UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

A Domain Adaptation Framework for Harmonized Representation Learning in Medical Datasets

Author:
Alejandro VARA MIRA

Supervisor:
Dr. Oriol PUJOL VILA &
Bárbara LOBATO DELGADO

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

January 17, 2026

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

A Domain Adaptation Framework for Harmonized Representation Learning in Medical Datasets

by Alejandro VARA MIRA

This Master's Thesis addresses the critical challenge of clinical data fragmentation and the prohibitive costs of medical data acquisition by proposing a deep learning architecture for cross-dataset knowledge transfer. While the medical community possesses vast amounts of data, it remains largely trapped in isolated silos characterized by structural heterogeneity and measurement bias. To bridge these gaps, this research introduces a multi-branch neural framework that leverages a large-scale auxiliary dataset, MIMIC-III, to enrich the latent representations of smaller, specialized target datasets.

The methodology centers on a dual-encoding strategy where a shared encoder extracts robust statistical patterns from common clinical attributes across populations, while independent private encoders preserve domain-specific niche variables. Empirical validation in the context of ICU mortality prediction demonstrates that this harmonized representation learning consistently improves Precision-Recall and AUC-ROC metrics. By employing a rigorous methodology upon sequential experiments, the study confirms that these performance gains are statistically significant and directly attributable to the enhanced feature representation, rather than artifacts of stochasticity or overfitting.

Ultimately, this work provides a scalable blueprint for clinical data codification, proving that common attributes can serve as a functional bridge to maximize the utility of existing medical records in data-constrained environments.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Oriol and Barbara, for their unwavering guidance and support. This project was completed during a period of significant personal challenge, and I cannot overstate how much their constant availability and kindness meant to me. Beyond their expertise, they provided a level of patience and encouragement that was essential to my progress. It is truly a rare privilege to work with people who are as exceptional in their professionalism as they are in their character; their presence made this journey not only possible but deeply rewarding.

Finally, I give profound thanks to my family. Your steady belief in me has been the backbone of my academic career. Thank you for standing by me through every hurdle and for constantly reminding me that dedication and hard work are the pillars of a meaningful life. I am incredibly fortunate to have your love as my foundation.

Contents

Chapter 1

Introduction

1.1 Motivation

The development of robust machine learning models in the clinical domain is frequently bottlenecked by the nature of medical data acquisition. Beyond the significant prohibitive costs, data collection often requires navigating complex regulatory landscapes, obtaining Institutional Review Board (IRB) approvals, and securing informed consent from patients. As a result, available medical datasets are often narrowly defined, purpose-built, and limited in scope. For instance, a dataset curated for the study of acute kidney injury in a surgical ICU may not easily generalize to general cardiac care units due to the niche selection of laboratory markers.

Furthermore, even when datasets address the same clinical outcome—such as 1-year mortality—they often suffer from structural heterogeneity such as feature inconsistency and measurement bias. This creates a paradox: while the medical community possesses vast amounts of total data, it remains fragmented in data silos. Given the difficulty and slow pace of new data collection, it is imperative to maximize the utility of existing records.

1.2 Goal

The primary objective of this research is to design, implement, and validate a deep learning architecture for tabular clinical data that enables cross-dataset knowledge transfer. The study aims to determine if a model can leverage the broad patterns found in a large, general-purpose clinical database to enhance the predictive performance on a smaller, more specialized, and structurally different dataset.

1.3 Contributions

The principal contribution of this thesis is the development of a multi-branch neural architecture designed to enhance the predictive power of specialized clinical datasets through shared feature codification. Unlike standard models that treat small datasets in isolation, this work proposes a framework that leverages a large-scale auxiliary dataset to inform the representation of common clinical attributes. By routing common variables through a shared encoder, the model effectively utilizes the statistical richness of a larger population to learn a more robust latent representation of shared traits, which is then transferred to the specialized target domain.

Furthermore, this research demonstrates a successful methodology for integrating datasets with mismatched feature spaces. The architecture allows for the simultaneous processing of domain-specific variables through private encoders, ensuring that the specialized knowledge of the smaller dataset is preserved while its common

attributes are enriched by the knowledge extracted from the larger dataset. This approach proves that common attributes can serve as a bridge for knowledge transfer, allowing the model to achieve superior performance metrics even when the datasets possess divergent internal statistics and varying measurement standards.

Finally, this work contributes an empirical validation of this encoding strategy within the context of ICU mortality prediction. By documenting consistent improvements in AUC-ROC and Precision-Recall, the thesis provides a blueprint for codifying small-scale clinical data using larger, public repositories like MIMIC-III. Furthermore, the implementation of a rigorous n-fold cross-validation and testing framework—structured upon a sequential experimental methodology that established firm baselines before introducing architectural complexity—ensures that the performance improvements are statistically significant and directly attributable to the enhanced feature representation, rather than artifacts of stochasticity or overfitting.

1.4 Structure

This thesis is structured to reflect the progressive development of the research. It first situates the work within the broader context of transfer and representation learning, then introduces the proposed harmonized architecture for clinical tabular data. The experimental methodology and evaluation framework are subsequently presented, followed by a detailed analysis of the results using standard clinical metrics. The thesis concludes with a summary of the main contributions, a discussion of limitations, and directions for future research.

Chapter 2

Background and state of the art

The problem of learning from multiple, non-identical data sources is commonly addressed within the broader field of transfer learning, which seeks to improve performance on a target task by leveraging knowledge acquired from a related source task. A closely related subfield, domain adaptation (DA), focuses on scenarios in which the source and target domains share the same predictive objective but differ in their underlying data distributions. This setting is particularly relevant in clinical machine learning, where datasets collected across institutions often reflect distinct patient populations, measurement protocols, and feature definitions.

The theoretical motivation for learning transferable representations is rooted in the broader field of representation learning. Y. Bengio, Courville, and Vincent, 2013, argued that deep models are effective because they learn hierarchical abstractions that disentangle underlying generative factors of variation. Such abstractions are more likely to generalize across domains, as they capture latent concepts that are invariant to superficial changes in data acquisition or measurement processes. This perspective provides a foundational justification for the use of shared latent representations in domain adaptation settings.

The foundation for using deep learning to align disparate data distributions was significantly advanced by Glorot, Bordes, and Yoshua Bengio, 2011. Although their primary application was natural language processing, they were among the first to demonstrate that stacked denoising autoencoders could be used to extract a shared representation that is robust across different domains. By forcing a model to reconstruct and encode data from multiple sources simultaneously, their approach encouraged the emergence of a harmonized feature space that captures the underlying structure of the data rather than domain-specific noise.

A pivotal advancement in this field was the introduction of Domain Separation Networks (DSN) in Bousmalis et al., 2016. While originally designed for image-to-image translation and object recognition, their core innovation—the partition of latent space into shared and private components—provides the theoretical foundation for this thesis. Bousmalis et al. argued that a shared representation alone might be contaminated by noise specific to the source domain. By explicitly modeling private components that capture domain-specific noise (such as camera angles in images or measurement bias in medical machines), the shared encoder is freed to learn truly domain-invariant features.

This principle is central to the architectural proposal of this thesis. By employing a shared encoder to codify common clinical attributes, we follow the precedent established by Glorot, Bordes, and Yoshua Bengio, 2011, under the assumption that there exists a higher-level physiological representation that remains consistent across institutions and measurement protocols. However, whereas Glorot et al. focused on learning a single unified latent representation, the architecture proposed in this study extends that approach by incorporating private encoders, as suggested by

Domain Separation Networks. This shared–private factorization is particularly critical in medical contexts, where feature mismatch is common due to heterogeneous data collection practices. As emphasized by Bengio et al. (2013), such architectures enable models to capture abstract physiological concepts that are independent of the specific modalities or measurement procedures through which they are observed.

Adapting these representation-learning paradigms—originally developed for images and text—to clinical tabular data introduces additional challenges. Unlike pixels or word embeddings, tabular clinical features are often heterogeneous, sparsely observed, and lack an inherent spatial or sequential structure. Recent work has explored Domain Adversarial Neural Networks (DANN) (Ganin et al., 2016) to align feature distributions across institutions by enforcing domain indistinguishability in the learned representations. While effective in reducing distributional divergence, DANN-based approaches implicitly assume that all domain-specific variation should be eliminated. In contrast, this thesis adheres to the DSN philosophy by explicitly acknowledging that certain information is inherently domain-specific and should be preserved within private representations rather than forcibly aligned. This separation improves both the interpretability and stability of the shared clinical representation.

The transition from traditional clinical risk scores to deep representation learning represents a significant shift in predictive modeling within healthcare. Classical methods for in-hospital mortality prediction, such as the Sequential Organ Failure Assessment (SOFA) score, have long served as standardized benchmarks. Studies such as Liu et al., 2019 demonstrate that combining physiological scoring systems with manually selected features can yield interpretable and reliable predictions. However, such scoring systems are typically designed as "one-size-fits-all" tools, limiting their adaptability to heterogeneous datasets drawn from distinct clinical environments.

While shared–private representation learning has shown considerable success in computer vision and natural language processing, its application to specialized clinical tabular datasets with heterogeneous feature spaces remains underexplored. Most existing transfer learning approaches in healthcare rely on fine-tuning, which presupposes identical feature schemas across source and target datasets. This constraint substantially limits their applicability in real-world clinical settings. This thesis addresses this gap by demonstrating how a DSN-inspired architecture can facilitate effective knowledge transfer even when the "language" (feature space) of medical datasets differs.

Chapter 3

Architectural proposal

The proposed architecture is designed as a multi-input, multi-output deep neural network that facilitates knowledge transfer through a shared latent space. The core of the model consists of three distinct encoding blocks: a shared encoder (E_c) and two independent private encoders (E_p^A and E_p^B), as illustrated in Figure 3.1.

We define two datasets, X^A and X^B, which possess different feature spaces and dimensionalities. Each dataset is partitioned into private and common feature sets:

- X_p^A and X_p^B represent the private (dataset-specific) features.
- X_c^A and X_c^B represent the common features shared between both domains.

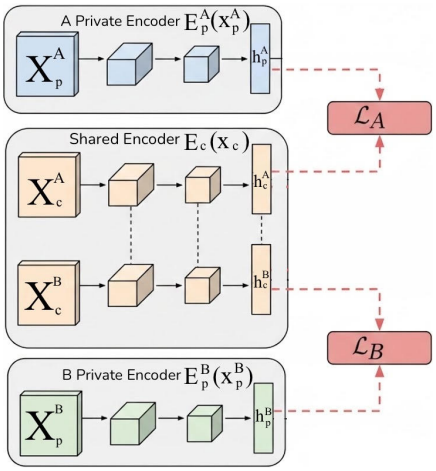


FIGURE 3.1: Overview of the proposed neural network architecture.

Let us consider X^A is significantly smaller than X^B. Our objective is to enhance the prediction capability of the model on X^A by improving the representation of its common attributes through the auxiliary information in X^B.

The shared encoder E_c serves as the primary mechanism for harmonized representation learning. It processes X_c^A and X_c^B to produce the shared latent representations h_c^A and h_c^B respectively. By routing shared variables through common weights, the smaller dataset X^A leverages the broader statistical patterns of the larger dataset X^B to improve the representation of common variables.

The private encoders $E_p^A(x_p^A)$ and $E_p^B(x_p^B)$ are designed to process features unique to each dataset, generating private latent vectors h_p^A and h_p^B. For the specialized dataset X^A, the private encoder ensures that niche variables—which do not exist in the auxiliary dataset—are preserved during the harmonization process.

All encoding blocks are implemented as feed-forward neural networks consisting of fully connected dense layers that progressively decrease in size. This funnel structure acts as a dimensionality reduction mechanism, forcing the encoders to extract the most salient features into a compressed latent space. To ensure robust training, penalize complexity, and reduce the risk of overfitting, the architecture incorporates batch normalization, dropout, and L2 regularization.

The latent representations are then concatenated (e.g., $h_p^A \oplus h_c^A$), followed by a single dense neural layer designed to fuse the shared and private information. The model is optimized against two separate losses simultaneously: $\mathcal{L}_A$ and $\mathcal{L}_B$. Both branches utilize a Softmax activation and Categorical Cross-Entropy with label smoothing.

This multi-task learning environment allows the shared encoder to receive gradients from both paths. The larger auxiliary dataset X^B provides a stable anchor for E_c, while gradients from X^A fine-tune the shared space and the private encoder E_p^A to leverage specific features. This dual-gradient flow is the fundamental mechanism for transferring knowledge from the larger dataset to the smaller one.

Chapter 4

Experimental settings

The complete implementation is publicly available in the following GitHub repository (Vara, 2026).

4.1 Data and Preprocessing

The primary data source for this study is the Medical Information Mart for Intensive Care III (MIMIC-III) database. MIMIC-III is a large, publicly available clinical database containing de-identified health records for more than 40,000 patients admitted to intensive care units between 2001 and 2012. It is widely regarded as a benchmark resource in clinical informatics research due to its high-resolution longitudinal data, including laboratory measurements, bedside monitoring signals, and administrative billing codes.

Due to the high dimensionality and volume of the raw relational data, a dedicated Extract–Transform–Load (ETL) pipeline was implemented to construct a structured tabular dataset suitable for downstream modeling. The data preparation process was extensive and non-trivial; therefore, a detailed description of the pipeline is provided in Appendix A.

4.1.1 Dataset Construction

The relational tables were flattened into a single tabular representation. To ensure statistical independence and prevent data leakage, whereby a model might implicitly learn outcomes from prior hospitalizations, only the first ICU admission for each patient was retained. The cohort was further restricted to patients aged between 1 and 80 years at the time of admission, resulting in a final sample of approximately 31,500 patients. Patient age was computed as the difference between the time of first admission and the recorded date of birth.

Feature Engineering

Clinical features were extracted by selecting the first recorded value of the admission. This ensures the model predicts based on the patient's baseline status upon arrival. The resulting feature space includes multiple categories: demographic and anthropometric variables; vital signs; neurological status; respiratory and metabolic laboratory panels; liver function tests; hematological and coagulation measures; and grouped comorbidity indicators.

Data Cleaning and Sparsity Handling To improve data quality and model efficiency, both feature and patient level filtering strategies were applied. Features exhibiting excessive sparsity (greater than 90% missing values), such as Central Venous

Pressure (CVP) and Pulmonary Artery Pressure (PAP), were removed. Admissions with more than 12 missing clinical variables were excluded from the dataset. Redundant features were also identified and eliminated where values could be reliably inferred from other measurements. For instance, osmolality can be approximated as $Osmolality \approx 2 \times [Na^+] + [Glucose]/18 + [BUN]/2.8$

Target Variable: 1-Year Mortality

The prediction objective for this research is the binary classification of 1-year mortality. This target variable was engineered by calculating the temporal difference between the patient's recorded date of death and their initial hospital admission:

$$\text{Mortality}_{1\text{yr}} = \begin{cases} 1, & \text{if DOD} \leq \text{ADMITTIME} + 365 \text{ days}, \\ 0, & \text{otherwise} \end{cases}$$

The resulting dataset cohort consists of approximately 29000 patients. Within this population, 23000 patients (80%) survived at least one year following their admission, while 6000 patients (20%) died within that one-year window. These figures represent the class distribution for the training and evaluation of the predictive models.

4.1.2 Domain Partitioning

To evaluate the proposed framework (Chapter 3), the full dataset was partitioned into two distinct experimental domains. This division is designed to simulate the heterogeneity of clinical environments, where patient data originates from disparate streams: continuous real-time monitoring (A,bedside vitals) and intermittent biochemical diagnostics (B,laboratory results).

For the experimental setup, the population was segmented into two subsets of equal size, containing approximately 14,500 records per domain. To ensure statistical validity and prevent learning bias, a stratified sampling approach was applied based on the target variable ($\texttt{Mortality}_{1\text{yr}}$). This ensures that both Domain A and Domain B preserve the original class distribution, maintaining the 80/20 survival-to-mortality ratio.

Feature Set Architecture The feature space is organized into three blocks:

- X_c **(Common Attributes):** These 14 features constitute the core of the harmonized representation. They include demographics, anthropometrics, neurological status (GCS), and the 10 Elixhauser comorbidity categories. Being present in both domains, they allow the framework to learn invariant patient characteristics.
- X_p^A **(Monitoring Domain):** Composed of 10 attributes specific to Domain A, focused on vital signs and respiratory parameters such as Heart Rate, Temperature, and SpO_2.
- X_p^B **(Laboratory Domain):** Composed of 16 attributes specific to Domain B, focused on biomarkers such as Creatinine, Electrolytes, and Blood Counts.

Experimental Dataset Feature Configuration The final dataset feature configuration for training and validating our framework are formally defined as follows:

- $X^A = \{X_c \cup X_p^A\}$
- $X^B = \{X_c \cup X_p^B\}$

This structure enables the required multi-input framework, in which dataset-specific private encoders process the unique information associated in each, while shared-weight components harmonize common information into a unified latent representation. The resulting representations are then optimized for one-year mortality prediction in both X^A and X^B, which in turn encourages the learning of more informative common features for the smaller dataset.

4.2 Evaluation Metrics

Due to the 80/20 class imbalance in the dataset, accuracy alone is an insufficient metric for clinical utility. Consequently, performance was evaluated through two primary lenses:

1. **Discrimination Capacity:** Measured via the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC).
2. **Precision-Recall Trade-off:** Measured via the Precision-Recall (PR) curve and the Average Precision (AP), which provides a more rigorous assessment of the model's ability to identify the minority class (mortality).

4.3 Experimental design

4.3.1 Experiment 1: Preliminary Analysis

Firstly, a comprehensive suite of preliminary experiments was conducted. The objective of these trials was to validate the predictive signal across the various engineered feature sets and to confirm that the relative performance hierarchy of the chosen models remained consistent across disparate data configurations. Furthermore, these trials aimed to determine the optimal modeling paradigm, linear against more complex non-linear architectures

The preliminary phase involved testing five distinct data configurations to evaluate how different combinations of common and domain-specific attributes contribute to mortality prediction. For each configuration, three classical models were compared: Logistic Regression (LR), Random Forest (RF), and a Multi-Layer Perceptron (MLP).

The configurations tested were:

- **Common-Only (A or B):** Models trained using only X_c (X_c^A, X_c^B) on the respective subsets to establish a preliminary analysis for demographic and comorbidity data.
- **Full Domain (A or B):** Models trained on the full available feature set for that domain (X^A, X^B).
- **Consolidated Dataset:** A "global" model trained on the entire population (approximately 29k patients) using all available features to identify the maximum possible predictive performance.

Hyperparameter Optimization and Training

Traditional Classifiers To ensure a fair comparison, both traditional classifiers underwent hyperparameter optimization using Grid Search with 5-fold Cross-Validation. This procedure systematically explored predefined parameter grids to identify the optimal configuration for each model. For the Random Forest classifier, we tuned the number of estimators (100 and 200) and the maximum tree depth (10 and 20), while fixing the minimum number of samples per leaf to 4 in order to mitigate overfitting. Logistic Regression was optimized by varying the regularization strength ($C \in \{0.01, 0.1, 1, 10\}$) and considering both $L1$ and $L2$ regularization penalties.

Neural Network The neural network consisted of a multi-layer perceptron (MLP) trained using the Adam optimizer. To promote stable convergence, three learning rates were evaluated ($10^{-2}, 10^{-3}, 10^{-4}$). Given the binary classification setting, the model was trained using categorical cross-entropy loss. The network architecture comprised an input layer followed by three fully connected hidden layers with 256, 128, and 64 neurons respectively, each employing ReLU activation functions. The output layer consisted of two neurons with a softmax activation to model the class probabilities.

4.3.2 Experiment 2: Performance Ceiling

Before evaluating the proposed framework in highly constrained scenarios where one dataset is significantly smaller than the other, it is necessary to establish a performance ceiling for the architecture. This experiment utilizes the full scope of X^A and X^B without the size reductions applied in subsequent sections. By training and testing within the complete statistical distribution of each domain, we identify the maximum predictive accuracy attainable under ideal conditions. This benchmark represents the model's capacity when free from the constraints of data scarcity.

By testing the model on the full datasets first, you confirm that the architecture is capable of learning the clinical patterns when given enough information. This ensures that if performance drops later during the specialized experiments, the cause is identified as the lack of available data rather than a flaw in the model's design or feature selection.

The configurations tested are:

Baseline The first configuration establishes the baseline performance by treating each domain, X^A and X^B, as an isolated entity. In this setup, the datasets are processed through two separate neural network branches (encoders) with no shared weights or information exchange. To ensure a fair comparison, each branch maintains the structural design of our final architecture, consisting of dedicated components for both common (X_c) and private (X_p) attributes. By optimizing these branches, we establish a reference point for predictive accuracy in a scenario where no transfer learning is attempted. This allows us to quantify the inherent difficulty of the mortality prediction task for each dataset in isolation before any cross-domain harmonization occurs.

Private The second configuration, evaluates the predictive signal contained within domain-specific attributes. We isolate the private features, X_p^A and X_p^B, and route them through their respective private encoders without any connection to a shared latent space. Similar to the baseline, these models are optimized independently.

This test identifies the contribution of niche clinical variables to the final prediction, ensuring we understand the value of site-specific data before common attributes are integrated.

Final The final configuration involves the full implementation of the harmonized architecture using both complete datasets. In this configuration, the shared encoder is activated to facilitate cross-domain knowledge transfer and shared feature codification. By comparing the results of this integrated framework against the isolated baseline and private tests, we can quantify the performance gains provided by the shared latent space. This configuration represents the absolute performance ceiling for the proposed framework under ideal data conditions.

These configurations are preserved across all subsequent experiments to ensure methodological consistency.

4.3.3 Experiment 3: Data Scarcity

The research shifts to evaluating the model's robustness under conditions of extreme data scarcity. In this phase, dataset B remains at its full scale, while dataset A is partitioned into forty, stratified by Y, random folds to simulate a specialized, small-scale clinical environment, with 365 rows. For each experimental configuration, ten of these folds are utilized, and the results are averaged to ensure statistical stability, not a "lucky" data split.

While each fold undergoes an internal process of training, validation, and testing within its specific subset—which we define as S_{test}—this small-scale evaluation alone is insufficient to prove generalizability. To rigorously test for overfitting, we introduce an external evaluation phase. For every fold trained, the model is subsequently evaluated against a significantly larger, unseen portion of X^A. This "large test" set is comprised of 50% of the remaining data not utilized in the current fold's training process, totaling approximately 7,000 rows. By comparing the performance on the limited S_{test} against this larger distribution, we can empirically demonstrate that the architecture is learning robust clinical patterns rather than merely memorizing the noise inherent in small-scale samples.

This experiment, and all subsequent in this study, have been replicated by alternating the roles of A and B, the results can be found in Appendix B.

4.3.4 Experiment 4: Data Scarcity and Statistical Shift

This experiment introduces a dual challenge to the architecture by combining extreme data scarcity with intentional shifts in the underlying statistics of the target population. While the previous experiment focused purely on sample size reduction, this configuration modifies the statistical distribution of dataset X^A to represent a more specialized and distinct clinical environment.

To simulate a highly specialized department, the target dataset is filtered based on specific clinical statistics that differ significantly from the general auxiliary distribution found in X^B. This filtering is applied across three distinct test scenarios:

- **Demographic Statistics:** Filtering the dataset by age and weight to represent specific populations, such as geriatric cohorts with obesity, where the statistical relationship between vitals and mortality may differ from the general population. The parameters utilized are age between 55 and 80, and weight between 90 and 250 kilograms.

- **Clinical Statistics by Comorbidity:** Restricting the dataset to patients with specific chronic conditions; one test focuses on cancer and the other on diabetes. This filtering inherently changes the statistical importance of certain features, as the presence of these comorbidities alters the baseline risk profile of the patient.

It is important to note that, for each filtering scenario, we again perform a 10-fold evaluation over the filtered dataset. Consequently, the data scarcity conditions remain present in all cases.

Chapter 5

Experimental results

5.1 Results Experiment 1: Preliminary Analysis

The discriminative power of the preliminary models is visualized through ROC and Precision-Recall curves in Figures 5.1 and 5.2.

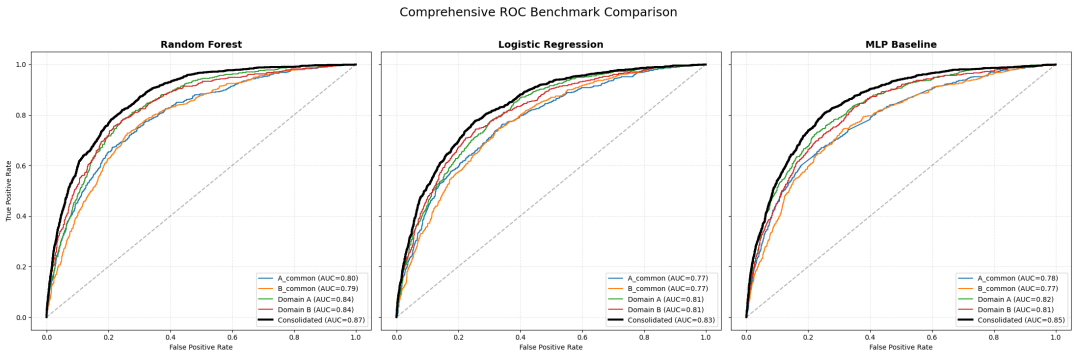


FIGURE 5.1: ROC of Preliminary Analysis:RF, LR, and MLP.

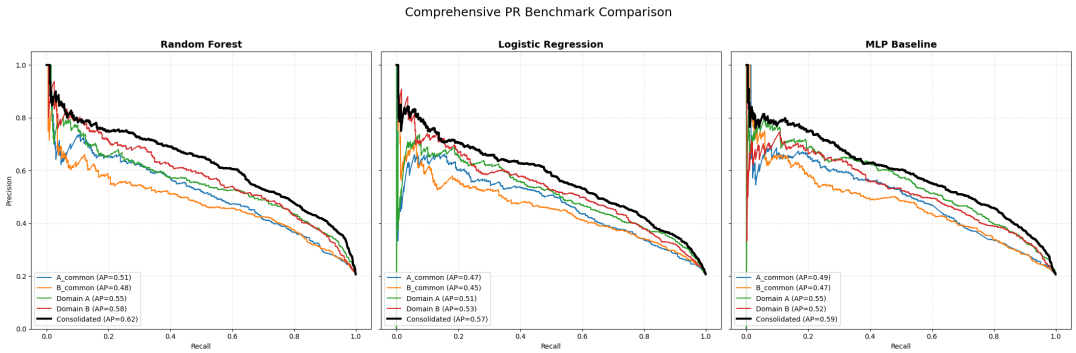


FIGURE 5.2: PR of Preliminary Analysis: RF, LR, and MLP.

This preliminary phase established a rigorous performance foundation for the research:

Architectural Hierarchy: Confirmation over the non-linear nature of clinical tabular data, benefiting from more complex modeling paradigms. Thereby justifying the use of complex neural frameworks as the foundational architecture.

Clinical Feature Synergy: A clear performance gain was observed when moving from models trained solely on demographic and comorbidity data to those incorporating domain-specific clinical sets. Both bedside monitoring vitals (Set A) and laboratory results (Set B) provided significant and complementary predictive signals.

5.2 Results Experiment 2: Performance Ceiling

The experimental results, see Figure 5.3, obtained from the full-scale datasets validate the theoretical framework of the harmonized architecture and stablish its performance ceiling. The findings prove that the proposed framework is not only structurally sound but superior to isolated parallel branches under ideal data conditions. Because the harmonized configuration consistently leads to improved metrics compared to the baseline. This ensures that any performance variations observed in subsequent experiments—where data is intentionally reduced or specialized—can be attributed to the constraints of the data environment rather than architectural flaws.

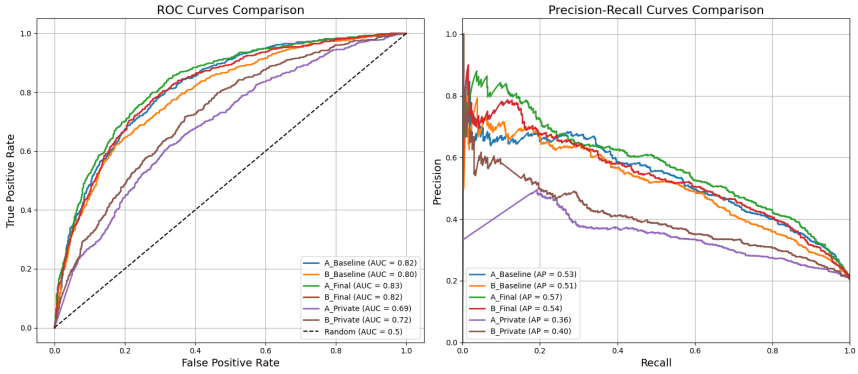


FIGURE 5.3: ROC and PR of Performance Ceiling Experiment.

Also a significant observation is the performance gap between the private-only models and the integrated final models in the precision-recall space. While the private models maintain high general discrimination, they exhibit substantially lower average precision when isolated. This highlights a key finding: although private features are predictive, they require the "anchoring" effect of the common attributes and the shared encoder to maintain high precision. This justifies the bifurcated design, as the shared encoder provides a stable statistical foundation that the private encoders alone cannot achieve.

5.3 Results Experiment 3: Data Scarcity

The results, see Figure 5.4, show that despite the overall reduction in metrics, the harmonized architecture demonstrates significant robustness compared to isolated frameworks.

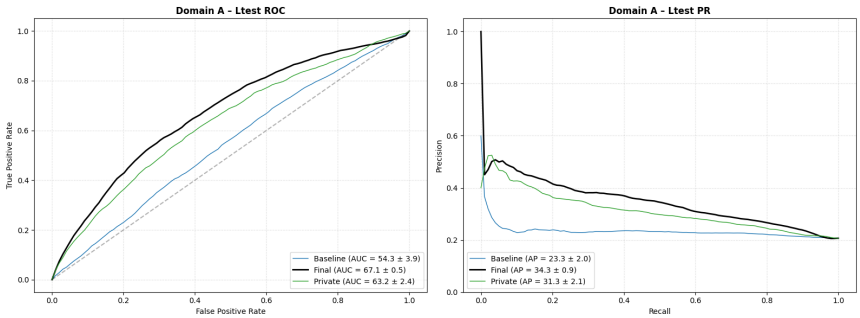
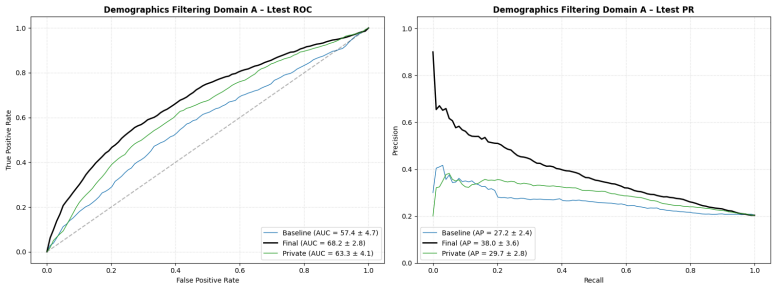


FIGURE 5.4: ROC and PR of Data Scarcity Experiment.

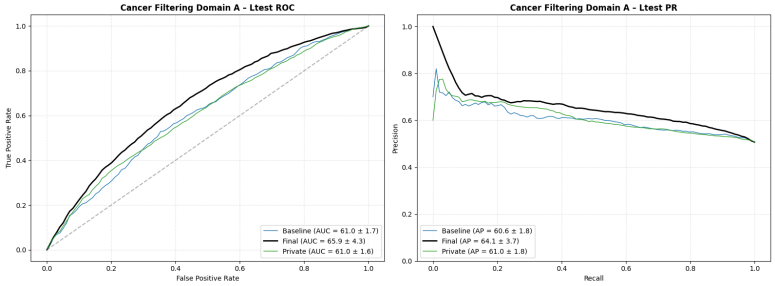
The higher metrics on the larger unseen set empirically demonstrate that the architecture has learned generalizable physiological patterns rather than memorizing the training folds. This validates the effectiveness of the shared-private bifurcated design in extracting universal clinical wisdom to support specialized, small-scale medical domains.

5.4 Results Experiment 4: Data Scarcity and Statistical Shift

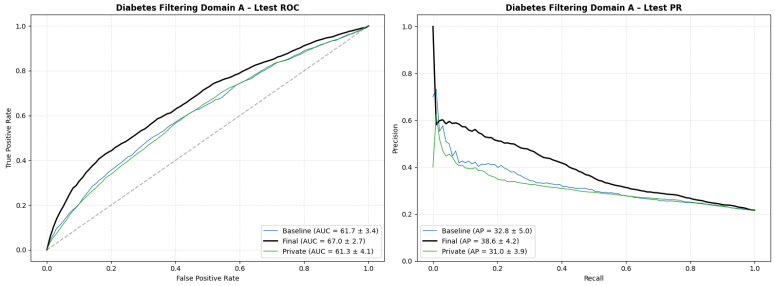
The results presented here in the following Figures 5.5a, 5.5b, and 5.5c demonstrate that the harmonized architecture consistently outperforms isolated frameworks across all specialized scenarios, effectively managing the combined challenges of data scarcity and statistical shift. By successfully adapting to the distinct clinical profiles of geriatric, oncological, and diabetic cohorts, the model proved that the private encoders can capture site-specific nuances while the shared encoder provides a stable, universal foundation. This consistent improvement confirms that the architecture is not only robust against small sample sizes but also capable of generalizing clinical patterns even when the target population's statistics diverge significantly from the general auxiliary data.



(A) ROC and PR of Experiment 4: Demographic Statistics Filtered.



(B) ROC and PR of Experiment 4: Cancer Statistics Filtered.



(C) ROC and PR of Experiment 4: Diabetes Statistics Filtered.

Chapter 6

Conclusion

Reflecting on the motivations presented in Chapter 1, and after following a methodological rigorous approach, the research presented in this thesis successfully demonstrates the effectiveness of a multi-branch neural architecture for harmonized representation learning in medical datasets. By leveraging shared–private encoding representations, the framework addresses a critical paradox of medical data: the existence of vast amounts of information trapped in fragmented and heterogeneous silos. By achieving consistent improvements in *AUC-ROC* and Precision-Recall, the research proves that deep learning models do not need to treat small datasets in isolation. The observed performance improvements are statistically significant and directly attributable to enhanced feature representation, rather than artifacts of stochasticity or overfitting.

- **Overcoming Data Scarcity:** This work shows that the high costs and regulatory hurdles associated with new data collection can be mitigated by maximizing the utility of existing large-scale repositories.
- **Handling Heterogeneity:** The dual-path architecture successfully manages feature inconsistency. The private encoders (E_p) ensure that niche, specialized variables are preserved, while the shared encoder harmonizes and improves the representation of common traits.
- **Statistical Shifts:** Despite differences in underlying statistical distributions, performance improvements were consistently observed.

While this study focused on ICU mortality prediction using tabular data, the modular nature of the architecture allows for future extensions:

- **Architectural Refinement:** Further optimization of the current encoding blocks and fusion layers to enhance the integration of shared and private latent
- **Expansion of Datasets and Cross-Domain Application:** Evaluation of the framework across a broader range of clinical repositories and testing it on other clinical outcomes, such as readmission rates, would further validate the generalizability of the knowledge transfer mechanism.
- **Multimodal Integration:** Future research could incorporate unstructured data, such as clinical notes or high-frequency waveforms, into the private encoding blocks.
- **Temporal Dynamics:** Implementing Recurrent Neural Networks (RNNs) or Transformers within the shared encoder could more effectively capture the time-series nature of bedside monitoring.

Appendix A

Preprocessing MIMIC-III

A.0.1 Data Acquisition and Infrastructure Setup

The raw data was obtained in CSV format. Due to the significant scale of the clinical data—specifically the `CHARTEVENTS` and `LABEVENTS` tables, which contain hundreds of millions of rows—standard ingestion methods were insufficient for local hardware constraints. `CHARTEVENTS` is particularly voluminous as it records every charted observation for a patient during their ICU stay.

- **File Partitioning and Cleaning:** A custom PowerShell script was developed to manage these large-scale files. Using a `StreamReader` and `StreamWriter`, the script split the files into 800MB chunks to stay within local hardware memory limits. Crucially, the script included a cleaning routine to fix structural errors in the raw CSVs, such as unescaped or malformed quotes (`""`) that frequently caused ingestion errors in standard SQL parsers.
- **Database Engine:** A **PostgreSQL** instance was initialized to host the relational schema. This allowed for the execution of complex joins and efficient filtering of the millions of clinical observations via indexed queries.

A.0.2 Relational Schema and Table Functions

After analyzing the database relationships in SchemaSpy, 2017, the following core tables were selected and integrated:

- `PATIENTS`: This is the root table for all individuals in the database. It stores the unique `subject_id`, gender, and date of birth (DOB). It also contains the date of death (DOD), which is essential for our mortality labeling.
- `ADMISSIONS`: Every time a patient is hospitalized, a unique `hadm_id` is generated. This table records the `admittime` and `dischtime`, allowing us to group clinical events by specific hospital stays and determine the chronological order of a patient's medical history.
- `CHARTEVENTS`: This is the largest table in MIMIC-III, containing all charted data for a patient's stay. It includes routine vital signs (heart rate, SpO2, temperature) and bedside assessments like the Glasgow Coma Scale (GCS). Each row links an `itemid` to a specific value and timestamp.
- `LABEVENTS`: This table contains the results of all laboratory measurements (blood, urine, etc.). Unlike `CHARTEVENTS`, these are typically processed in a lab rather than at the bedside. It provides critical biochemical markers such as creatinine, electrolytes, and blood cell counts.

- `D_ITEMS`: This serves as the master dictionary. Since clinical events are stored as numeric IDs, this table is used to map those IDs (e.g., `50818`) to their human-readable labels (e.g., `pCO2`).
- `DIAGNOSES_ICD`: This table contains the ICD-9 billing codes assigned at the end of a stay. These codes were used to reconstruct the patient's medical context through comorbidity indices.

A.0.3 Data Transformation and Dataset Construction

The relational data was flattened into a single tabular view titled `first_admission_data` to prepare it for neural network training.

Cohort Selection and Age Filtering

The study focuses exclusively on the **first admission** of each patient to ensure statistical independence and avoid "data leakage," where a model might learn a patient's future outcome from a previous stay.

- **Inclusion Criteria:** Only patients aged between 1 and 80 years were included, leaving a total of 31.5k patients.
- **Age Calculation:** Age was calculated as the difference between the first `ADMITTIME` and the `DOB`.

Feature Engineering

Clinical features were extracted by selecting the first recorded value of the admission. This ensures the model predicts based on the patient's baseline status upon arrival.. The following categories define the feature space:

Demographics and Anthropometrics These attributes define the patient's basic profile and physical measurements.

Variable	Table	ITEMID(s) or Source Notes
Gender	patients	gender field
Age	admissions	date of birth and admission time (calculate age)
Weight (kg)	chartevents	763, 226512 $\sim$25k measurements

Vital Signs and Neurological Status These variables track immediate physiological and cognitive function.

Variable	Table	ITEMID(s) or Source Notes
Temperature	chartevents	223761 (F$\rightarrow$C), 677
Respiratory Rate	chartevents	618, 220210
Heart Rate	chartevents	211, 220045
NBP Systolic	chartevents	455, 220179
NBP Diastolic	chartevents	456, 220180
GCS (Glasgow Coma Scale)	chartevents	CV: 198; MV: 223900, 223901, 220739

Respiratory and Metabolic Panels These laboratory and bedside markers monitor oxygenation and internal chemical balance.

Variable	Table	ITEMID(s) or Source Notes
SOFA (Respiratory)	chartevents	SpO_2: 646, 220277; FiO_2: 190, 223835
pO_2	labevents	50821-Arterial blood gas panel (∼23k)
pCO_2	labevents	50818
pH	labevents	50820
Sodium	labevents	50983
Potassium	labevents	50971
Calcium	labevents	50893
Glucose	labevents	50931
Creatinine	labevents	50912
BUN	labevents	51006 (Kidney function)
Anion Gap	labevents	50868 (Electrolyte balance)

Liver, Hematology, and Coagulation These markers assess hepatic health and blood profiles.

Variable	Table	ITEMID(s) or Source Notes
Bilirubin	labevents	50885
Albumin	labevents	50862
WBC	labevents	51301
Hemoglobin	labevents	51222
Hematocrit	labevents	51221
Platelet Count	labevents	51265
INR (PT)	labevents	51237 (Coagulation profile)
PT	labevents	51274 (Coagulation profile)
PTT	labevents	51275 (Coagulation profile)

Comorbidity Grouping Chronic conditions were identified from the `diagnoses_icd` table and grouped into the following categories using ICD-9 Elixhauser groupings.

Group Name	Included Comorbidities
Cardiovascular	congestive heart failure, cardiac arrhythmias, valvular disease, pulmonary circulation, peripheral vascular, hypertension
Neurological	paralysis, other neurological
Pulmonary	chronic pulmonary
Diabetes	diabetes uncomplicated, diabetes complicated
Renal	renal failure
Liver	liver disease, peptic ulcer
Cancer	metastatic cancer, solid tumor, lymphoma
Mental Health	psychoses, depression, alcohol, drug
Hematologic	coagulopathy, fluid electrolyte, blood loss anemia, deficiency anemia, obesity, weight loss, hypothyroidism
Autoimmune	rheumatoid arthritis

Data Cleaning and Sparsity Handling Several attributes and patients were excluded to ensure data quality and model efficiency, for example:

- **Feature Pruning:** Features with excessive missingness (over 90% null values), such as Central Venous Pressure (CVP), Pulmonary Artery Pressure (PAP) or Height, were dropped due to low quality or excessive null values.
- **Row Deletion:** Admissions with more than 12 missing clinical attributes were excluded. This ensures that the remaining samples have enough feature density for the domain adaptation model to learn meaningful representations.
- Redundancy is minimized by recognizing that values like Osmolality can be estimated: $Osmolality \approx 2 \times [Na^+] + [Glucose]/18 + [BUN]/2.8$.

Target Variable: 1-Year Mortality

The prediction objective for this research is the binary classification of 1-year mortality. This target variable was engineered by calculating the temporal difference between the patient's recorded date of death (`DOD`) and their initial hospital admission time (`ADMITTIME`):

$$\text{Mortality}_{1\text{yr}} = \begin{cases} 1, & \text{if DOD} \leq \text{ADMITTIME} + 365 \text{ days}, \\ 0, & \text{otherwise} \end{cases}$$

The resulting dataset cohort consists of approximately 29k patients. Within this population, 23k patients (80%) survived at least one year following their admission, while 6k patients (20%) died within that one-year window. These figures represent the class distribution for the training and evaluation of the predictive models.

Appendix B

Supplementary Experimental Results: Dataset B as Target

To further validate the robustness of the proposed framework, Experiments 3 (Section 4.3.3) and 4 (Section 4.3.4) were replicated by alternating the roles of the datasets, specifically utilizing Dataset B as the primary target for knowledge transfer.

The resulting performance metrics and visualization curves are presented below. These results align with the primary findings and confirm the architectural consistency observed in the main study; consequently, no additional commentary is provided beyond the analysis already established in Chapter 5.

B.1 Results Experiment 3: Data Scarcity

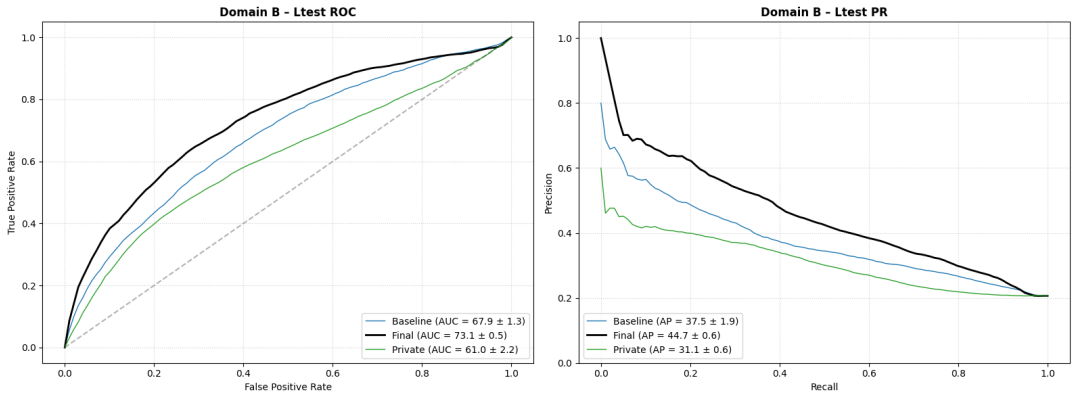
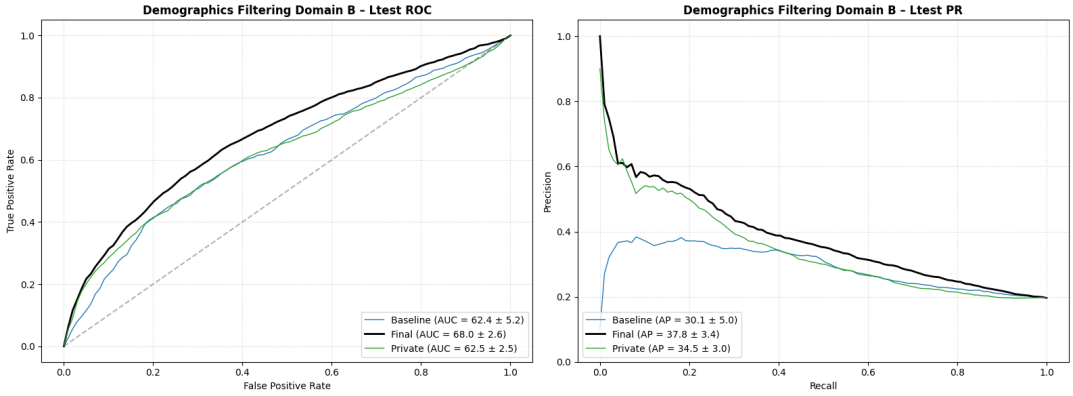
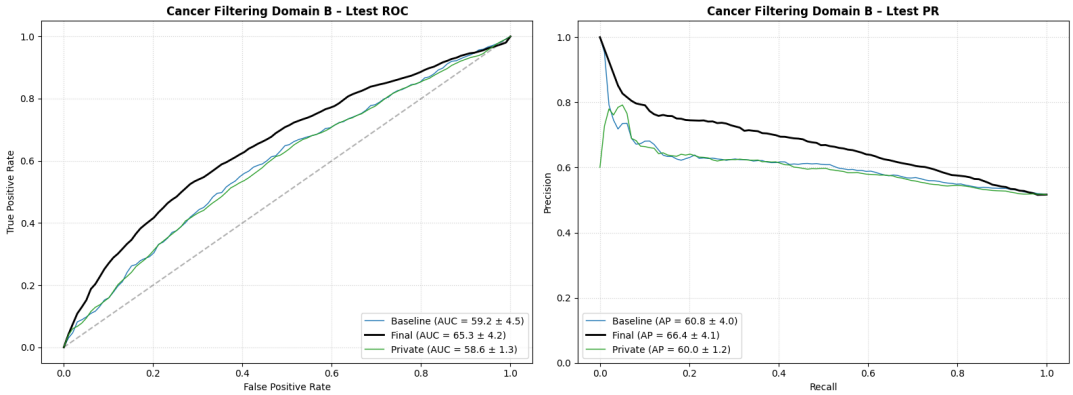


FIGURE B.1: ROC and PR of Data Scarcity Experiment.

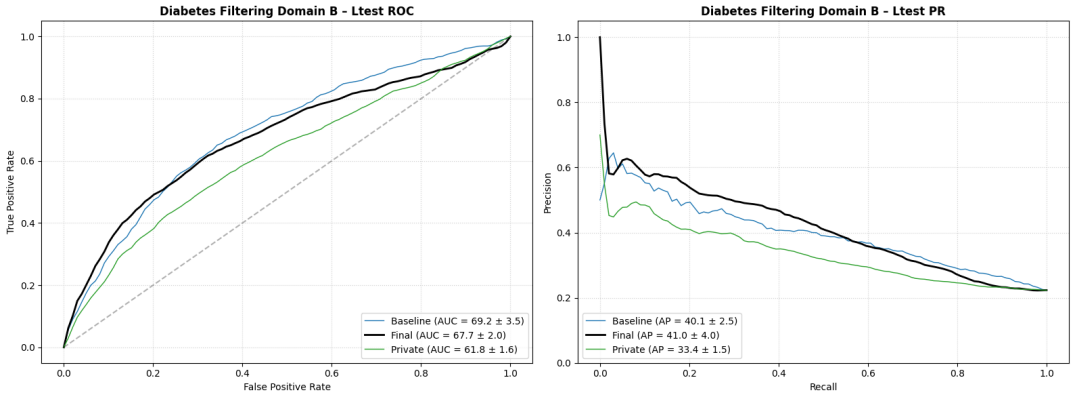
B.2 Results Experiment 4: Data Scarcity and Statistical Shift



(A) ROC and PR of Experiment 4: Demographic Statistics Filtered.



(B) ROC and PR of Experiment 4: Cancer Statistics Filtered.



(C) ROC and PR of Experiment 4: Diabetes Statistics Filtered.

Bibliography

Bengio, Y., Aaron Courville, and Pascal Vincent (2013). "Representation Learning: A Review and New Perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8, pp. 1798–1828. URL: https://arxiv.org/pdf/1206.5538.

Bousmalis, K. et al. (2016). "Domain Separation Networks". In: *Advances in Neural Information Processing Systems*. Vol. 29, pp. 343–351. URL: https://arxiv.org/pdf/1608.06019.

Ganin, Y. et al. (2016). "Domain-Adversarial Training of Neural Networks". In: *Journal of Machine Learning Research* 17.59, pp. 1–35. URL: https://arxiv.org/pdf/1505.07818.

Glorot, X., Antoine Bordes, and Yoshua Bengio (2011). "Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach". In: *Proceedings of the 28th International Conference on Machine Learning*, pp. 513–520. URL: http://link.aip.org/link/?RSI/62/1/1.

Liu, R. et al. (2019). "Predicting in-hospital mortality for MIMIC-III patients: A nomogram combined with SOFA score". In: *Frontiers in Medicine* 6, pp. 1–10. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC9592355/pdf/medi-101-e31251.pdf.

SchemaSpy (2017). *MIMIC Schema Spy: Table Relationships*. https://lcp.mit.edu/mimic-schema-spy/relationships.html. Accessed: January 2026.

Vara, Alejandro (2026). *TFM-Alejandro-Vara: A Domain Adaptation Framework for Harmonized Representation Learning in Medical Datasets*. URL: https://github.com/AlejandroVara/TFM-Alejandro-Vara.