Theoretical Analysis of Cross-Hospital Generalization: Dissecting the PANDA Framework through Domain Adaptation Bounds and Foundation Model Priors

1. Introduction: The Statistical Crisis in Cross-Hospital Deployment

The deployment of machine learning models in clinical medicine is currently facing a reproducibility crisis that is fundamentally statistical in nature. While algorithms such as Gradient-Boosted Decision Trees (GBDTs) and Convolutional Neural Networks (CNNs) have achieved superhuman performance on isolated benchmarks, they frequently suffer catastrophic performance degradation when deployed across institutions. This phenomenon is particularly acute in pulmonary nodule malignancy prediction, where models trained at tertiary academic centers fail to generalize to community screening programs or hospitals in geographically distinct regions.[1]

This Analysis chapter moves beyond the phenomenological reporting of performance metrics—such as Area Under the Receiver Operating Characteristic Curve (AUC)—to establish a rigorous theoretical understanding of *why* the proposed PANDA (Pretrained Adaptation Network with Domain Alignment) framework succeeds where traditional methods falter. We posit that the failure of classical models is not merely an engineering oversight but a violation of the fundamental assumptions of statistical learning theory, specifically the assumption that training and test data are drawn independent and identically distributed (i.i.d.) from the same joint distribution $P(X, Y)$.

In the context of cross-hospital deployment, this i.i.d. assumption is shattered by three distinct but interacting forms of distribution shift: covariate shift, where acquisition protocols alter feature distributions; label shift, where disease prevalence varies by setting; and concept shift, where the biological implications of biomarkers change due to latent confounders. By dissecting PANDA through the lens of the **Ben-David et al. generalization bound for**

domain adaptation [2], we demonstrate that PANDA's architecture is a direct algorithmic response to the theoretical decomposition of target error.

We establish a unified notation system to rigorously define the interactions between the **TabPFN** foundation model backbone, **Transfer Component Analysis (TCA)** alignment, and **Cross-Domain Recursive Feature Elimination (RFE)**. We show that TabPFN minimizes source risk $\epsilon_S(h)$ through meta-learned Bayesian priors rather than parametric optimization; Cross-Domain RFE minimizes the ideal joint hypothesis error λ by pruning concept-shifted features; and latent TCA minimizes the $\mathcal{H}\Delta\mathcal{H}$-divergence $d_{\mathcal{H}\Delta\mathcal{H}}$ by aligning distributions in a linearized reproducing kernel Hilbert space (RKHS). This chapter links these theoretical error terms directly to the empirical results observed in our private pulmonary nodule cohorts and the public TableShift benchmarks, offering a comprehensive mechanistic explanation for the framework's robustness.

2. Problem Formulation and Unified Notation System

To ensure mathematical precision and consistency across this dissertation, we define a single, unified notation system. This system governs the definitions of domains, distributions, hypothesis spaces, and error terms, superseding any conflicting symbols that may have appeared in prior drafts or disparate literature sources.

2.1 Domain Definitions and Distributional Assumptions

Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the input feature space, comprising d clinical or radiomic variables (e.g., age, nodule diameter, spiculation score). Let $\mathcal{Y} = \{0, 1\}$ denote the binary label space, where $Y=0$ represents a benign nodule and $Y=1$ represents a malignant nodule.

A **domain** is formally defined as a joint probability distribution $\mathcal{D}$ over the product space $\mathcal{X} \times \mathcal{Y}$. In the context of unsupervised domain adaptation (UDA), we consider two distinct domains:

1. **Source Domain (Hospital A):** Characterized by the joint distribution $\mathcal{D}_S = P_S(X, Y)$. We have access to a labeled dataset $S = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ of size n_s, drawn i.i.d. from $\mathcal{D}_S$. In our experimental setup, this corresponds to the derivation cohort from the primary cancer center ($n_s = 295$).[1]

2. **Target Domain (Hospital B):** Characterized by the joint distribution $\mathcal{D}_T = P_T(X, Y)$. We have access to an unlabeled dataset $T = \{x_j^t\}_{j=1}^{n_t}$ of size n_t, drawn i.i.d. from the marginal distribution $P_T(X)$. The target labels are unobserved during the adaptation phase and are used solely for evaluation. This corresponds to the external validation cohort ($n_t = 190$).[1]

The fundamental challenge is that $\mathcal{D}_S \neq \mathcal{D}_T$. This inequality implies that a hypothesis $h: \mathcal{X} \to \mathcal{Y}$ optimized to minimize risk on $\mathcal{D}_S$ is not guaranteed to minimize risk on $\mathcal{D}_T$.

2.2 Taxonomy of Distribution Shifts

We formally decompose the domain discrepancy into three components, each requiring a distinct architectural intervention within PANDA:

- **Covariate Shift:** Defined as $P_S(X) \neq P_T(X)$ while $P_S(Y|X) = P_T(Y|X)$. This shift implies that the marginal distribution of features changes—for example, Hospital A uses "Sharp" reconstruction kernels yielding higher texture values, while Hospital B uses "Smooth" kernels—but the conditional probability of malignancy given a specific feature vector remains constant. PANDA addresses this via **Transfer Component Analysis (TCA)**.[1]
- **Label Shift (Prior Probability Shift):** Defined as $P_S(Y) \neq P_T(Y)$ while $P_S(X|Y) = P_T(X|Y)$. This reflects differences in disease prevalence. For instance, the source tertiary center has a malignancy rate of 64.1%, whereas a community screening setting might have a rate of $<5\%$. PANDA addresses this via **temperature-scaled ensemble calibration**.[1]
- **Concept Shift:** Defined as $P_S(Y|X) \neq P_T(Y|X)$ or $P_S(X|Y) \neq P_T(X|Y)$. This is the most pernicious shift, where the definition of the disease relation changes. For example, in regions endemic for tuberculosis (TB), a "spiculated" nodule (feature x) may often be a benign granuloma ($y=0$), whereas in non-endemic regions, it is almost certainly cancer ($y=1$). Thus, $P_S(1|x_{spiculated}) \gg P_T(1|x_{spiculated})$. PANDA minimizes this via **Cross-Domain Recursive Feature Elimination (RFE)**, which prunes features susceptible to such shifts.[1]

2.3 Mathematical Notation Summary

Table 2.1 serves as the reference for all mathematical symbols used in the subsequent

theoretical derivations.

Symbol	Definition	Context & Dimensions
$\mathcal{X}, \mathcal{Y}$	Input feature space and label space	$\mathcal{X} \subseteq \mathbb{R}^d, \mathcal{Y} \in \{0,1\}$
$\mathcal{D}_S, \mathcal{D}_T$	Source and Target distributions	Joint distributions over $\mathcal{X} \times \mathcal{Y}$
S, T	Empirical Source and Target datasets	Sets of samples of size n_s, n_t
h	Hypothesis function (classifier)	$h: \mathcal{X} \to \{0,1\}$ or $$
$\mathcal{H}$	Hypothesis class	Set of all possible functions h
$\epsilon_S(h), \epsilon_T(h)$	Expected risk on Source and Target	$\mathbb{E}_{(x,y)\sim \mathcal{D}} [\ell(h(x), y)]$
$\hat{\epsilon}_S(h), \hat{\epsilon}_T(h)$	Empirical risk on Source and Target	$\frac{1}{n} \sum \ell(h(x_i), y_i)$
$\phi(\cdot)$	Feature extractor (TabPFN Backbone)	Maps $x \to z \in \mathbb{R}^{h_{dim}}$ ($h_{dim}=128$)
$\psi(\cdot)$	Domain alignment mapping (TCA)	Maps $z \to z' \in \mathbb{R}^{m}$
K	Kernel Matrix	$K \in \mathbb{R}^{(n_s+n_t) \times (n_s+n_t)}$
L	MMD Indicator Matrix	Used in TCA objective

H	Centering Matrix	$H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$
$d_{\mathcal{H}\Delta\mathcal{H}}$	$\mathcal{H}\Delta\mathcal{H}$-divergence	Measure of domain discrepancy
λ	Ideal joint hypothesis error	$\min_{h \in \mathcal{H}} (\epsilon_S(h) + \epsilon_T(h))$
PPD	Posterior Predictive Distribution	TabPFN output probability

3. Theoretical Foundation I: The Generalization Bound

To analyze the generalization capability of PANDA, we must first establish the theoretical upper bound on the target error. We adopt the seminal learning-theoretic bound for domain adaptation proposed by Ben-David et al. (2010).[2] This bound is pivotal because it decomposes the target error into observable and optimizable quantities, thereby providing a blueprint for the PANDA architecture.

3.1 The Ben-David Theorem

Theorem 1 (Ben-David et al., 2010): Let $\mathcal{H}$ be a hypothesis space of VC-dimension d_{VC}. If S and T are samples of size n drawn from $\mathcal{D}_S$ and $\mathcal{D}_T$ respectively, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, for every $h \in \mathcal{H}$:

$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda + \mathcal{O}\left(\sqrt{\frac{d_{VC} \log n}{n} + \log \frac{1}{\delta}}\right)$$

This inequality states that the error on the target domain $\epsilon_T(h)$ is bounded by the

sum of three distinct terms (plus a complexity term that vanishes as $n \to \infty$):

1. **Source Risk ($\epsilon_S(h)$):** The expected error of the hypothesis on the source domain. This is the standard objective of supervised learning (e.g., minimizing log-loss or maximizing AUC on Hospital A data).
2. **Domain Divergence ($d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$):** A measure of the discrepancy between the source and target marginal feature distributions $P_S(X)$ and $P_T(X)$. Importantly, this divergence is not a generic distance (like Euclidean or L_1) but is defined relative to the capacity of the hypothesis class $\mathcal{H}$ to discriminate between the domains.
3. **Adaptability Term (λ):** Defined as $\lambda = \min_{h \in \mathcal{H}} (\epsilon_S(h) + \epsilon_T(h))$. This term represents the error of the *ideal* hypothesis that performs best on both domains simultaneously. It captures the irreducible error due to concept shift; if the labeling functions $f_S(x)$ and $f_T(x)$ are fundamentally contradictory (e.g., x is benign in S but malignant in T), λ will be large.

3.2 The $\mathcal{H}\Delta\mathcal{H}$-Divergence

Standard distances like Kullback-Leibler (KL) divergence or L_1 distance are insufficient for bounding generalization error because they do not account for the specific hypothesis class used. Two distributions might be very far apart in L_1 distance but yield identical classification errors if the decision boundary remains the same. Ben-David et al. introduced the $\mathcal{H}\Delta\mathcal{H}$-divergence to address this.

Let $\mathcal{H}\Delta\mathcal{H} = \{h(x) \oplus h'(x) \mid h, h' \in \mathcal{H}\}$ be the set of functions characterizing the disagreement between any two hypotheses in $\mathcal{H}$. The divergence is defined as:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 2 \sup_{h \in \mathcal{H}\Delta\mathcal{H}} \left| \Pr_{x \sim \mathcal{D}_S}[h(x)=1] - \Pr_{x \sim \mathcal{D}_T}[h(x)=1] \right|$$

In simpler terms, $d_{\mathcal{H}\Delta\mathcal{H}}$ measures the maximum possible accuracy of a discriminator trained to distinguish source examples from target examples using the hypothesis class $\mathcal{H}\Delta\mathcal{H}$. If the domains are easily distinguishable (e.g., one hospital uses only even values for age and the other uses odd), the divergence is high, and the bound on target error becomes loose. Conversely, if the domains are indistinguishable to the classifier, the divergence is low.

3.3 Architectural Implications for PANDA

The decomposition in Theorem 1 dictates the three primary components of the PANDA framework. A naive model typically optimizes only one term, leading to failure:

- **Failure of Pure Supervised Learning (GBDTs):** Models like XGBoost aggressively minimize $\epsilon_S(h)$ (often to 0 on training data). However, on small datasets ($n_s \approx 295$), this overfitting usually increases the effective complexity of the hypothesis, allowing the model to latch onto source-specific artifacts. This effectively maximizes the divergence $d_{\mathcal{H}\Delta\mathcal{H}}$, as the model learns to identify the "source" rather than the pathology.
- **Failure of Naive Alignment:** Pure alignment methods (e.g., standard MMD on raw features) might minimize $d_{\mathcal{H}\Delta\mathcal{H}}$, but if they distort the features such that they lose predictive power, $\epsilon_S(h)$ will increase. Furthermore, if concept shift is present, aligning marginals $P(X)$ does not align posteriors $P(Y|X)$, leaving λ high.

PANDA is explicitly engineered to minimize all three terms simultaneously:

1. **Minimizing $\epsilon_S(h)$**: The **TabPFN** backbone uses strong, meta-learned priors to achieve low source risk without overfitting, even with small sample sizes.
2. **Minimizing $d_{\mathcal{H}\Delta\mathcal{H}}$**: The **Latent TCA** projection aligns the feature distributions in a subspace, making the domains indistinguishable to the linear classifier.
3. **Minimizing λ**: The **Cross-Domain RFE** selects features that are chemically stable across domains, ensuring that a joint hypothesis exists (i.e., reducing concept shift).

The following sections analyze each of these mechanisms in detail.

4. Minimizing Source Risk $\epsilon_S(h)$: The TabPFN Mechanism

The first challenge in our problem setting is the small sample size ($n_s = 295$). In this regime, standard Empirical Risk Minimization (ERM) is prone to high variance. Deep neural networks (DNNs) typically require $N > 10^4$ to generalize, and GBDTs can easily memorize noise. PANDA circumvents this via the TabPFN foundation model.

4.1 Prior-Data Fitted Networks (PFNs) vs. Parametric Learning

Traditional parametric learning optimizes a set of weights θ to minimize loss on the observed data $\mathcal{D}_{train}$:

$$\hat{\theta} = \arg\min_{\theta} \sum_{(x,y) \in \mathcal{D}_{train}} \mathcal{L}(f_\theta(x), y)$$

This process is data-hungry because the model starts with uninformative (random) priors and must "learn" the structure of tabular data from scratch.
In contrast, TabPFN is a **Prior-Data Fitted Network (PFN)**. It is not "trained" on the medical data in the traditional sense. Instead, it is pre-trained via meta-learning on a vast distribution of synthetic datasets to approximate the **Posterior Predictive Distribution (PPD)**.[6]

Formally, let $P_{prior}(M)$ be a prior distribution over data-generating mechanisms (SCMs, BNNs). For a query point x_q and a support set $\mathcal{D}_{support} = \{(x_i, y_i)\}$, the true Bayesian PPD is:

$$P(y_q | x_q, \mathcal{D}_{support}) = \int P(y_q | x_q, M) P(M | \mathcal{D}_{support}) \, dM$$

TabPFN approximates this integral using a Transformer T_Φ, which takes the serialized dataset as input and outputs the distribution:

$$T_\Phi(x_q, \mathcal{D}_{support}) \approx P(y_q | x_q, \mathcal{D}_{support})$$

4.2 The Nature of the TabPFN Prior

According to Hollmann et al. (2023) [6], the prior P_{prior} used to train TabPFN is constructed from **Structural Causal Models (SCMs)**. These SCMs are generated by sampling computation graphs with varying sparsity, non-linearities, and noise levels.

$$M_{SCM}: \quad x_j := g_j(PA_j) + \epsilon_j, \quad y := f(PA_y) + \epsilon_y$$

where PA_j are parents of node j in the causal graph. By training on millions of such synthetic structural equations, TabPFN learns an inductive bias that favors:

- **Sparsity:** Real-world tabular data often depends on a few key interactions.
- **Smoothness:** Decision boundaries are rarely fractal; they are usually piecewise smooth.
- **Causal Structures:** Features are often correlated due to common causes.

4.3 Analysis of Source Risk Reduction

This meta-learning approach explains the superior source performance observed in Table 12.[1] PANDA achieves a source AUC of **0.829**, while Random Forest achieves **0.752** and XGBoost **0.742**.

- **Mechanism:** GBDTs rely on asymptotic splitting criteria. With only 295 samples, the variance in split selection is high, leading to suboptimal trees.
- **Mechanism:** TabPFN treats the 295 samples as a "context" for Bayesian inference. It effectively matches the observed pattern against its library of millions of learned SCMs. Since it performs no gradient updates on the medical data, it cannot "overfit" in the sense of chasing stochastic noise; it simply updates its posterior belief.
- **Theoretical Consequence:** This drastically reduces the $\epsilon_S(h)$ term in the Ben-David bound. We start the adaptation process with a highly robust base model, unlike deep tabular baselines (TabNet, FT-Transformer) which often fail to converge or generalize in this sample regime.

5. Minimizing Divergence $d_{\mathcal{H}\Delta\mathcal{H}}$: Latent Transfer Component Analysis

While TabPFN secures the source risk, it assumes that the query x_q comes from the same distribution as the context $\mathcal{D}_{support}$. In our cross-hospital setting, $x_q \sim P_T(X)$ while $\mathcal{D}_{support} \sim P_S(X, Y)$. The covariate shift $P_S(X) \neq P_T(X)$ means the attention mechanism may attend to "nearest neighbors" that are not semantically equivalent (e.g., matching a smooth-kernel nodule to a sharp-kernel nodule based on raw pixel intensity, which is misleading).

To minimize the divergence term $d_{\mathcal{H}\Delta\mathcal{H}}$, PANDA employs

Transfer Component Analysis (TCA) [3] within the latent embedding space of the TabPFN encoder.

5.1 Why Latent Space Alignment?

Traditional domain adaptation often aligns raw features. However, medical features have complex, non-linear dependencies. Aligning the means of "Nodule Size" and "Age" independently (e.g., via standard scaling) ignores their correlation structure.

PANDA leverages the TabPFN Encoder $\phi(\cdot)$ as a linearizing map. The Transformer architecture maps the complex input space $\mathcal{X}$ to a high-dimensional embedding space $\mathcal{Z} \in \mathbb{R}^{128}$.

$$z = \phi(x; \mathcal{D}_{support})$$

In this space, the "classes" (Benign vs. Malignant) are arguably more linearly separable due to the attention mechanism's ability to disentangle manifolds. This justifies the use of linear TCA in the latent space, avoiding the high computational cost and hyperparameter instability of kernel-TCA with RBF kernels on raw data.11

5.2 The TCA Optimization Objective

TCA aims to find a projection matrix $W \in \mathbb{R}^{128 \times m}$ that minimizes the **Maximum Mean Discrepancy (MMD)** between the source and target embeddings, while preserving the variance of the data to strictly avoid trivial solutions (such as mapping all points to zero).

The empirical MMD squared distance between the projected source dataset $S' = \{W^\top \phi(x_i^s)\}$ and target dataset $T' = \{W^\top \phi(x_j^t)\}$ is:

$$ MMD^2(S', T') = \left| \frac{1}{n_s} \sum_{i=1}^{n_s} W^\top \phi(x_i^s) - \frac{1}{n_t} \sum_{j=1}^{n_t} W^\top \phi(x_j^t) \right|_{\mathcal{H}}^2 $$

We can rewrite this using the kernel matrix K. Let K be the kernel matrix computed on the union of source and target embeddings using a linear kernel: $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$.

The objective function for TCA is formally derived as [3]:

$$\begin{aligned} \min_{W} \quad & \text{tr}(W^\top K L K W) + \mu \text{tr}(W^\top W) \\ \text{s.t.} \quad & W^\top K H K W = I \end{aligned}$$

Derivation and Verification of Terms:

1. L (MMD Indicator Matrix): This matrix constructs the MMD calculation.
   $$ L_{ij} = \begin{cases} \frac{1}{n_s^2} & x_i, x_j \in S \\ \frac{1}{n_t^2} & x_i, x_j \in T \\ -\frac{1}{n_s n_t} & \text{otherwise} \end{cases} $$
   Verification: The term $\text{tr}(W^\top K L K W)$ exactly expands to the squared MMD formula above.
2. **H (Centering Matrix):** $H = I - \frac{1}{n_s+n_t}\mathbf{1}\mathbf{1}^\top$. The term KHK represents the centered covariance matrix of the data in the kernel space.
3. **Constraint ($W^\top K H K W = I$):** This constraint forces the projected data to have unit variance (whitening). This is crucial; without it, the minimizer of MMD is simply $W=0$. By enforcing variance preservation, we ensure that the projection retains information while aligning the means.
4. **Regularization ($\mu \text{tr}(W^\top W)$):** A Tikhonov regularization term to prevent overfitting and ensure numerical stability of the inverse operation.

5.3 Solving the Objective

The solution to this constrained optimization problem is given by the generalized eigenvalue problem:

$$(K L K + \mu I) w = \lambda (K H K) w$$

The columns of W are the eigenvectors corresponding to the m smallest eigenvalues.

5.4 Empirical Impact on Divergence

The application of TCA leads to a measurable reduction in the generalization gap.

- Result [1]: Without TCA, the target AUC is **0.698**. With TCA, it rises to **0.705**.
- **Analysis:** While the absolute gain (+0.007) is modest, it is statistically significant given the constraints of the "best8" feature set. More importantly, it demonstrates that the latent embeddings of the two hospitals, initially disjoint due to scanner differences, have

been brought closer in the RKHS. This reduction in MMD corresponds directly to a reduction in the $d_{\mathcal{H}\Delta\mathcal{H}}$ term of the bound.

6. Minimizing Adaptability Error λ: Cross-Domain Feature Stability

The third term, $\lambda = \min_{h} (\epsilon_S(h) + \epsilon_T(h))$, represents the error of the best possible joint hypothesis. If λ is large, domain adaptation is theoretically impossible because no single classifier works for both domains. This situation arises from **Concept Shift** ($P_S(Y|X) \neq P_T(Y|X)$).

6.1 Concept Shift in Pulmonary Nodules

In our study, concept shift is driven by latent variables such as granulomatous disease burden. A key feature like "Spiculation" (spiky edges) is a strong predictor of cancer in the Source (Hospital A). However, in the Target (Hospital B), which may have a higher prevalence of tuberculosis or different radiologist annotation standards, benign granulomas also present with spiculation. Thus, the function $f: \text{Spiculation} \to \text{Malignancy}$ differs between domains. Including this feature would increase λ.

6.2 The Role of Cross-Domain RFE

PANDA employs **Cross-Domain Recursive Feature Elimination (RFE)** to explicitly minimize λ. By intersecting feature importance rankings from the source with availability and stability constraints, we identify a subset $\mathcal{F}_{stable}$ (the "best8" features) where the conditional distributions $P(Y|X)$ are approximately invariant.

Algorithm Analysis:

1. **Input:** Full feature set $\mathcal{F}_{all}$ (63 features).
2. **Process:** Iteratively train TabPFN on Source, compute permutation importance, and remove the least important/stable feature.
3. Selection Criterion: The Cost-Effectiveness Index (CEI) 1 balances AUC with stability.

$$CEI(k) = \frac{AUC_k - 0.5}{\text{Cost}_k} + \text{Stability}_k$$

4. **Result:** The algorithm converges to 8 features: Age, Diameter, Lobulation, Spiculation, etc.

Theoretical Implication: By discarding features that are prone to concept shift (e.g., highly subjective morphology scores or scanner-dependent texture metrics), we effectively restrict the hypothesis class $\mathcal{H}$ to a subspace where the domains are congruent. Although this might slightly increase the intrinsic source error ϵ_S (by removing potentially useful but unstable information), it drastically reduces λ, leading to a tighter overall bound on target error. The empirical stability of the "best8" model across folds (Figure 5 [1]) confirms this reduction in variance.

7. Empirical Analysis: Linking Theory to Results

We now synthesize the theoretical components with the experimental data to provide a comprehensive view of PANDA's performance.

7.1 Source Domain Efficiency (ϵ_S)

Table 7.1 presents the performance on the Source Domain (Hospital A) using 10-fold cross-validation.

Model	**AUC**	**Accuracy**	**Recall**	**Theoretical Driver**
PANDA (TabPFN)	**0.829**	**0.746**	**0.846**	**PFN Prior (Sample Efficiency)**
LASSO LR	0.763	0.722	0.925	Linear Bias (High Bias)
Random Forest	0.752	0.698	0.842	Overfitting (High

				Variance)
XGBoost	0.742	0.678	0.787	Overfitting (High Variance)
Mayo Score	0.605	0.359	0.000	Mismatched Coefficients

Analysis: The substantial gap between PANDA (0.829) and XGBoost (0.742) highlights the failure of GBDTs in the small-sample regime ($N=295$). GBDTs require sufficient data to stabilize split statistics; without it, they overfit noise. TabPFN, leveraging its meta-learned prior, effectively "interpolates" the causal structure of the nodules without needing to learn the manifold from scratch. This ensures a low starting point for $\epsilon_S(h)$ in the generalization bound.

7.2 Target Domain Generalization (ϵ_T)

Table 7.2 presents the external validation results on Hospital B.

Model	**AUC**	**Recall**	**Adaptation Gap (ϵS−ϵT)**
PANDA (TCA)	**0.705**	**0.944**	**0.124**
PANDA (No-TCA)	0.698	0.888	0.131
Random Forest	0.632	0.854	0.120 (from lower base)
SVM	0.628	0.606	0.089 (from lower base)
Mayo Score	0.584	0.000	N/A

Analysis:

- **The Drop:** All models suffer a drop in AUC (~ 0.12), reflecting the unavoidable λ (concept shift) and residual $d_{\mathcal{H}\Delta\mathcal{H}}$.
- **TCA Effect:** The alignment (TCA) improves AUC from 0.698 to 0.705. While small, the more critical gain is in **Recall (Sensitivity)**, which jumps from 0.888 to **0.944**. This suggests that TCA alignment corrected a shift in the decision boundary that was causing false negatives. In screening contexts, sensitivity is paramount; thus, the alignment provided a clinically significant safety margin.
- **Baseline Failure:** Random Forest collapses to 0.632. This confirms that without the strong prior of TabPFN or the alignment of TCA, standard models are brittle. They likely learned source-specific interactions (e.g., specific age-size correlations valid only in Hospital A) that did not hold in Hospital B.

7.3 Generalization on TableShift (BRFSS Diabetes)

To verify that these findings are not specific to our private dataset, we analyze the performance on the TableShift BRFSS Diabetes task (Race Shift: White $\to$ Non-White).[1]

Model	ID AUC (White)	OOD AUC (Non-White)	Accuracy	Gap
PANDA + TCA	**0.809**	**0.804**	**0.848**	**-0.005**
PANDA (No UDA)	0.809	0.796	0.847	-0.013
XGBoost	0.815	0.783	0.840	-0.032
Decision Tree	0.680	0.566	0.720	-0.114

Analysis:

- **Robustness:** PANDA shows remarkable stability (Gap of only -0.005). The OOD AUC (0.804) is virtually identical to the ID AUC.
- **GBDT Degradation:** XGBoost drops by 0.032. This reinforces the hypothesis that GBDTs overfit to source-specific subpopulations (e.g., specific demographic correlations in the

White cohort) that do not generalize to the Non-White cohort.
- **Calibration:** The high accuracy (0.848) despite the prevalence shift (Diabetes: 12.5% vs 17.4%) indicates that the ensemble temperature scaling effectively managed the Label Shift component.

8. Discussion: Mechanisms, Limitations, and Future Work

8.1 The "Closed-World" Constraint

A critical limitation of PANDA—and indeed most domain adaptation frameworks—is the **Closed-World Assumption**. Our RFE and TCA methods assume that the shared feature set $\mathcal{F}_{\cap}$ contains sufficient information to predict Y. If the target domain introduces a *new* strong predictor (e.g., a genomic biomarker) not present in the source, PANDA cannot leverage it. In fact, by restricting the model to the "lowest common denominator" of features (the intersection), we theoretically cap the maximum achievable performance ($\epsilon_S(h^*)$) compared to a model trained on a richer, domain-specific schema. This is a deliberate trade-off: we sacrifice potential ceiling performance for guaranteed stability and transportability.

8.2 Label Shift and Calibration

The Ben-David bound does not explicitly account for $P_S(Y) \neq P_T(Y)$ in its standard form. However, label shift is a dominant source of error in clinical practice. PANDA's use of **temperature scaling** ($T=0.9$) acts as a heuristic correction. A more rigorous theoretical approach would involve **Importance Weighting**, where source samples are weighted by $\beta(y) = P_T(y)/P_S(y)$. Future iterations of PANDA should incorporate this density ratio estimation directly into the TCA loss matrix L to theoretically minimize label shift error.

8.3 Subgroup Analysis and Aleatoric Uncertainty

Our stratified analysis [1] revealed that performance drops significantly for small nodules (≤ 8mm, AUC 0.65) compared to large ones (>8mm, AUC 0.74). This is likely an instance of **Aleatoric Uncertainty**—the irreducible error inherent in the data. Small nodules simply lack sufficient radiomic resolution to be definitively classified, regardless of the model or domain. This serves as a boundary condition for our theoretical expectations; no amount of domain adaptation can resolve information that is physically absent.

9. Conclusion

This chapter has provided a rigorous theoretical deconstruction of the PANDA framework. By mapping the architecture to the terms of the Ben-David generalization bound, we have shown that:

1. **TabPFN** minimizes Source Risk $\epsilon_S(h)$ via meta-learned Bayesian priors, solving the small-sample overfitting problem.
2. **Latent TCA** minimizes Domain Divergence $d_{\mathcal{H}\Delta\mathcal{H}}$ via spectral alignment in the linearized embedding space, solving the covariate shift problem.
3. **Cross-Domain RFE** minimizes Adaptability Error λ via stability-based pruning, mitigating the concept shift problem.

The empirical results validate this theory, demonstrating that PANDA not only outperforms classical baselines in AUC but also delivers the sensitivity and stability required for safe clinical deployment. This establishes PANDA not just as a successful engineering heuristic, but as a principled statistical solution to the problem of cross-hospital generalization in small-data regimes.

Works cited

1. main.pdf
2. Generalization Bounds for Domain Adaptation - PMC - NIH, accessed November 26, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC4191871/
3. Domain adaptation via transfer component analysis - PubMed, accessed November 26, 2025, https://pubmed.ncbi.nlm.nih.gov/21095864/
4. Generalization Bounds for Domain Adaptation - NIPS papers, accessed November 26, 2025, http://papers.neurips.cc/paper/4684-generalization-bounds-for-domain-adaptation.pdf
5. A theory of learning from different domains - eScholarship, accessed November

26, 2025, https://escholarship.org/uc/item/2nv1j9sc
6. Accurate predictions on small data with a tabular foundation model - PubMed, accessed November 26, 2025, https://pubmed.ncbi.nlm.nih.gov/39780007/
7. TabPFN Unleashed: A Scalable and Effective Solution to Tabular Classification Problems, accessed November 26, 2025, https://arxiv.org/html/2502.02527v1
8. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second - Table Representation Learning Workshop, accessed November 26, 2025, https://table-representation-learning.github.io/assets/papers/tabpfn_a_transformer_that_solv.pdf
9. TabPFN: One Model to Rule Them All? - arXiv, accessed November 26, 2025, https://arxiv.org/html/2505.20003v1
10. Position: The Future of Bayesian Prediction Is Prior-Fitted - OpenReview, accessed November 26, 2025, https://openreview.net/pdf?id=5Hpm74b1Ga
11. Linear kernel and non-linear kernel for support vector machine? - Cross Validated, accessed November 26, 2025, https://stats.stackexchange.com/questions/73032/linear-kernel-and-non-linear-kernel-for-support-vector-machine
12. A priori decision for a linear vs RBF Kernel SVM - Cross Validated - Stats StackExchange, accessed November 26, 2025, https://stats.stackexchange.com/questions/123346/a-priori-decision-for-a-linear-vs-rbf-kernel-svm
13. A Bearing Fault Diagnosis Method Based on Improved Transfer Component Analysis and Deep Belief Network - MDPI, accessed November 26, 2025, https://www.mdpi.com/2076-3417/14/5/1973