Problem Formulation and Theoretical Framework: Cross-Hospital Adaptation in Small-Sample Tabular Medical Regimes

1. Introduction: The Stochastic Nature of Clinical Deployment

The central challenge addressed in this dissertation is the fragility of predictive modeling when transported across the rigid boundaries of healthcare institutions. While contemporary machine learning has achieved parity with human experts in specific, closed-world diagnostic tasks, these systems frequently suffer catastrophic performance degradation when deployed in environments that differ even slightly from their training distributions. This phenomenon is particularly acute in the domain of tabular medical data—structured clinical records comprising demographics, laboratory values, and radiomic features—where the interplay of **sample scarcity**, **distributional shift**, and **feature heterogeneity** creates a hostile landscape for standard statistical learning.

The problem formulation presented here is motivated by the specific clinical use case of pulmonary nodule malignancy prediction. A model trained on a curated cohort from a tertiary academic center (Hospital A) must be deployed to a community screening site or an external hospital (Hospital B) to assist in early cancer detection. However, privacy regulations such as HIPAA and GDPR creates a "blind" deployment scenario: the target domain data is unlabeled, and patient records cannot be centralized. This necessitates an **Unsupervised Domain Adaptation (UDA)** framework capable of aligning distributions without direct supervision in the target domain.[1]

This chapter rigorously formulates the mathematical and theoretical landscape of this problem. We move beyond the simplifying assumptions of Independent and Identically Distributed (i.i.d.) data to model the generative processes of clinical shifts. We analyze the specific failure modes of Gradient Boosted Decision Trees (GBDTs) and Deep Neural Networks

(DNNs) in this regime, and we formally derive the PANDA (Pretrained Adaptation Network with Domain Alignment) framework as a principled solution that integrates Tabular Foundation Models (TabPFN) with kernel-based statistical alignment (Transfer Component Analysis) and stability-driven feature selection (Cross-Domain RFE).

2. Formal Notation and Problem Setup

To establish a precise vocabulary for the analysis, we define the following notation governing domains, tasks, and distributions.

2.1 Domain and Task Definitions

Let $\mathcal{X}$ denote the input feature space and $\mathcal{Y}$ denote the label space. For the binary classification task of pulmonary nodule malignancy, $\mathcal{Y} = \{0, 1\}$, where $y=1$ represents a malignant nodule (cancer) and $y=0$ represents a benign nodule (e.g., granuloma, hamartoma).

A **domain** $\mathcal{D}$ is defined as a tuple consisting of a marginal probability distribution on the inputs $P(X)$ and the feature space itself $\mathcal{X}$. A **task** $\mathcal{T}$ consists of a conditional probability distribution $P(Y|X)$ and the label space $\mathcal{Y}$.

In the cross-institutional setting, we define two distinct domains:

1. **The Source Domain ($\mathcal{D}_s$):** This corresponds to the training institution (Hospital A). We have access to a labeled dataset $S = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$, where n_s is the number of labeled samples. These samples are drawn i.i.d. from the joint source distribution $P_s(X, Y)$. In our empirical setting, n_s is small ($n_s \approx 295$), characteristic of specialized medical registries.[1]
2. **The Target Domain ($\mathcal{D}_t$):** This corresponds to the deployment institution (Hospital B). We have access to an unlabeled dataset $T = \{x_j^t\}_{j=1}^{n_t}$, where n_t is the number of observed samples ($n_t \approx 190$). These samples are drawn from the target marginal $P_t(X)$. The labels y_j^t are unobserved during the adaptation phase, reflecting the privacy constraint that prevents sharing outcomes.

2.2 The Fundamental Discrepancy

The core problem is that the joint distributions differ between hospitals:

$$P_s(X, Y) \neq P_t(X, Y)$$
This inequality invalidates the guarantees of Empirical Risk Minimization (ERM). If we train a hypothesis h to minimize the source risk $\epsilon_s(h) = \mathbb{E}_{(x,y)\sim P_s}[\ell(h(x), y)]$, there is no guarantee that the target risk $\epsilon_t(h) = \mathbb{E}_{(x,y)\sim P_t}[\ell(h(x), y)]$ will be low. In fact, clinical literature suggests that $\epsilon_t(h)$ is often substantially higher than $\epsilon_s(h)$, with AUC drops of 0.1–0.3 being common.[1]

2.3 Feature Space Heterogeneity and Schema Mismatch

Unlike standard computer vision tasks where the input is a fixed grid of pixels, tabular medical data is characterized by **schema mismatch**. Hospitals utilize different Laboratory Information Systems (LIS), purchase assay kits from different vendors, and adhere to different clinical guidelines for ordering tests.

Let $\mathcal{F}_s$ be the set of feature indices recorded in the source domain, and $\mathcal{F}_t$ be the set of feature indices recorded in the target domain. We define three subspaces:

Subspace	**Definition**	**Clinical Implication**
Intersection	$\mathcal{F}_{\cap} = \mathcal{F}_s \cap \mathcal{F}_t$	The core biomarkers available at both sites (e.g., Age, Spiculation, Diameter). This is the *admissible* feature space for the model.
Source-Specific	$\mathcal{F}_{s\setminus t} = \mathcal{F}_s \setminus \mathcal{F}_t$	Variables unique to Hospital A (e.g., a specific PET-CT metabolic parameter). These must be discarded or imputed to

		avoid domain dependence.
Target-Specific	$\mathcal{F}_{t\setminus s} = \mathcal{F}_t \setminus \mathcal{F}_s$	Variables unique to Hospital B. These cannot be used by a model trained on A.

The problem formulation requires a mapping $\Phi_{schema}: \mathbb{R}^{|\mathcal{F}_s|} \rightarrow \mathbb{R}^{d_{shared}}$ that projects the source data onto the shared subspace $\mathcal{F}_{\cap}$. PANDA addresses this via **Cross-Domain Recursive Feature Elimination (RFE)**, which not only identifies this intersection but actively selects a subset $\mathcal{F}^* \subset \mathcal{F}_{\cap}$ that maximizes stability.[1]

3. Taxonomy of Distributional Shifts in Medicine

To solve the adaptation problem, we must decompose the generic distributional shift $P_s \neq P_t$ into its constituent causal mechanisms. In pulmonary nodule malignancy prediction, the shift is a complex superposition of Covariate Shift, Label Shift, and Concept Shift.

3.1 Covariate Shift: The Acquisition Gap

Covariate shift occurs when the marginal distribution of inputs changes, but the causal relationship between features and labels remains stable:

$$P_s(X) \neq P_t(X) \quad \text{and} \quad P_s(Y|X) = P_t(Y|X)$$
In our context, this is driven by **technological and demographic heterogeneity**:

1. **Scanner Protocols:** CT scanners vary in their reconstruction kernels (e.g., "Lung" vs. "Standard" vs. "Smooth"). A nodule scanned with a sharp kernel will exhibit higher values for texture features like "entropy" or "spiculation" compared to the same nodule scanned with a smooth kernel. This shifts the probability density function $P(x_{texture})$ along the real number line.[1]
2. **Demographics:** As seen in the TableShift BRFSS Diabetes benchmark, source populations (e.g., White respondents) may have different distributions of Age, BMI, and

Income compared to target populations (e.g., non-White respondents). This changes the support and density of the input space.[1]

Mathematically, covariate shift implies that the importance weight $\beta(x) = P_t(x) / P_s(x)$ is not uniform. Standard ERM effectively integrates over $P_s(x)$, meaning regions of high density in the target domain that have low density in the source domain are under-weighted, leading to poor generalization.

3.2 Label Shift: The Prevalence Gap

Label shift, or prior probability shift, is characterized by a change in the class balance:

$$P_s(Y) \neq P_t(Y) \quad \text{and} \quad P_s(X|Y) = P_t(X|Y)$$
This is endemic to the structure of healthcare referrals:

- **Source (Hospital A):** A tertiary cancer center receives referrals of suspicious, high-risk nodules. The prevalence of malignancy might be $P_s(y=1) \approx 0.64$.[1]
- **Target (Hospital B):** A community screening program or general hospital encounters a broader population with many benign incidental findings. The prevalence might be significantly lower, e.g., $P_t(y=1) \approx 0.10 - 0.20$.

The implication for a probabilistic classifier $f(x)$ is severe. If trained on the balanced/high-prevalence source, the model learns a prior π_s. When applied to the low-prevalence target, the posterior probabilities $P(y=1|x)$ will be systematically calibrated upwards (over-estimated). A threshold of 0.5 might yield excellent sensitivity but catastrophic specificity (many False Positives), necessitating unnecessary biopsies.

3.3 Concept Shift: The Definition Gap

Concept shift is the most pernicious form of divergence, where the conditional distribution changes:

$$P_s(Y|X) \neq P_t(Y|X)$$
In pulmonary medicine, this arises from latent confounders:

- **Geographic Pathology:** In regions with endemic tuberculosis (TB) or fungal infections

(e.g., the Ohio River Valley for Histoplasmosis, or parts of East Asia for TB), granulomas are common. These benign lesions often mimic the radiographic appearance of malignancy (spiculated, upper lobe). In a Western cohort (e.g., Mayo Clinic data), an upper-lobe spiculated nodule is highly likely to be cancer ($P(y=1|x_{upper}, x_{spic}) \approx 0.9$). In a TB-endemic Asian cohort, the same features might only imply a 40% probability of cancer ($P(y=1|x_{upper}, x_{spic}) \approx 0.4$).[1]

- **Drift in Definitions:** In the BRFSS dataset, survey questions regarding "smoking status" or "general health" may change wording slightly between years, altering the semantic meaning of the feature vector x relative to the label y.

PANDA addresses Concept Shift fundamentally through **Feature Selection**. By identifying and pruning features that exhibit unstable importance rankings across domains (via Cross-Domain RFE), we attempt to isolate a subspace $\mathcal{X}_{stable}$ where the conditional assumption $P_s(Y|X_{stable}) \approx P_t(Y|X_{stable})$ holds true.[1]

4. Theoretical Constraints of Existing Models

To justify the architecture of PANDA, we must formally analyze why existing state-of-the-art models fail in this specific problem formulation ($N \approx 300$, Unlabeled Target, Tabular Data).

4.1 The Failure of Tree Ensembles (XGBoost, LightGBM)

Gradient Boosted Decision Trees (GBDTs) are the standard for tabular data due to their ability to handle irregular decision boundaries and mixed data types. However, they suffer from two critical limitations in the adaptation setting:

1. **Non-Differentiability:** GBDTs partition the feature space using hard, axis-aligned splits ($x_j < \theta$). The resulting function is piecewise constant and non-differentiable. This precludes the use of gradient-based domain adaptation techniques (like Adversarial Training or Gradient Reversal Layers) which require backpropagating a domain loss $\mathcal{L}_{domain}$ into the feature encoder.
2. **Inability to Extrapolate:** Tree models cannot extrapolate beyond the range of the training data. If covariate shift moves the target distribution $P_t(x)$ outside the support of $P_s(x)$, the tree will map all such points to the value of the nearest leaf node, which is often statistically invalid.
3. **Small-Sample Overfitting:** While robust on medium data ($N > 10k$), GBDTs can easily

memorize noise in small datasets ($N < 500$), especially with high-dimensional features ($d > 50$).

4.2 The Failure of Deep Tabular Models (TabNet, FT-Transformer)

Deep Learning models offer differentiability, allowing for domain alignment. However, they lack the appropriate **Inductive Bias** for tabular data.

- **Rotational Invariance:** Standard MLPs are rotationally invariant (with appropriate weight matrices), but tabular features are not rotationally interchangeable (e.g., rotating "Age" and "Creatinine" creates a nonsensical feature).
- **Data Hunger:** Neural networks typically require large datasets to converge to a generalizable solution. With $N \approx 300$, deep tabular models (like TabNet) are prone to severe overfitting or convergence to local minima, often underperforming simple logistic regression.
- **Hyperparameter Sensitivity:** These models require extensive tuning of learning rates, decay, and dropout, which is perilous in an unsupervised DA setting where we lack a labeled target validation set to guide the tuning.[1]

4.3 The Solution: Tabular Foundation Models (TabPFN)

PANDA employs **TabPFN** (Tabular Prior-Data Fitted Network) to resolve the tension between the performance of trees and the differentiability of networks. TabPFN is a Transformer trained via **Meta-Learning** on millions of synthetic datasets generated from Structural Causal Models (SCMs).

Formally, TabPFN approximates the posterior predictive distribution (PPD) of a Bayesian inference process. Instead of learning weights θ on the source dataset $\mathcal{D}_s$, it performs In-Context Learning (ICL). The model takes a sequence of tokens representing the training set and the query:

$$\text{Input} = [x_1^s, y_1^s, x_2^s, y_2^s, \dots, x_{n_s}^s, y_{n_s}^s, x_{query}]$$

and outputs $P(y_{query} | x_{query}, \mathcal{D}_s)$ in a single forward pass.
This formulation provides three key advantages for our problem:

1. **Strong Priors for Small N:** The meta-learned priors allow the model to generalize from

as few as $N=20$ samples, essentially transferring knowledge from the synthetic pre-training to the medical task.
2. **Differentiable Embeddings:** Internally, TabPFN maps inputs to a continuous latent space $\Phi(x) \in \mathbb{R}^{h}$. This embedding space captures the semantic structure of the data and, unlike tree leaves, is amenable to mathematical alignment operations like TCA.
3. **No Training Required:** Inference is a forward pass, eliminating the risk of overfitting via gradient descent on the small source dataset.

5. The PANDA Framework Formulation

We now formally derive the components of the PANDA framework. The framework is a sequential composition of operators designed to minimize the target risk upper bound.

5.1 Optimization Objective: The Adaptation Bound

We appeal to the domain adaptation theory of Ben-David et al. (2010). The target risk $\epsilon_t(h)$ is bounded by:

$$\epsilon_t(h) \leq \epsilon_s(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(P_s, P_t) + \lambda$$

Where:
- $\epsilon_s(h)$ is the source risk (minimized by the TabPFN classifier).
- $d_{\mathcal{H}\Delta\mathcal{H}}(P_s, P_t)$ is the divergence between source and target distributions (minimized by TCA).
- λ is the ideal joint error (minimized by Feature Selection, removing concept-shifted features).

PANDA minimizes this bound via a multi-stage pipeline:
$$ f(x) = \text{Ensemble} \left(h \circ \mathcal{A}{TCA} \circ \Phi{TabPFN} \circ \mathcal{T}_{RFE} (x) \right) $$

5.2 Step 1: Cross-Domain Recursive Feature Elimination (RFE)

The first operator $\mathcal{T}_{RFE}: \mathbb{R}^{d_{raw}} \rightarrow \mathbb{R}^{k}$ reduces the feature space to a stable subset. We formulate this as an optimization of a **Cost-Effectiveness Index**.

Let $\text{Imp}(f_j)$ be the permutation importance of feature j. We explicitly penalize features that are unstable across folds or domains. The objective function for subset selection is:
$$ \mathcal{F}^* = \arg\max_{k} \left(w_1 S_{perf}(k) + w_2 S_{eff}(k) + w_3 S_{stab}(k) + w_4 S_{simp}(k) \right) $$
Where:

- $S_{perf}(k)$ measures AUC/Accuracy.
- $S_{eff}(k)$ measures computational efficiency (1 - normalized time).
- $S_{stab}(k)$ measures the stability (1 - coefficient of variation across folds).
- $S_{simp}(k)$ encourages sparsity via $\exp(-\alpha k)$.

This step is crucial for **Concept Shift**. By enforcing stability and sparsity (yielding subsets like "Best-7" or "Best-8"), we remove features that are highly predictive in the source but noisy or uncorrelated in the target (e.g., site-specific biomarkers), effectively lowering the λ term in the error bound.[1]

5.3 Step 2: Foundation Model Encoding

The stable features are passed to the frozen TabPFN encoder.

$$\Phi(x) = \text{Transformer}_{\theta^*}(\text{Tokenize}(x)) \in \mathbb{R}^{h}$$

where $h=128$. This step linearizes the decision boundary. The transformer attention mechanism computes interactions between features x_i and context examples x_{ctx}, creating a representation that contextualizes the query point within the manifold of the training data.

5.4 Step 3: Transfer Component Analysis (TCA)

This is the central alignment engine. We assume that there exists a latent space

transformation W such that the marginal distributions are matched: $P(W^\top \Phi(X_s)) \approx P(W^\top \Phi(X_t))$.

We employ Transfer Component Analysis (TCA), which minimizes the Maximum Mean Discrepancy (MMD) in a Reproducing Kernel Hilbert Space (RKHS).
Let K be the kernel matrix computed on the TabPFN embeddings $\Phi(X)$. We use a Linear Kernel $K_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle$, justified by the assumption that the heavy lifting of non-linear disentanglement is already performed by the TabPFN transformer.
The optimization problem is:

$$\min_{W} \text{tr}(W^\top K L K^\top W) + \mu \text{tr}(W^\top W)$$

Subject to: $W^\top K H K^\top W = I$ (variance constraint).
Here, L is the MMD matrix:
$$ L_{ij} = \begin{cases} \frac{1}{n_s^2} & x_i, x_j \in \mathcal{D}_s \ \frac{1}{n_t^2} & x_i, x_j \in \mathcal{D}t \ -\frac{1}{n_s n_t} & \text{otherwise} \end{cases} $$
Minimizing $\text{tr}(W^\top K L K^\top W)$ is equivalent to minimizing the squared distance between the empirical means of the source and target in the RKHS. This explicitly shrinks the $d{\mathcal{H}\Delta\mathcal{H}}$ term in the generalization bound.
The solution W is obtained analytically via generalized eigendecomposition, which is computationally feasible for the small sample sizes ($N \approx 500$) typical of this domain.

5.5 Step 4: Ensemble Aggregation and Calibration

To handle the variance inherent in small-sample alignment and the label shift issue, PANDA employs a Multi-Branch Ensemble.
We generate $B=4$ views of the data for each sample:

1. **Raw:** Original feature distribution.
2. **Rotated:** Features permuted cyclically (to mitigate positional bias in the Transformer).
3. **Quantile Transformed:** Maps features to a Gaussian $N(0,1)$. This is critical for **Covariate Shift**, as it normalizes the marginal scales of features (e.g., handling different units of measurement across hospitals).
4. **Ordinal Encoded:** Handles categorical shift.

For each view, we run the TCA + TabPFN pipeline with $S=8$ random seeds. The final probability is the temperature-scaled average:

$$\hat{p}(y=1|x) = \frac{1}{32} \sum_{m=1}^{32} \sigma\left(\frac{z_m(x)}{T}\right)$$

where $T=0.9$. This temperature scaling acts as a soft calibration mechanism, smoothing the overly confident predictions often produced by foundation models when samples are out-of-distribution.

6. Constraints, Assumptions, and Safety

6.1 The "Open World" Challenge

While TabPFN is a "closed world" model (assuming context covers query support), medical deployment is "open world". New scanners may produce values (e.g., negative pixel values) never seen in training. The **Quantile Transformation** branch of the ensemble provides a safeguard by mapping arbitrary input distributions to a standard normal, ensuring the foundation model always receives inputs within its expected numerical range.

6.2 Privacy Compliance

The formulation strictly respects the constraint $\mathcal{D}_t = \{x_j^t\}$. The alignment (TCA) depends only on the kernel matrix K, which is computed from features. No target labels are required. This allows the model to be "aligned" at Hospital B without any patient data leaving the firewall or requiring local annotation efforts.

6.3 Sample Size Limits

The computational complexity of the TCA eigendecomposition is $O((n_s+n_t)^3)$. This limits the approach to datasets where $N < 2000-3000$. This constraint matches the reality of "High-Value" medical datasets (like biopsy-confirmed nodules), which rarely exceed a few thousand cases. For larger EHR datasets (e.g., MIMIC-III), Nyström approximations would be required, but for the specific problem of specialized diagnostic registries, the exact solution is feasible and preferred.

7. Conclusion

The "Problem Formulation" for PANDA is a rigorous response to the failure of standard ML in clinical settings. It acknowledges that **Distribution Shift** is not an anomaly but the default state of medical data. It rejects the "Big Data" solution (Deep Learning from scratch) in favor of a "Smart Data" approach: leveraging **Foundation Model Priors** to handle sample scarcity, **Kernel Alignment** to handle covariate shift, and **Stability Selection** to handle concept shift. This formulation provides the theoretical guarantee that, provided the shared feature subspace contains predictive signal, the target risk can be minimized without access to target labels.

Table 1: Mathematical Mapping of Clinical Problems to PANDA Components

Clinical Challenge	**Statistical Mechanism**	**PANDA Component Solution**	**Theoretical Justification**
Scanner Variance (Sharp vs Smooth Kernels)	**Covariate Shift:** $P_s(X) \neq P_t(X)$	**TCA (Transfer Component Analysis)**	Minimizes MMD Divergence $d_{\mathcal{H}\Delta\mathcal{H}}$ in RKHS.
Referral Patterns (Cancer Center vs Screening)	**Label Shift:** $P_s(Y) \neq P_t(Y)$	**Ensemble Aggregation & Temperature Scaling**	Calibrates posteriors; smooths overconfidence from prior mismatch.
Biological Confounders (TB vs Cancer)	**Concept Shift:** $P_s(Y | X) \neq P_t(Y | X)$		

Small Cohorts ($N \approx 300$)	**High Variance / Overfitting**	**TabPFN (Foundation Model)**	Transfers meta-learned priors; effectively increases sample efficiency.
Heterogeneous LIS (Different Lab Vendors)	**Schema Mismatch:** $\mathcal{F}_s \neq \mathcal{F}_t$	**Schema Intersection & Quantile Transform**	Aligns support and scales of input features.

This table summarizes the direct link between the clinical reality, the mathematical formulation, and the engineered solution, defining the scope and logic of the PANDA framework.

Works cited

1. main.pdf