use.

The concept of domain can be defined as the set of conditions characterizing the data generating process. In biomedical applications, a domain may correspond to a hospital, a clinical centre, an imaging scanner, a genomic sequencing platform, a geographical region, or a specific patient population. When a model is developed in one or more source domains and subsequently applied to a different target domain, differences in population characteristics, acquisition procedures, measurement protocols, or clinical practices may introduce domain shift.

Domain shift is common in biomedical research because the biological process of interest often remains the same while the way it is measured changes. Examples include differences in imaging scanners, acquisition protocols, reconstruction algorithms, preprocessing pipelines, laboratory procedures, or sequencing technologies. In these situations, the underlying biological signal remains largely unchanged, whereas its representation within the data varies across domains. [6]

Closely related to domain shift is the concept of out-of-distribution (OOD) generalization. In conventional machine learning, predictive performance is typically evaluated on test samples drawn from the same distribution as the training data. In practice deployed models frequently encounter data originating from different environments. OOD generalization therefore refers to the ability of a model to maintain its predictive performance when evaluated on data generated under conditions that differ from those observed during training. [7]

From this perspective, cross-site transportability can be viewed as a form of OOD generalization. Each clinical centre, imaging facility, or sequencing laboratory may represent a distinct domain characterized by its own patient population, acquisition protocols, and technical conditions. Evaluating transportability therefore requires assessing not only how well a model performs within the environments used during development, but also how effectively it generalizes to previously unseen sites. [7]

The consequences of domain shift are not limited to a reduction in predictive accuracy. Models may become systematically biased, produce unreliable probability estimates, or exhibit different performance across patient subgroups. These issues are important in medicine, where model outputs may influence diagnostic, prognostic, or therapeutic decisions. Ignoring domain shift can therefore compromise the safety, fairness, and clinical utility of machine learning systems [5, 4, 2] .

Cross-site transportability refers to the ability of a machine learning model developed using data from one or more source sites to maintain its predictive performance when applied to data collected at a different target site. This concept is especially relevant in biomedical applications because models are often developed using data from a limited number of institutions but are expected to operate reliably across a much broader range of clinical settings.

In this context, cross-site transportability can be viewed as a practical form of domain generalization. A model is trained on labelled data from several source environments and then evaluated on a target environment that was not available during training [8]. This setting reflects many real biomedical applications, where models developed in one group of institutions are expected to remain reliable when applied to new clinical sites.

Several studies have shown that machine learning models frequently experience substantial performance degradation when evaluated on external datasets collected at different institutions [8]. These findings demonstrate that strong internal validation performance alone is insufficient to guarantee successful deployment across sites.

The evaluation of transportability extends beyond predictive accuracy alone. A model may retain acceptable discrimination performance while becoming unstable across domains or producing poorly calibrated probability estimates. For this reason, robustness and calibration have emerged as important complementary dimensions of model evaluation.

Robustness refers to the ability of a model to maintain stable performance despite variations in the data generating conditions. In cross-site prediction, robust models should exhibit limited sensitivity to differences in population composition, acquisition procedures, and sampling variability. Calibration describes the agreement between predicted probabilities and observed outcome frequencies. Well calibrated models provide probability estimates that can be interpreted as reliable measures of uncertainty, an aspect that is particularly important in clinical decision making.

The importance of transportability has increased considerably with the growing availability of large multi centre biomedical datasets. Although the inclusion of multiple sites can improve data diversity and potentially enhance generalization, it also introduces additional sources of variability that must be explicitly assessed and reported. In multi site biomedical machine learning, data sharing alone is not sufficient if differences in data quality and source variability are not characterized, since these factors may reduce the representativeness of

training data and introduce bias in model development [4]. Consequently, several strategies have been proposed to improve transportability, including harmonization procedures, domain adaptation methods, domain generalization approaches, sample reweighting techniques, and calibration strategies. Their effectiveness remains highly dependent on the characteristics of the data and the nature of the domain shift [8, 4].

Understanding the factors that influence cross-site transportability is therefore essential for the successful deployment of machine learning models in healthcare. A model that performs well only within the environment in which it was developed has limited practical value, whereas a transportable model can support decision making across different institutions, populations, and clinical settings. Transportability has become a central component of trustworthy biomedical machine learning, complementing traditional assessments of predictive performance and internal validation.

1.3 Objectives

The main objective of this thesis is to investigate whether adaptation and harmonization strategies can improve the cross-site transportability of machine learning models across heterogeneous biomedical datasets.

To achieve this goal, the following specific objectives are addressed:

1. Evaluate the ability of ML models to maintain predictive performance when transferred across different biomedical domains.
2. Compare the effectiveness of different adaptation and harmonization strategies in improving cross-site transportability.
3. Assess transportability using complementary evaluation criteria, including discrimination, calibration, and robustness under out-of-distribution conditions.
4. Identify the dataset and domain characteristics that influence the effectiveness of different adaptation strategies.
5. Provide practical recommendations for selecting adaptation strategies according to the characteristics of the target domain and the intended biomedical application.

This thesis contributes primarily to the United Nations Sustainable Development Goal 3, Good Health and Well being, by investigating methods that improve the reliability and transportability of artificial intelligence models across healthcare institutions, supporting the development of more robust clinical decision support systems. It also contributes to

Sustainable Development Goal 9, Industry, Innovation and Infrastructure, by advancing methodological approaches for trustworthy and transferable biomedical machine learning in multi site environments.

The overall workflow adopted in this thesis is summarized in Figure 1.1.

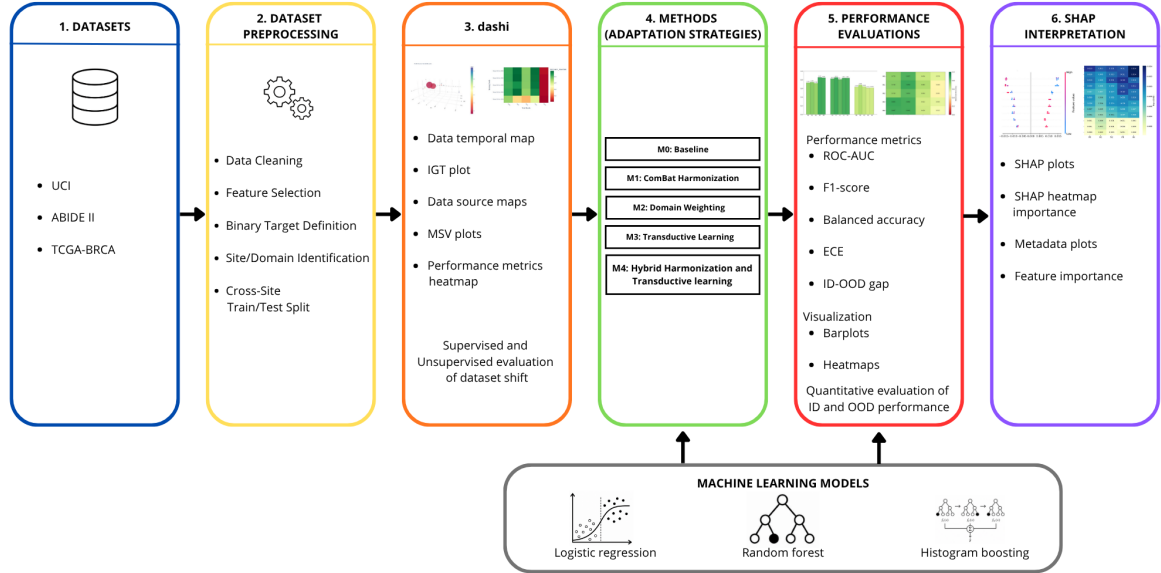


Figure 1.1: Overall workflow of the proposed framework for evaluating cross-site transportability in biomedical machine learning. The benchmark compares five methodological settings (M0–M4): baseline (M0), ComBat harmonization (M1), domain weighting (M2), transductive learning (M3), and hybrid harmonization with transductive learning (M4). Dataset characterization is performed using *dashi*, while model performance is assessed using the area under the receiver operating characteristic curve ROC-AUC, F1-score, balanced accuracy, expected calibration error ECE, and the in-distribution ID out-of-distribution OOD performance gap. Model interpretation is performed using Shapley additive explanations SHAP.

Chapter 2

Background and Related Work

2.1 Cross-Site Transportability and Out-of-Distribution Generalization

Out-of-distribution (OOD) generalization refers to the ability of a predictive model to maintain its performance when applied to data generated under conditions that differ from those observed during training. Formally, the problem arises whenever the training distribution $P_{tr}(X, Y)$ differs from the test distribution $P_{ts}(X, Y)$, a situation in which the statistical assumptions underlying model development are no longer fully satisfied [9]. OOD evaluation assesses the ability of a model to generalize beyond the environments used during development, while in-distribution (ID) evaluation is based on training and test samples originate from the same distribution.

This problem is relevant in biomedical machine learning where data collected at different hospitals, imaging centres, laboratories, or research institutions often differ in both population characteristics and data generation process, creating substantial sources of variability across sites. As a consequence, models that perform well during internal validation may experience a substantial reduction in performance when deployed in a new environment.

Within this context, cross-site transportability can be regarded as a practical form of OOD generalization. A model is developed using data from one or more source sites and subsequently evaluated on a target site that was not used during training. Good transportability is achieved when predictive performance remains stable despite the presence of differences between the source and target domains. Conversely, large discrepancies between ID and OOD performance indicate limited transportability and suggest that the model has learned relationships that are specific to the training environments rather than robust patterns that generalize across sites.

Performance degradation under OOD conditions is often related to dataset shift, namely a change in the data distribution between training and deployment. In cross-site biomedical applications, this dataset shift is usually observed as domain shift, because each site, hospital,

acquisition setting, or cohort can be interpreted as a different domain. In other words, domain shift represents the practical manifestation of dataset shift across domains.

Dataset shift can arise from multiple sources and may affect different components of the data. Covariate shift occurs when the distribution of the input variables changes across domains. Concept shift refers to changes in the relationship between the available predictors and the target outcome across environments. In biomedical datasets, these forms of shift may coexist and can emerge from differences across populations, clinical settings, and data collection procedures [7].

The domain shift is not always directly observable from the predictive features alone. Differences between domains also be reflected in auxiliary variables describing the characteristics of the data collection process. These differences are referred to as metadata shift and involve demographic variables, clinical descriptors, phenotypic characteristics, or technical information related to data acquisition. Although metadata shift does not fully capture all aspects of domain heterogeneity, it can provide insight into how source and target domains differ and may therefore help explain variations in transportability.

The magnitude of the discrepancy between domains can be described through measures of domain or site distance. These measures quantify how different two domains are by comparing their distributions across relevant data characteristics, such as feature values, metadata, or class composition. They provide a numerical estimate of the degree of heterogeneity between the source and target domains. Larger distances indicate greater differences between domains and may increase the difficulty of OOD generalization, although their impact depends on the nature of the underlying shift. For this reason, characterizing and identifying domain shift has become an important step in the development and evaluation of transportable machine learning models.

The OOD generalization challenge is that machine learning models are optimized to minimize the average prediction error observed during training. They may exploit predictive correlations that are valid within the source domains but unstable across environments. Such correlations can improve performance under ID conditions, but they may become unreliable when the model is transferred to a different site. High internal validation performance should therefore not be interpreted as evidence of successful deployment in unseen environments.

Evaluating transportability requires going beyond conventional measures of predictive accuracy. It is also necessary to assess robustness across domains, calibration of probabilistic predictions, and the relationship between predictive performance and characteristics of the target environment. Understanding these factors is essential for identifying the conditions under which machine learning models can be safely and reliably transferred across heterogeneous biomedical settings.

2.2 Cross-Site Adaptation and Harmonization Strategies

Different strategies have been proposed to reduce domain shift and improve OOD generalization.

- **Harmonization methods** aim to reduce site related variability directly in the feature space by removing systematic differences introduced by data acquisition or processing while preserving the underlying biological signal [10].
- **Weighting methods** modify the contribution of training samples according to their relevance to the target domain, with the objective of reducing the mismatch between source and target data distributions [11].
- **Adaptation methods** exploit information available from the target domain, even in the absence of target labels, to facilitate model adaptation before deployment [12].
- **Hybrid methods** combine two or more adaptation mechanisms within a single framework in order to address different sources of domain shift simultaneously [12].

Adaptation strategies are compared against a baseline. This comparison helps determine whether an adaptation procedure provides a real advantage over standard model.

The methods evaluated in this work are ComBat harmonization, domain weighting, transductive learning, and a hybrid strategy combining harmonization and transductive adaptation. Together, these approaches represent different ways of addressing cross-site distribution shift and provide a useful framework for studying the conditions under which adaptation improves out-of-distribution performance.

2.2.1 ComBat Harmonization

ComBat is a harmonization method originally developed to correct batch effects in high dimensional biological data through an Empirical Bayes framework [13]. Batch effects are systematic variations introduced by differences across sites or processing procedures. In multi site biomedical datasets, these effects can make samples collected at different centres only partially comparable and may reduce the transportability of predictive models.

ComBat assumes that site related variability can be represented through additive and multiplicative effects acting on each feature. In its original formulation, the observed value of a feature is decomposed into a biological component that should be preserved and a batch specific component that represents unwanted variability, as shown in Equation 2.1:

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg} \quad (2.1)$$

where the biological covariates of interest are represented by $X\beta_g$, while γ_{ig} and δ_{ig} model additive and multiplicative batch effects, respectively [13].

ComBat uses Empirical Bayes estimation, which improves the stability of site effect estimates by sharing information across features instead of estimating each correction parameter separately. This approach is advantageous in biomedical datasets, where the number of features is large and the number of samples available within each site may be limited. By borrowing statistical strength across features, ComBat can provide robust harmonization even in relatively small datasets.

Although ComBat was originally developed for gene expression studies, it has later been extended to multi site neuroimaging, where it has been shown to reduce scanner and site related variability while preserving biologically meaningful information.[10]. Their study showed that harmonization improved measurement consistency across acquisition sites, increased statistical power, and maintained associations with relevant biological variables, supporting the use of ComBat in large multi centre studies. These findings contributed to establishing ComBat as a reference harmonization approach for neuroimaging applications and, more broadly, for biomedical datasets affected by site dependent variability. [10]

ComBat acts before model training by reducing systematic differences between domains. The aim is to make feature distributions more comparable across sites, so that the model can focus on relevant patterns. During deployment, data from new sites must also be harmonized using the same harmonization procedure before being provided to the trained model. This approach is most effective when the observed variability mainly reflects unwanted technical differences rather than information genuinely associated with the outcome.

2.2.2 Domain Weighting

Domain weighting adjusts the importance assigned to each source domain during model training. The idea is that, in a multi site setting, some source domains may be more informative than others for generalizing to a target site. Some domains may contain patterns that are stable across sites and therefore more transferable, whereas others may be more strongly influenced by local characteristics and site specific biases.

The method is conceptually related to importance weighting, which is commonly used to address situations in which the training and test distributions differ. Under covariate shift, assigning different weights to training samples can reduce the mismatch between the empirical training distribution and the distribution encountered during deployment [11]. Rather than

assuming that all observations contribute equally to the learning process, weighting allows the model to place greater emphasis on the samples that are expected to be more informative for generalization.

In the context of multi site biomedical studies, weighting can be performed at the domain level rather than at the individual sample level. The central idea is to estimate how transferable each source domain is with respect to the remaining source domains and to use this information to guide model training. Domains that exhibit stronger cross domain transferability receive a larger influence, whereas domains that appear less representative contribute less to the final model.

Domain weighting does not modify the feature space, it acts directly on the learning process by assigning a different weight to each training sample. Samples belonging to source domains with higher estimated transferability have a greater influence during model training, whereas samples from less transferable domains contribute less to the optimization process. As a consequence, domain weighting aims to improve robustness to domain shift without altering the input data.

From the perspective of cross-site transportability, domain weighting can be interpreted as a source only adaptation strategy. From the target domain no information is required during the estimation of the weights. The method therefore attempts to improve generalization by exploiting the internal structure of the source domains and by prioritizing those domains that appear more transferable to other environments.

2.2.3 Transductive Learning

Transductive learning is a transfer learning setting in which information from the target domain is available during model development, while the predictive task remains unchanged across source and target domains. The objective is to exploit the available target domain information in order to reduce the mismatch between training and deployment conditions and improve out-of-distribution generalization [12].

Within the transfer learning framework, a distinction is made between the predictive task and the domain in which the task is performed. In transductive transfer learning, the source and target tasks are assumed to be identical, whereas the corresponding domains differ because they are characterized by different data distributions:

$$T_S = T_T, \quad D_S \neq D_T.$$

where T_S and T_T denote the source and target tasks, respectively, D_S and D_T denote the source and target domains. The condition $T_S = T_T$ means that the prediction task is the same in both

settings. The condition $D_S \neq D_T$ means that the data come from different domains.

This situation is frequently in biomedical studies, where the same clinical outcome is predicted across multiple institutions, but populations, acquisition protocols, or measurement procedures may vary between sites.

The characteristic of transductive learning is the use of information from the target domain during the adaptation process. These data provide useful information about the structure of the deployment environment and can help the model adjust to target specific characteristics. The learning process, by incorporating this information, becomes partially aware of the target distribution rather than relying exclusively on the source domains.

In this work, the target data used for transductive learning were never incorporated into the training of the source predictive model. Instead, the source model was trained exclusively on the source domains, while the target *train_1* subset was used only for the adaptation component. During internal validation, the same protocol was reproduced using only the source domains, the held out source domain was further divided into an adaptation subset and an independent evaluation subset. The adaptation subset was used to fit the transductive component, whereas performance was computed exclusively on the evaluation subset. During OOD evaluation, the adaptation step was instead performed using the target *train_1* subset and evaluated on the independent target *test_1* subset.

In contrast to source only approaches, transductive learning starts from the idea that the target domain may not follow the same distribution as the training data. For this reason, it uses the available target information to reduce the effect of domain shift, rather than trying to build a model that is completely independent of the domain.

This is the main advantage of transductive learning, as the adaptation process can be guided by the target domain. However, the same aspect can also become a limitation if the model becomes too dependent on the specific target sample used for adaptation.

2.2.4 Hybrid Harmonization and Transductive Learning

The hybrid adaptation strategy combines harmonization and transductive learning in a single framework. The hybrid strategy is based on the observation that domain shift rarely has a single cause and may therefore require more than one adaptation mechanism.

Harmonization methods such as ComBat aim to reduce systematic site related variability in the feature space by correcting differences associated with acquisition procedures, technical settings, or batch effects [13]. By reducing variability among the source domains, harmonization seeks to improve the consistency of the data representation used for model training.

Transductive learning addresses a different aspect of the problem. Rather than focusing on the source domains, it incorporates information from the target domain in order to adapt the predictive model to the characteristics of the deployment environment [12]. This adaptation can be beneficial when the target site differs from the source domains in ways that cannot be fully corrected through harmonization alone.

The hybrid approach combines these two perspectives. Harmonization is first used to reduce source domain heterogeneity, thereby encouraging the model to learn patterns that are less dependent on specific sites. Subsequently, target aware adaptation is used to incorporate information about the target domain and improve alignment between the learned model and the deployment environment.

This approach can be viewed as a two level adaptation strategy. The first level acts on the feature representation by reducing site related variability, whereas the second level acts on the predictive process by exploiting information from the target domain. The objective is to determine whether combining feature level harmonization and target aware adaptation provides more robust transportability than either strategy alone.

However, the effectiveness of the hybrid approach cannot be assumed a priori. If harmonization removes predictive information or if target domain adaptation introduces instability, the combined strategy may not necessarily outperform simpler alternatives. For this reason, the value of M4 must be assessed empirically and compared with both the source only baseline and the individual adaptation methods.

2.3 Calibration, Robustness and Model Reliability

In biomedical machine learning, predictive performance alone is often insufficient to determine whether a model is suitable for real world deployment. Two models may achieve similar classification performance while differing substantially in the reliability of their probability estimates, their sensitivity to sampling variability, or their ability to generalize across heterogeneous populations. For this reason, model evaluation should consider not only discrimination metrics, but also calibration, robustness, and overall reliability.

Calibration refers to the agreement between predicted probabilities and observed outcomes. A model is well calibrated when predictions assigned a given probability correspond, on average, to the same empirical frequency of the event. This aspect is important in healthcare, where model outputs may be interpreted as risk estimates and can support diagnostic, prognostic, or therapeutic decisions. A model may therefore achieve good discrimination while still producing overconfident or underconfident probability estimates [14].

Several tools can be used to assess calibration, including reliability diagrams, calibration

curves, the Brier score, and the Expected Calibration Error (ECE). ECE is commonly used because it summarizes the discrepancy between predicted confidence and observed outcome frequency across groups of predictions. Lower ECE values indicate better probabilistic reliability, whereas higher values suggest poorer calibration [14].

Robustness refers to the ability of a model to maintain stable performance when applied to data that differ from those encountered during training. In multi site biomedical studies, these differences may arise from variations in patient populations, acquisition protocols, measurement devices, clinical practices, or other site specific factors. A robust model should therefore avoid large performance degradations when transferred to new domains and should remain reliable under moderate changes in the data distribution.

Closely related to robustness is the concept of stability, which describes the consistency of model performance under different sampling conditions. Performance estimates obtained from a single train–test split may be influenced by random variation in the composition of the data. Resampling procedures such as the bootstrap provide a general framework for assessing the variability of statistical estimates and are widely used to evaluate uncertainty and stability in model assessment [15].

Together, discrimination, calibration, robustness, and stability provide complementary perspectives on model quality. A model that performs well according to one criterion may still be unreliable according to another. This is particularly relevant in cross-site transportability, where a method may improve OOD discrimination but worsen calibration or increase variability across target sites.

These concepts are directly related to the objectives of this thesis. The adaptation and harmonization strategies investigated in this work are evaluated not only according to their ability to improve predictive performance, but also according to their impact on calibration, robustness, and reliability under out-of-distribution conditions. The specific metrics and experimental procedures used to assess these properties are described in Chapter 3.

Chapter 3

Materials and Methods

3.1 Overview of the Selected Datasets

Multiple biomedical datasets were initially explored for this thesis, with the main selection criterion being the availability of a clear site or domain identifier suitable for cross-site evaluation. Based on this, three datasets were selected and are described in detail below: UCI Heart Disease, ABIDE II, and TCGA-BRCA. This choice allowed transportability to be evaluated across heterogeneous biomedical settings. Their main characteristics are summarized in Table 3.1.

Dataset	Data type	Sample size and sites	Variables
UCI Heart Disease	Clinical	920 samples, 4 domains	Clinical and demographic
ABIDE II	Neuroclinical	1114 subjects, 19 sites	Phenotypic and diagnostic
TCGA-BRCA	Genomic	829 samples, 13 valid sites	Gene expression and clinical metadata

Table 3.1: Overview of the biomedical datasets used in this thesis.

A requirement for the present study was the availability of a site or domain identifier. This information was preserved throughout preprocessing and used to define the cross-site evaluation protocol. Each dataset was analysed not only as a standard binary classification problem, but also as a transportability problem, in which the central question was how well a model trained on a set of source sites could generalize to a different target site.

The three datasets provide progressively different test cases for cross-site generalization: a small and interpretable clinical dataset, an intermediate multi site neuroclinical dataset, and a high dimensional genomic dataset. This structure allowed the evaluated methods to be compared under heterogeneous biomedical conditions and provided a broader view of when adaptation strategies may or may not improve transportability.

3.1.1 Dataset Preprocessing

All datasets were preprocessed before applying the adaptation strategies. In this thesis, preprocessing refers to the set of operations used to convert each dataset into a coherent analysis ready table. The objective was to define the samples included in the study, the binary target outcome, the site or domain identifiers, and the feature set used throughout the experimental pipeline.

Particular attention was given to avoiding information leakage. Transformations requiring the estimation of parameters from the data, such as scaling, imputation, harmonization, or model dependent feature processing, were not fitted during the initial dataset construction. These operations were instead applied later within the appropriate training pipelines, using only the data available at each stage of the experimental protocol. This distinction is essential in a cross-site setting, because fitting transformations on the full dataset would allow information from the target site or the final test set to influence the training process.

After preprocessing, the datasets were organized with a binary outcome, consistent sample identifiers, site or domain information, and a set of features ready to be used in the classification pipeline. Quality control summaries were also generated to document sample size, class distribution, missing values, site composition, and metadata availability. These checks provided the starting point for the comparison of the adaptation strategies and helped ensure that the following analyses were conducted under a consistent and leakage safe cross-site evaluation framework.

3.1.2 UCI Heart Disease

UCI Heart Disease is the smallest dataset considered in this work and provides a useful first benchmark for evaluating the cross-site transportability pipeline in a relatively interpretable setting.

The UCI Heart Disease dataset was originally collected from four medical centres and made publicly available through the UCI Machine Learning Repository [16].

After preprocessing the dataset contains 920 samples, 17 columns, and four domains: Cleveland, Hungarian, Long Beach VA, and Switzerland. These domains were preserved through the `domain_id` variable. The predictive feature set includes five numerical variables, namely `age`, `trestbps`, `chol`, `thalach`, and `oldpeak`, and eight categorical variables, namely `sex`, `cp`, `fbs`, `restecg`, `exang`, `slope`, `ca`, and `thal`. Additional columns include `subject_id`, `domain_id`, `target_original`, and the final binary target.

Variable	Type	Description
age	Numerical	Age of the patient
trestbps	Numerical	Resting blood pressure
chol	Numerical	Serum cholesterol
thalach	Numerical	Maximum heart rate achieved
oldpeak	Numerical	ST depression induced by exercise relative to rest
sex	Categorical	Patient sex
cp	Categorical	Chest pain type
fbs	Categorical	Fasting blood sugar indicator
restecg	Categorical	Resting electrocardiographic results
exang	Categorical	Exercise induced angina
slope	Categorical	Slope of the peak exercise ST segment
ca	Categorical	Number of major vessels observed by fluoroscopy
thal	Categorical	Thalassemia related test result

Table 3.2: Predictive variables included in the UCI Heart Disease dataset.

The original UCI target was defined on a five level scale from 0 to 4. In this thesis, it was converted into a binary classification task: class 0 indicates absence of heart disease, whereas class 1 groups all levels associated with disease presence. The overall prevalence of the positive class is approximately 55.3%, but the distribution is not uniform across domains.

The four domains differ in both sample size and outcome prevalence. Cleveland contains 303 samples with a positive prevalence of 45.9%, Hungarian contains 294 samples with a prevalence of 36.1%, Long Beach VA contains 200 samples with a prevalence of 74.5%, and Switzerland contains 123 samples with a prevalence of 93.5%. Therefore, the cross-site problem is not limited to differences in the predictor variables, but also includes a substantial shift in the distribution of the target outcome.

Although UCI is relatively small and clinically interpretable, it presents several data quality and intersite variability issues [17]. The final dataset contains 1748 missing values, with particularly high missingness for `ca` (66.3%), `thal` (51.9%), and `slope` (33.5%). Missingness is also domain dependent: Cleveland is almost complete, whereas the other domains show more substantial gaps in some variables. This makes the dataset useful for transportability analysis, because sites differ not only in class prevalence but also in the availability of clinical information.

The quality report also identified several non standard or suspicious values. The variable `chol` contains 172 zero values, which are strongly site dependent and unlikely to represent physiologically plausible cholesterol measurements. Other examples include one zero value

in `trestbps`, twelve negative values in `oldpeak`, one value equal to 9 in `ca`, and one value equal to 0 in `slope`. The variable `thal` also shows heterogeneous coding patterns. These issues were documented rather than automatically corrected, since they reflect part of the real data quality complexity of the dataset.

A further quality control note concerns duplicate records. Four exact duplicate rows were identified during the preprocessing audit, but no exact duplicates were present in the final version. One duplicated `subject_id` remained and was documented as a warning. Overall, UCI Heart Disease is therefore not only a simple tabular benchmark, but also a useful example of how class imbalance, missingness, and site specific data quality issues can affect cross-site generalization.

3.1.3 ABIDE II

ABIDE II was selected as the multi site neuroclinical case study because it provides a more heterogeneous cross-site setting than UCI Heart Disease, while remaining less high dimensional than TCGA-BRCA.

ABIDE II was specifically designed to facilitate large scale studies of autism by aggregating data collected across multiple international sites using a common sharing framework [18].

The final strict dataset contains 1114 subjects, 9 columns, and 5 predictive features collected across 19 sites. Among these sites, 17 were eligible to be used as target domains in the cross-site evaluation protocol. Two sites, `ABIDEII-KUL_3` and `ABIDEII-NYU_2`, contained only one class and were therefore used only as source domains, not as independent target sites.

The predictive variables included in the strict dataset are `age_at_scan`, `sex_code`, `handedness_category_code`, `eye_status_at_scan_code`, and `fiq`. The dataset also includes `subject_id`, `site`, `disease_level`, and `disease_level_code`. The binary target was derived from `disease_level_code`, with Autism Spectrum Disorder (ASD) encoded as 1 and Control encoded as 0. At the global level, the prevalence of ASD was approximately 46.8%, suggesting an almost balanced dataset overall.

Variable	Type	Description
age_at_scan	Numerical	Age of the subject at the time of acquisition
sex_code	Categorical	Sex of the subject encoded as a numerical category
handedness_category_code	Categorical	Handedness category of the subject
eye_status_at_scan_code	Categorical	Eye status during the scan acquisition
fiq	Numerical	Full scale intelligence quotient
subject_id	Identifier	Unique subject identifier
site	Domain identifier	Acquisition site used as domain label
disease_level	Label descriptor	Diagnostic group label
disease_level_code	Target code	Numerical code used to derive the binary target

Table 3.3: Variables included in the ABIDE II strict dataset.

This global balance, however, does not fully describe the dataset. ABIDE II showed substantial variability across sites in terms of sample size, class composition, age distribution, and completeness of clinical information. The largest site, ABIDEII-KKI_1, contained 211 subjects, whereas several sites contained fewer than 30 subjects. Among target eligible sites, ASD prevalence ranged from approximately 26.5% to 61.5%. When monaclass sites were also considered, the range extended up to 100%. This made ABIDE II a useful dataset for studying transportability in a setting where sites reflected partially different subject populations.

From a data quality perspective, ABIDE II strict was cleaner than UCI in terms of identifiers: no exact duplicated rows and no duplicated `subject_id` values were found. However, some variables contained missing values, especially `fiq`, which had 99 missing entries, corresponding to 8.89% of the dataset. Missing values were preserved during dataset construction and handled later within the machine learning pipeline. In particular, missing values in `fiq` were imputed using a median based strategy fitted only on the training data, while categorical variables were imputed using the most frequent category. The variables `handedness_category_code` and `eye_status_at_scan_code` had 23 and 1 missing values, respectively, while `age_at_scan` and `sex_code` were complete.

The missingness of `fiq` was especially relevant because it was not evenly distributed across sites. ABIDEII-EMC_1 showed complete missingness for this variable, while ABIDEII-IP_1 had more than half of its `fiq` values missing. Overall, 121 subjects had at least one missing

feature in the strict phenotype bundle. This indicates that site differences concerned not only the values of the variables, but also the availability of the measurements themselves.

Age distribution represented another important source of heterogeneity. The difference between the highest and lowest site level mean age was approximately 31.74 years, suggesting that the sites did not simply represent technical replications of the same population. Some centres included mainly younger subjects, whereas others included older participants. Even without imaging derived features, this produced a meaningful phenotypic shift across sites.

The categorical variables in ABIDE II, such as `sex_code`, `handedness_category_code`, and `eye_status_at_scan_code`, were stored as numerical codes but were not interpreted as continuous measurements. This distinction was taken into account in the modelling pipelines. Overall, ABIDE II represented an intermediate and informative benchmark: more complex and genuinely multi site than UCI, but less high dimensional than TCGA-BRCA.

3.1.4 TCGA-BRCA

TCGA-BRCA was used as the high dimensional genomic case study and represents the most demanding dataset considered in this thesis. Unlike UCI Heart Disease and ABIDE II, which are based on a relatively small number of clinical and phenotypic variables, TCGA-BRCA is characterized by tens of thousands of gene expression measurements obtained through RNA sequencing. Consequently, the transportability problem is influenced not only by site heterogeneity, but also by high dimensionality, sparsity, class imbalance, and technical variability associated with molecular data generation.

TCGA-BRCA includes transcriptomic profiles generated through RNA sequencing together with extensive clinical annotations, making it one of the most widely used resources for cancer genomics research [19].

The final dataset was constructed at the subject level from primary tumour samples belonging to the TCGA breast cancer cohort. The expression matrix was organized with samples as rows and genes as columns, while TCGA barcodes were parsed to recover subject level and site level information. The cohort was restricted to primary tumours, and subjects with multiple primary tumour samples were resolved through a deterministic quality aware selection procedure. Clinical information on pathological American Joint Committee on Cancer (AJCC) stage was subsequently integrated and converted into a binary classification target, distinguishing Stage I–II tumours from Stage III–IV tumours.

Only sites with sufficient representation of both outcome classes were retained for the cross-site benchmark. This filtering step was necessary because a site can serve as a meaningful target domain only if the binary classification task can be evaluated within that site. The final analysis ready dataset contains 829 samples, 60 697 columns, 60 660 genomic features,

and 37 metadata variables. The positive class, corresponding to Stage III–IV disease, has an overall prevalence of 24.37%, indicating a substantial imbalance toward early stage tumours. The final benchmark includes 13 valid sites identified through the TCGA Tissue Source Site (TSS) code.

The dataset combines several sources of information. The main predictive block consists of the gene expression matrix, whose features are represented by ENSG gene identifiers. In addition, the table contains clinical and barcode derived metadata, including `clinical_gender`, `clinical_race`, `clinical_ethnicity`, `clinical_primary_diagnosis`, `clinical_tissue_or_organ_of_origin`, `sample_type_label`, `vial`, `tss`, and `site`. Expression quality control summaries were also retained, including `total_signal`, `nonzero_gene_count`, `zero_fraction`, and `low_total_signal_outlier`, providing additional information on the quality and structure of the molecular measurements.

Variable	Type	Description
<code>clinical_gender</code>	Clinical metadata	Sex of the patient
<code>clinical_race</code>	Clinical metadata	Reported race category
<code>clinical_ethnicity</code>	Clinical metadata	Reported ethnicity category
<code>clinical_primary_diagnosis</code>	Clinical metadata	Primary cancer diagnosis
<code>clinical_tissue_or_organ_of_origin</code>	Clinical metadata	Tissue or organ where the tumour originated
<code>sample_type_label</code>	Sample metadata	Type of biological sample
<code>vial</code>	Barcode metadata	TCGA vial identifier
<code>tss</code>	Site metadata	Tissue Source Site code used to identify the domain
<code>site</code>	Domain identifier	Site/domain label used in the cross-site analysis
<code>total_signal</code>	QC summary	Total expression signal of the sample
<code>nonzero_gene_count</code>	QC summary	Number of genes with non zero expression
<code>zero_fraction</code>	QC summary	Fraction of genes with zero expression
<code>low_total_signal_outlier</code>	QC summary	Indicator of unusually low total expression signal

Table 3.4: Clinical, barcode derived, and quality control variables retained in the TCGA-BRCA dataset.

The final table was clean at the subject identification level, with no exact duplicated rows and no duplicated `subject_id` values. This result followed an earlier consolidation step, during

which 11 duplicated primary subject entries were identified and resolved. Removing these duplicates was important to avoid including multiple correlated observations from the same patient in the cross-site analyses.

One of the defining characteristics of TCGA-BRCA is the combination of high dimensionality and sparsity. Although the expression matrix contains no missing values, it includes 34 269 989 zero entries, corresponding to a zero entry rate of 46.08%. Furthermore, 2 775 genes exhibit zero variance across the final cohort. At the sample level, the mean total expression signal is approximately 200 744, the mean number of expressed genes is approximately 32 514, and the mean zero fraction is approximately 0.464. These properties illustrate the complexity of the dataset and the challenges associated with modelling genomic data across sites.

Some metadata fields also require careful interpretation. Barcode derived variables such as `analyte`, `center`, `plate`, and `portion` are completely missing and were retained only as provenance metadata, not as predictors. Among the clinical variables, `clinical_ethnicity` is missing for 146 subjects (17.61%), whereas `clinical_race` is missing for 77 subjects (9.29%). These variables were used only for dataset characterization and not as predictive covariates, so no imputation was applied. The variable `clinical_gender` is complete but highly imbalanced, containing 819 female subjects and only 10 male subjects. Although this distribution is expected in a breast cancer cohort, it remains relevant for the characterization of the dataset.

The valid sites also differ substantially in both size and class composition. The largest site, BH, contains 147 samples, whereas the smallest valid sites contain fewer than 40 samples. The prevalence of advanced stage tumours ranges from approximately 15.3% in site E2 to approximately 38.5% in site AN. Site level quality indicators, including total expression signal, number of expressed genes, and zero fraction, also vary across domains, suggesting that the cross-site problem may reflect both biological heterogeneity and technical differences in data acquisition or processing.

TCGA-BRCA represents the most challenging test case for the transportability framework. Its complexity is not only to the presence of multiple sites, but also to the high dimensional nature of genomic data and to the variability introduced by molecular heterogeneity, class imbalance, and data quality differences. For this reason, it complements UCI Heart Disease and ABIDE II by testing the proposed adaptation strategies in a more complex biomedical scenario.

3.2 Machine Learning Models

To investigate cross-site transportability, three machine learning algorithms were selected: Logistic Regression, Random Forest, and Histogram Gradient Boosting. These models were chosen because they represent different learning paradigms and different levels of model complexity, while remaining widely used in biomedical predictive modelling.

The objective was to evaluate whether the effects of the proposed adaptation strategies were consistent across classifiers with different assumptions. Logistic Regression was included as a linear and interpretable baseline model, while Random Forest and Histogram Gradient Boosting were selected as non linear tree based methods capable of capturing more complex relationships between predictors and the target outcome.

Using models with different characteristics makes it possible to distinguish effects that are specific to a particular algorithm from patterns that remain consistent across learning approaches. This is important in a cross-site setting, because the impact of domain shift and adaptation strategies may depend not only on the dataset, but also on the model used to learn the predictive function.

3.2.1 Logistic Regression

Logistic Regression is a supervised learning model commonly used for binary classification. In biomedical applications, it is particularly useful because many clinical endpoints can be expressed as the presence or absence of a condition, disease status, or response to treatment. Rather than directly predicting a class label, logistic regression estimates the probability that an observation belongs to the positive class [20, 21].

Given a feature vector $X = (x_1, x_2, \dots, x_p)$, logistic regression models the relationship between the predictors and the log odds of the positive class as a linear function, as shown in Equation 3.1.

$$\log \left(\frac{p}{1-p} \right) = \beta_0 + \sum_{j=1}^p \beta_j x_j \quad (3.1)$$

where $p = P(Y = 1|X)$, β_0 is the intercept, and β_j represents the coefficient associated with the j -th predictor. The linear predictor is then transformed into a probability bounded between 0 and 1:

$$P(Y = 1|X) = \frac{1}{1 + \exp [-(\beta_0 + \beta^T X)]}.$$

Model parameters are commonly estimated by maximum likelihood. For a set of n independent observations $\{(x_i, y_i)\}_{i=1}^n$, with $y_i \in \{0, 1\}$, the likelihood function can be written as:

$$L(\beta) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i},$$

where $p_i = P(Y_i = 1 | X_i = x_i)$ denotes the predicted probability for the i -th observation [20]. Equivalently, the log likelihood is:

$$\ell(\beta) = \sum_{i=1}^n [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)].$$

Maximum likelihood estimation selects the coefficients β that maximize this log likelihood, corresponding to the parameter values that make the observed class labels most probable under the model. Once probabilities have been estimated, class labels can be assigned by applying a decision threshold.

Logistic Regression was included in this thesis as a representative linear classifier. Its role is not only to provide a simple and interpretable reference model, but also to assess whether the proposed cross-site adaptation strategies can improve transportability when the underlying predictive model has a relatively constrained structure. It also provides a useful comparison with more flexible algorithms, such as tree based ensemble methods, which can capture more complex nonlinear relationships in the data [20, 21].

3.2.2 Random Forest

Random Forest is a supervised machine learning algorithm based on an ensemble of decision trees. The model builds a collection of trees and combines their predictions to obtain a more stable and robust classifier. This approach is closely related to bagging, where multiple trees are trained on bootstrap samples of the original dataset and their outputs are aggregated [20, 21].

A decision tree partitions the feature space into a set of regions by applying a sequence of binary splitting rules. At each internal node, the algorithm selects a predictor and a threshold that divide the data into two subsets, with the aim of improving class separation. For classification problems, the final prediction of a tree is usually given by the most frequent class among the training observations that fall into the corresponding terminal node [20].

Random Forest introduces two sources of randomness. First, each tree is trained on a bootstrap sample drawn from the original training set. Second, at each split only a random

subset of predictors is considered as candidate splitting variables. This second step reduces the correlation between individual trees, making the ensemble more effective than a simple aggregation of highly similar trees [20, 21].

For a classification task, the final prediction is obtained by aggregating the predictions of all trees in the forest. If B trees are grown and $\hat{C}_b(x)$ denotes the class predicted by the b -th tree for an observation x , the Random Forest classifier can be written as:

$$\hat{C}_{RF}(x) = \text{majority vote } \{\hat{C}_1(x), \hat{C}_2(x), \dots, \hat{C}_B(x)\}.$$

Class probabilities can be estimated as the proportion of trees voting for each class:

$$\hat{P}(Y = k|X = x) = \frac{1}{B} \sum_{b=1}^B I(\hat{C}_b(x) = k),$$

where $I(\cdot)$ is an indicator function equal to 1 when the condition is true and 0 otherwise [20].

In this work, Random Forest was included as a representative tree based ensemble method. It provides a flexible modelling framework that can capture nonlinear relationships and interactions among predictors without requiring them to be specified in advance. Its role in the experimental comparison is therefore to assess whether cross-site adaptation strategies behave differently when applied to a nonlinear ensemble classifier rather than to a simpler linear model.

3.2.3 Histogram Gradient Boosting

Histogram Gradient Boosting is an ensemble machine learning method based on the gradient boosting framework. Boosting methods construct a sequence of weak learners, where each new model focuses on correcting the errors made by the previous ones. The final prediction is obtained by combining the contribution of all learners in the ensemble [20, 21].

The central idea of gradient boosting is to iteratively minimize a loss function by adding decision trees that improve the current model. Starting from an initial prediction $F_0(x)$, the model is updated at each iteration by adding a new tree $h_m(x)$ weighted by a learning rate ν :

$$F_m(x) = F_{m-1}(x) + \nu h_m(x),$$

where m denotes the boosting iteration. Each tree is fitted to the negative gradient of the loss function with respect to the current predictions, which can be interpreted as the direction of greatest improvement in model performance [20].

Histogram Gradient Boosting introduces a computational optimization. Before constructing the trees, continuous features are discretized into a finite number of bins, creating histogram based representations of the data. Instead of evaluating all possible split points, the algorithm searches for optimal splits using these histograms. This reduces computational cost and memory usage while preserving most of the predictive performance of conventional gradient boosting methods, making it particularly suitable for large datasets and high dimensional problems [21].

The final prediction is obtained by combining the contribution of all trees in the ensemble:

$$F(x) = F_0(x) + \nu \sum_{m=1}^M h_m(x),$$

where M is the total number of boosting iterations. For binary classification tasks, the resulting score is transformed into a probability through a logistic link function [20].

Histogram Gradient Boosting was included in this thesis as a representative boosting based ensemble method. Compared with Logistic Regression and Random Forest, Histogram Gradient Boosting provides a more flexible way to model complex relationships in the data. This can lead to strong predictive performance, but it may also make the model more sensitive to differences between source and target domains. For this reason, it is useful in this thesis to test whether the adaptation strategies remain effective even with a highly flexible model.

3.3 dashi Benchmark

Exploratory characterization of each dataset was performed using *dashi*, a Python library designed for the analysis of dataset shift and transportability across data sources and over time in artificial intelligence applications, with particular emphasis on healthcare data [22]. In this thesis, *dashi* was not used as an adaptation method, but as a complementary framework for describing the structure of the datasets and the relationships among domains.

The motivation for including *dashi* is that predictive performance alone does not fully explain why models transfer successfully to some domains while showing reduced performance in others. Differences in transportability may arise from multiple sources, including changes in feature distributions, variations in class prevalence, acquisition related effects, or more complex forms of domain heterogeneity. To investigate these aspects, *dashi* combines unsupervised analyses of dataset variability with supervised analyses of cross-domain predictive performance [22].

The unsupervised component is based on the Multi Source Variability framework originally proposed by Sáez et al. [17]. Within this framework, each source is represented through its

empirical probability distribution and compared with the remaining sources using information theoretic measures of dissimilarity. In this work, the source corresponds to the site or domain identifier available in each dataset.

Pairwise comparisons between sources were performed using the Jensen–Shannon distance, a symmetric and bounded measure derived from the Kullback–Leibler divergence. Compared with other divergence measures, the Jensen–Shannon distance is particularly suitable for heterogeneous biomedical data because it remains finite and provides an interpretable measure of distributional dissimilarity between domains [17].

Based on these pairwise distances, *dashi* generates a low dimensional representation of the dataset structure, allowing the relative position of the different sources to be visualized. In addition, two summary metrics are computed:

- Global Probabilistic Deviation (GPD), which quantifies the overall degree of variability among all sources in the dataset.
- Source Probabilistic Outlyingness (SPO), which quantifies the relative deviation of an individual source with respect to the global multi source structure.

Together, these measures describe how heterogeneous the dataset is and how much each source domain differs from the others.

The supervised part of *dashi* then evaluates how well models transfer across domains. A model is trained on one domain and tested on another, producing pairwise transferability matrices. In this way, it is possible to observe which domains generalize well to each other, rather than relying only on average performance values.

The outputs generated by *dashi* include data source maps, Multi Source Variability visualizations, GPD and SPO summaries, and source to source transferability matrices. Since the supervised *dashi* analyses require complete feature matrices, missing values in the numeric predictors were imputed using a median based strategy before running the analyses. In practice, this affected the UCI and ABIDE datasets, whereas the selected TCGA gene expression features were already complete.

dashi provides quantitative and visual descriptions of domain heterogeneity and inter site relationships. These analyses helped interpret the transportability results presented in Chapter 4 and explore whether performance differences were associated with characteristics of the underlying domains.

3.4 Cross-Site Experimental Protocol

3.4.1 Source and Target Domains

The experimental framework was designed to reproduce a realistic scenario in which a predictive model is applied to a previously unseen domain. Each dataset included an explicit site or domain identifier, corresponding to different clinical centres, acquisition sites, sequencing centres, or cohorts depending on the specific biomedical context.

For each experimental configuration, one domain was selected as the target site, while all remaining domains were treated as source sites. The source domains were combined to create the initial training set, whereas the target domain was kept separate from the model development process. This design reproduces a realistic deployment scenario in which a model developed using data collected at a set of institutions is subsequently applied to a new site characterized by its own population and data distribution.

Formally, let

$$D = \{D_1, D_2, \dots, D_S\}$$

denote the set of available domains, where S is the total number of sites. For a given iteration, one domain D_t is selected as the target domain and the remaining domains are combined to form the source set:

$$D_{\text{source}} = \bigcup_{s \neq t} D_s.$$

The selected domain D_t is then treated as the target site for that experimental configuration.

This procedure was repeated for all eligible domains within each dataset, so that every site could be evaluated as an independent target domain. The resulting protocol follows a leave one site out design and provides a systematic assessment of transportability across the full range of available domains.

The source domains were used to train the initial predictive models and to estimate all source based adaptation procedures. For the source only methods (M0, M1, and M2), no information from the target domain was used during model training or adaptation. Methods M3 and M4 additionally incorporated information from a dedicated subset of the target domain, as described in Section 3.4.3. In all cases, the final evaluation was performed on target samples that remained completely unseen during model development.

3.4.2 Target Split Strategy

Once the target domain had been selected, its samples were divided into two subsets, denoted as *train_I* and *test_I*. This split was applied across all datasets and adaptation strategies (M0–M4) in order to maintain a clear separation between adaptation and final evaluation.

A 50/50 split was used for all target domains. This choice represented a compromise between providing enough target samples for the adaptation strategies while retaining a sufficiently large and representative subset for evaluation. Whenever possible, the split preserved the original class proportions in both subsets through stratified sampling. When stratification was not feasible because one or more classes contained too few samples within a target domain, a random split was used instead. A fixed random seed was adopted to ensure reproducibility.

The *test_I* subset was used only for the final OOD evaluation. It was kept separate from the entire model development process, meaning that it was not used for training, adaptation, calibration, or any other intermediate step. In this way, the reported OOD results reflect performance on target samples that remained truly unseen until the final evaluation.

The role of *train_I* depended on the adaptation strategy considered. For the source only methods (M0, M1, and M2), *train_I* was not used at any stage of the learning process. Model training, harmonization, weighting procedures, and internal validation were performed exclusively using the source domains. For the target aware methods (M3 and M4), *train_I* provided the only information available from the target domain and was used to perform the transductive adaptation step. In these methods, adaptation was applied after the initial model had been trained on the source domains, while *test_I* was kept completely separate until the final evaluation.

Internal performance was estimated separately from the target domain split. For UCI Heart Disease and ABIDE II, validation was performed using a leave one source domain out procedure. In each validation iteration, one source domain was temporarily held out for validation, while the remaining source domains were used for training. This allowed internal performance to be estimated while preserving the domain structure of the data. For TCGA-BRCA, folds were instead defined at the site level. In all cases, validation used only source domain data and never included the target evaluation subset.

The same separation between adaptation and evaluation was maintained in the bootstrap stability analyses described later in this chapter. For each bootstrap iteration, the target domain split was repeated and the entire adaptation pipeline was executed again. This procedure allowed the variability of OOD performance estimates to be quantified while preserving a leakage safe evaluation framework.

3.4.3 Adaptation Strategies (M0–M4)

In this work, the adaptation strategies are referred to using the labels M0–M4. These labels are used throughout the thesis to simplify the comparison between methods, while their implementation is described in the following sections.

Method	Description
M0	Baseline
M1	ComBat harmonization
M2	Domain weighting
M3	Transductive learning
M4	ComBat harmonization + transductive learning

Table 3.5: Adaptation strategies evaluated in the cross-site benchmark.

Five methodological settings were tested to understand how different adaptation strategies affect cross-site transportability. Since all methods were applied within the same source target framework, their performance could be compared under the same experimental conditions.

M0 was used as the baseline. The model was trained only on the source domains and then tested directly on the target domain, without applying any adaptation. This represents the standard source only scenario and serves as the reference for comparing the other methods.

M1 applied ComBat harmonization before model training. The harmonization parameters were estimated from the available training data and then used to adjust the feature values. The aim was to make the data more comparable across sites while retaining information useful for the prediction task.

M2 implemented domain weighting. Domain specific weights were estimated using only the source domains. In particular, each source domain was evaluated according to its pairwise predictive transferability to the other source domains, rather than according to its distance from the target site or from a source domain centroid. A reference Logistic Regression model was trained on each source domain and evaluated on the remaining source domains, the resulting transfer scores were then converted into domain weights. These weights were incorporated into model training, increasing the contribution of source domains that showed higher transferability and reducing the influence of less transferable domains.

M3 implemented transductive learning. A predictive model was first trained using the source domains. Subsequently, the target domain subset *train_1* was used to construct a target specific adaptation model. The predictions of the source model and the target adapter were then combined through probability blending. No samples from the target *test_1* subset were used during this process.

M4 combined ComBat harmonization and transductive learning. Harmonization was first applied to reduce source domain heterogeneity, after which the transductive adaptation procedure was performed using the target *train_1* subset. This configuration was designed to evaluate whether feature level harmonization and target aware adaptation provide complementary benefits when used together.

Each adaptation strategy was combined with the three machine learning models considered in this study, Logistic Regression, Random Forest, and Histogram Gradient Boosting. Consequently, every target domain experiment was evaluated under fifteen model adaptation combinations. This design makes it possible to assess not only the overall effectiveness of each adaptation strategy, but also whether its behaviour remains consistent across predictive models characterized by different levels of complexity and flexibility.

3.4.4 ID and OOD Evaluation

The evaluation framework was designed to distinguish between performance obtained within the source domains and performance achieved when transferring the model to a previously unseen target domain. For this reason, two complementary evaluation settings were considered: in-distribution (ID) evaluation and out-of-distribution (OOD) evaluation.

ID performance was estimated exclusively using source domain data. After selecting the target site, all remaining domains were treated as source domains. Within this source set, internal validation was performed by creating train validation partitions that preserved the domain structure. For UCI Heart Disease and ABIDE II, one source domain was held out at a time for validation, while the remaining source domains were used for training. For TCGA-BRCA, grouped validation was used, with folds defined according to the site identifier. In all cases, target domain samples were excluded from the ID evaluation process.

OOD performance was measured on the *test_1* subset of the target domain, which was kept separate from the entire model development process. This provided an independent evaluation of how well the model transferred to a new site. For the target aware methods, M3 and M4, only *train_1* was used during adaptation, while *test_1* was used only at the final evaluation stage.

The distinction between ID and OOD performance allowed the effects of domain shift to be quantified. While ID performance reflects the predictive ability of a model within environments similar to those used for training, OOD performance measures the extent to which this predictive ability is preserved when the model is applied to a new site characterized by a different data distribution.

To summarize this effect, an ID–OOD gap was computed as the difference between the corresponding performance estimates. Smaller gaps indicate greater stability across domains

and therefore better transportability, whereas larger gaps suggest a stronger degradation of performance when transferring the model to the target site.

The primary evaluation metrics were the area under the receiver operating characteristic curve (ROC-AUC), F1-score, and balanced accuracy. Model calibration was additionally assessed through the Expected Calibration Error (ECE). The definition and motivation of these metrics are described in Section 3.5.

The same evaluation framework was maintained in the bootstrap stability analyses. For each bootstrap iteration, the target domain split was repeated, the complete adaptation pipeline was re executed, and both ID and OOD performance estimates were recomputed. The resulting distributions were subsequently summarized through their mean and standard deviation to quantify performance variability across repeated target domain splits.

3.5 Evaluation Metrics

Model performance was assessed using a set of complementary evaluation metrics, since no single measure is sufficient to capture all relevant aspects of predictive behaviour [21]. The evaluation framework considered discrimination, classification performance, calibration, and transportability across domains. The main comparison between the adaptation strategies M0–M4 was therefore based on ROC-AUC, F1-score, Balanced Accuracy, Expected Calibration Error (ECE), and the ID–OOD gap.

ROC-AUC The area under the receiver operating characteristic curve (ROC-AUC) was used as the primary threshold independent measure of discrimination. ROC-AUC quantifies the ability of a model to rank positive samples above negative samples across all possible classification thresholds [21]. A value of 0.5 corresponds to random discrimination, whereas a value of 1 indicates perfect separation between classes.

Because ROC-AUC is computed directly from predicted probabilities, it is independent of any specific classification threshold and provides a robust estimate of the intrinsic discriminative ability of a model.

F1-score The F1-score was used to evaluate classification performance after converting predicted probabilities into binary labels. It combines precision and recall through their harmonic mean:

$$F_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Precision and recall are defined as:

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN}.$$

The F1-score is particularly useful when class frequencies are unbalanced because it penalizes both false positives and false negatives [21].

Balanced Accuracy Balanced Accuracy was included to account for potential class imbalance across sites. Unlike standard accuracy, which can be dominated by the majority class, Balanced Accuracy gives equal importance to both classes and is defined as:

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}.$$

Sensitivity and specificity are defined as:

$$\text{Sensitivity} = \frac{TP}{TP + FN},$$

$$\text{Specificity} = \frac{TN}{TN + FP}.$$

This metric is particularly relevant in biomedical applications, where disease prevalence often varies across populations and sites [21].

Expected Calibration Error Model calibration was assessed using the Expected Calibration Error (ECE), a widely used measure of probabilistic reliability [14]. ECE measures calibration by comparing predicted confidence with observed outcome frequency within M bins:

$$ECE = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|,$$

where B_m denotes the set of samples belonging to the m -th confidence bin, $\text{acc}(B_m)$ is the empirical accuracy within the bin, and $\text{conf}(B_m)$ is the mean predicted confidence.

Lower ECE values indicate better agreement between predicted probabilities and observed

outcomes, whereas larger values indicate overconfident or underconfident predictions.

ID–OOD Gap To quantify the impact of domain shift, an ID–OOD gap was computed as the difference between in-distribution and out of distribution performance:

$$\text{ID–OOD Gap} = \text{Performance}_{ID} - \text{Performance}_{OOD}.$$

The gap was calculated separately for ROC-AUC, F1-score, and Balanced Accuracy. Smaller values indicate that performance is preserved when transferring the model from the source domains to a previously unseen target site, whereas larger values indicate a stronger degradation in predictive performance and therefore reduced transportability.

Additional Metrics in dashi The exploratory *dashi* analyses produced additional supervised metrics, including accuracy, precision, recall, PR-AUC, and log-loss [22]. These measures were used to characterize source to source transferability and dataset heterogeneity, but they were not considered part of the primary benchmark used to compare the M0–M4 adaptation strategies.

3.6 Transportability Characterization Analyses

In addition to the main benchmark metrics, a set of complementary analyses was performed to better characterize the conditions under which cross site transportability was more or less successful. These analyses were not intended to define new adaptation strategies, but to provide additional information about the stability of the results, the characteristics of the target domains, and the relationship between domain heterogeneity and model performance.

The analyses included bootstrap stability assessment, target domain entropy and sample size summaries, site distance measures, metadata shift analysis, clustered heatmaps, and SHAP based meta learning interpretation. Together, these components were used to move beyond average OOD performance and to examine whether the behaviour of the M0–M4 strategies could be related to measurable properties of the datasets and target sites.

3.6.1 Bootstrap Stability Analysis

A bootstrap based stability analysis was performed to assess whether the observed OOD performance patterns were robust to different random partitions of the target domain. Bootstrap methods provide a general resampling framework for estimating the variability of statistical

quantities when their sampling distribution is not directly known [15]. In this thesis, the same principle was adapted to the cross-site evaluation setting by repeatedly re splitting the target domain into adaptation and evaluation subsets.

For each dataset, 100 bootstrap iterations were performed to assess the stability of OOD performance. The original 50/50 target domain split into *train_I* and *test_I* was kept fixed. At each iteration, the *train_I* subset was resampled with replacement, while *test_I* remained unchanged and was used only for final OOD evaluation. The complete M0–M4 pipeline was then re executed using the same source target structure described in Section 3.4. For the source only methods M0, M1, and M2, the resampled *train_I* subset was not used during training or adaptation. For the target aware methods M3 and M4, it was used only for the transductive adaptation step. In all cases, *test_I* remained fully held out until final evaluation.

The main quantity summarized in this analysis was OOD balanced accuracy. For each combination of dataset, target site, machine learning model, and adaptation strategy, the mean and standard deviation across bootstrap iterations were computed. The mean reflected the average OOD performance across repeated target splits, while the standard deviation indicated how much the results varied depending on the specific samples included in the target evaluation set.

This analysis helped determine whether the observed performance differences were consistent across repeated splits or mainly driven by a specific random partition. This was especially important for target domains with few samples or imbalanced classes, where OOD estimates can be less stable. Bootstrap stability provided an additional check of reliability alongside the main ID/OOD metrics.

3.6.2 Entropy, Sample Size and Site Distance

To investigate whether the effectiveness of the adaptation strategies depended on target domain characteristics, three additional descriptors were computed for each target site: class entropy, target sample size, and site distance. These variables were used to characterize the structure of the target domain and to explore whether specific conditions were associated with larger or smaller gains over the source only baseline.

Class entropy was used to summarize the balance of the binary target distribution within each site. Entropy was originally introduced by Shannon as a measure of uncertainty in information theory [23]. For a target domain with class proportions p_0 and p_1 , entropy was computed as:

$$H(Y) = - \sum_{c \in \{0,1\}} p_c \log(p_c).$$

Higher entropy values indicate a more balanced class distribution, whereas lower values indicate stronger class imbalance. This measure was included because class imbalance can affect both model training and performance evaluation, especially in biomedical classification problems where one class may be under represented [24].

Target sample size was defined as the number of samples available in the selected target site. This quantity was considered because adaptation methods may require a sufficient amount of target domain information to estimate stable target specific patterns. In particular, target aware methods such as M3 and M4 use the *train_1* subset during adaptation, which means that very small target domains may provide limited or unstable information for this step.

Site distance was used as a proxy for the degree of distributional mismatch between source and target domains. In cross-site biomedical studies, differences between training and test distributions may involve changes in feature distributions, class proportions, or the relationship between predictors and outcomes [25]. For each target site, distance measures were computed by comparing the target domain with the pooled source domains. These measures were then used to describe how far the target domain was from the data available during model development.

The purpose of this analysis was exploratory. Class entropy, sample size, and site distance were not used to train the main predictive models, but were used to interpret the behaviour of the adaptation strategies after the benchmark had been completed. Performance differences between each adaptation method and M0 were analysed as a function of these target domain descriptors. This made it possible to examine whether adaptation was more beneficial under specific conditions, such as larger target sample sizes, more balanced class distributions, or smaller source target distances.

These analyses were interpreted descriptively rather than causally. Since the number of target domains was limited, especially in UCI Heart Disease and TCGA-BRCA, the objective was not to derive definitive decision rules for method selection. Instead, the aim was to identify recurring patterns that could help explain why some adaptation strategies performed better in certain target sites and worse in others.

3.6.3 Metadata Shift Analysis

Metadata shift analysis was performed to describe how the target site differed from the source sites in terms of demographic, clinical, phenotypic, or technical variables. The objective was not to train the predictive models with these variables, but to use them as an interpretative tool for understanding possible sources of domain heterogeneity. [25]

For each target site, the distribution of each metadata variable was compared with the corresponding distribution in the pooled source domains. Continuous variables were summarized

using standardized mean differences, while categorical variables were summarized using distributional differences between category proportions. These comparisons provided a site level description of how different each target domain was from the available source domains.

To obtain compact indicators of metadata shift, the variable level differences were aggregated into three summary scores: `mean_continuous_metadata_shift`, `mean_categorical_metadata_shift`, and `overall_metadata_shift_score`. Higher values indicate stronger differences between the metadata profile of the target site and that of the source domains.

The type of metadata differed across datasets. In UCI Heart Disease, metadata mainly described clinical and demographic characteristics. In ABIDE II, it included phenotypic and acquisition related information such as age, sex, handedness, eye status, and IQ. In TCGA-BRCA, metadata included clinical variables and technical quality control information related to sequencing and sample processing.

Overall, metadata shift analysis was used to complement the main transportability benchmark by describing observable differences between source and target sites. The analysis was interpreted together with class entropy, sample size, site distance, calibration, and stability results, since metadata differences alone cannot fully explain cross-site transportability.

3.6.4 Clustered Heatmaps

Clustered heatmaps were used to visually explore whether target sites and adaptation strategies showed similar transportability patterns. The analysis was intended as an exploratory tool rather than a formal statistical test, with the aim of highlighting recurring patterns in OOD performance, ID–OOD gaps, bootstrap variability, metadata shift, and domain distance measures.

A heatmap represents a matrix of values through a color scale, making it useful for visualizing structured patterns across rows and columns. Hierarchical clustering was applied to reorder rows and columns according to their similarity, so that sites or methods with comparable profiles appeared close to one another in the visualization. This approach is commonly used in exploratory data analysis to identify groups of observations or variables with similar behaviour [26].

In this thesis, clustered heatmaps were applied to several matrices derived from the benchmark results. For performance based heatmaps, rows generally represented adaptation strategies or methods, while columns represented target domains. For metadata and domain shift heatmaps, columns represented target sites and rows represented the corresponding descriptors. When variables were measured on different scales, standardization was applied before clustering in order to make the values comparable.

Hierarchical clustering was used to group target sites with similar profiles and, when appropriate, to group adaptation strategies with similar behaviour. The resulting clusters were interpreted descriptively, with the aim of identifying whether difficult target sites tended to share common patterns, whether some methods behaved similarly across domains, and whether transportability problems were associated with specific domain characteristics.

The clustered heatmaps were therefore used to complement the numerical summaries reported in the main benchmark. While average performance values provide an overall comparison between methods, clustered heatmaps preserve the site level structure of the results and help reveal patterns that may be hidden by global averages. For this reason, they were particularly useful for interpreting cross-site transportability as a domain dependent phenomenon.

3.6.5 SHAP Meta Learning Analysis

A SHAP based meta learning analysis was performed to investigate which target domain characteristics were associated with differences in OOD performance. SHAP values are based on concepts from cooperative game theory and provide a way to estimate the contribution of each feature to a model prediction [27]. In this thesis, SHAP was not used to explain individual clinical predictions, but to interpret the behaviour of the transportability benchmark at the domain level.

The analysis was based on a meta learning table constructed from the experimental results. In this setting, each row represented one experimental configuration, defined by the dataset, target site, machine learning model, adaptation strategy, target domain descriptors, source target distance measures, metadata shift summaries, and the corresponding OOD ROC-AUC. Target domain descriptors included variables such as target sample size, class entropy, class balance, and minority class ratio. OOD ROC-AUC was used as the outcome of the meta learning analysis because the objective was to investigate which domain level characteristics were associated with better or worse out-of-distribution discrimination.

Meta learning approaches are commonly used to relate the performance of algorithms to characteristics of datasets or tasks [28]. Following this idea, the SHAP analysis was used to examine whether the success or failure of each adaptation strategy could be associated with measurable properties of the target domain, such as class entropy, sample size, metadata shift, or source target distance.

To avoid information leakage, the meta learning models did not include variables that were directly computed from the performance results. Metrics such as OOD performance, ID performance, ID–OOD gaps, calibration scores, and related quantities were excluded. The predictors kept in the analysis described the target site and its relationship with the source domains, rather than the outcome that the SHAP model was trying to explain.

Separate meta learning models were trained for the different adaptation strategies M0–M4. This allowed the analysis to investigate whether different methods were sensitive to different target domain characteristics. SHAP values were then computed for each model and summarized using feature importance rankings, comparative heatmaps, and beeswarm plots.

The purpose of this analysis was exploratory and interpretative. The SHAP models were not used to select the best method during the benchmark and did not influence the training of the main predictive models. They were used after the main benchmark to better understand when each adaptation strategy was more or less effective. Since the number of target domains was limited, these results were interpreted as descriptive patterns rather than causal explanations.

Chapter 4

Results

4.1 Descriptive analysis of the databases

Figure 4.1 summarizes the distribution of samples and target classes across the sites included in the three study datasets. The figure highlights substantial differences in both the number of subjects available at each site and the class composition of the corresponding domains.

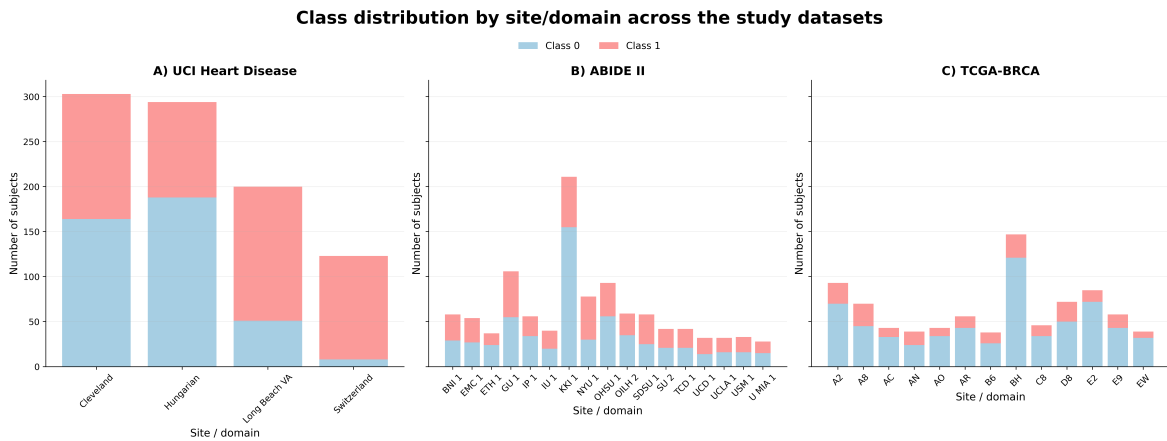


Figure 4.1: Site wise sample size and class distribution across the three study datasets. For UCI Heart Disease, Class 0 indicates absence of heart disease and Class 1 indicates presence of heart disease. For ABIDE II, Class 0 corresponds to Control subjects and Class 1 to Autism Spectrum Disorder (ASD). For TCGA-BRCA, Class 0 corresponds to Stage I–II tumours and Class 1 to Stage III–IV tumours.

UCI Heart Disease included four domains with relatively large sample sizes, although the class distribution varied considerably across sites. Cleveland and Hungarian showed a more balanced composition, whereas Long Beach VA and Switzerland contained a higher proportion of positive cases. Overall, the dataset presented moderate heterogeneity in both sample size and class prevalence.

ABIDE II comprised the largest number of domains, with nineteen acquisition sites contributing subjects. Site sizes were highly variable, ranging from small cohorts to substantially larger centres. The proportion of Autism Spectrum Disorder (ASD) and control subjects also

differed across sites, providing a heterogeneous multi site setting for evaluating cross-site transportability.

TCGA-BRCA contained thirteen valid sites with a more homogeneous distribution of sample sizes than ABIDE II, although differences between centres were still evident. Most sites contained a larger proportion of Stage I–II tumours than Stage III–IV tumours, reflecting the overall class imbalance of the dataset.

These descriptive characteristics provide the context for the subsequent analyses. In particular, the observed differences in site size and class composition motivate the use of the DASHI characterization framework before evaluating the adaptation strategies.

4.2 dashi Analysis

dashi was applied to each dataset to obtain an initial characterization of domain heterogeneity and source to target transferability patterns. Since the selected datasets do not contain a temporal structure, the analysis focused on the Multi Source Variability (MSV) representations and on the supervised transferability heatmaps. Additional visualizations and performance summaries were generated for multiple evaluation metrics, although only the most representative results are reported here.

4.2.1 UCI Heart Disease

The unsupervised analysis indicated that the four UCI domains were characterized by distinct feature distributions. The global dataset variability index (GPD) was equal to 0.6752, suggesting that the sites were characterized by partially distinct feature distributions. This was also reflected in the source specific overlap scores (SPO), which varied across domains, ranging from 0.4809 for Long Beach VA to 0.6469 for Cleveland.

Figure 4.2 shows the two dimensional Multi Source Variability (MSV) representation. The projection highlights differences between domains, with Cleveland and Switzerland occupying opposite regions of the latent space, while Hungarian appears more separated along the second dimension. Long Beach VA is positioned closer to the centre of the representation. The plot confirms that measurable distributional differences were already present before supervised model training.

The supervised transferability analysis highlighted that generalization across sites was not symmetric. Figure 4.3 reports the source to target accuracy heatmap, where rows represent the training domain and columns represent the target domain. This representation shows that performance changed substantially depending on the specific source target pair considered.

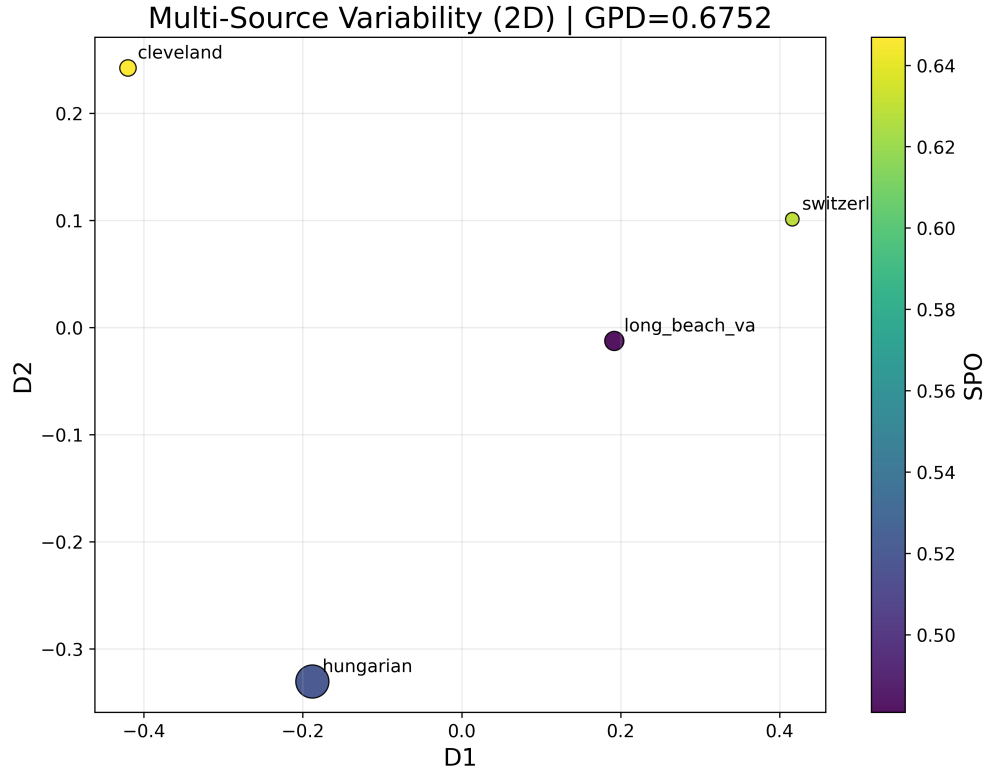


Figure 4.2: Two dimensional Multi Source Variability representation for the UCI Heart Disease dataset. Point size reflects domain sample size, while color indicates the source overlap score (SPO).

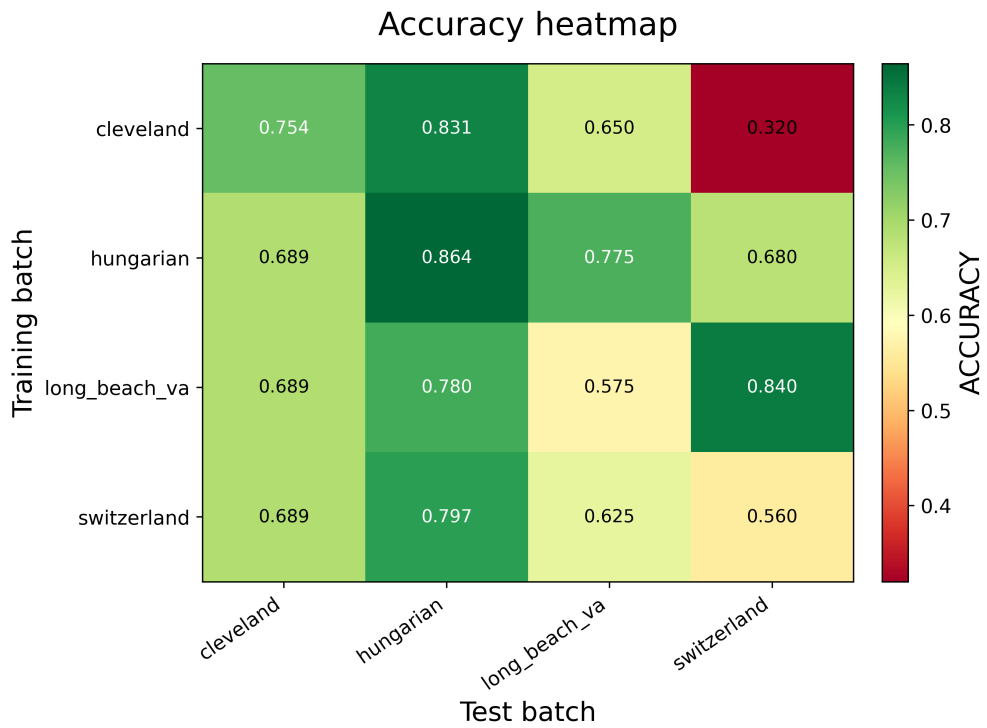


Figure 4.3: Source to target accuracy heatmap generated by dashi for the UCI Heart Disease dataset. Rows represent the training domain and columns represent the target domain.

Several patterns emerged from the supervised results. Hungarian appeared as the easiest target domain, showing the highest average performance across multiple metrics, including AUC macro, F1 macro and accuracy. Cleveland also showed relatively stable transferability, with good performance across most source target combinations.

Long Beach VA was more challenging as a target domain, while Switzerland represented the most problematic case. This behaviour was consistent with the strong class imbalance observed in the Switzerland domain, where negative samples were much less represented than positive ones. This imbalance was associated with lower F1 macro values, unstable AUC estimates in some source target combinations, and high log loss values.

Overall, these findings indicate that raw, source to target transportability was not equally distributed across the four domains. Some source target combinations generalized better than others, suggesting that variations in class composition and feature distributions already affected performance before any adaptation strategy was applied. This variability was also reflected in the later analyses across target sites and underlines the importance of considering domain heterogeneity when evaluating cross-site generalization.

4.2.2 ABIDE II

The unsupervised analysis highlighted substantial heterogeneity across the ABIDE II sites. The global dataset variability index (GPD) was equal to 0.6856, indicating a level of multi site variability comparable to that observed in UCI. The source overlap scores (SPO) also showed considerable variation across sites. Some domains, such as EMC_1, SU_2 and IP_1, exhibited relatively high overlap scores, whereas others, including KKI_1, USM_1 and UCLA_1, showed noticeably lower values. These differences suggest that the position and overlap of the sites within the multivariate feature space were not uniform.

Figure 4.4 shows the two dimensional Multi Source Variability (MSV) representation. Several sites clustered together on the left side of the plot, while others appeared more separated. In particular, SU_2 and U_MIA_1 were positioned in the upper right part of the latent space, whereas UCD_1 and IP_1 were located in the lower right area. KKI_1, although it was one of the largest sites in terms of sample size, remained closer to the central group. This visualization suggests that the retained ABIDE II sites were not fully homogeneous and that site related differences were already visible at the level of feature distributions.

The supervised transferability analysis revealed a similarly heterogeneous picture. Figure 4.5 reports the source to target accuracy heatmap. Compared with UCI, the matrix appears considerably more irregular, with substantial variation across source target combinations and no obvious pattern of uniformly high or uniformly low performance.

Several patterns emerged from the supervised results. Unlike UCI, no single site consistently

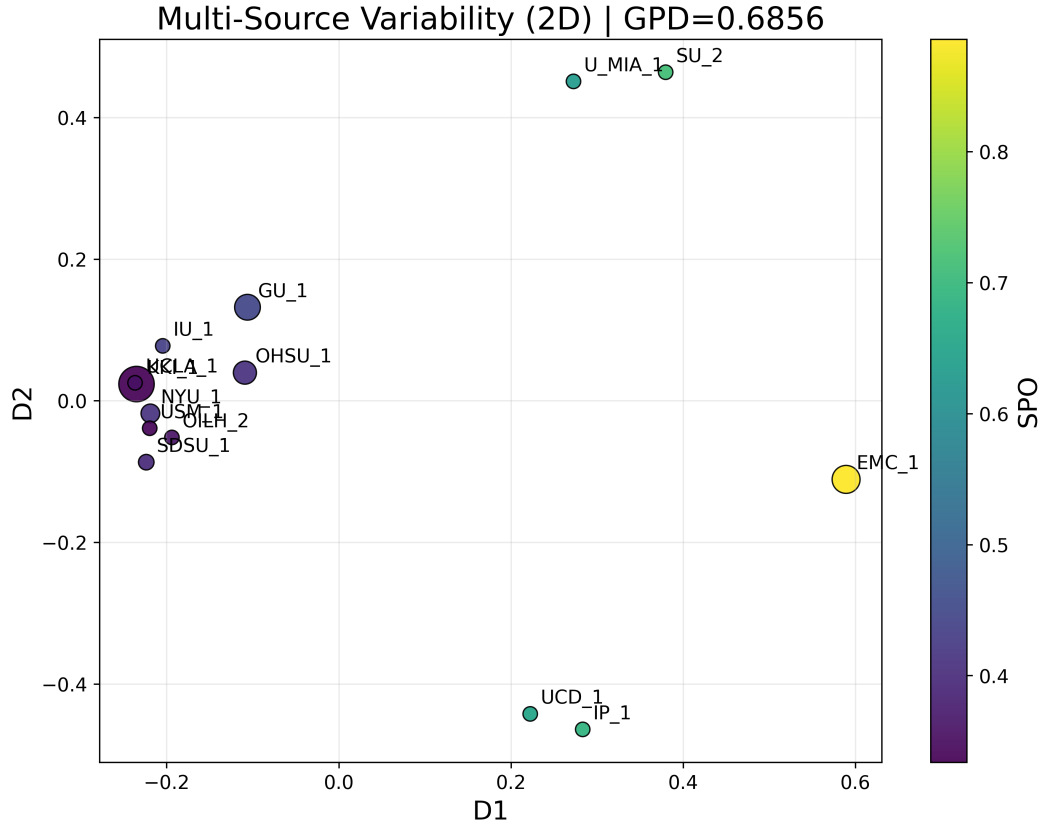


Figure 4.4: Two dimensional Multi Source Variability representation for the ABIDE II dataset. Point size reflects domain sample size, while color indicates the source overlap score (SPO).

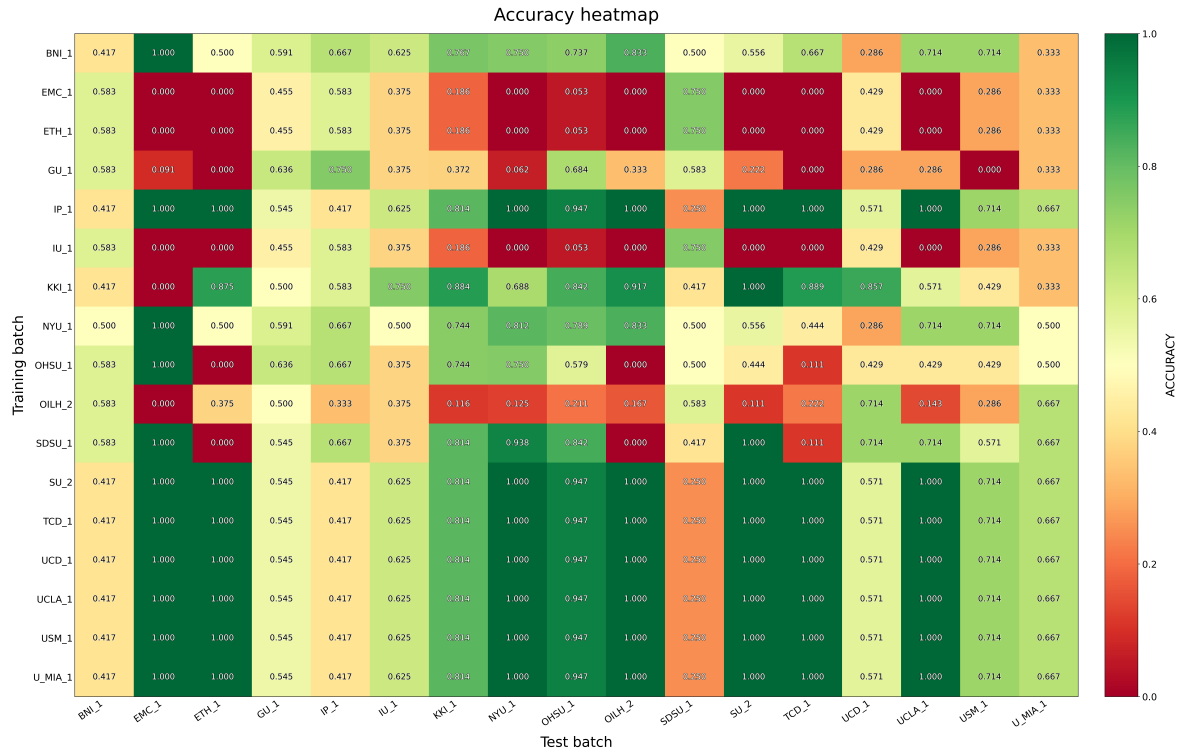


Figure 4.5: Source to target accuracy heatmap generated by dashi for the ABIDE II dataset. Rows represent the training domain and columns represent the target domain.

acted as the best source domain or the easiest target domain across all evaluation metrics. Instead, the observed behaviour depended on both the source target pair and the metric under consideration.

When discrimination oriented metrics such as AUC macro and PR-AUC macro were considered, sites such as KKI_1 and NYU_1 tended to produce stronger results when used as training sources. In contrast, domains including IP_1, UCD_1, U_MIA_1 and USM_1 generally achieved better performance for threshold dependent measures such as accuracy and F1 macro. Similar differences were observed from the target perspective. Sites such as IP_1, KKI_1, OHSU_1 and UCD_1 tended to be associated with higher AUC- and PR-AUC based performance, whereas EMC_1, NYU_1, OILH_2 and SU_2 appeared comparatively more favourable when accuracy and F1 score were considered.

An additional aspect of the supervised analysis was the presence of a substantial number of missing metric values, particularly for AUC macro. Overall, 357 metric values were unavailable, including 119 cells in the AUC macro matrix alone. For this reason, the interpretation of the supervised results should not rely exclusively on AUC based measures. Rather, the different metrics should be considered together to obtain a more complete picture of transferability patterns across sites.

Taken together, these results indicate that ABIDE II was a more heterogeneous transportability scenario than UCI. No single source or target site clearly dominated, suggesting that transferability was shaped by several factors acting together. Differences in sample size, class distribution, data completeness, and feature distributions likely contributed to the observed variability. This complexity helps explain the larger variability later observed in the adaptation experiments and shows why consistent cross-site generalization is difficult in multi site neuroimaging data.

4.2.3 TCGA-BRCA

The unsupervised analysis suggested a more moderate level of domain heterogeneity than that observed in UCI and ABIDE. The global dataset variability index (GPD) was equal to 0.3610, indicating that the retained sites were not completely overlapping in terms of their multivariate feature distributions, although the overall degree of separation was less pronounced than in the other datasets. The source overlap scores (SPO) nevertheless showed noticeable variability across sites. Domains such as C8, A8 and AC displayed relatively high overlap scores, whereas E2, D8 and AO were characterized by lower values, suggesting that some sites occupied more distinct positions within the feature space.

Figure 4.6 shows the two dimensional Multi Source Variability (MSV) representation obtained from the selected set of highly variable genes. Most sites were distributed within a relatively

compact region of the latent space, consistent with the lower global GPD value. At the same time, some domains occupied more peripheral positions. In particular, C8 and A8 appeared separated from the central cluster, while sites such as AC and B6 were located in a different region of the projection. Overall, the visualization indicates that differences between sites were present, although less extreme than those observed in ABIDE.

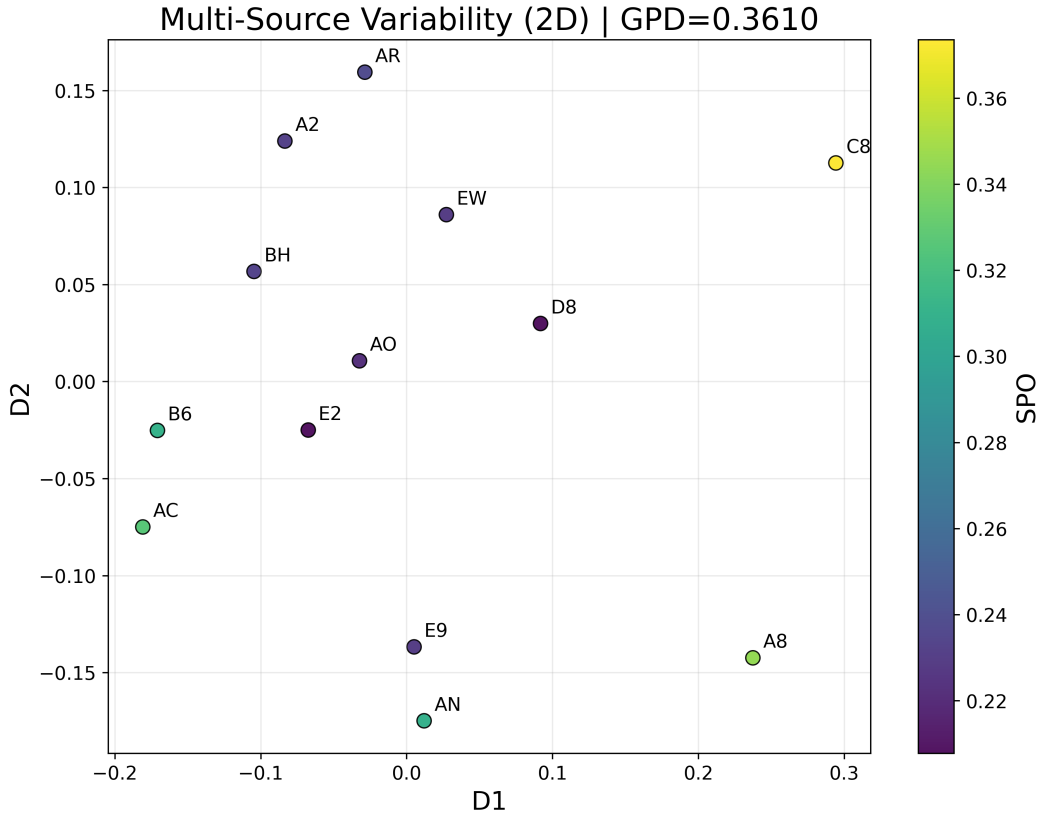


Figure 4.6: Two dimensional Multi Source Variability representation for the TCGA-BRCA dataset. Point size reflects domain sample size, while color indicates the source overlap score (SPO).

The supervised transferability analysis provided a complementary perspective on the relationships between sites. Figure 4.7 reports the source to target accuracy heatmap. Compared with ABIDE, the matrix appears more regular and easier to interpret, with fewer extreme values and a more coherent pattern across source target combinations. In addition, no missing metric values were observed in the supervised outputs, allowing a more stable comparison across domains and evaluation measures.

Several patterns emerged from the supervised results. No site was clearly the best across all metrics, but some domains showed a stronger overall behaviour. A8 was among the most effective source domains, with the highest average AUC macro and PR-AUC macro, and good results also for accuracy, F1 macro, and log-loss. B6 and C8 performed well in terms of accuracy when used as training sources, while E2 and D8 showed competitive results for F1 macro.

From the target domain perspective, AC emerged as the most favourable site in a remarkably

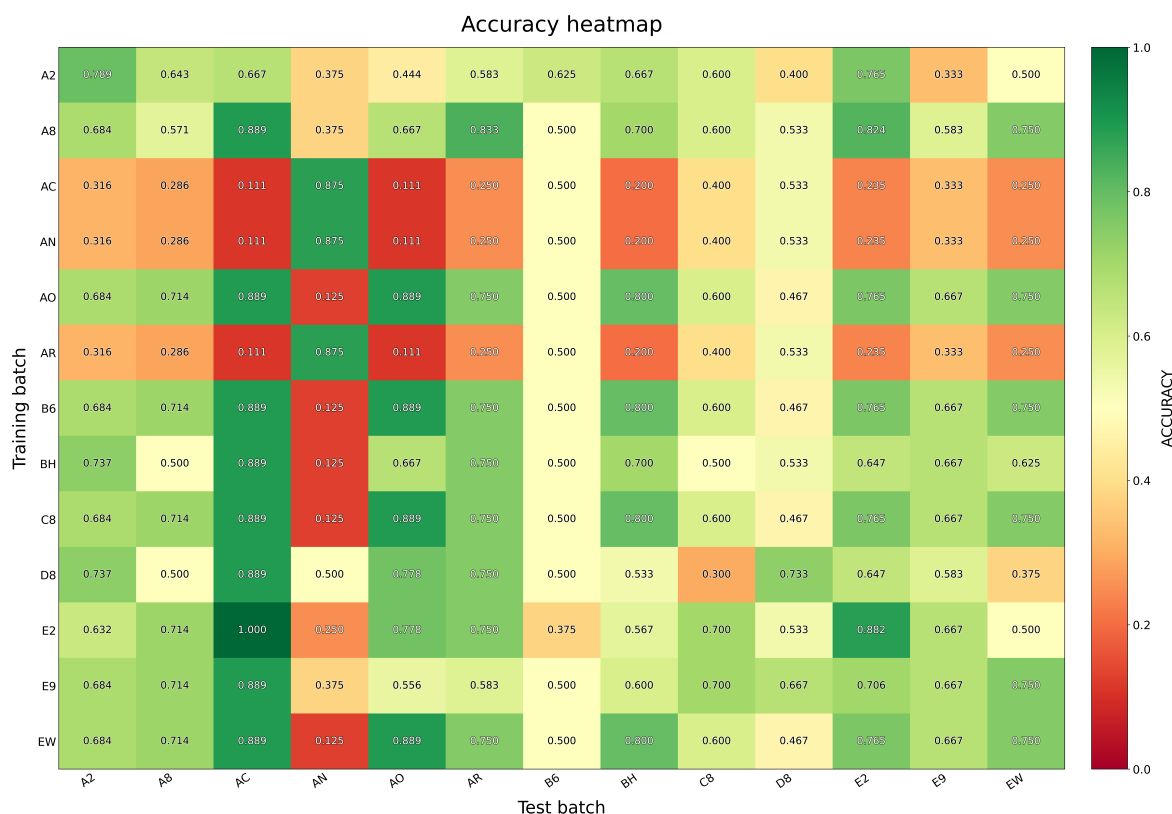


Figure 4.7: Source to target accuracy heatmap generated by dashi for the TCGA-BRCA dataset. Rows represent the training domain and columns represent the target domain.

consistent manner. It achieved the highest average values for multiple metrics, including AUC macro, accuracy and PR-AUC macro, while also showing the lowest average log-loss. Other domains such as AR, E2 and, to a lesser extent, A2 also tended to produce comparatively favourable target side performance. In contrast, AN represented the most challenging target domain across several evaluation measures, showing lower accuracy and F1 scores together with less favourable log-loss values. Less favourable profiles were also observed for EW, E9, B6 and D8 depending on the metric considered.

Taken together, the supervised results indicate that source to target transferability within TCGA-BRCA was not fully symmetric. Even though the overall multivariate separation between sites appeared relatively limited according to the MSV analysis, measurable differences in predictive performance were still observed across source target pairs. This finding suggests that a moderate degree of global distributional variability does not necessarily translate into uniform transportability.

Overall, the dashi analysis showed an important contrast. Even though the unsupervised analysis suggested limited global variability, the supervised results still revealed clear differences in how well sites transferred to each other. This means that transportability in TCGA-BRCA cannot be explained only by overall distributional overlap. More subtle factors, such as site composition, class prevalence, and high dimensional molecular profiles, likely played a role.

These factors may create transferability differences even when the global separation between sites appears small, helping to explain why adaptation methods later showed less consistent improvements than in the other datasets.

4.3 Performance of Adaptation Strategies

The main analysis compared the five adaptation strategies considered in this study: the baseline (M0), ComBat harmonization (M1), domain weighting (M2), transductive adaptation (M3), and the hybrid ComBat + transductive approach (M4).

Although several evaluation metrics were computed, this section focuses on OOD balanced accuracy and on the corresponding ID–OOD gap, in order to provide a clear and comparable view across datasets. OOD balanced accuracy is reported as the mean performance across 100 bootstrap iterations, with error bars representing the standard deviation across iterations. The ID–OOD gap is reported using heatmaps.

In the barplots, each group on the x axis corresponds to a target site, and the bars within each group represent the five adaptation strategies. For each target site, all remaining sites were merged and used as source domains, as described in Chapter 3. The target domain was divided into *train_1* and *test_1*, and the reported value was computed on *test_1*. Therefore, each bar represents the OOD performance obtained for a specific combination of model, method, and target site, summarized across the bootstrap iterations. In some figures, target sites are arranged over multiple rows to improve readability; this is only a visual choice and does not imply any grouping of sites.

The heatmaps follow the same target site structure, but focus on the ID–OOD gap. Rows correspond to the adaptation strategies (M0–M4), while columns represent the target domains. Each cell reports the gap observed for a given method and target site, using the same merged source setting.

4.3.1 UCI Heart Disease

In the UCI Heart Disease dataset, OOD performance was generally higher than in the other datasets. However, the relative behaviour of the methods still depended on the model, the metric, and the target site.

Figures 4.8–4.10 report the mean OOD balanced accuracy across the 100 bootstrap iterations for the four UCI target sites. Cleveland and Hungarian showed the highest values across models, whereas Long Beach VA and Switzerland were more challenging. The transductive methods, especially M3 and M4, often improved performance on Cleveland and Hungarian,

while the differences between methods were smaller or less consistent for the more difficult target sites. Variability across bootstrap iterations remained generally limited for most model method combinations.

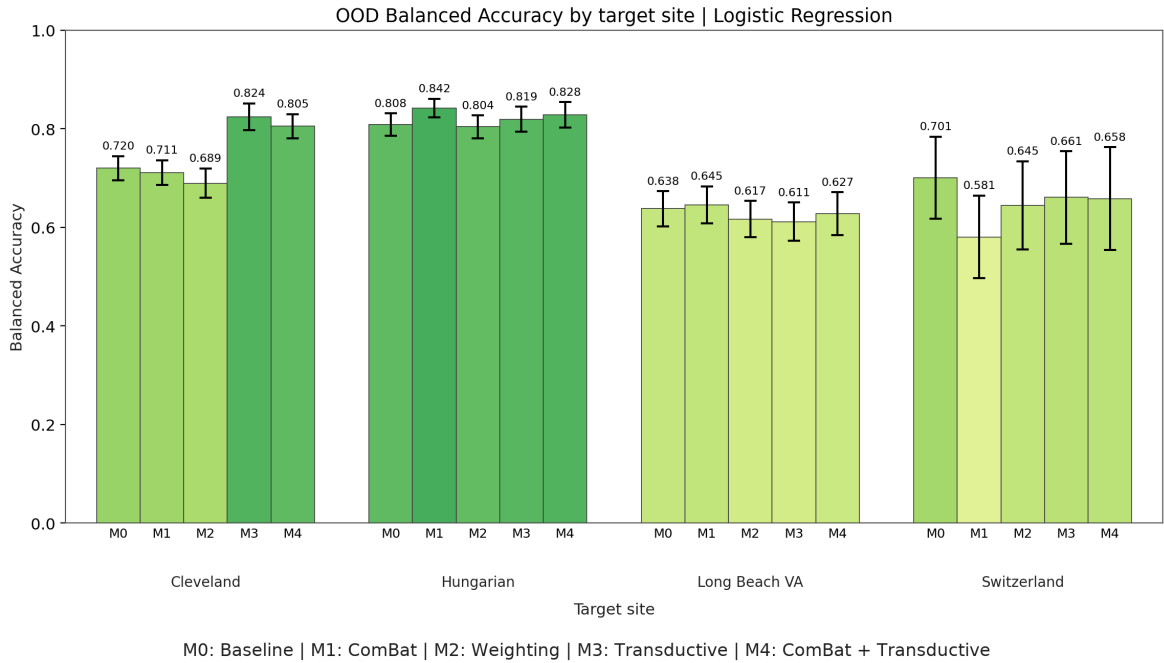


Figure 4.8: Bootstrap mean OOD balanced accuracy across UCI target sites for Logistic Regression. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations.

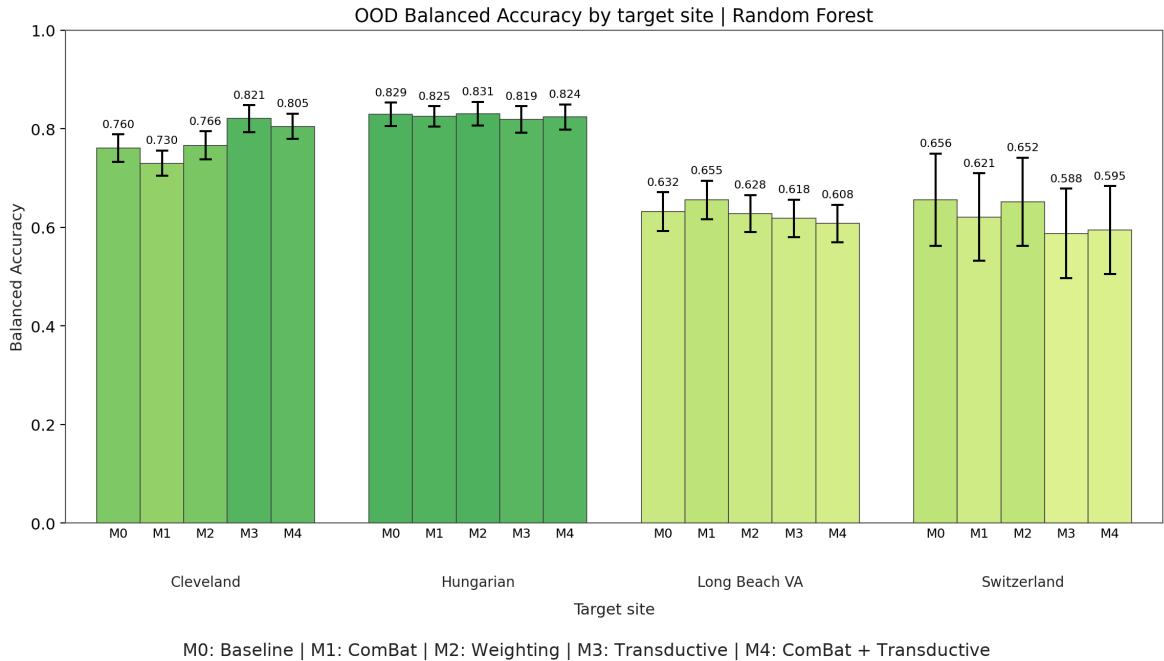


Figure 4.9: Bootstrap mean OOD balanced accuracy across UCI target sites for Random Forest. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations.

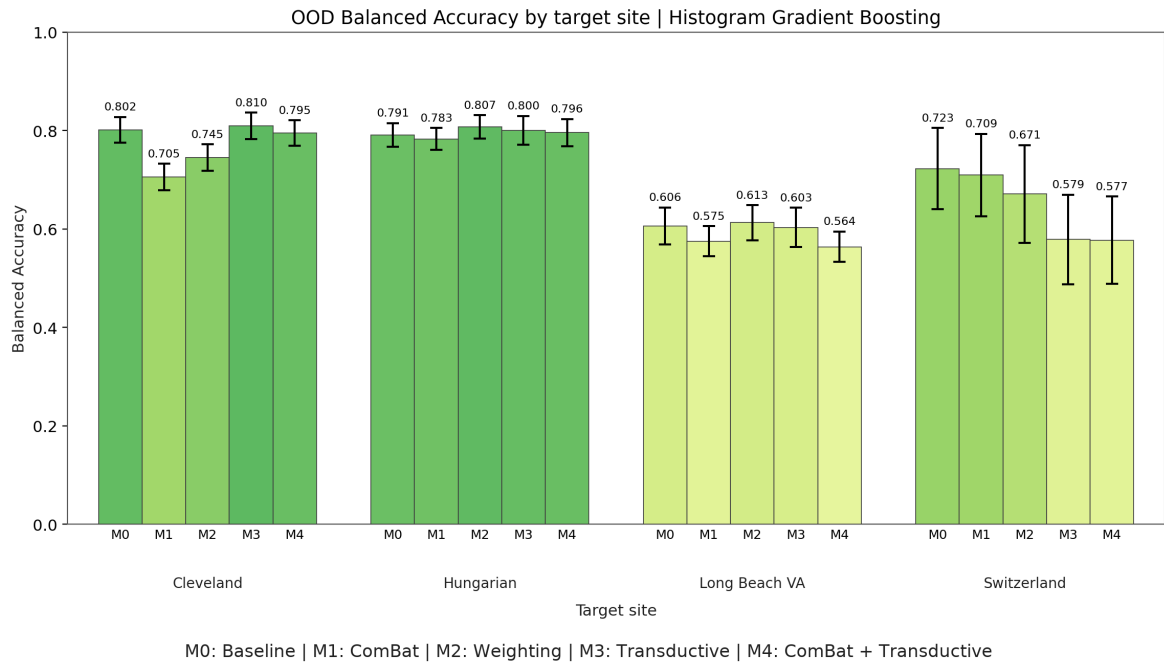


Figure 4.10: Bootstrap mean OOD balanced accuracy across UCI target sites for Histogram Gradient Boosting. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations.

Figures 4.11–4.13 report the ID–OOD balanced accuracy gap. Cleveland and Hungarian often showed negative or small gap values, indicating that OOD performance was close to, or in some cases higher than, ID performance. In contrast, Long Beach VA and Switzerland showed larger positive gaps, especially for some transductive configurations, suggesting a stronger loss of performance when moving from source domain validation to target domain testing.

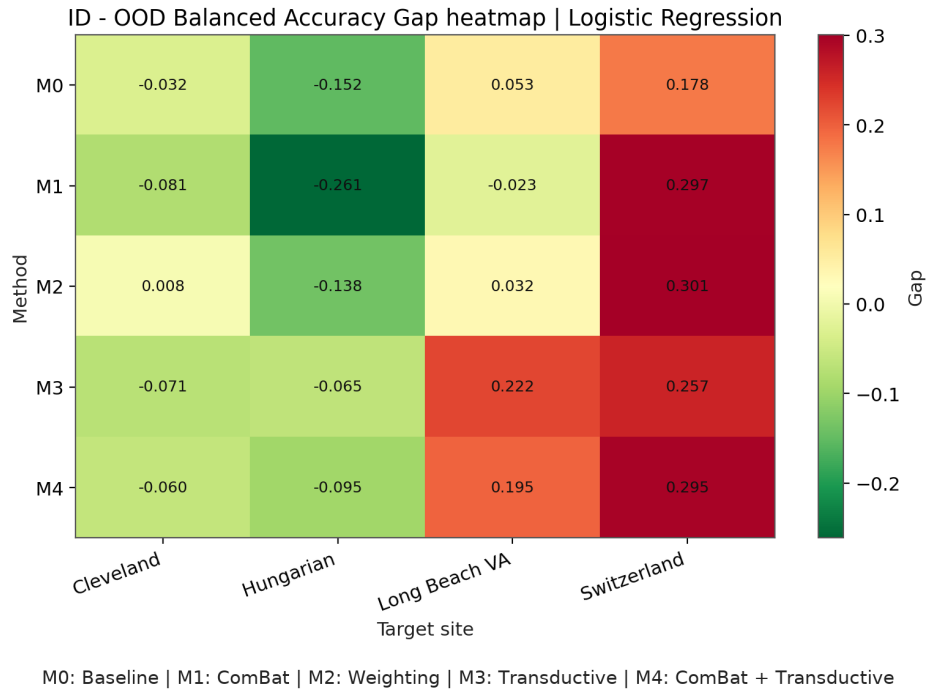


Figure 4.11: ID–OOD balanced accuracy gap across UCI target sites for Logistic Regression. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.

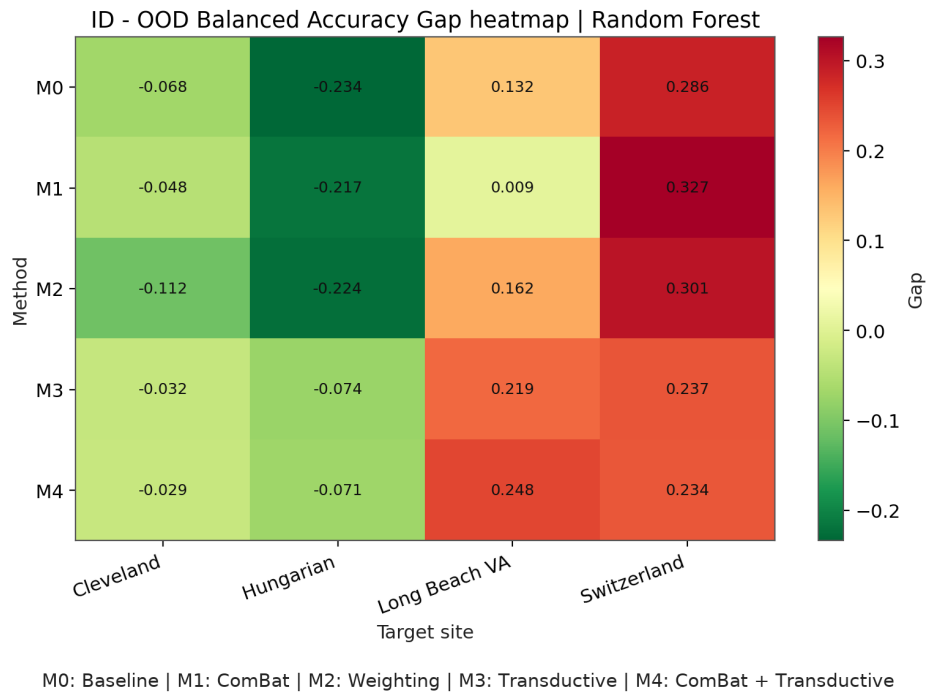


Figure 4.12: ID–OOD balanced accuracy gap across UCI target sites for Random Forest. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.

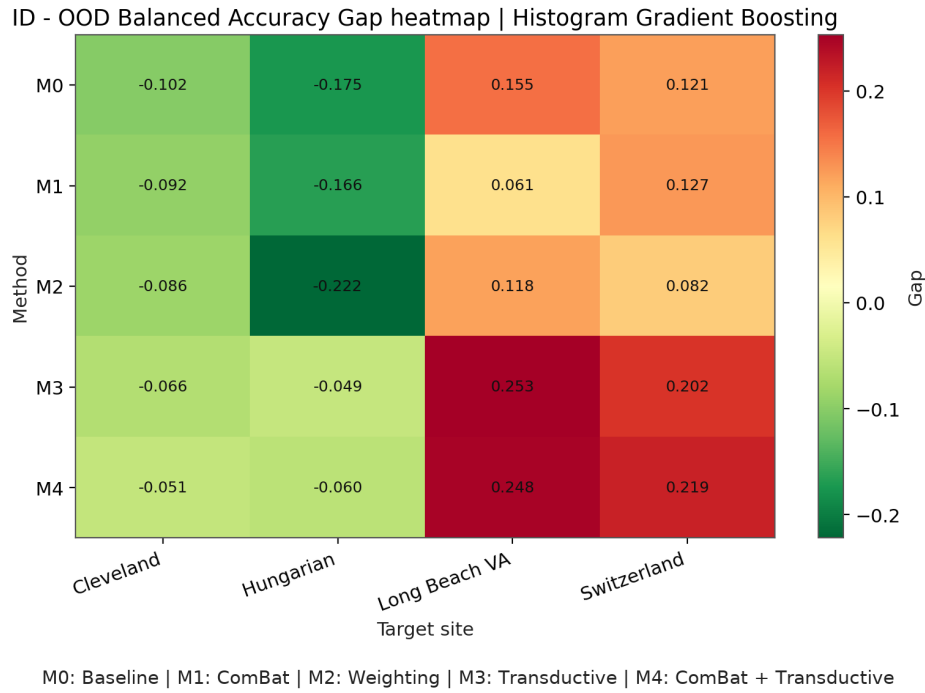


Figure 4.13: ID–OOD balanced accuracy gap across UCI target sites for Histogram Gradient Boosting. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.

In summary, UCI Heart Disease showed higher OOD balanced accuracy than the other datasets, with visible differences across target sites and classifiers. M3 and M4 often achieved high values in Cleveland and Hungarian, while Long Beach VA and Switzerland remained more variable across methods.

4.3.2 ABIDE II

In ABIDE II, the results were less consistent than in UCI. The relative performance of the adaptation strategies varied more noticeably across classifiers and target sites, indicating that their effectiveness was strongly influenced by the specific site being considered.

Figures 4.14–4.16 show the mean OOD balanced accuracy across the 100 bootstrap iterations for the 17 eligible ABIDE II target sites. Some domains showed relatively stable performance across methods, whereas others displayed larger differences between adaptation strategies. Compared with UCI, the results showed wider variability across target sites and classifiers, with some configurations presenting larger error bars.

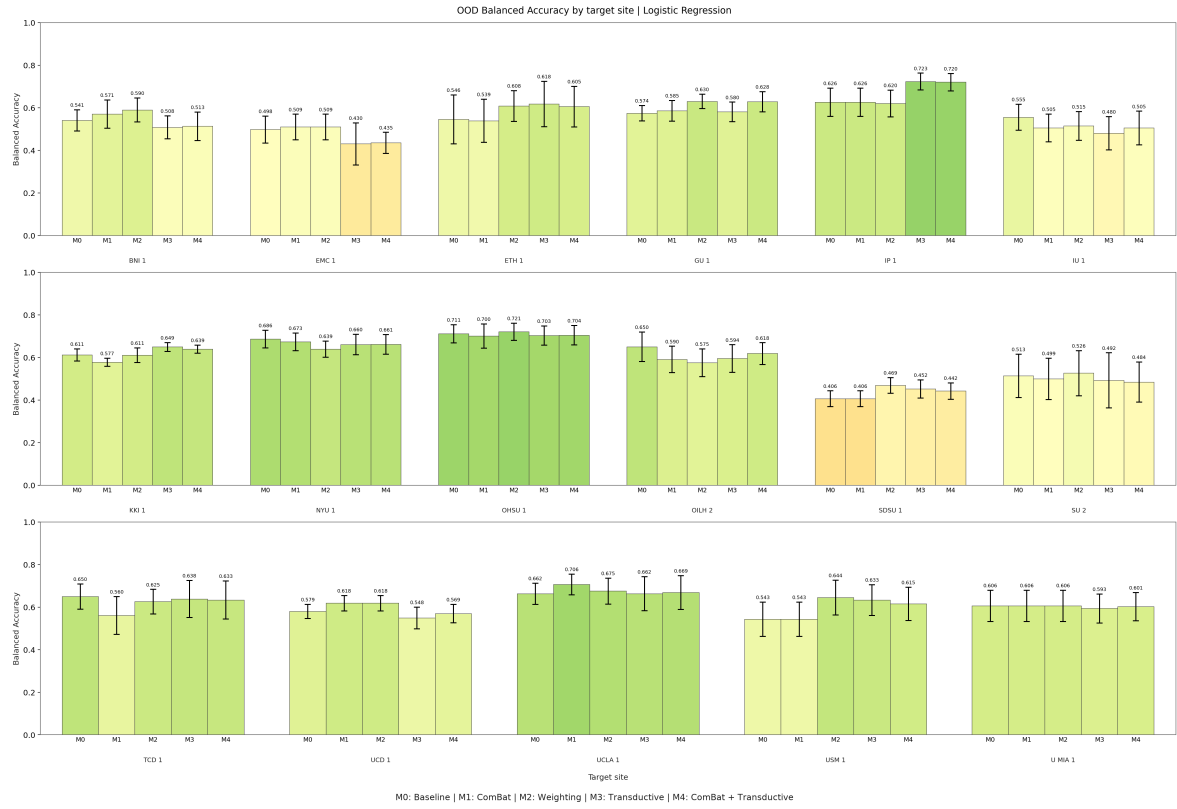


Figure 4.14: Bootstrap mean OOD balanced accuracy across ABIDE II target sites for Logistic Regression. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations.

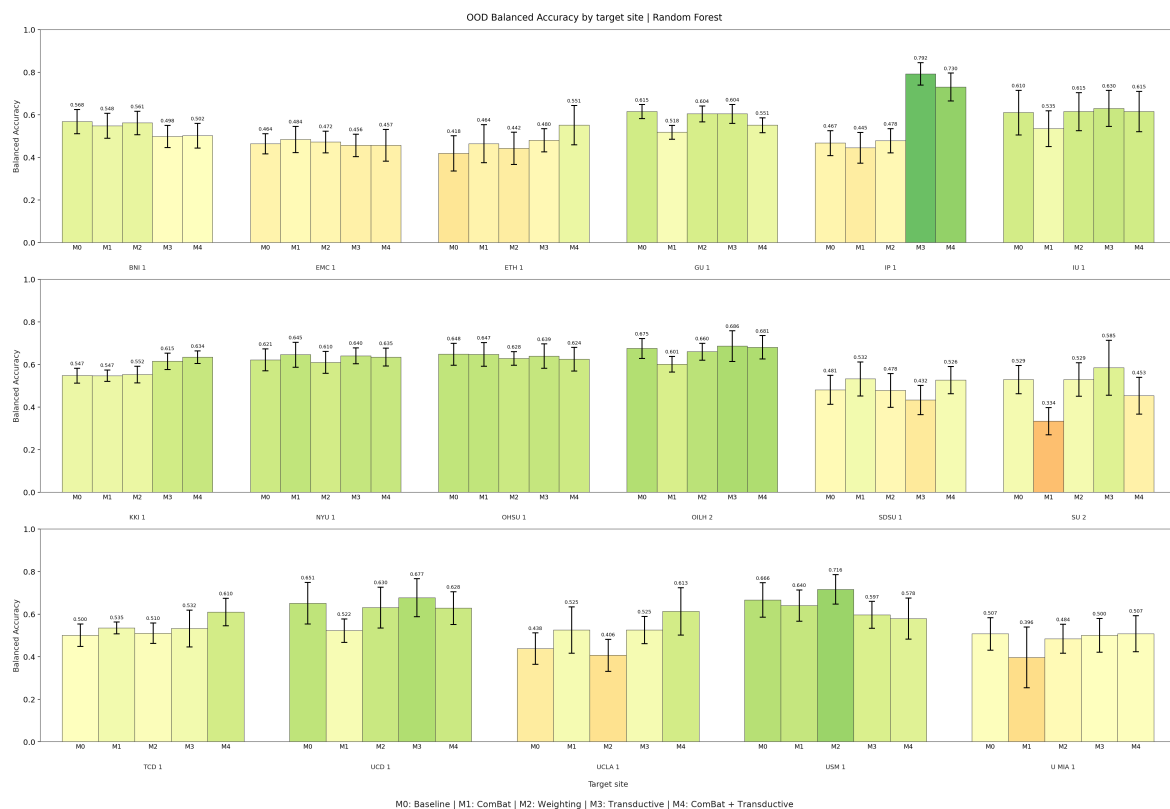


Figure 4.15: Bootstrap mean OOD balanced accuracy across ABIDE II target sites for Random Forest. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations.

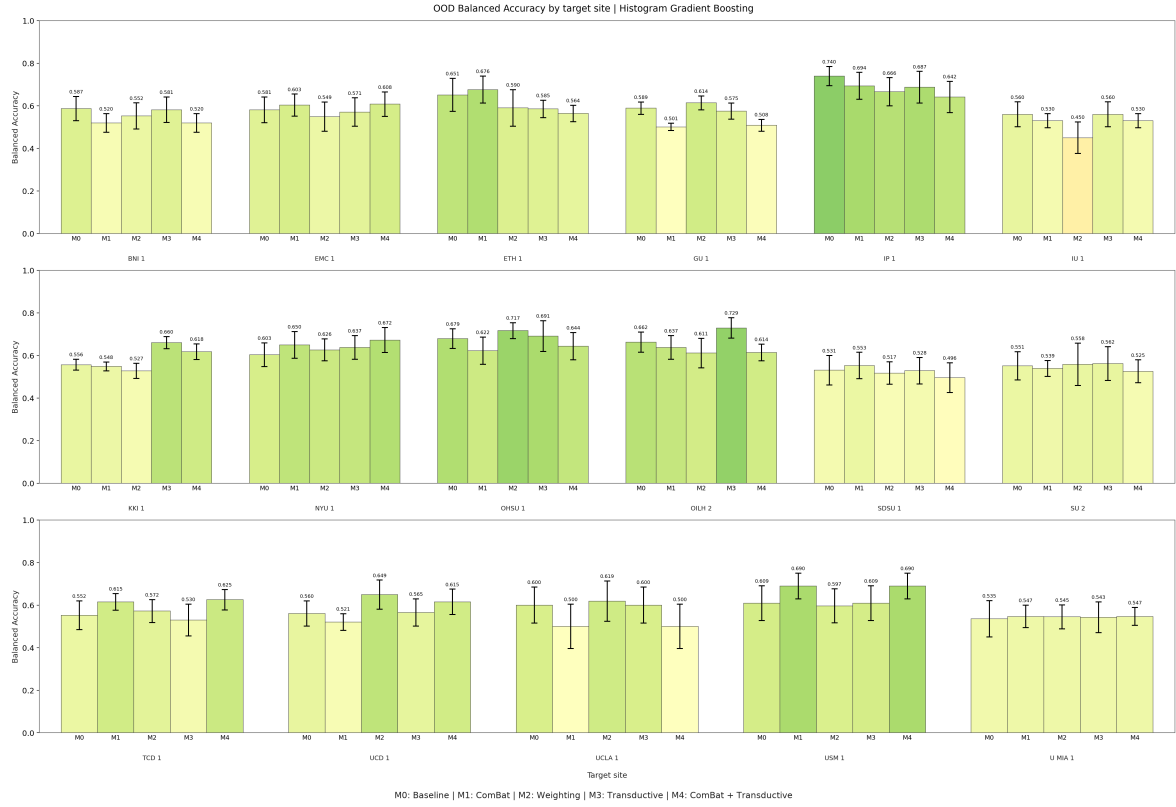


Figure 4.16: Bootstrap mean OOD balanced accuracy across ABIDE II target sites for Histogram Gradient Boosting. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations.

Figures 4.17–4.19 report the ID–OOD balanced accuracy gap. The heatmaps confirm that the transferability gap was not uniform across sites. Some target domains showed small or negative gaps, whereas others showed larger losses from ID to OOD evaluation, depending on the model and method used.

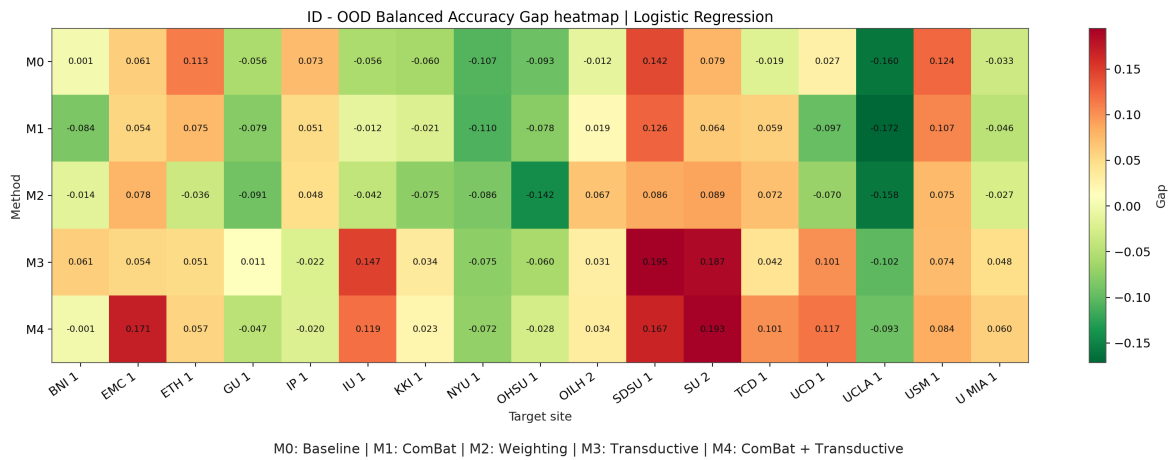


Figure 4.17: ID–OOD balanced accuracy gap across ABIDE II target sites for Logistic Regression. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.

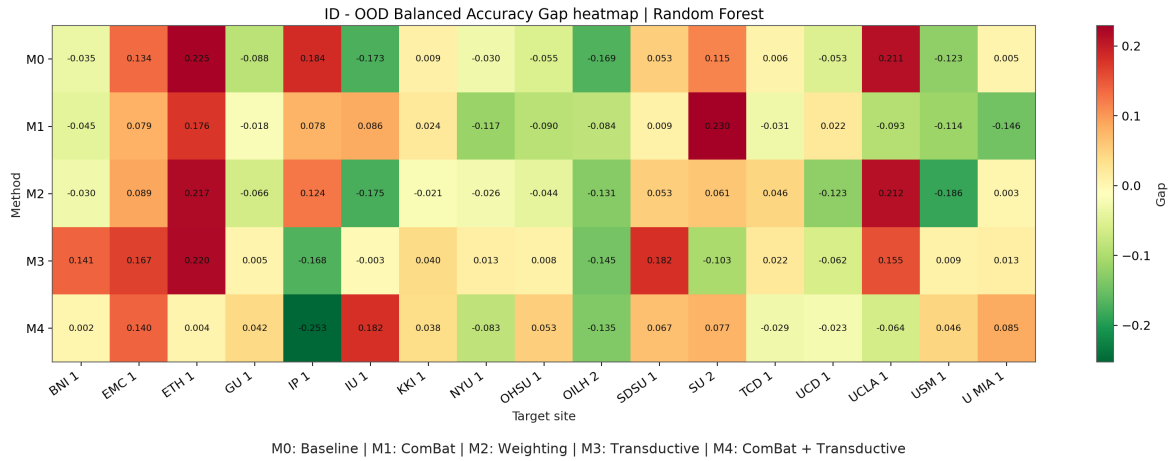


Figure 4.18: ID–OOD balanced accuracy gap across ABIDE II target sites for Random Forest. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.

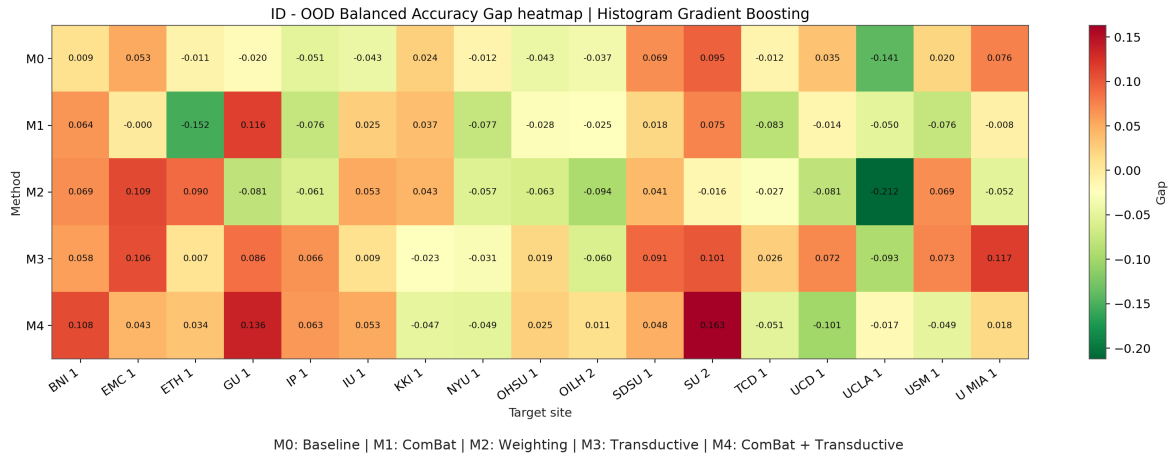


Figure 4.19: ID–OOD balanced accuracy gap across ABIDE II target sites for Histogram Gradient Boosting. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.

Overall, the results for ABIDE II showed substantial variability across target sites, machine learning models, and adaptation strategies. The ranking of the methods changed across domains and performance metrics, with no single approach consistently achieving the highest OOD performance.

4.3.3 TCGA-BRCA

TCGA-BRCA produced the most challenging transportability scenario among the three datasets. Overall OOD performance was lower, and the adaptive methods did not systematically outperform the baseline.

Figures 4.20–4.22 report the mean OOD balanced accuracy across the 100 bootstrap iterations for the thirteen TCGA-BRCA target sites. Several sites showed only modest differences

between methods, while others displayed stronger variability depending on the classifier and adaptation strategy. Sites such as A2, AO, and B6 generally showed higher balanced accuracy values, whereas AN appeared among the most challenging targets. Several target sites also displayed non negligible variability across bootstrap iterations, as reflected by wider error bars.

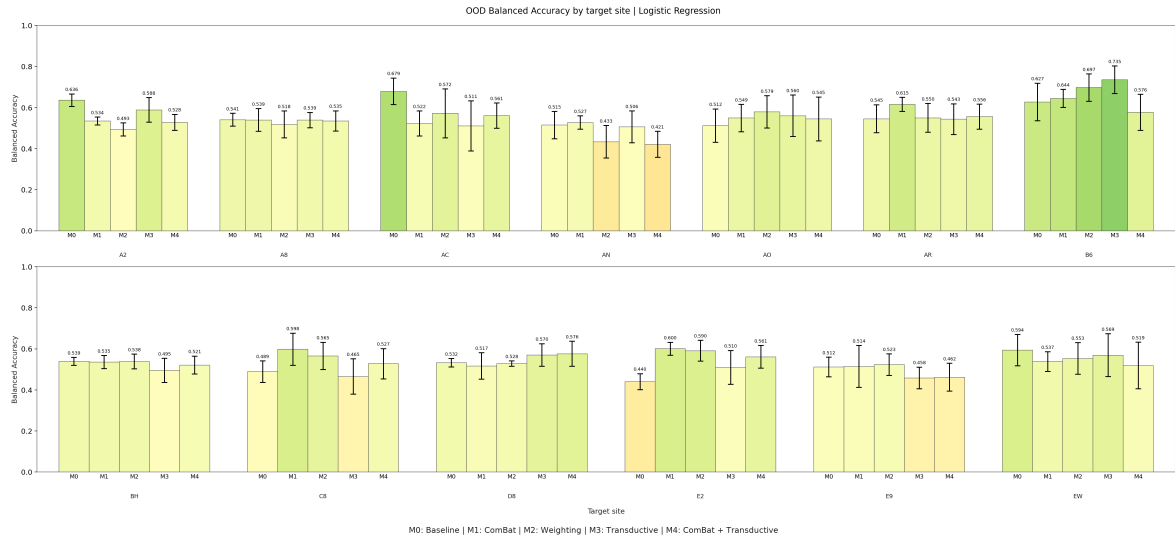


Figure 4.20: Bootstrap mean OOD balanced accuracy across TCGA-BRCA target sites for Logistic Regression. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations.

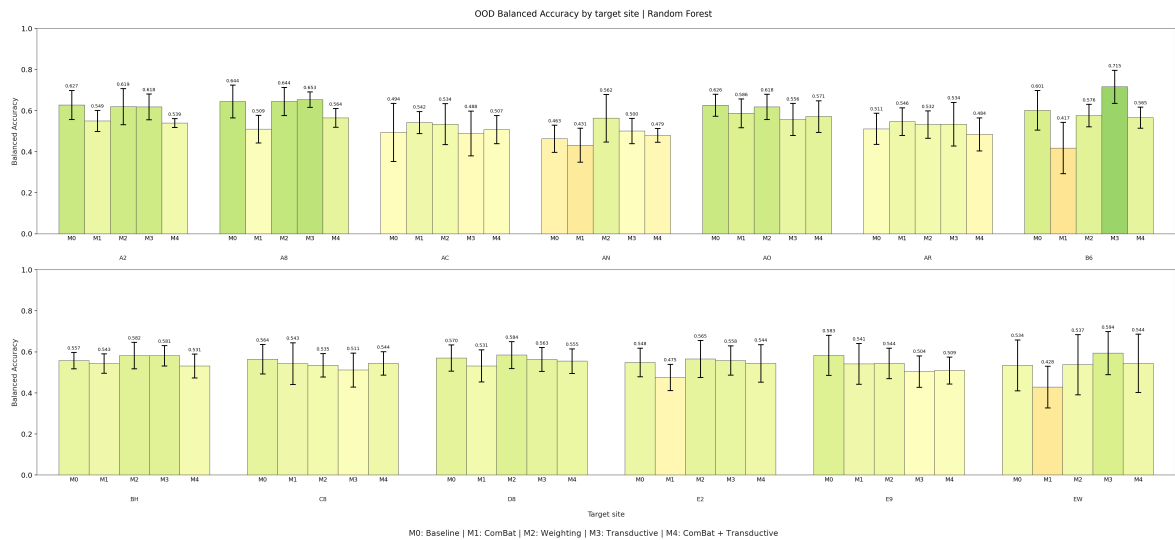


Figure 4.21: Bootstrap mean OOD balanced accuracy across TCGA-BRCA target sites for Random Forest. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations.

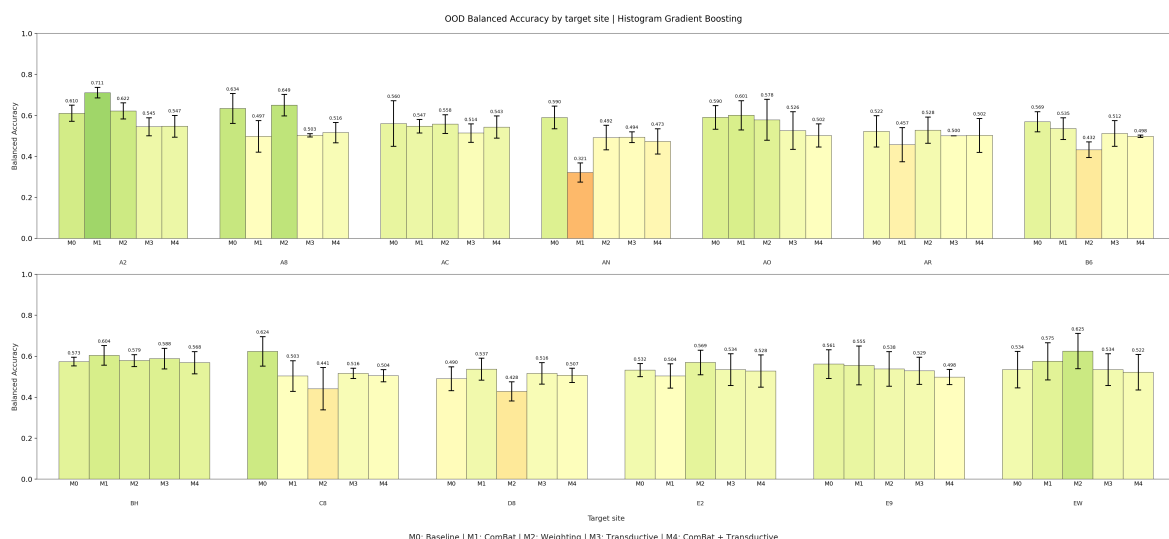


Figure 4.22: Bootstrap mean OOD balanced accuracy across TCGA-BRCA target sites for Histogram Gradient Boosting. Bars represent the five adaptation strategies M0–M4, and error bars indicate the standard deviation across 100 bootstrap iterations.

Figures 4.23–4.25 report the ID–OOD ROC-AUC gap. The heatmaps show a heterogeneous pattern, with some target sites showing negative gaps and others showing stronger losses when moving from source domain validation to target domain evaluation. No method consistently minimized the gap across all models and target sites.

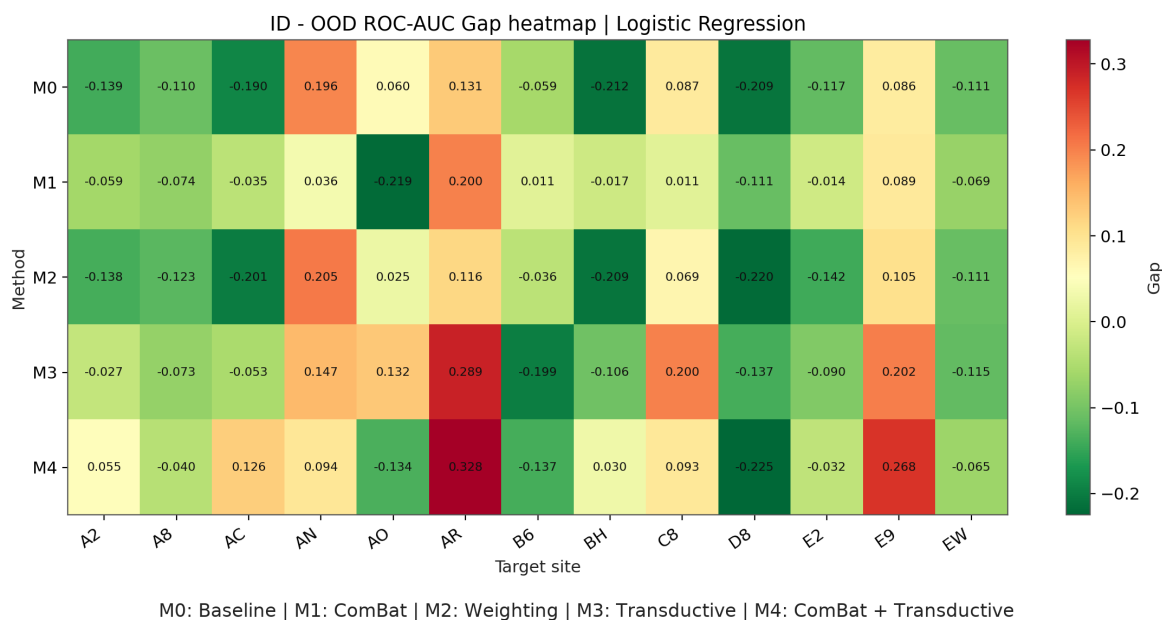


Figure 4.23: ID–OOD ROC-AUC gap across TCGA-BRCA target sites for Logistic Regression. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.

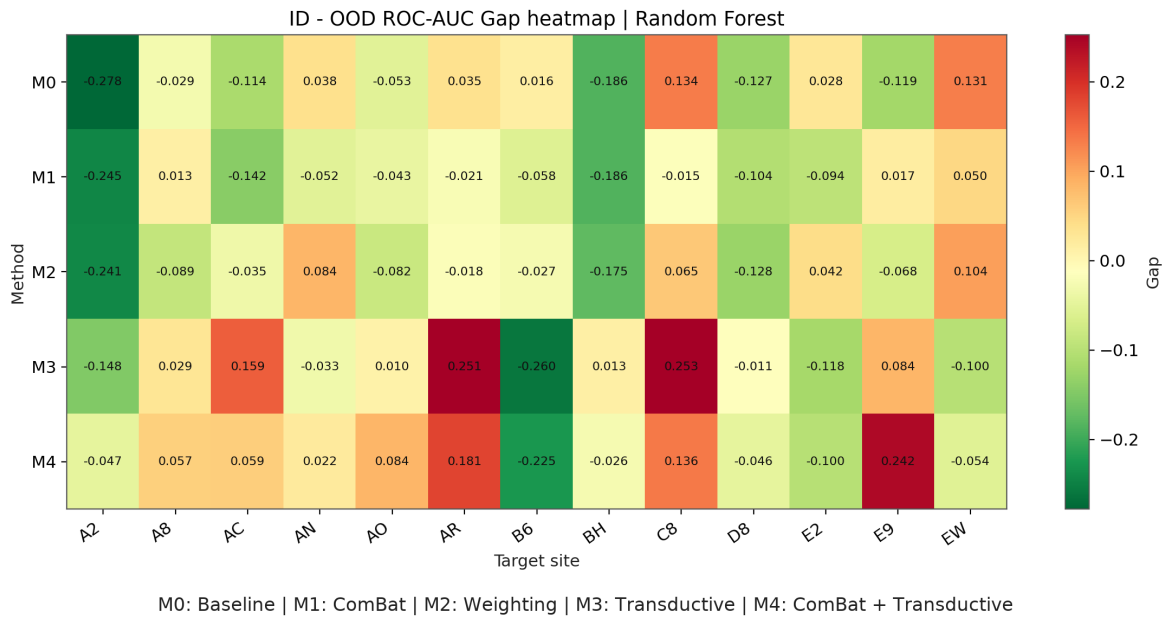


Figure 4.24: ID–OOD ROC-AUC gap across TCGA-BRCA target sites for Random Forest. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.

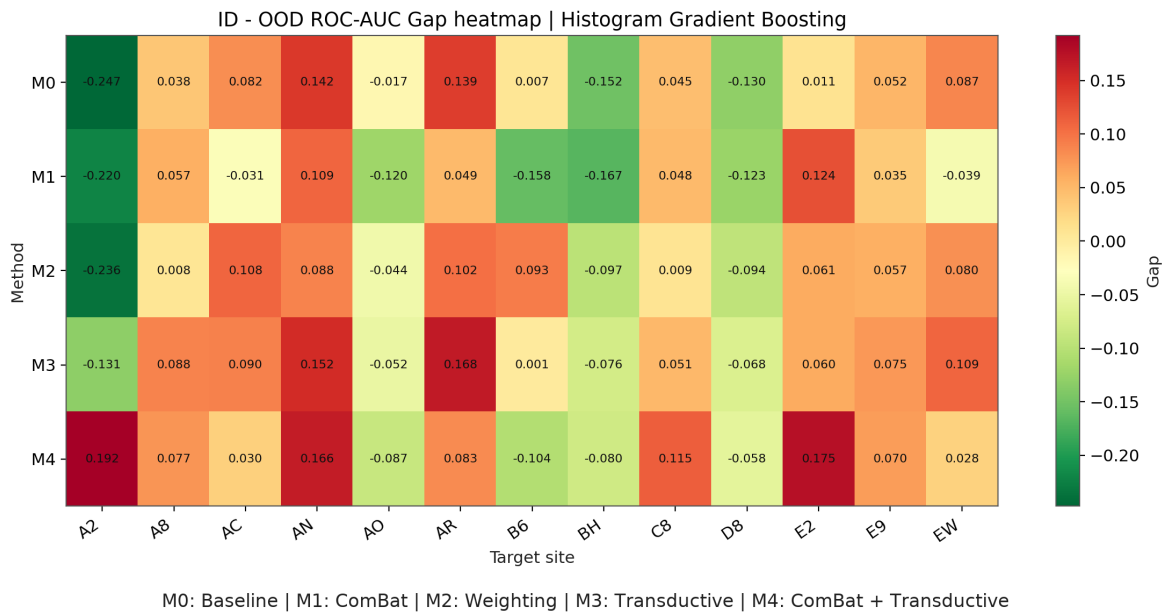


Figure 4.25: ID–OOD ROC-AUC gap across TCGA-BRCA target sites for Histogram Gradient Boosting. Rows represent the adaptation strategies M0–M4, while columns represent the target sites.

Overall, the TCGA-BRCA results showed limited differences in OOD discrimination across adaptation strategies. In several target sites, the baseline M0 achieved performance comparable to or higher than the adaptation methods, whereas M3 and M4 more frequently produced improvements in calibration than in discrimination.

Target site effects were also evident. AO and A2 appeared as the most favourable domains in terms of ROC-AUC, while AN was the most difficult target and also showed poor balanced

accuracy and calibration. The ranking of target difficulty changed across metrics, indicating that a site could appear less problematic according to F1-score but still be difficult in terms of ROC-AUC or balanced accuracy.

4.4 Cross Dataset Comparison

After analysing each dataset separately, a direct comparison was performed to evaluate how the adaptation strategies behaved across different transportability scenarios. Figure 4.26 reports the mean OOD ROC-AUC obtained by the five methods for each machine learning model and dataset.

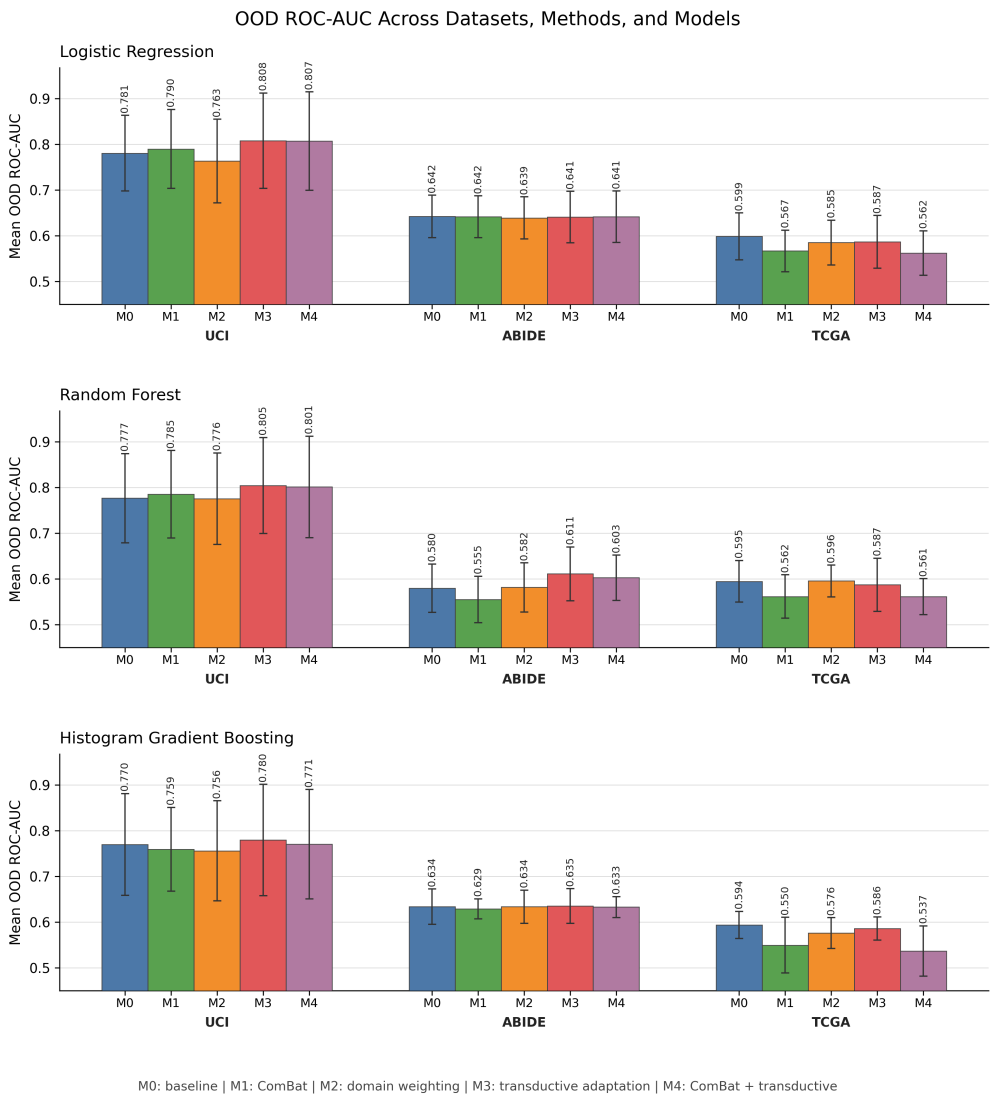


Figure 4.26: Mean OOD ROC-AUC across datasets, adaptation strategies, and machine learning models. Error bars represent variability across target sites.

A clear performance gradient emerged across the three datasets. UCI Heart Disease con-

sistently achieved the highest OOD ROC-AUC values, with all models reaching average performances around 0.75–0.81. ABIDE II showed intermediate results, with most configurations concentrated around 0.63–0.64. TCGA-BRCA represented the most challenging scenario, with OOD ROC-AUC values generally below 0.60 regardless of the adaptation strategy adopted. These findings indicate that transportability becomes progressively more difficult when moving from structured clinical variables to neuroimaging and high dimensional genomic data.

The behaviour of the adaptation methods also changed noticeably across datasets. In UCI, the transductive methods M3 and M4 generally achieved the highest average ROC-AUC across the three models. In ABIDE II, M3 remained one of the strongest methods, but the differences between strategies were smaller. In TCGA-BRCA, adaptive methods did not show a clear advantage. The baseline M0 often remained highly competitive and frequently obtained the highest average ROC-AUC, while M2 and M3 improved performance only in some specific configurations.

Another important point concerns how large the differences between methods were. In UCI, the gap between the best and worst strategies was more evident, especially for Logistic Regression and Random Forest. In ABIDE II, the methods were closer to each other, and in TCGA-BRCA the differences became even smaller. This suggests that adaptation is not useful simply because domain shift is present. Its effectiveness depends on whether the available target information can actually help the model generalize better without adding further instability.

The error bars provide further insight into this behaviour. Variability across target sites was generally limited in UCI, increased in ABIDE, and remained substantial in TCGA. This observation is consistent with the dataset specific analyses presented earlier and supports the view that site level heterogeneity plays an increasingly important role as the transportability problem becomes more complex.

4.5 Robustness and Reliability Analyses

4.5.1 Calibration

Calibration was assessed using the Expected Calibration Error (ECE), where lower values indicate better agreement between predicted probabilities and observed outcomes. Differently from metrics such as ROC-AUC or balanced accuracy, ECE evaluates how reliable the predicted probabilities are, providing complementary information about model performance.

Across the three datasets, calibration showed a more consistent behaviour than OOD dis-

crimination. Although no adaptation strategy was best across all metrics, M3 and M4 almost always achieved the lowest ECE values. This suggests that using target domain information may be especially useful for improving probability estimates, even when the improvement in classification performance is limited.

In UCI Heart Disease, M3 obtained the lowest mean OOD ECE (0.1176), followed by M4 (0.1282), whereas the source only approaches showed higher calibration error. Hungarian and Cleveland were the best calibrated target sites, while Switzerland consistently displayed poorer calibration. Interestingly, these results broadly mirror the transportability patterns observed in the previous analyses, where Switzerland also emerged as one of the most challenging domains.

A similar tendency was observed in ABIDE II. M3 and M4 achieved almost identical mean ECE values, 0.1834 and 0.1833, respectively, clearly outperforming M0, M1 and M2. Nevertheless, calibration remained strongly site dependent, with some domains showing substantially more reliable probability estimates than others. This suggests that, even when adaptation improves calibration on average, local site characteristics continue to influence the quality of uncertainty estimation.

The difference between calibration and discrimination was especially clear in TCGA-BRCA. M0 generally showed the best OOD discrimination performance, but M3 and M4 achieved the lowest ECE values, equal to 0.1494 and 0.1569. This means that the methods producing the most accurate classifications were not necessarily the same methods producing the most reliable probability estimates. From a calibration perspective, AN was the most problematic target site, as it consistently showed higher calibration error than the other domains.

Overall, calibration was one of the clearest aspects in which transductive adaptation provided a stable benefit. However, better calibration did not always correspond to better discrimination. These results show that discrimination and calibration capture different aspects of transportability and should be evaluated together. In particular, the TCGA-BRCA results suggest that adaptation may still be useful even when ROC-AUC or balanced accuracy do not improve substantially, because better calibrated probabilities can support more reliable downstream decisions.

4.5.2 Entropy, Sample Size and Site Distance

An exploratory analysis was conducted using three target domain meta features to better understand the conditions under which each adaptation strategy was most effective: class entropy, target sample size, and source–target distance. In this work, source–target distance refers to a score summarizing the dissimilarity between the target domain and the pooled source domains. The score was obtained by averaging the standardized values of five comple-

mentary distance measures: Euclidean centroid distance, Mahalanobis distance, Wasserstein distance, Jensen–Shannon divergence, and Maximum Mean Discrepancy MMD. The results of this analysis should be interpreted as descriptive associations rather than as evidence of causal relationships.

Figures 4.27–4.35 examine whether these target domain characteristics were associated with the effectiveness of the adaptation strategies. For each target site, machine learning model, and adaptation strategy M_i , with $i \in \{1, 2, 3, 4\}$, the change in OOD ROC-AUC relative to the baseline $M0$ was computed as:

$$\Delta\text{ROC-AUC} = \text{ROC-AUC}_{M_i} - \text{ROC-AUC}_{M0}. \quad (4.1)$$

Each point corresponds to the OOD performance gain or loss obtained for one combination of target site and machine learning model under a given adaptation strategy. The colour identifies the adaptation strategy. The horizontal coordinate represents the characteristic of the corresponding target domain, such as class entropy, target sample size, or source–target distance, while the vertical coordinate represents the associated $\Delta\text{ROC-AUC}$ with $M0$. Positive values indicate that the adaptation strategy achieved a higher OOD ROC-AUC than the baseline, whereas negative values indicate lower performance.

Because the same target site was evaluated using multiple machine learning models, several points may share the same value on the horizontal axis while presenting different values of $\Delta\text{ROC-AUC}$. The solid lines represent separate linear regression fits computed across all target site and machine learning model evaluations for each adaptation strategy. They therefore summarize the overall association between the selected target domain characteristic and the benefit of adaptation, rather than the average performance of the machine learning models. The shaded areas around the regression lines indicate the corresponding confidence intervals.

4.5.2.1 UCI Heart Disease

Figures 4.27, 4.28, and 4.29 summarize the relationships between OOD performance gains relative to $M0$ and the three target domain meta features considered for UCI Heart Disease.

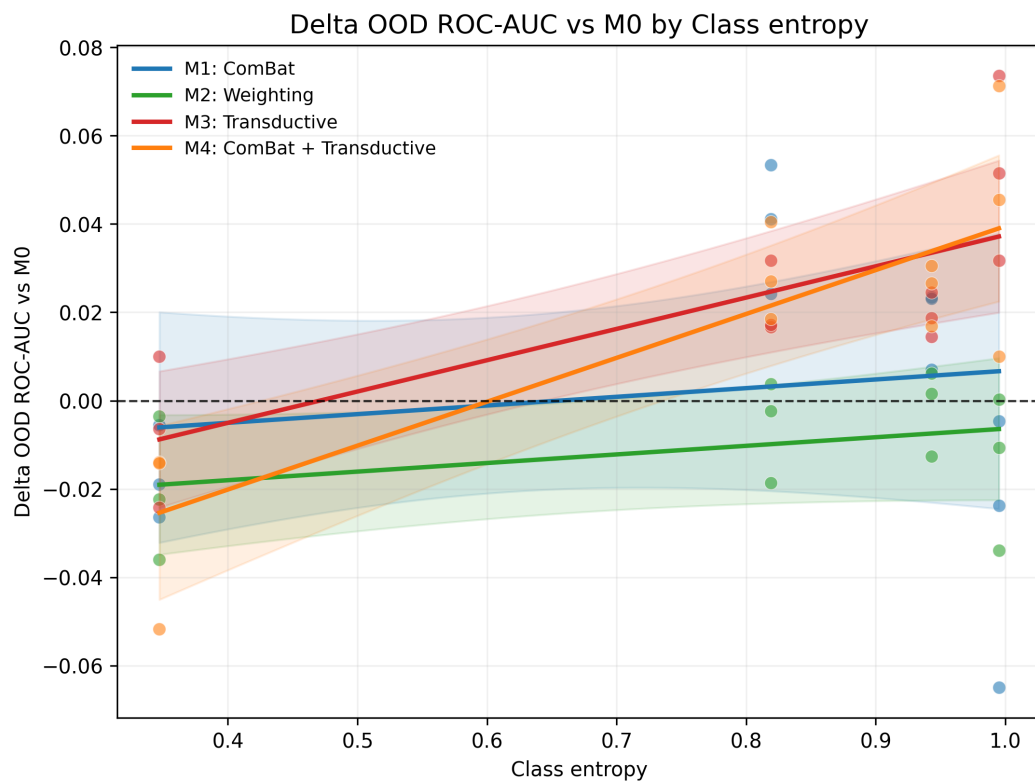


Figure 4.27: Delta OOD ROC-AUC relative to $M0$ as a function of class entropy in UCI Heart Disease.

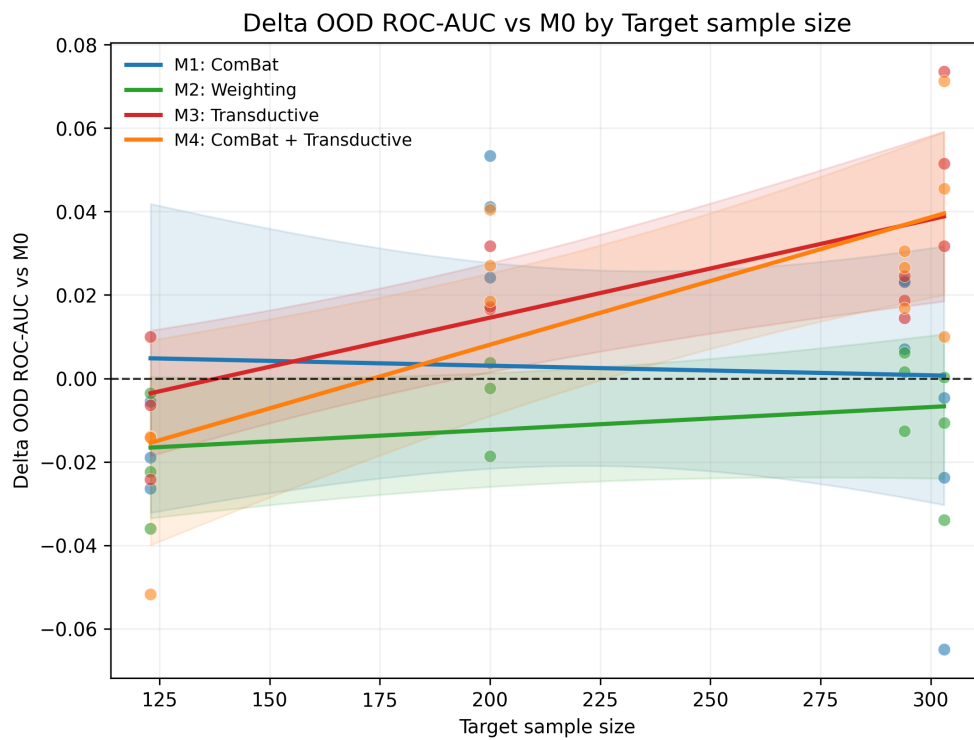


Figure 4.28: Delta OOD ROC-AUC relative to $M0$ as a function of target sample size in UCI Heart Disease.

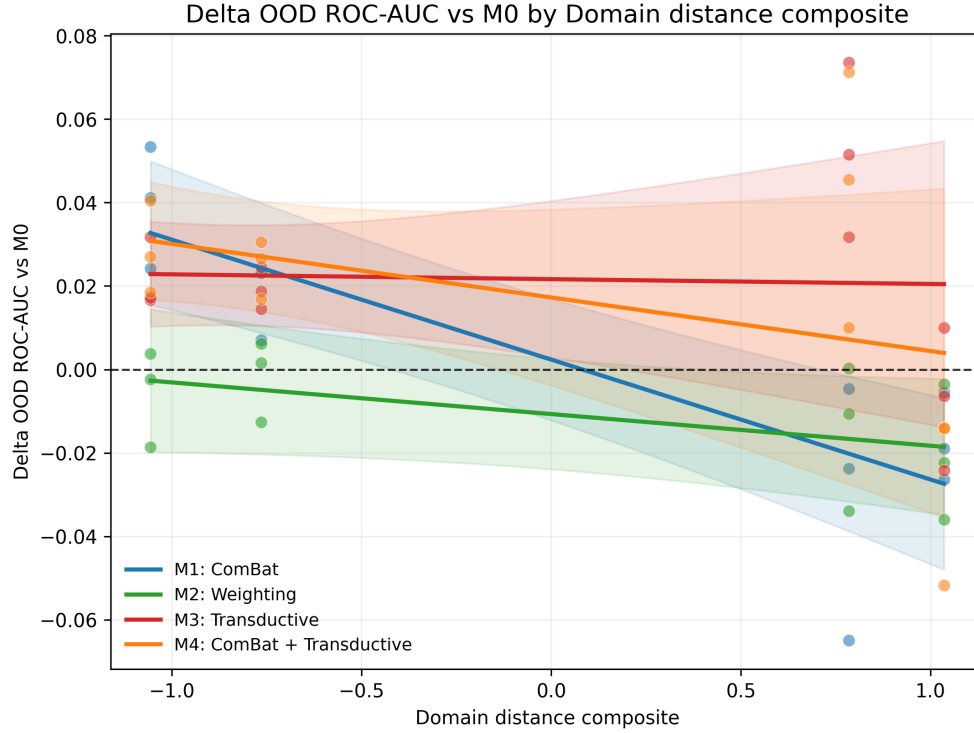


Figure 4.29: Delta OOD ROC-AUC relative to $M0$ as a function of source–target distance in UCI Heart Disease.

Among the three datasets, UCI showed the clearest descriptive associations between target domain characteristics and the relative effectiveness of adaptation. Larger target sites and more balanced class distributions were generally associated with greater performance gains for the target aware methods, particularly $M3$. Under the most favourable conditions, $M3$ achieved the highest average OOD ROC-AUC, equal to 0.894, corresponding to an improvement of approximately 0.05 ROC-AUC points over the baseline.

Source–target distance showed a different pattern. Adaptation gains were generally larger when the target domain was relatively close to the pooled source domains and tended to decrease as the distance increased. At higher distance values, the advantage over $M0$ was reduced and, in some target site and model combinations, disappeared altogether. This descriptive pattern suggests that transductive adaptation may be more beneficial when the target domain differs from the source domains but still preserves sufficient similarity for the available target information to support adaptation.

4.5.2.2 ABIDE II

The corresponding analyses for ABIDE II are reported in Figures 4.30, 4.31, and 4.32.

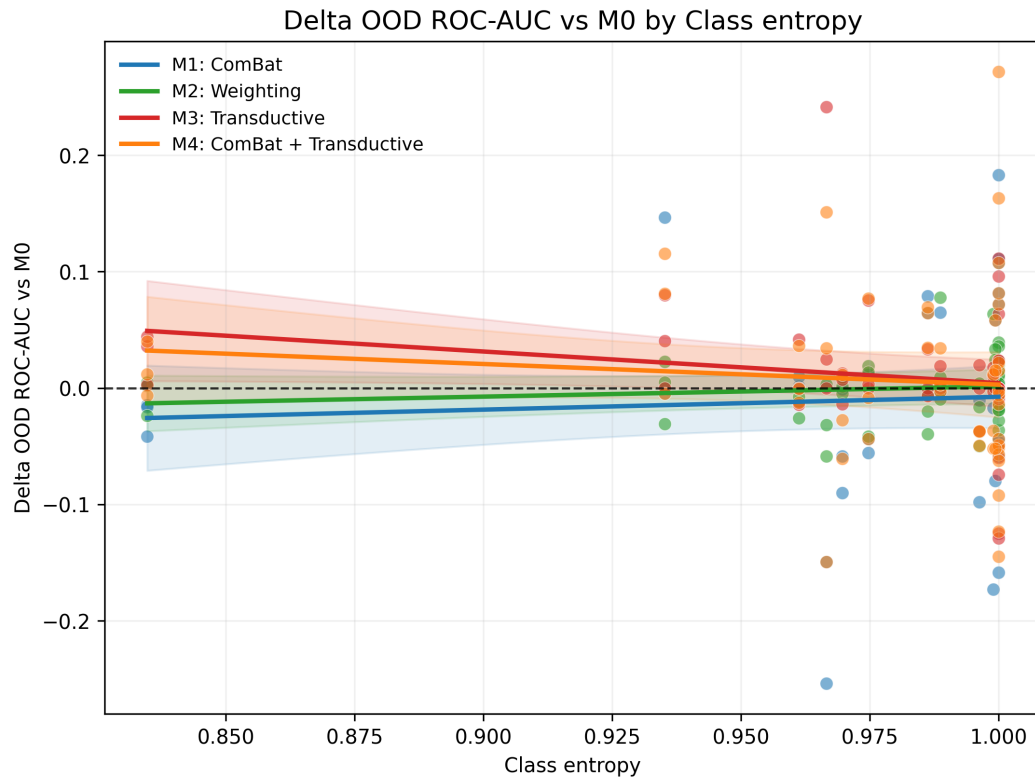


Figure 4.30: Delta OOD ROC-AUC relative to $M0$ as a function of class entropy in ABIDE II.

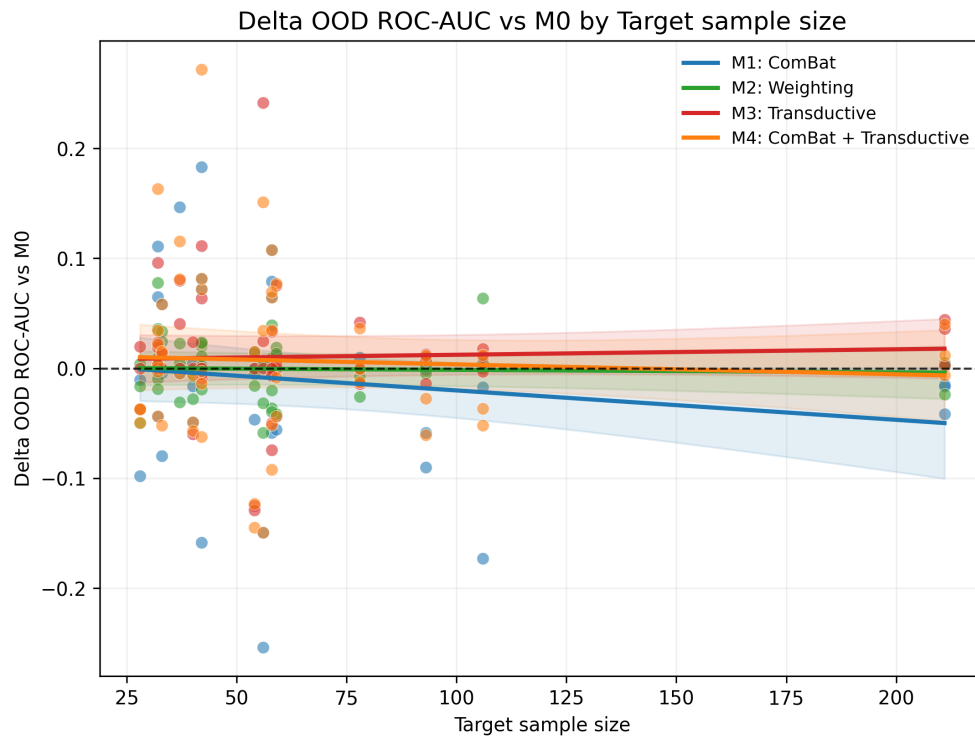


Figure 4.31: Delta OOD ROC-AUC relative to $M0$ as a function of target sample size in ABIDE II.

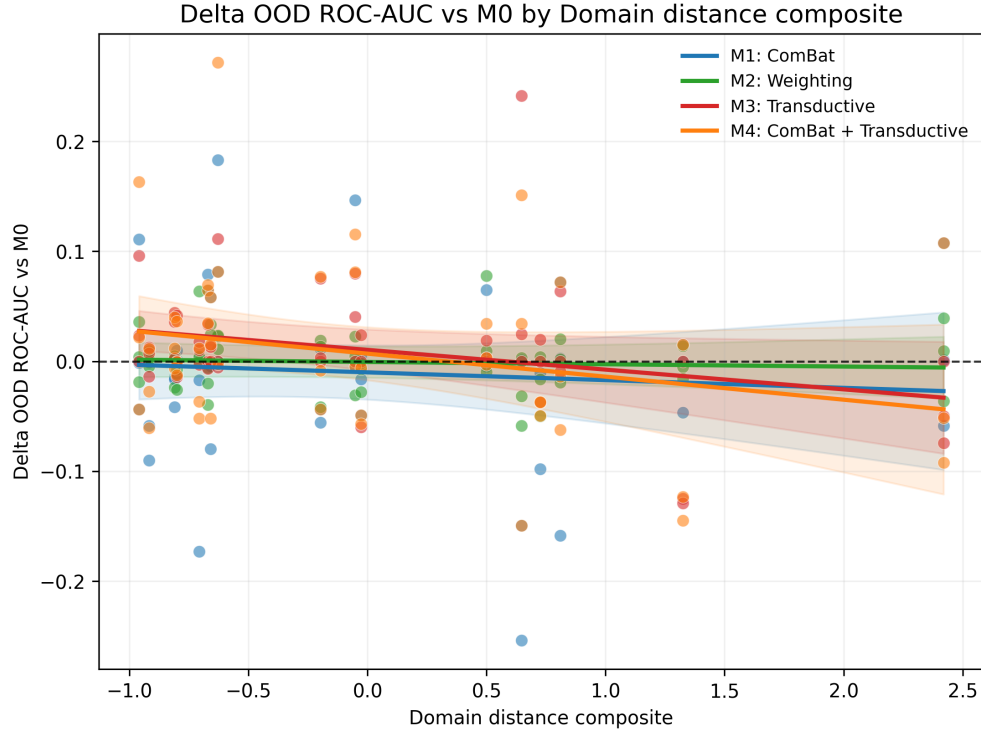


Figure 4.32: Delta OOD ROC-AUC relative to $M0$ as a function of source–target distance in ABIDE II.

Compared with UCI, ABIDE II showed more heterogeneous relationships between the target domain characteristics and adaptation gains. $M3$ remained among the strongest methods across both smaller and larger target domains, suggesting that its usefulness was not restricted to a narrow range of target sample sizes.

The relationship with class entropy was less uniform. At lower entropy values, $M3$ often produced the largest improvements over the baseline, whereas at higher entropy values the differences among methods tended to become smaller and $M2$ was competitive in some target site and model combinations. Source–target distance also appeared to influence the relative benefit of adaptation. Gains were generally more favourable when source and target sites were relatively similar, while larger distances were associated with smaller and more variable improvements. Overall, these results indicate that adaptation can be beneficial in ABIDE II, although its effectiveness depends strongly on the specific target site and machine learning model.

4.5.2.3 TCGA-BRCA

Figures 4.33, 4.34, and 4.35 summarize the same analyses for TCGA-BRCA.

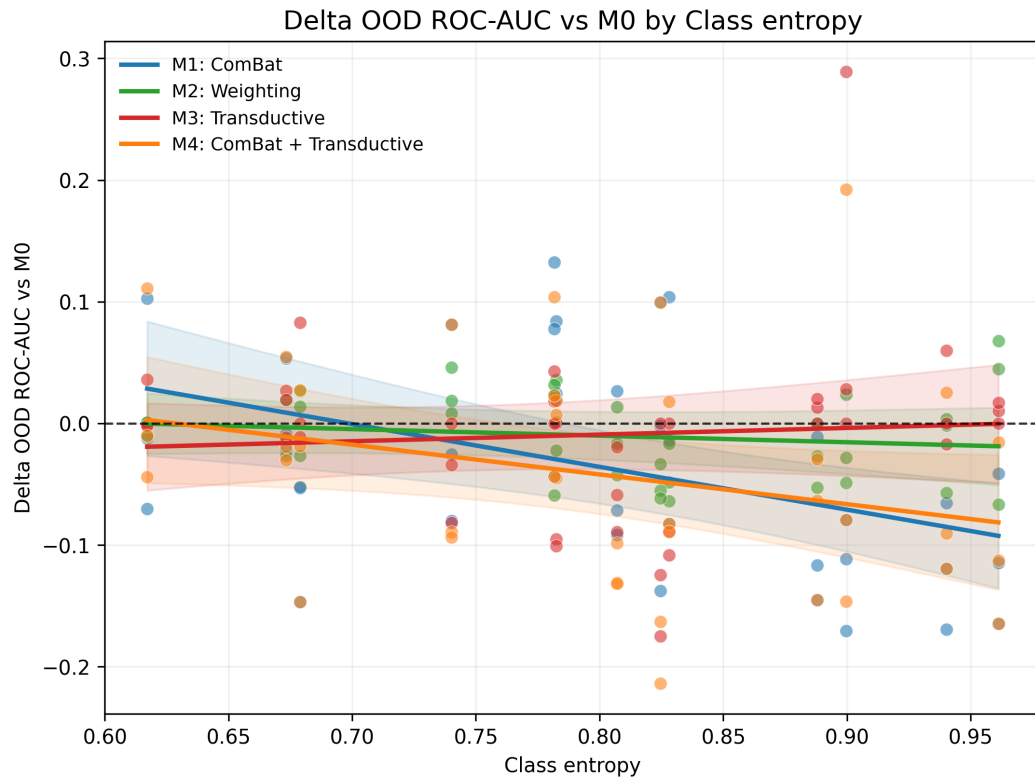


Figure 4.33: Delta OOD ROC-AUC relative to $M0$ as a function of class entropy in TCGA-BRCA.

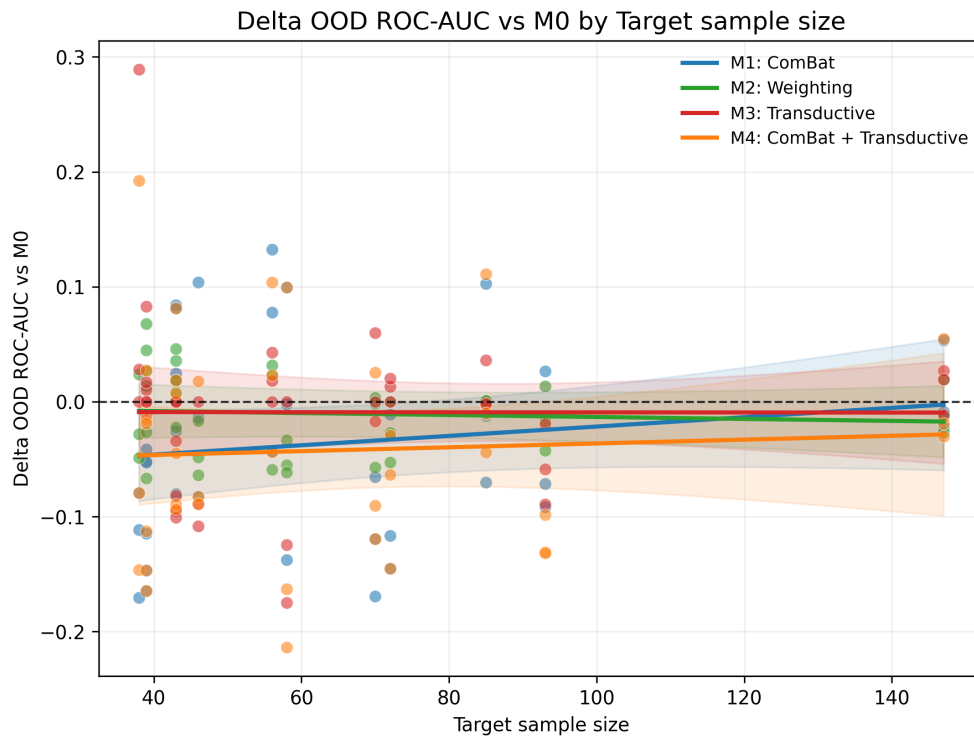


Figure 4.34: Delta OOD ROC-AUC relative to $M0$ as a function of target sample size in TCGA-BRCA.

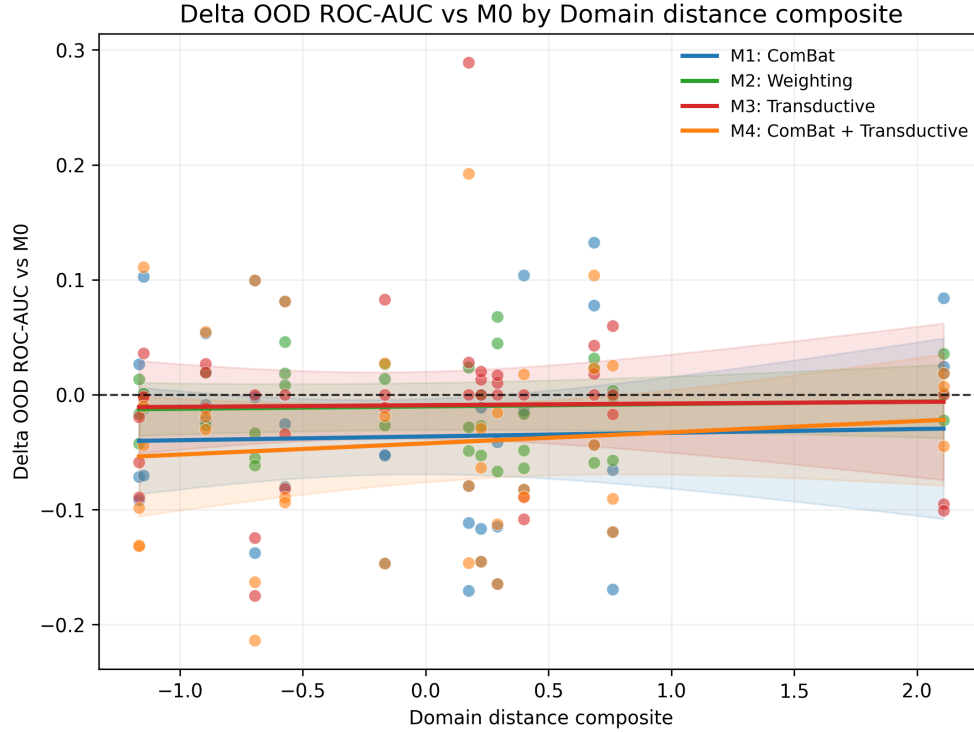


Figure 4.35: Delta OOD ROC-AUC relative to $M0$ as a function of source–target distance in TCGA-BRCA.

TCGA-BRCA presented the most conservative scenario among the three datasets. Across most target domain conditions, $M0$ remained competitive and neither harmonization nor transductive adaptation produced systematic improvements in OOD ROC-AUC. This pattern is consistent with the results reported in the previous sections, where adaptation methods only occasionally outperformed the baseline in terms of discrimination performance.

Class entropy was the meta feature showing the clearest, although still limited, association with adaptation gains. At higher entropy values, $M3$ showed modest positive improvements over $M0$ in several target site and model evaluations. However, these gains were smaller and less consistent than those observed in UCI and ABIDE II. Target sample size and source–target distance did not reveal equally clear trends. The results indicate that, within the TCGA-BRCA setting considered in this study, the evaluated adaptation strategies provided limited discrimination benefits over the source only baseline.

Across the three datasets, class entropy, target sample size, and source–target distance were not associated with adaptation performance in the same way. UCI showed the clearest conditional patterns, with adaptation tending to provide larger gains for larger and more balanced target domains. ABIDE II showed similar but more heterogeneous tendencies, whereas in TCGA-BRCA the baseline remained competitive across most conditions. A recurring descriptive pattern in UCI and ABIDE II was that adaptation gains tended to decrease as source–target distance increased, suggesting that large discrepancies between source and target domains

remain an important obstacle to transportability.

These findings indicate that the effectiveness of adaptation is context dependent. The analyses do not identify a strategy that is consistently optimal across all datasets, target sites, and machine learning models. Instead, the relative benefit of each method appears to depend on the characteristics of the target domain and on the structure of the dataset. This supports the broader conclusion of the thesis that transportability is not an absolute property of a model, but should be evaluated in relation to the specific deployment context.

4.5.3 Source Target Covariate Shift Analysis

To better understand the sources of heterogeneity between domains, an additional analysis focused on metadata shift. The aim was to examine whether differences in demographic, clinical, phenotypic, and technical variables were related to the transportability patterns observed across sites.

Metadata shift was present in all datasets, although its magnitude and meaning changed depending on the dataset. These differences did not fully explain OOD performance, but they helped describe how source and target domains differed beyond the main predictive features.

4.5.3.1 UCI Heart Disease

The metadata analysis for UCI Heart Disease included eleven clinical and demographic variables. The strongest source–target differences involved cholesterol, maximum heart rate, age, ST depression and resting ECG.

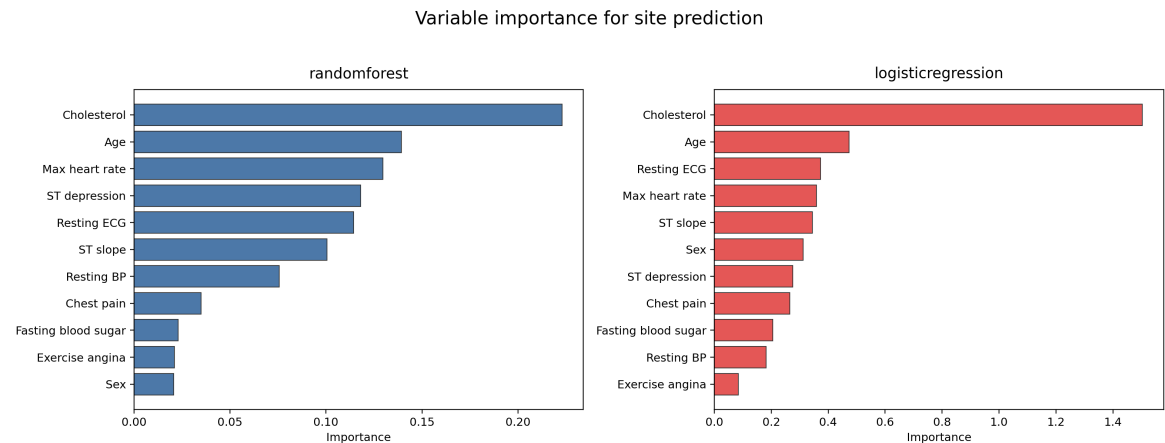


Figure 4.36: Variable importance for site prediction in UCI Heart Disease .

The variable importance analysis (Figure 4.36) further showed that several variables contributing to site heterogeneity were also strongly associated with disease prediction. This

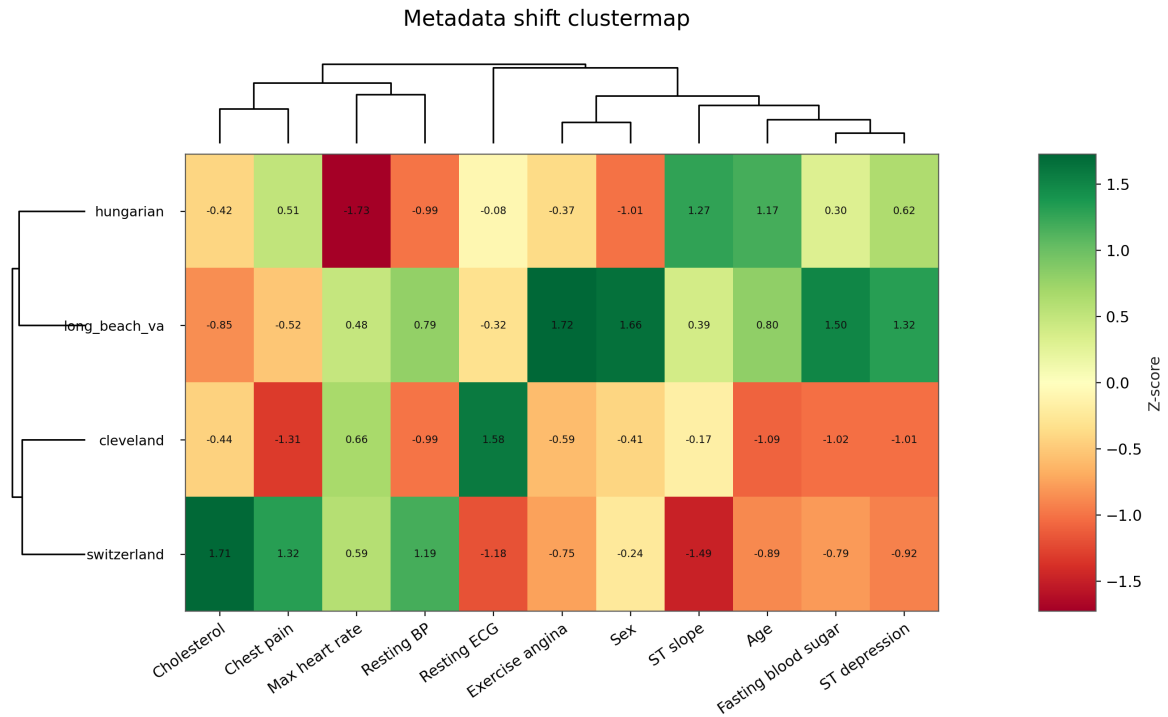


Figure 4.37: Metadata shift clustermap across UCI Heart Disease target sites. Values are reported as standardized Z-scores.

overlap between site related and disease related information was consistent with the strong association observed between site membership and target label ($\chi^2 = 157.38$, $p = 6.73 \times 10^{-34}$, Cramer’s $V = 0.410$).

As shown in Figure 4.37, Long Beach VA, Switzerland and Hungarian exhibited the largest metadata discrepancies relative to the pooled source domains. Importantly, the shifts involved multiple variables simultaneously, suggesting meaningful differences in patient composition across centers.

In terms of transportability, UCI showed the clearest link between metadata shift and OOD performance. Sites with larger metadata differences generally had lower transportability, suggesting that clinical and demographic differences between cohorts contributed to the drop in performance. Among the three datasets, UCI was therefore the case in which metadata heterogeneity appeared most closely related to cross site generalization.

4.5.3.2 ABIDE II

For ABIDE II, the metadata analysis focused on age at scan, eye status during acquisition, full scale IQ, handedness and sex.

Although the overall metadata shift was smaller than in UCI, substantial site dependent heterogeneity remained evident. The largest shifts were observed for ABIDEII-BNI_1,

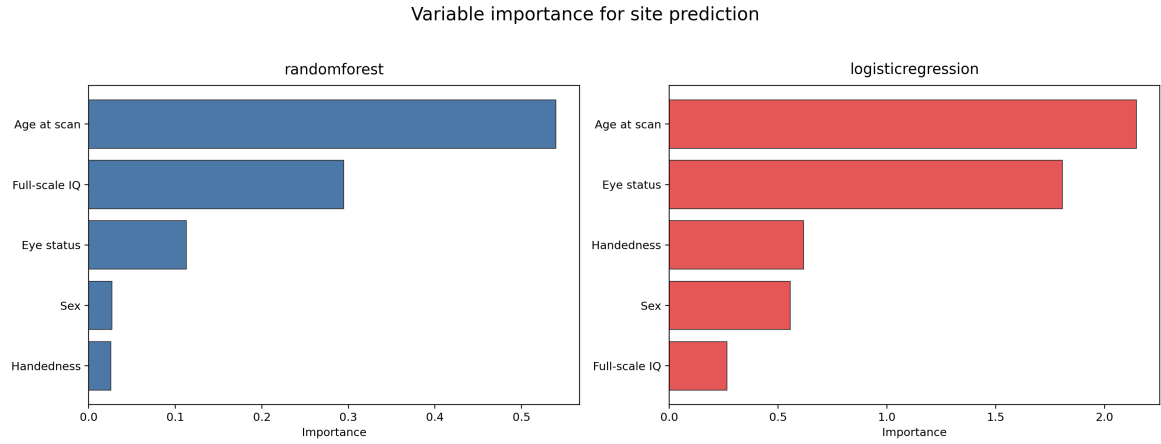


Figure 4.38: Metadata importance for site prediction in ABIDE II.

ABIDEIII-IP_1, and ABIDEIII-EMC_1, while age and eye status emerged as the variables most strongly associated with site membership.

The association between site and target label remained significant ($\chi^2 = 115.13$, $p = 3.46 \times 10^{-16}$, Cramer's $V = 0.295$), but it was weaker than in UCI. This indicates that differences in participant composition contributed to site variability, although they were less directly related to disease prevalence.

Compared with UCI, the link between metadata shift and transportability was more moderate. Metadata heterogeneity appears to be only one part of the broader domain shift affecting ABIDE II. Other factors, such as site specific acquisition protocols and neuroimaging related variability, likely played an important role in shaping the observed transportability patterns.

4.5.3.3 TCGA-BRCA

The TCGA-BRCA metadata analysis included both clinical and technical descriptors related to patient characteristics, sequencing quality and sample processing.

Unlike UCI and ABIDE, metadata heterogeneity in TCGA-BRCA reflected a combination of biological and technical factors. The largest source target differences involved variables related to sequencing quality, signal intensity and clinical composition, with C8, AN and AC showing the strongest metadata shifts.

A notable difference from the previous datasets was the much weaker association between site membership and target label ($\chi^2 = 20.39$, $p = 0.0601$, Cramer's $V = 0.101$). This indicates that site related variability was only weakly connected to the prediction target at a global level.

Nevertheless, some metadata variables remained associated with transportability. In particular, shifts in clinical stage appeared related to reductions in OOD performance, suggesting

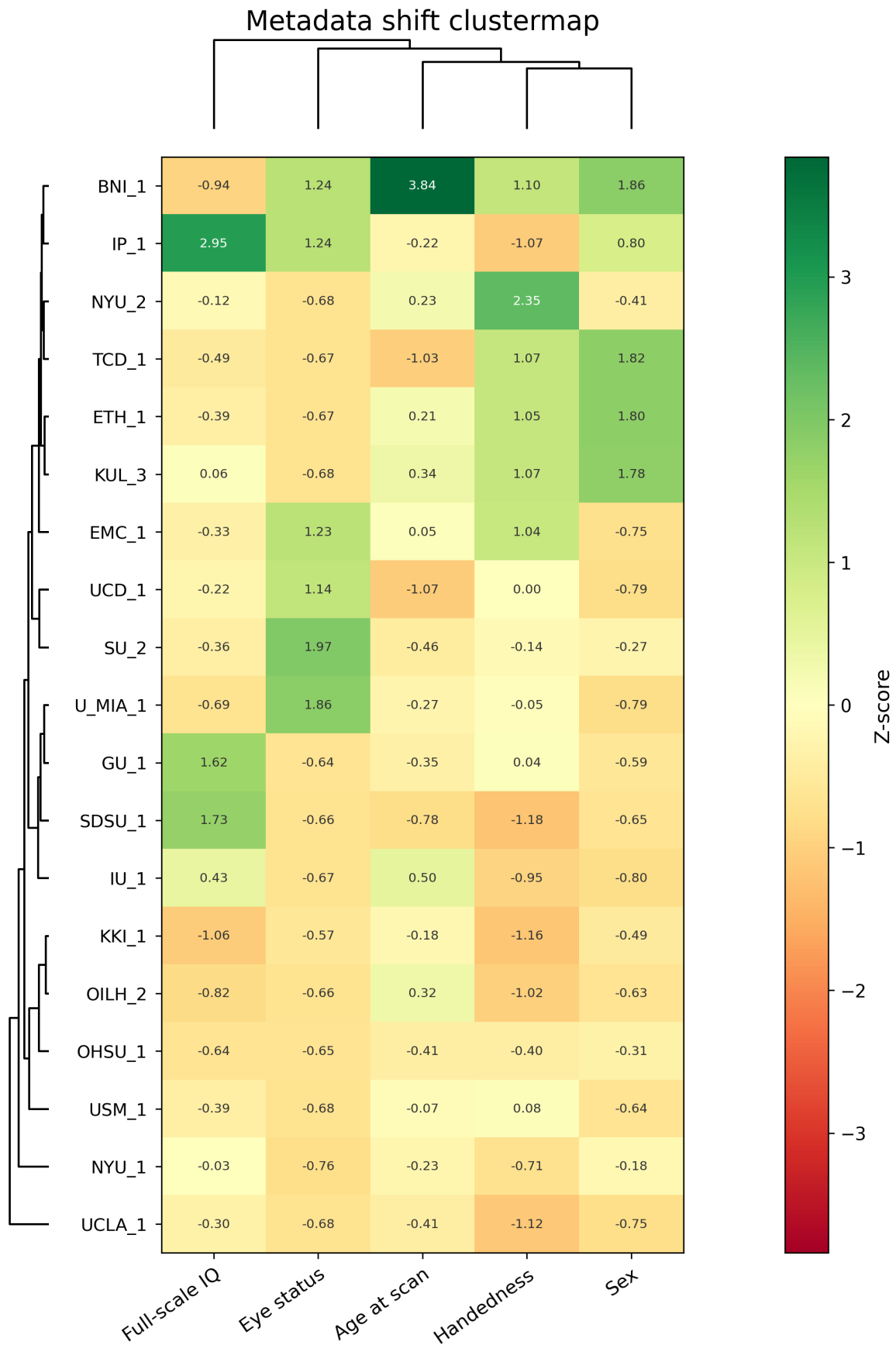


Figure 4.39: Metadata shift clustermap across ABIDE II target sites. Values are reported as standardized Z-scores.

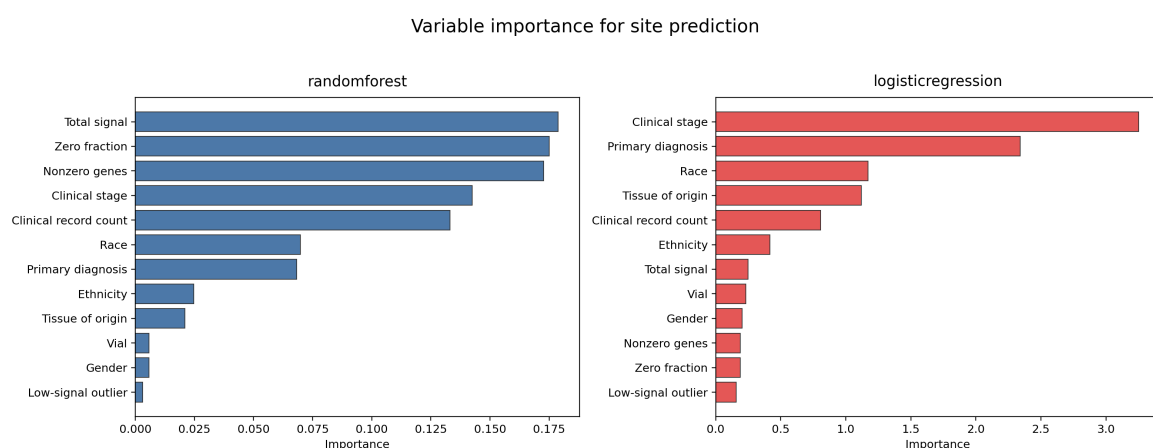


Figure 4.40: Metadata importance for site prediction in TCGA-BRCA.

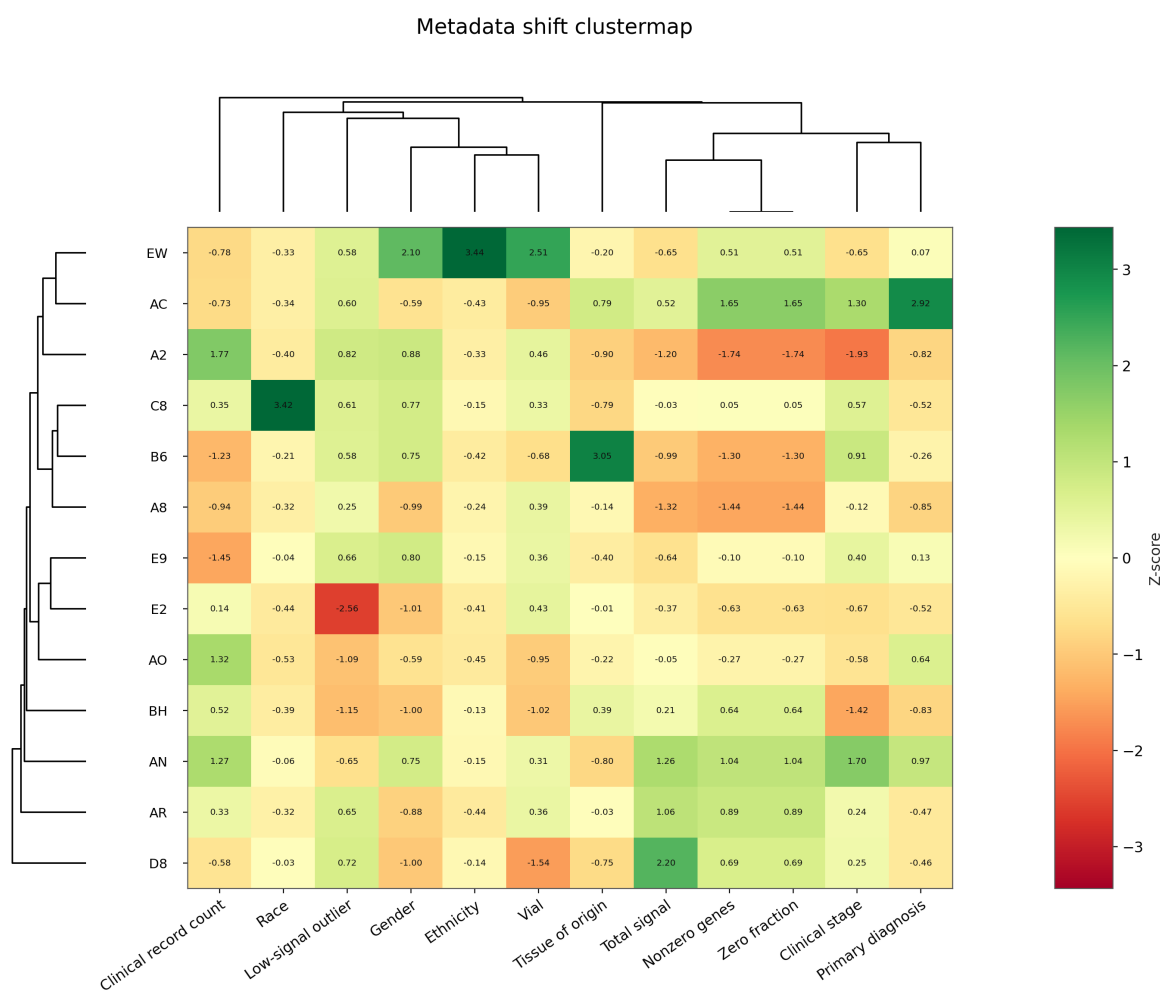


Figure 4.41: Metadata shift clustermap across TCGA-BRCA target sites. Values are reported as standardized Z-scores.

that differences in patient composition can still affect generalization even when overall site label associations are weak. More broadly, the TCGA results highlight that metadata shift in genomic datasets cannot be interpreted solely as demographic or clinical heterogeneity, but also reflects technical variation introduced during sample collection and sequencing.

Across the three datasets, the metadata analyses showed that sites were not simply repeated versions of the same population. Site membership could be predicted from metadata better than chance, confirming the presence of domain heterogeneity. However, the meaning of this shift changed across datasets. In UCI, metadata differences were mainly clinical and were closely related to transportability. In ABIDE II, they reflected a combination of phenotypic and acquisition related variability. In TCGA-BRCA, they resulted from the interaction between clinical characteristics and technical factors.

4.6 Pattern Discovery and Meta learning Analysis

4.6.1 Correlation Heatmaps

To complement the performance results, correlation heatmaps were generated for each dataset. These figures describe the pairwise relationships between site distance measures and aggregated transportability indicators. The objective was to summarize how different measures of source target dissimilarity were related to OOD performance, ID–OOD gap, calibration error, and bootstrap variability.

In these heatmaps, the domain shift descriptors included five source target distance measures: `euclidean_centroid_distance`, `mahalanobis_distance`, `wasserstein_mean_distance`, `jensen_shannon_mean_divergence`, and `mmd_rbf`. The transportability indicators included `mean_ood_performance`, `std_ood_performance`, `id_ood_gap`, and `mean_ece`. Pearson correlation coefficients were computed between all variables included in the matrix.

Group		Variables included in the heatmaps
Site distance descriptors		<code>euclidean_centroid_distance</code> , <code>mahalanobis_distance</code> , <code>wasserstein_mean_distance</code> , <code>jensen_shannon_mean_divergence</code> , <code>mmd_rbf</code>
Transportability	indicators	<code>mean_ood_performance</code> , <code>std_ood_performance</code> , <code>id_ood_gap</code> , <code>mean_ece</code>

Table 4.1: Variables included in the site distance correlation heatmaps.

Each cell in the heatmap represents a Pearson correlation coefficient between two numerical variables. Positive values indicate that two quantities tend to increase together, whereas negative values indicate an inverse relationship. The intensity of the colour reflects the magnitude of the association. These figures do not report performance for individual methods or individual target sites, but provide a compact summary of the relationships between domain distance measures and transportability indicators.

4.6.1.1 UCI Heart Disease

For UCI Heart Disease, the correlation heatmap is shown in Figure 4.42.

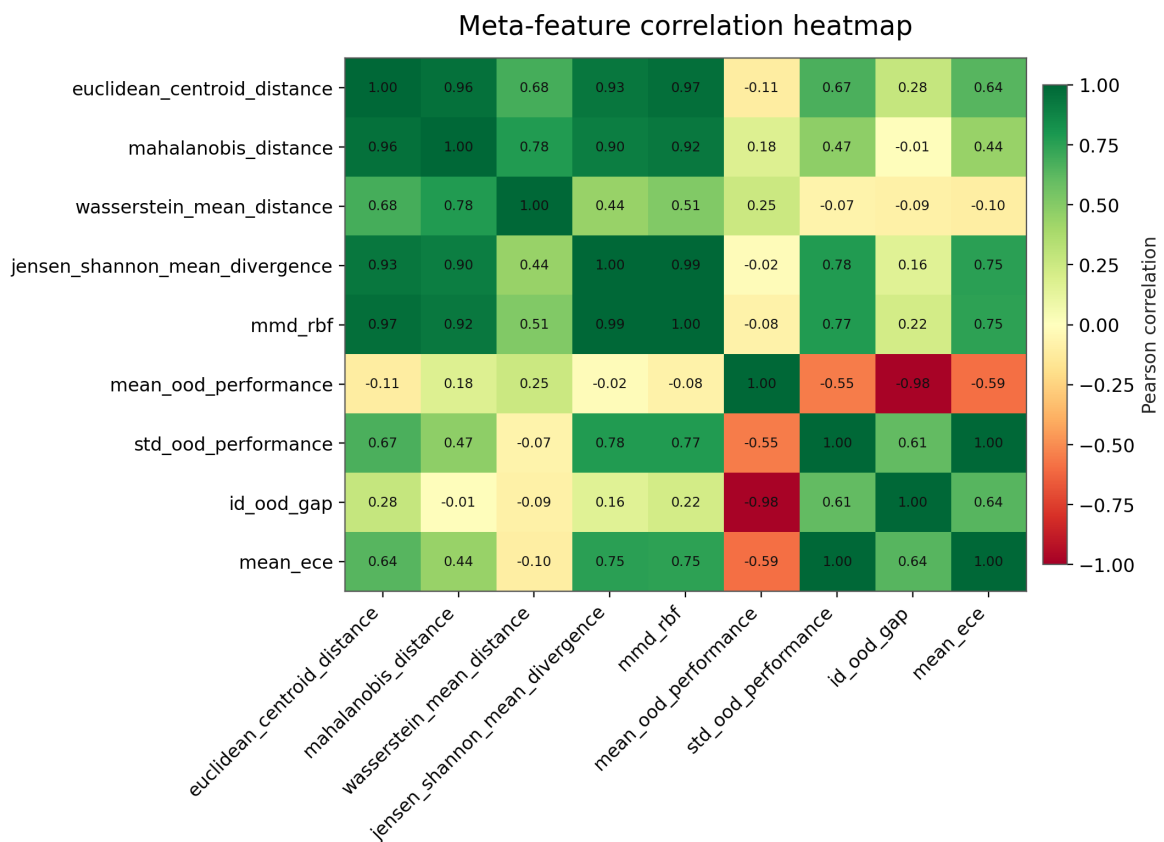


Figure 4.42: Pearson correlation heatmap between site distance descriptors and transportability indicators in UCI Heart Disease.

In UCI Heart Disease, the five site distance descriptors showed strong positive correlations with one another. The heatmap also showed associations between distance based descriptors and the transportability indicators, including OOD performance, ID–OOD gap, calibration error, and bootstrap variability.

4.6.1.2 ABIDE II

For ABIDE II, the corresponding correlation heatmap is reported in Figure 4.43.

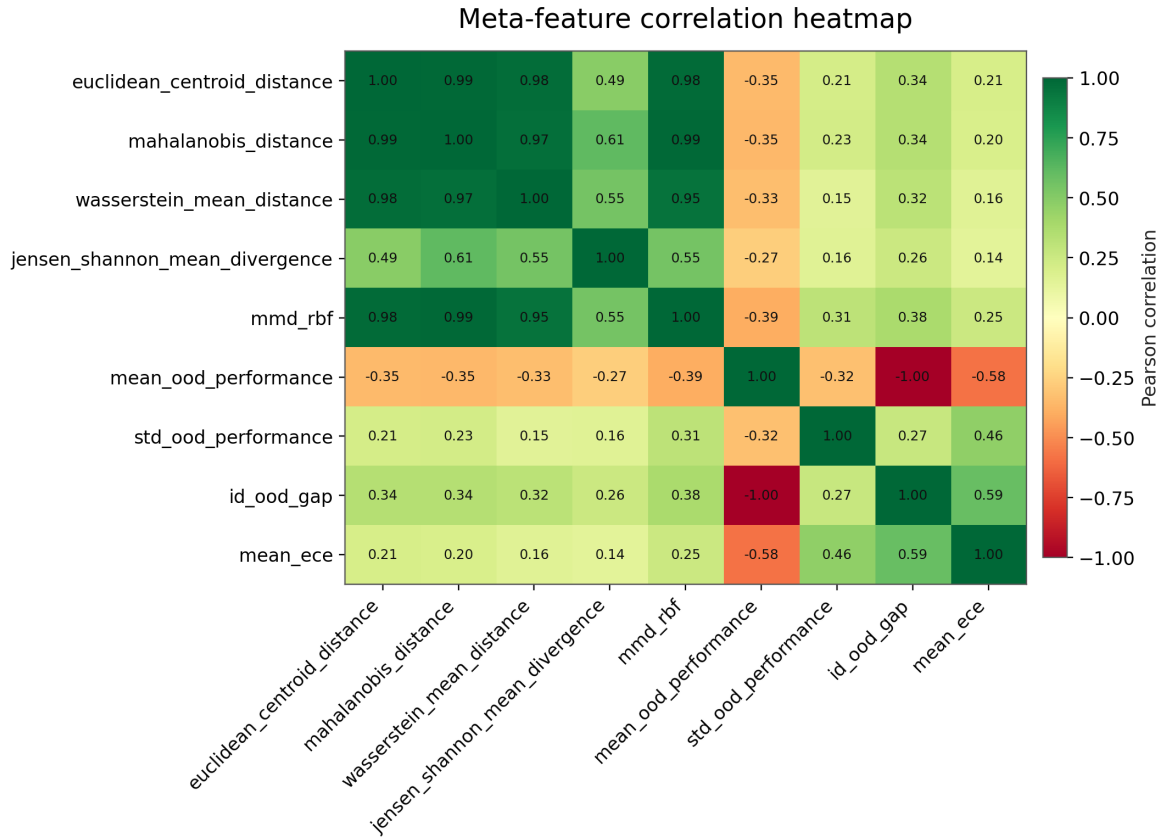


Figure 4.43: Pearson correlation heatmap between site distance descriptors and transportability indicators in ABIDE II.

In ABIDE II, the site distance descriptors were again positively correlated with one another. The correlations between these descriptors and the transportability indicators were weaker and less uniform than in UCI Heart Disease.

4.6.1.3 TCGA-BRCA

For TCGA-BRCA, the correlation heatmap is shown in Figure 4.44.

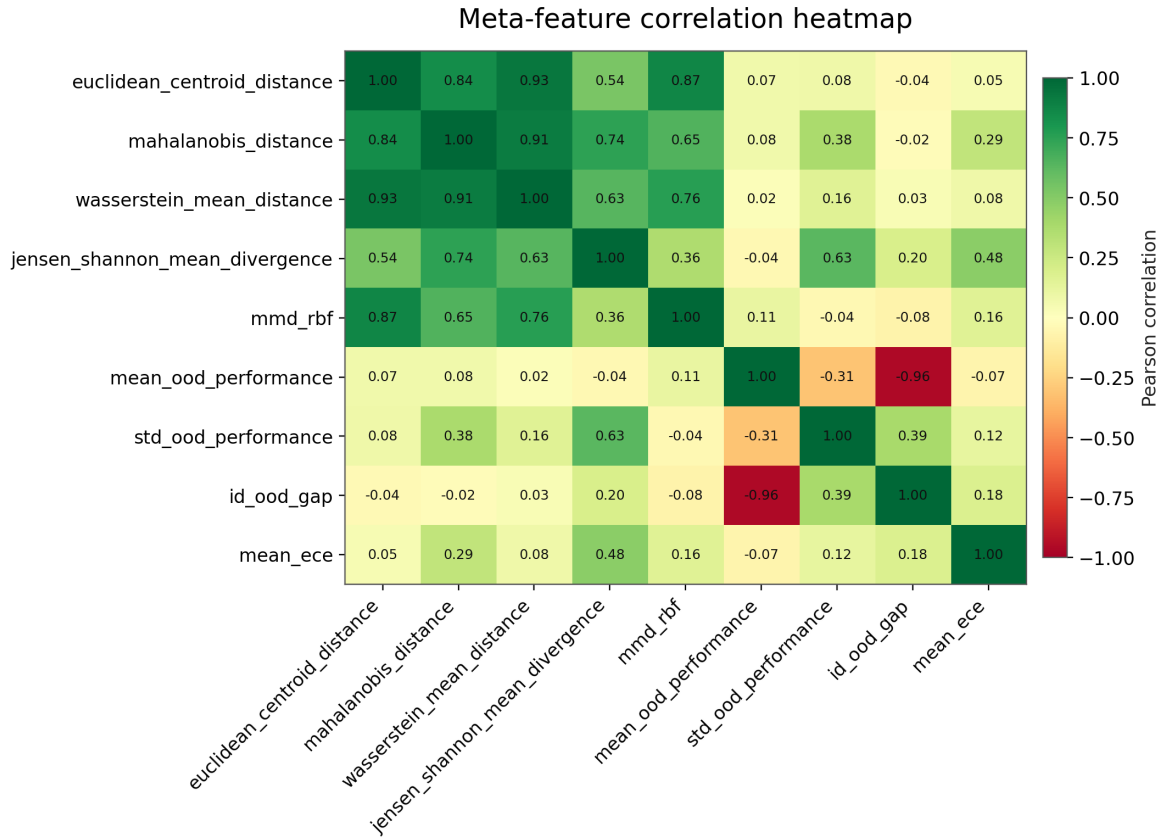


Figure 4.44: Pearson correlation heatmap between site distance descriptors and transportability indicators in TCGA-BRCA.

In TCGA-BRCA, the distance based descriptors also showed positive correlations with one another. Compared with UCI Heart Disease, their correlations with OOD discrimination were less consistent, while visible associations were observed with calibration error and bootstrap variability.

Overall, the correlation heatmaps provided a compact descriptive view of how site distance descriptors and transportability indicators were related in each dataset.

4.6.2 SHAP Meta Learning Analysis

A SHAP based meta learning analysis was performed to examine which domain level metafeatures were associated with the observed OOD ROC-AUC. In this analysis, SHAP was not used to explain individual clinical predictions. Instead, it was applied to meta learning models in which each row represented one experimental configuration, defined by dataset, target site, machine learning model, adaptation strategy, metafeatures, and the corresponding OOD ROC-AUC.

Separate meta learning models were trained for each adaptation strategy from M0 to M4.

This allowed the contribution of the metafeatures to be examined separately for each method. Variables directly derived from model performance, such as OOD metrics, ID metrics, ID–OOD gaps, calibration scores, and related quantities, were excluded to avoid information leakage. The retained predictors described properties of the target domain and of the source target configuration, including sample size, class distribution, covariate shift summaries, and domain distance measures.

For each dataset, the results are reported using three complementary visualizations. The heatmap summarizes the mean absolute SHAP importance for each metafeature and method. The barplot shows the feature importance ranking separately for M0–M4. The beeswarm plot shows the direction of the SHAP contributions, where positive SHAP values increase the predicted OOD ROC-AUC and negative values decrease it.

4.6.2.1 UCI Heart Disease

For UCI Heart Disease, the SHAP results showed a relatively consistent pattern across adaptation strategies. As shown in Figure 4.45, several metafeatures related to covariate shift, source sample size, and target class composition contributed to the prediction of OOD ROC-AUC across multiple methods.

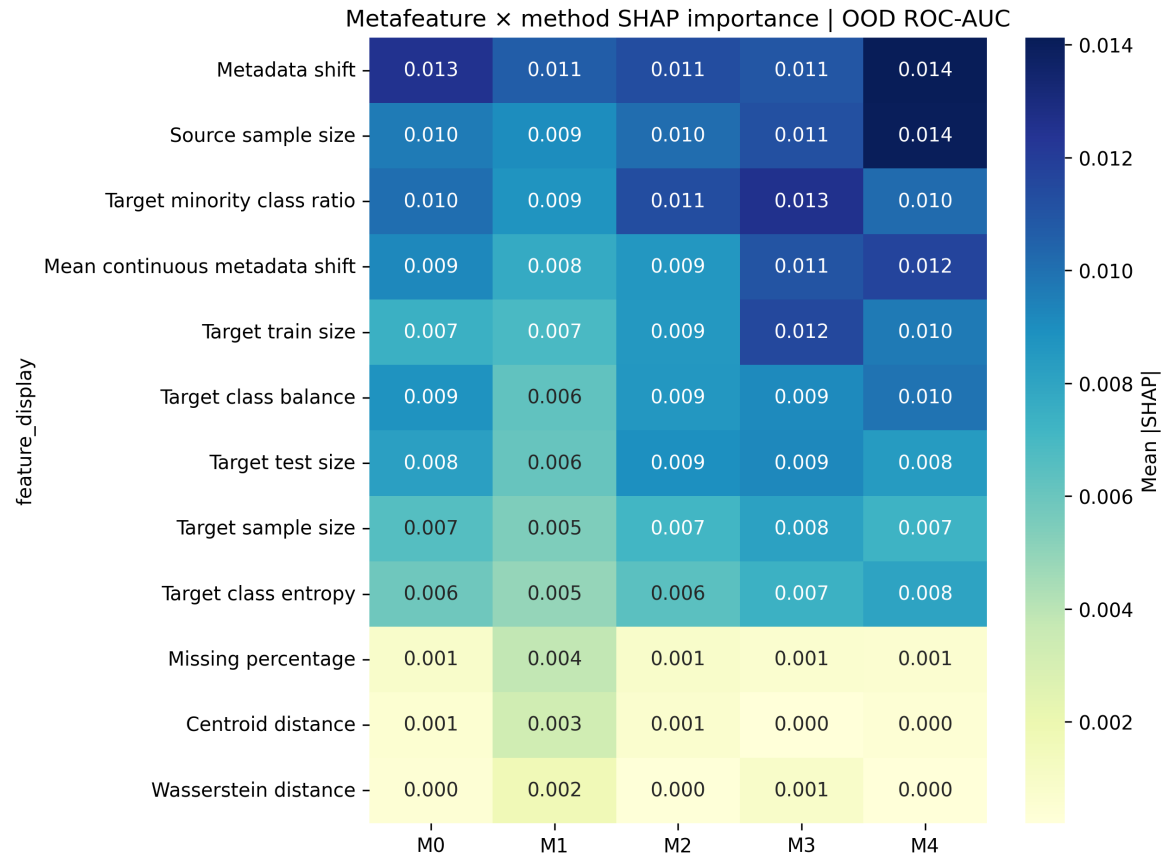


Figure 4.45: Mean absolute SHAP importance for UCI Heart Disease across methods M0–M4.

The method specific SHAP rankings are shown in Figure 4.46. Metadata shift, source sample size, target minority class ratio, mean continuous covariate shift, target train size, target class balance, and target test size appeared among the most relevant metafeatures across several methods. In contrast, individual geometric distance measures, such as centroid distance and Wasserstein distance, showed lower importance.

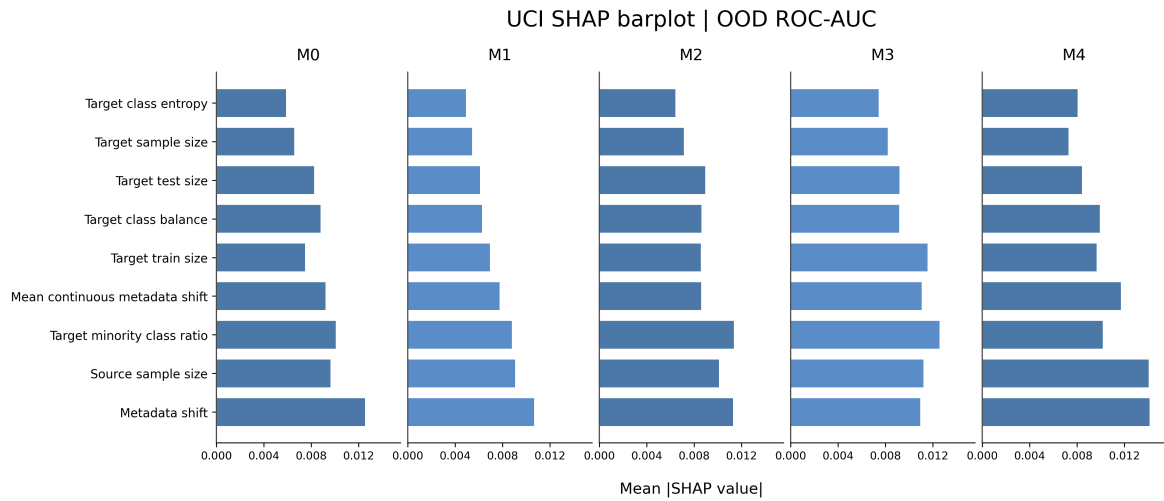


Figure 4.46: SHAP feature importance rankings for OOD ROC-AUC in UCI Heart Disease across methods M0–M4.

The beeswarm plot in Figure 4.47 shows the direction of the SHAP contributions. Higher values of global shift related variables were often associated with negative SHAP contributions, whereas lower values were more frequently associated with positive contributions. Class composition and sample size variables also contributed across methods, although their direction varied according to the specific method and configuration.

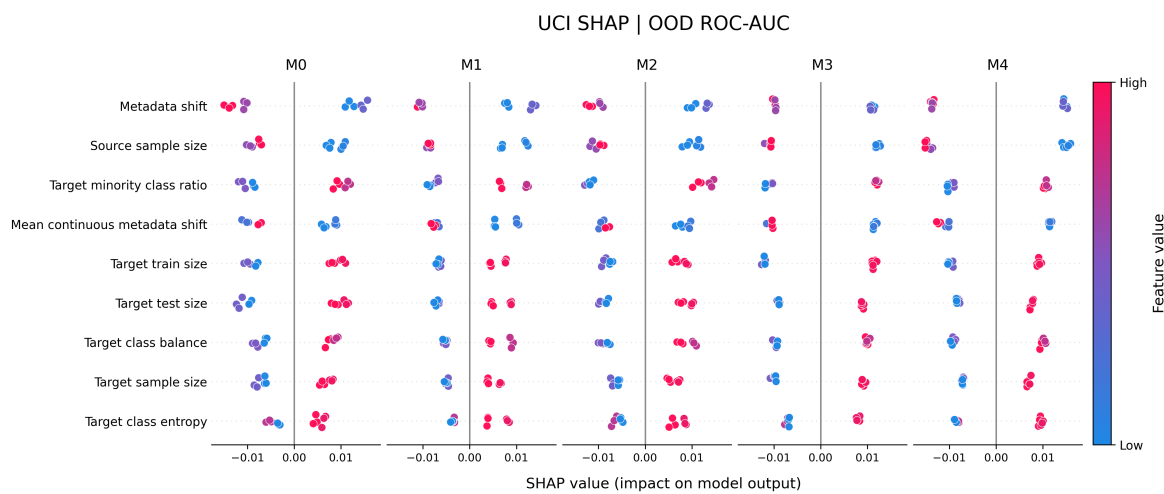


Figure 4.47: SHAP beeswarm plot for OOD ROC-AUC in UCI Heart Disease across methods M0–M4.

4.6.2.2 ABIDE II

In ABIDE II, the SHAP analysis showed greater variability across methods. Figure 4.48 shows that metafeatures related to target class distribution, categorical covariate shift, and domain distance contributed differently depending on the adaptation strategy.

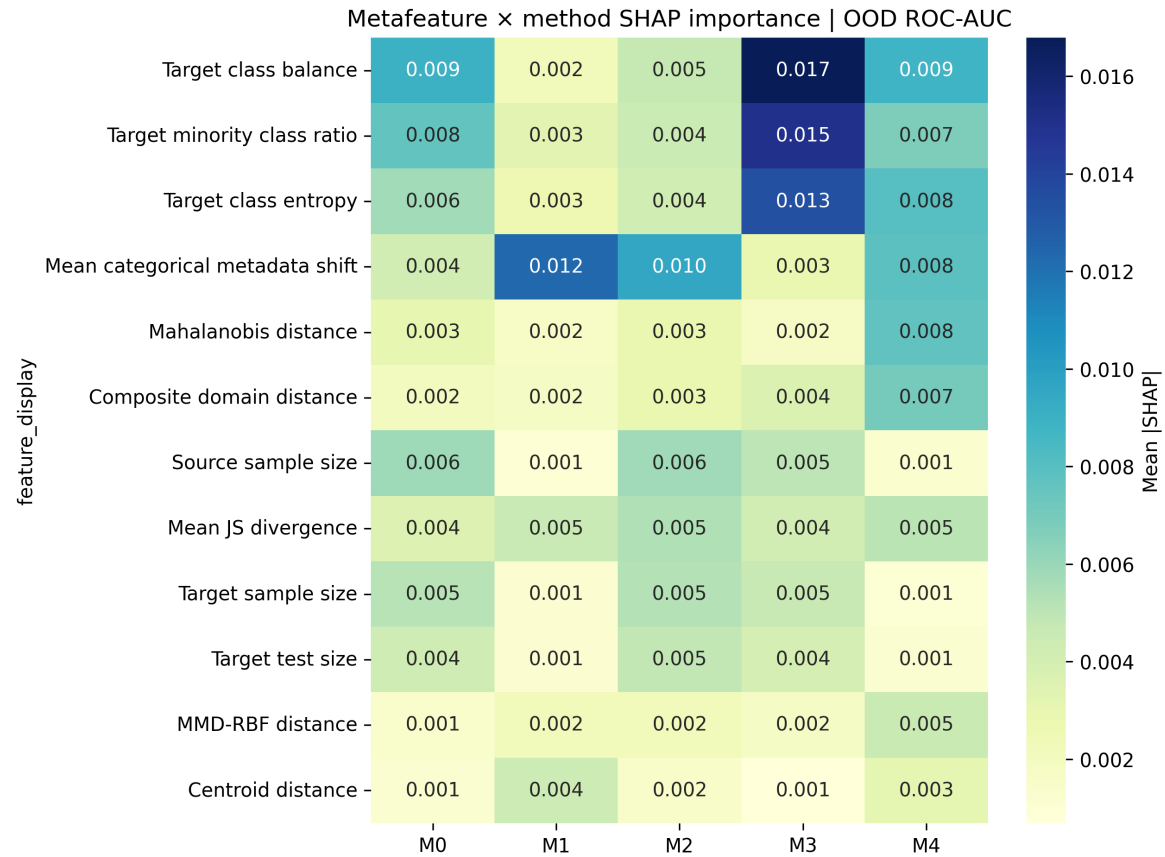


Figure 4.48: Mean absolute SHAP importance for ABIDE II across methods M0–M4.

The method specific rankings in Figure 4.49 show that M0 was mainly associated with target class balance, target minority class ratio, source sample size, and target class entropy. M1 and M2 showed higher importance for mean categorical covariate shift, while M3 was mainly driven by target class balance, target minority class ratio, and target class entropy. M4 showed a more mixed profile, combining class distribution variables, categorical covariate shift, and domain distance measures.

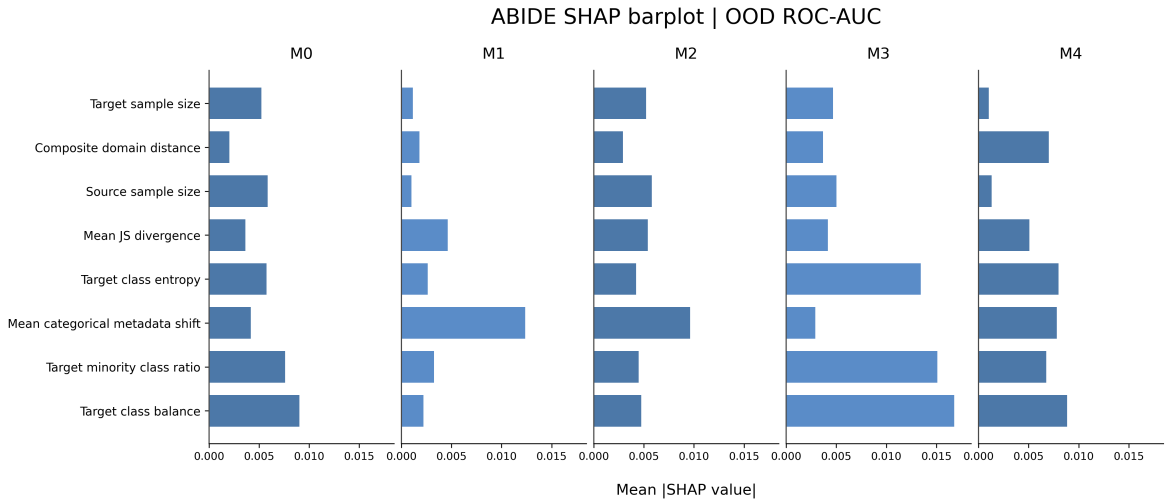


Figure 4.49: SHAP feature importance rankings for OOD ROC-AUC in ABIDE II across methods M0–M4.

The beeswarm plot in Figure 4.50 further shows that the direction of the SHAP contributions differed across methods. For M3, class distribution variables showed the clearest spread of SHAP values, while for M1 and M2 the largest variations were more visible for categorical covariate shift and divergence related variables.

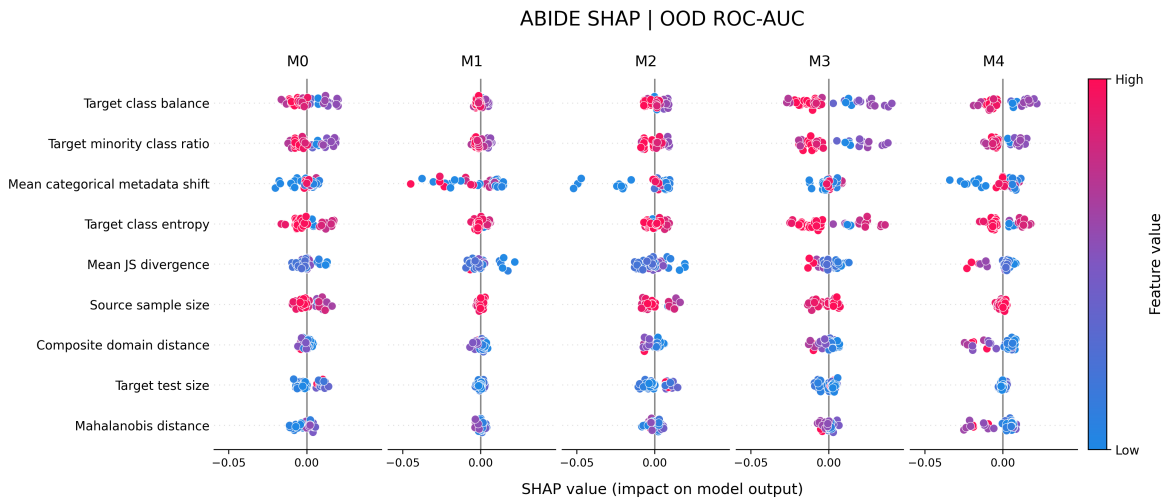


Figure 4.50: SHAP beeswarm plot for OOD ROC-AUC in ABIDE II across methods M0–M4.

4.6.2.3 TCGA-BRCA

For TCGA-BRCA, the SHAP results showed a stronger role of covariate shift and technical quality descriptors. As shown in Figure 4.51, metadata shift, mean continuous covariate shift, and mean categorical covariate shift appeared among the most relevant metafeatures for several methods.

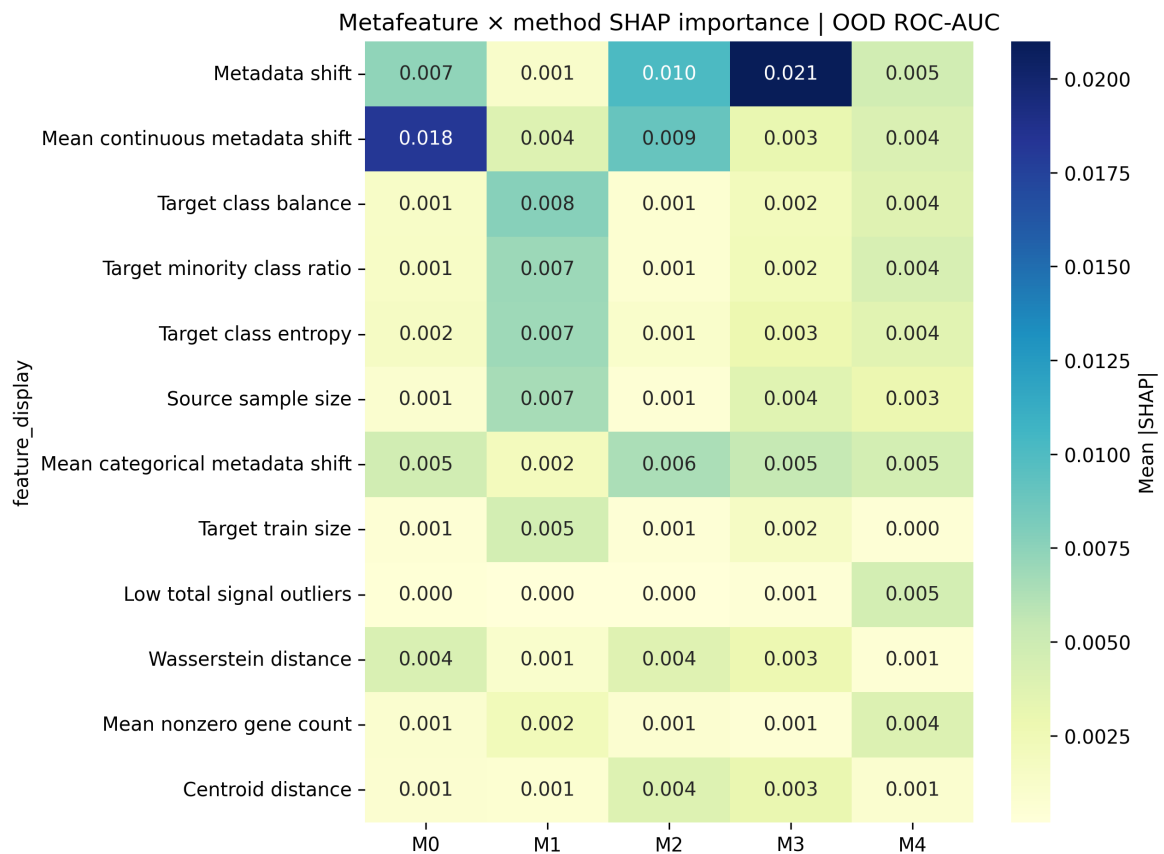


Figure 4.51: Mean absolute SHAP importance for TCGA-BRCA across methods M0–M4.

The method specific rankings in Figure 4.52 show different profiles across the adaptation strategies. M0 was mainly associated with mean continuous covariate shift, whereas M1 showed higher importance for target class balance, target minority class ratio, target class entropy, and source sample size. M2 was again influenced by shift related variables, while M3 showed a dominant contribution of metadata shift. For M4, the ranking was more distributed and included metadata shift, mean categorical covariate shift, low total signal outliers, target minority class ratio, and mean nonzero gene count.

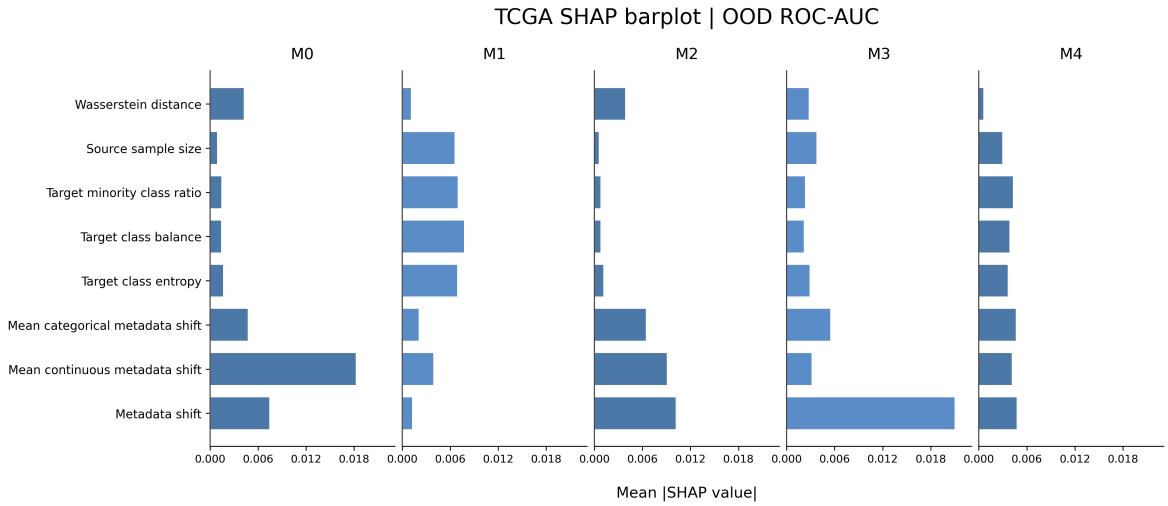


Figure 4.52: SHAP feature importance rankings for OOD ROC-AUC in TCGA-BRCA across methods M0–M4.

The beeswarm plot in Figure 4.53 shows that higher values of metadata shift and mean continuous covariate shift often contributed negatively to the predicted OOD ROC-AUC, especially for M0, M2, and M3. In TCGA-BRCA, some dataset specific technical descriptors, such as low total signal outliers and mean nonzero gene count, also appeared among the relevant variables for selected methods.

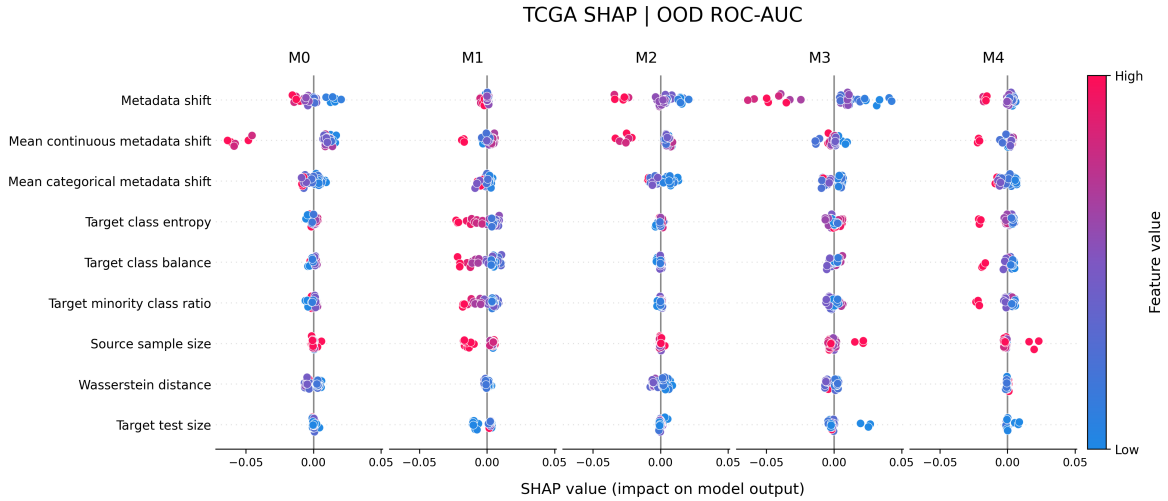


Figure 4.53: SHAP beeswarm plot for OOD ROC-AUC in TCGA-BRCA across methods M0–M4.

Overall, the SHAP meta learning analysis showed that the metafeatures associated with OOD ROC-AUC differed across datasets and adaptation strategies. UCI Heart Disease showed a relatively stable profile across methods, ABIDE II showed the strongest method specific differences, and TCGA-BRCA showed a stronger contribution of covariate shift and technical quality descriptors.

Chapter 5

Discussion

5.1 Overall Interpretation

The results of this thesis show that cross-site transportability is not determined by a single factor. Instead, it depends on the interaction between the dataset, the target domain, the machine learning model, the adaptation strategy, and the evaluation criterion considered. This was visible across all analyses: the same adaptation method could be useful in one dataset, neutral in another, and less effective in a third one.

A first important point is that transportability varied substantially across the three datasets. UCI Heart Disease represented the most favourable setting for cross-site prediction. This may be related to the relatively low dimensionality of the dataset and to the structured nature of its clinical variables, which made the differences between source and target domains easier to characterize and, in some cases, easier to compensate for. ABIDE II showed an intermediate behaviour, with moderate transportability but greater variability across target sites and metrics. In this case, the multisite neuroclinical setting introduced several sources of heterogeneity, including differences in participant characteristics, class composition, age distribution, and missing information. TCGA-BRCA represented the most challenging setting. The high dimensionality of gene expression data, together with biological variability, class imbalance, sparsity, and technical quality differences, likely generated a more complex form of domain shift that was harder to correct with the adaptation strategies considered in this work.

Taken together, these results suggest that transportability should not be interpreted as a fixed property of a model. A model that generalizes well in one biomedical setting may not behave in the same way when the data modality, number of features, site structure, or target population changes. For this reason, high internal validation performance alone is not sufficient to support cross-site deployment. Explicit OOD evaluation remains necessary to determine whether the learned relationships are stable across domains.

5.2 Dataset Specific Interpretation

5.2.1 UCI Heart Disease

UCI Heart Disease was the dataset in which adaptation strategies showed the clearest benefits. The transductive approaches, especially M3 and M4, frequently achieved strong OOD performance, particularly in target sites that were already more favourable in the descriptive and DASHI analyses. At the same time, the baseline M0 also remained competitive in several configurations, indicating that the underlying transportability of this dataset was already relatively high.

One possible explanation is that the UCI dataset combines two favourable characteristics. First, the feature space is relatively small and clinically interpretable. Second, the main differences between sites are visible in terms of class prevalence, missingness, and a limited number of clinical variables. Under these conditions, target aware adaptation can exploit useful information from the target domain without being overwhelmed by excessive dimensionality or highly complex feature interactions. This may explain why M3 and M4 were more effective here than in the other datasets.

UCI results also showed that adaptation was not equally useful for all target sites. Cleveland and Hungarian generally behaved as more favourable domains, whereas Long Beach VA and Switzerland were more problematic. Switzerland, in particular, combined small sample size, strong class imbalance, and larger covariate differences. This suggests that adaptation can provide additional gains when domain differences are moderate and the target sample contains enough useful information, but may become less reliable when the target domain is too small, too imbalanced, or affected by stronger data quality issues.

5.2.2 ABIDE II

ABIDE II showed a more heterogeneous pattern. No adaptation strategy was consistently best across all domains and metrics, and the ranking of the methods changed across target sites. This site dependent behaviour is coherent with the structure of the dataset. Although ABIDE II was less high dimensional than TCGA-BRCA, it included many acquisition sites with different sample sizes, age distributions, missingness patterns, and class compositions.

The repeated competitiveness of M3 suggests that using target domain information can be useful in multisite neuroclinical data. The size of this benefit was highly variable. In some sites, target aware adaptation helped, while in other sites the baseline or domain weighting remained similarly competitive. This indicates that the usefulness of adaptation in ABIDE II depended less on the general complexity of the method and more on whether the method

matched the specific characteristics of the target site.

The SHAP and conditional analyses further support this interpretation. In ABIDE II, target class balance, minority class ratio, and class entropy were among the most relevant drivers for several methods, especially M3. This suggests that the composition of the target population played a central role in determining OOD performance. Therefore, in this dataset, transportability was not only affected by generic source target distance, but also by how the diagnostic groups were represented within each target site.

5.2.3 TCGA-BRCA

TCGA-BRCA most clearly showed that adaptation does not always improve transportability. Although M3 and M4 are more complex methods, the baseline M0 often remained difficult to outperform for OOD discrimination. This result suggests that, in high dimensional genomic data, source target differences may be too complex, sparse, or structurally heterogeneous to be corrected by relatively simple adaptation strategies.

The difficulty of TCGA-BRCA is likely related to several factors acting together. The dataset contains tens of thousands of gene expression features, a relatively small number of samples per site, class imbalance toward early stage tumours, and technical quality control differences across domains. In such a setting, adaptation may introduce additional variability instead of reducing the relevant domain shift. This may explain why the source only baseline often remained robust for discrimination performance.

At the same time, the TCGA-BRCA results also showed that adaptation can still be useful even when it does not improve ROC-AUC or balanced accuracy. Transductive methods often produced better calibration, suggesting that target domain information may help adjust probability estimates even when it does not improve class separation. This distinction is important because, in biomedical applications, predicted probabilities may be used as risk estimates and not only as binary classification outputs.

5.3 Adaptation Strategies and Evaluation Criteria

Across the three datasets, no adaptation strategy was universally superior. This is one of the central findings of the thesis. The relative usefulness of M0–M4 changed according to the dataset, the target site, the classifier, and the metric considered.

The source only baseline M0 remained a strong reference throughout the benchmark. In simpler settings, it was sometimes improved by adaptation, but in more complex scenarios, especially TCGA-BRCA, it often remained one of the most competitive strategies for OOD

discrimination. This result is important because it shows that adaptation should not be assumed to improve performance by default. A well defined baseline remains necessary to determine whether an adaptation method provides a real advantage.

ComBat harmonization generally behaved as a conservative strategy. It did not consistently produce the highest OOD performance, but in several settings it helped reduce the ID–OOD gap or limited performance degradation. This is coherent with the objective of harmonization, which is not necessarily to maximize predictive performance, but to reduce systematic site related variability in the feature space. If the site related variability also contains information associated with the outcome, harmonization may remove useful signal together with unwanted variability.

Domain weighting showed a more context dependent profile. Since the weights were estimated from source to source transferability rather than from the target domain, M2 acted as a source only adaptation strategy. Its behaviour was therefore influenced by how informative the source domains were with respect to each other. In some settings, especially when label distribution characteristics were relevant, M2 remained competitive. Because it did not directly use target domain information, it could not adapt specifically to the structure of the selected target site.

The transductive methods were most useful when the target domain contained information that could guide adaptation. Their advantage was clearest in UCI and remained visible in parts of ABIDE II. However, their performance was not uniformly better in TCGA-BRCA, suggesting that target aware adaptation is not sufficient when the target data are limited, noisy, high dimensional, or affected by complex biological and technical heterogeneity.

A second important point is that discrimination, calibration, stability, and ID–OOD gap did not always provide the same ranking of methods. Some strategies improved OOD ROC-AUC or balanced accuracy without improving calibration. Others produced more reliable probability estimates without improving classification performance. This means that transportability is a multidimensional property. A method can be useful because it maximizes discrimination, because it reduces the gap between internal and external performance, because it improves calibration, or because it remains stable across repeated target splits.

For this reason, the choice of an adaptation strategy should depend on the objective of the application. If the priority is OOD discrimination, the best method may be different from the one preferred when calibration or robustness is more important. In clinical settings where predicted probabilities are interpreted as risk estimates, a slightly lower ROC-AUC may be acceptable if the model provides more reliable probabilities.

5.4 Calibration and Reliability

Calibration was one of the clearest aspects in which target aware adaptation provided a stable benefit. This was especially evident in TCGA-BRCA, where M0 often remained strongest for OOD discrimination, but M3 and M4 showed more favourable calibration behaviour. This indicates that the methods producing the best classification performance were not always the same methods producing the most reliable probability estimates.

This distinction is relevant for biomedical machine learning. In many clinical applications, model outputs are not only used to assign a class label, but also to support decisions based on estimated risk. Poorly calibrated probabilities in this situation can be misleading even when discrimination is acceptable. Evaluating only ROC-AUC or balanced accuracy may provide an incomplete view of model reliability.

The results of this thesis support the idea that calibration should be considered together with discrimination when evaluating transportability. A model may transfer reasonably well in terms of ranking patients, while becoming overconfident or underconfident in the target domain. Conversely, an adaptation strategy may not improve class separation but may still make predicted probabilities more reliable. This was one of the most important findings emerging from the TCGA-BRCA results.

5.5 Transportability

The characterization analyses showed that transportability was associated with several domain level factors, including class distribution, sample size, source target distance, covariate shift, and technical quality descriptors. These factors did not have the same importance in all datasets.

In UCI Heart Disease, covariate shift and sample size related variables were important across several methods. This suggests that performance differences were linked to measurable clinical and demographic differences between sites, as well as to the amount of available data. The relatively structured nature of the dataset made these relationships more visible and easier to interpret.

In ABIDE II, the most relevant factors were more strongly related to the target class distribution. Class balance, minority class ratio, and entropy repeatedly appeared as important drivers, particularly for transductive adaptation. This indicates that, in this dataset, the composition of the target population was a major component of transportability. The fact that site distance alone did not fully explain performance differences also suggests that multiple sources of heterogeneity acted simultaneously.

In TCGA-BRCA, covariate shift and technical quality descriptors played a stronger role. Variables related to continuous covariate differences, global shift, signal quality, and gene expression structure were among the drivers identified by the SHAP meta learning analysis. This is coherent with the nature of genomic data, where technical variability and high dimensionality can strongly affect the stability of predictive models.

The SHAP meta learning analysis therefore supports a condition dependent interpretation of transportability. The objective of this analysis was not to provide causal explanations, but to identify recurring patterns linking target domain characteristics to OOD performance. The results showed that different methods were sensitive to different metafeatures. UCI displayed a relatively stable profile across methods, ABIDE II showed stronger method specific differences, and TCGA-BRCA was more influenced by covariate shift and technical quality descriptors.

These findings taken together suggest that adaptation strategies do not fail or succeed for the same reason in every dataset. When the main difficulty is related to target class composition, methods using target information or weighting may become more useful. When the main difficulty is broader source target mismatch or technical variability, simpler or more conservative methods may remain preferable.

5.6 Implications for Cross-Site Biomedical Machine Learning

The practical implication of these results is that adaptation should not be applied automatically. The presence of domain shift does not necessarily mean that a more complex adaptation strategy will improve performance. In some cases, adaptation can improve OOD discrimination, while in others, it may mainly improve calibration or reduce instability. In more complex settings, the source only baseline may remain the most reliable option.

Therefore, a transportability oriented workflow should include three steps. First, the target domain should be characterized in terms of sample size, class distribution, covariate shift, and source target distance. Second, adaptation strategies should be compared against a strong source only baseline using explicit OOD evaluation. Third, the final choice should consider the objective of the application, including discrimination, calibration, robustness, and stability.

This thesis supports a cautious and context aware approach to biomedical machine learning. Models intended for use across institutions should not be evaluated only under internal validation conditions. They should be tested under explicit OOD settings, compared with simple baselines, and assessed using complementary metrics. From this perspective, transportability

is not something that can be assumed from high validation performance, but an empirical property that must be demonstrated in the deployment context.

5.7 Limitations and Future Directions

Several limitations should be considered when interpreting the results of this thesis. These limitations do not invalidate the main findings, but they define the scope within which the conclusions should be understood.

A first limitation concerns the number and type of datasets considered. The analysis included three biomedical datasets selected to represent different data modalities. This design allowed heterogeneous scenarios to be compared, but it does not cover the full range of biomedical machine learning applications. Other settings, such as medical imaging, radiomics, longitudinal clinical data, wearable sensor data, or multimodal datasets, may present different forms of domain shift and may lead to different adaptation profiles.

A second limitation concerns the number of available target domains. Some analyses, especially those based on covariate shift, site distance, entropy, and SHAP meta learning, depend on the number of sites available in each dataset. In datasets with a limited number of domains, such as UCI Heart Disease, the resulting patterns should be interpreted cautiously. These analyses were designed to provide descriptive and exploratory insight rather than definitive causal explanations.

Another limitation concerns the adaptation strategies evaluated. The methods were selected to represent different families of approaches while many other methods could be explored. The conclusions of this thesis therefore apply to the strategies evaluated here and should not be interpreted as a complete comparison of all possible transportability methods.

The implementation of transductive learning also represents a simplified target aware setting. In this thesis, target adaptation was performed using the target *train_1* subset and probability blending between source and target specific predictions. This design provided a controlled and leakage safe way to use target domain information, but more advanced semi supervised or transductive approaches could be investigated in future work.

The evaluation was restricted to supervised binary classification tasks. This allowed consistent comparison across datasets, but biomedical prediction problems often involve multiclass classification, regression, survival analysis, longitudinal prediction, or risk stratification. Extending the same framework to these tasks would help determine whether the conditional behaviour observed in this thesis also applies to broader biomedical modelling scenarios.

Future work should also investigate method selection before deployment. The results suggest that the characteristics of the target domain can provide useful information on the expected

behavior of adaptation strategies. A natural extension would be to develop meta learning models or decision support rules capable of recommending an adaptation strategy based on measurable properties of the source and target domains. Such approaches would require validation on a larger collection of biomedical datasets, but could represent a step toward more informed and reliable model deployment across sites.

Future research should further examine the relationship between discrimination, calibration, and robustness. An important finding of this thesis was that the methods improving OOD discrimination were not always the same methods improving calibration or reducing the ID–OOD gap. Future studies could therefore combine adaptation strategies with explicit recalibration, uncertainty estimation, and clinical utility analysis, in order to better align transportability evaluation with real world biomedical decision making.

Chapter 6

Conclusions

6.1 Main Findings

The main objective of this thesis was to evaluate whether adaptation and harmonization strategies can improve the cross-site transportability of machine learning models in biomedical datasets. To address this objective, five methodological settings were compared across three datasets representing different biomedical scenarios: UCI Heart Disease, ABIDE II, and TCGA-BRCA.

The main findings of this thesis can be summarized as follows:

1. Cross-site transportability varied substantially across datasets. UCI Heart Disease showed the most favourable OOD behaviour, ABIDE II showed more heterogeneous and site dependent results, and TCGA-BRCA represented the most challenging setting.
2. No adaptation strategy was consistently superior across all datasets, models, target sites, and evaluation metrics. The relative performance of M0–M4 changed according to the dataset and the target domain considered.
3. Transductive approaches, especially M3 and M4, provided the clearest benefits in UCI Heart Disease and remained competitive in ABIDE II, while in TCGA-BRCA the source only baseline M0 often remained difficult to outperform for OOD discrimination.
4. Calibration and discrimination did not always follow the same pattern. In several cases, adaptation improved the reliability of predicted probabilities even when ROC-AUC or balanced accuracy did not improve substantially.
5. Transportability varied according to the characteristics of the target domain. Sites with more imbalanced class distributions, smaller sample sizes, larger train test differences, stronger covariate shifts, or lower technical data quality tended to show poorer predictive performance. These factors therefore helped characterize why some target sites were more difficult to generalize to than others, although their effects were not always consistent across datasets and models.

6. The SHAP meta learning analysis showed that the drivers of OOD performance differed across datasets and methods. UCI showed a relatively stable profile across methods, ABIDE II showed stronger method specific differences, and TCGA-BRCA was more influenced by covariate shift and technical quality descriptors.
7. The results indicate that transportability should be considered a context dependent property. Adaptation can be useful, but its value depends on the dataset, the available target domain information, and the evaluation objective.

6.2 Final Remarks

This thesis investigated cross-site transportability in biomedical machine learning by comparing adaptation and harmonization strategies across three heterogeneous datasets. The results show that good internal performance is not sufficient to guarantee reliable generalization to new sites.

The main conclusion is that transportability is conditional rather than absolute. No adaptation strategy was universally superior across all datasets, models, and evaluation criteria. Instead, the usefulness of each method depended on the dataset, the target domain, and the evaluation objective.

These findings suggest that adaptation should not be applied automatically. In some cases, adaptation improved OOD performance, while in others it mainly improved calibration or stability. In more complex settings, a simpler source only baseline remained competitive.

Overall, this work supports a cautious and context aware approach to biomedical machine learning. Models intended for use across institutions should be evaluated under explicit OOD conditions, compared with strong baselines, and assessed using complementary criteria beyond discrimination alone. Transportability should therefore be treated as an empirical property that must be tested in the specific domain in which the model will be used.

Bibliography

- [1] VA Binson et al. “A review of machine learning algorithms for biomedical applications”. In: *Annals of Biomedical Engineering* 52.5 (2024), pp. 1159–1183.
- [2] Carlos Sáez, Pablo Ferri, and Juan M. García-Gómez. “Resilient Artificial Intelligence in Health: Synthesis and Research Agenda Toward Next-Generation Trustworthy Clinical Decision Support”. In: *Journal of Medical Internet Research* 26 (2024), e50295. DOI: 10.2196/50295. URL: <https://www.jmir.org/2024/1/e50295/>.
- [3] Joaquin Quiñonero-Candela et al., eds. *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press, 2009.
- [4] Carlos Sáez et al. “Potential Limitations in COVID-19 Machine Learning Due to Data Source Variability: A Case Study in the nCov2019 Dataset”. In: *Journal of the American Medical Informatics Association* 28.2 (2021), pp. 360–364. DOI: 10.1093/jamia/ocaa258. URL: <https://doi.org/10.1093/jamia/ocaa258>.
- [5] Gabriel Ferreira dos Santos Silva et al. “Strategies for detecting and mitigating dataset shift in machine learning for health predictions: A systematic review”. In: *Journal of Biomedical Informatics* (2025), p. 104902.
- [6] Rongguang Wang, Pratik Chaudhari, and Christos Davatzikos. “Embracing the Disharmony in Medical Imaging: A Simple and Effective Framework for Domain Adaptation”. In: *Medical Image Analysis* 76 (2022), p. 102309. DOI: 10.1016/j.media.2021.102309.
- [7] Jiashuo Liu et al. “Towards Out-Of-Distribution Generalization: A Survey”. In: *arXiv preprint arXiv:2108.13624* (2021).
- [8] Haoran Zhang et al. “An Empirical Framework for Domain Generalization in Clinical Settings”. In: *Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL)*. ACM, 2021, pp. 279–289. DOI: 10.1145/3450439.3451878.

- [9] Jiashuo Liu et al. *Towards Out-Of-Distribution Generalization: A Survey*. 2023. arXiv: 2108.13624 [cs.LG]. URL: <https://arxiv.org/abs/2108.13624>.
- [10] Jean-Philippe Fortin et al. “Harmonization of multi-site diffusion tensor imaging data”. In: *NeuroImage* 161 (2017), pp. 149–170. DOI: 10.1016/j.neuroimage.2017.08.047.
- [11] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. “Covariate Shift Adaptation by Importance Weighted Cross Validation”. In: *Journal of Machine Learning Research* 8.35 (2007), pp. 985–1005. URL: <http://jmlr.org/papers/v8/sugiyama07a.html>.
- [12] Sinno Jialin Pan and Qiang Yang. “A Survey on Transfer Learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359. DOI: 10.1109/TKDE.2009.191.
- [13] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. “Adjusting batch effects in microarray expression data using empirical Bayes methods”. In: *Biostatistics* 8.1 (Jan. 2007), pp. 118–127. DOI: 10.1093/biostatistics/kxj037.
- [14] Chuan Guo et al. “On Calibration of Modern Neural Networks”. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. Vol. 70. PMLR, 2017, pp. 1321–1330.
- [15] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.
- [16] Andras Janosi et al. *Heart Disease Database*. 1988.
- [17] Carlos Sáez, Manuel Robles, and Juan M. García-Gómez. “Stability Metrics for Multi-Source Biomedical Data Based on Simplicial Projections from Probability Distribution Distances”. In: *Statistical Methods in Medical Research* 26.5 (2017), pp. 2257–2274. DOI: 10.1177/0962280214545122.
- [18] Adriana Di Martino et al. “Enhancing studies of the connectome in autism using the Autism Brain Imaging Data Exchange II”. In: *Scientific Data* 4 (2017), p. 170010.
- [19] Katarzyna Tomczak, Paulina Czerwinska, and Maciej Wiznerowicz. “The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge”. In: *Contemporary Oncology* 19.1A (2015), A68–A77.

- [20] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer, 2009. DOI: 10.1007/978-0-387-84858-7.
- [21] Gareth James et al. *An Introduction to Statistical Learning: With Applications in Python*. Springer Texts in Statistics. Springer, 2023. DOI: 10.1007/978-3-031-38747-0.
- [22] David Fernández-Narro et al. “dashi: A Python Library for Dataset Shift Characterization to Support Trustworthy AI Development and Deployment”. In: *arXiv preprint* (2026). arXiv:2605.31360. URL: <https://arxiv.org/abs/2605.31360>.
- [23] Claude E. Shannon. “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423.
- [24] Haibo He and Edwardo A. Garcia. “Learning from imbalanced data”. In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284.
- [25] José G. Moreno-Torres et al. “A unifying view on dataset shift in classification”. In: *Pattern Recognition* 45.1 (2012), pp. 521–530.
- [26] Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. “Data Clustering: A Review”. In: *ACM Computing Surveys* 31.3 (1999), pp. 264–323. DOI: 10.1145/331499.331504.
- [27] Scott M. Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.
- [28] Joaquin Vanschoren. “Meta-Learning”. In: *Automated Machine Learning* (2019), pp. 35–61. DOI: 10.1007/978-3-030-05318-5_2.