Transforming Diagnosis through Advanced Machine Learning and Data Analytics

Qingyuan Liu[1]

[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China

Abstract

Accurate diagnostic risk prediction increasingly depends on machine learning models that can operate reliably across hospitals, populations, and data-collection protocols. In practice, however, clinical prediction models are often trained on small, imbalanced tabular cohorts with heterogeneous feature schemas and substantial distribution shift, so performance deteriorates once they leave the development site. To address this gap, we propose PANDA (Pretrained Adaptation Network with Domain Alignment), a framework that combines a pre-trained tabular foundation model with domain-aware feature selection and unsupervised alignment to stabilize diagnostic decision support. PANDA uses a TabPFN-style transformer backbone meta-trained on synthetic tabular tasks to provide data-efficient priors; a cross-cohort recursive feature elimination step identifies a compact set of shared biomarkers that remain predictive across sites; and Transfer Component Analysis (TCA) projects source and target cohorts into a shared latent space using unlabeled target data, mitigating covariate shift under strict privacy constraints.

We evaluate PANDA in two representative diagnostic settings. For cross-hospital pulmonary nodule malignancy prediction using structured clinical data from two Chinese cancer centers, PANDA achieves an internal AUC of 0.811 and, with TCA, attains an external AUC of 0.705 and recall of 0.944 on the target hospital, outperforming tree ensembles and classical clinical scores such as Mayo and PKUPH (AUC < 0.64). In the TableShift BRFSS Diabetes benchmark with a race-driven shift from White to non-White respondents, PANDA with TCA reaches an out-of-distribution AUC of 0.804 with only a -0.009 gap relative to in-distribution performance, while standard gradient boosting and random forests exhibit larger degradation. Together, these results indicate that pairing tabular foundation-model priors with cross-domain feature pruning and lightweight statistical alignment can make diagnostic machine learning systems more robust and data-efficient in realistic cross-hospital and cross-population deployments.

INSERT_TOC_HERE

1 Introduction

Accurate diagnostic risk prediction is a canonical setting in which advances in machine learning (ML) and data analytics can have immediate clinical impact. In pulmonary nodule screening, classical clinical risk scores such as the Mayo Clinic, Veterans Affairs, Brock (PanCan), PKUPH, and Li models achieve strong internal discrimination (AUC $\approx$ 0.80–0.94) by fitting logistic regressions to carefully curated, single-center cohorts [1, 2, 3, 4, 5, 6]. However, meta-analyses and external validations show that their performance can decline to AUCs of 0.60–0.75 when transported to community-screening sites, TB-endemic regions, or demographically distinct populations [7, 5, 6]. These degradations, driven by shifts in disease prevalence, acquisition protocols, and background pathology (e.g., granulomas versus tuberculosis), illustrate how non-adaptive risk calculators can become unreliable in cross-hospital practice.

From an AI perspective, these failures reflect a mismatch between the complexity of real-world deployment and the simplifying assumptions of classical supervised learning. Clinical tabular datasets are typically small, imbalanced, and heterogeneous: even high-value registries often contain only a few hundred labeled patients, with malignant nodules representing a minority class. Features are high-dimensional and only partially overlapping across sites, as institutions log different biomarker panels, coding schemes, and acquisition protocols. This combination of small sample size, distribution shift, and feature-space mismatch violates the closed-world assumptions underlying many standard models and exposes the limits of purely local training.

The algorithmic trajectory for structured data in healthcare mirrors this tension. Gradient-boosted decision trees (GBDTs), led by XGBoost, LightGBM, and related ensembles, remain the workhorses of tabular ML because they tolerate heterogeneous scales, missingness, and noisy categorical codes [8, 9]. Neural "deep tabular" architectures—including TabNet, TabTransformer, SAINT, FT-Transformer, NODE, and other attention- or gating-based variants—extend differentiability to structured data and enable multimodal fusion, but they require substantial data, are sensitive to hyperparameters, and often lag well-tuned trees on clinical benchmarks when effective sample sizes are small [10, 11, 12, 13, 9]. Radiomics pipelines engineer thousands of texture descriptors from CT volumes, and 3D convolutional neural networks (CNNs) achieve strong internal performance on datasets such as NLST and LIDC, yet their scanner sensitivity and propensity for shortcut learning often negate cross-site gains: external validations reveal double-digit AUC drops when voxel spacing, reconstruction kernels, or case mix shift, and models can latch onto hospital-specific artifacts rather than biological signals [14, 15, 7, 16].

More recently, tabular foundation models and large tabular language models have emerged as promising directions. TabPFN meta-learns a transformer that approximates Bayesian posterior predictions across millions of synthetic tabular tasks, delivering hyperparameter-free, small-sample inference via in-context learning [17, 18]. Successors such as TabPFN-2.5 and drift-resilient TabPFN extend context length, relax attention bottlenecks, and incorporate simulated drifts into the prior [19, 20, 21]. Other work explores more realistic priors and cross-domain training curricula [22, 20], and "TabLLM"-style approaches serialize rows into prompts to reuse general-purpose reasoning from large language models [23]. Parallel efforts investigate federated optimization and continual or on-device learning so that models can absorb new hospital evidence without breaching privacy constraints [24, 25]. Collectively, these developments define a new generation of AI systems for tabular healthcare data.

However, cross-hospital transfer remains fragile because three dominant pathologies of medical tabular data co-occur. First, sample scarcity: most pulmonary nodule cohorts contain only a few hundred labeled patients, which limits the stability of purely supervised training and amplifies overfitting [13]. Second, distribution shift: label prevalence, scanner kernels, demographics, and clinical workflows alter the marginal $P(X)$ and even the conditional $P(Y \mid X)$ between hospitals [26, 27]. Third, feature heterogeneity: sites log disjoint biomarker panels, adopt different measurement units, and follow distinct coding policies, which invalidates naive feature alignment and introduces missingness shifts [28]. Domain adaptation research in imaging and wearables shows that adversarial training, cycle-consistent style transfer, optimal transport, and statistical moment matching can recover some performance under shift [24, 29], but these methods are rarely specialized for structured clinical data. Benchmarks such as TableShift and Wild-Time demonstrate that off-the-shelf robustness mechanisms still incur large out-of-distribution (OOD) gaps even when in-distribution accuracy is high [30, 29]. Large-scale regulators and hospital governance boards increasingly regard shift detection, recalibration, and drift monitoring as core AI-safety requirements rather than optional post hoc checks [26].

Tabular foundation models partially alleviate data scarcity, yet they inherit a closed-world assump-

tion: the context set used during in-context learning is assumed to reflect the same joint distribution and feature schema as the query samples [18]. When shifts in biomarkers, acquisition settings, or schemata emerge, attention weights may anchor on non-comparable neighbors, yielding overconfident yet incorrect predictions [20, 21]. Emerging variants such as TabPFN-2.5 and drift-resilient TabPFN extend context length and inject synthetic drifts into the prior [19, 21], but they remain sensitive to mismatched feature spaces and unlabeled target domains in the absence of explicit alignment. Tabular LLM approaches add reasoning capacity but incur substantial latency, quantization error for numerical values, and lack built-in clinical calibration, especially when lab panels or race-specific prevalences deviate from training distributions [23]. Consequently, bridging the gap between high internal accuracy and safe cross-site deployment requires combining foundation models with principled unsupervised domain adaptation and feature selection that respect clinical realities.

Pulmonary nodule malignancy prediction is an archetypal stress test for these issues. Traditional clinical scores and their LASSO or GBDT successors were derived from narrowly defined cohorts with fixed demographic profiles and scanner protocols, so their coefficients silently encode source-specific prevalence, upper-lobe priors, and calcification heuristics [1, 2, 3, 4, 5]. Meta-analyses across Asian screening programs and European cancer centers show that the same score threshold yields widely varying sensitivities (50–90

Similar tensions arise in population-health settings such as the BRFSS race-shift diabetes task. Demographic composition, socioeconomic exposures, and survey-year wording alter the marginal distribution of risk factors, while diabetes prevalence rises from roughly 12.5

Feature engineering and feature selection choices are therefore as important as model class. Clinical tables mix continuous laboratory values, ordinal scores, sparse categorical codes, and structured missingness; naive one-hot encoding can expand dimensionality and encode site-specific artifacts. Stability-driven feature pruning, hierarchical encoding of categorical variables, and unit-aware normalization reduce spurious site signatures and focus attention on shared, clinically interpretable signals [31]. Recursive feature elimination (RFE) across domains further enforces schema overlap, trading a slight drop in ceiling accuracy for substantial gains in portability when hospitals differ, and it is particularly helpful in small-sample, high-dimensional, and imbalanced regimes such as pulmonary nodules and radiomics panels [31, 13].

Taken together, the research gap is stark. Tree ensembles and deep tabular networks struggle with small, heterogeneous cohorts and typically require retraining when schemas change [8, 9, 13]. Foundation models improve small-sample performance but assume matched domains and aligned schemas [17, 18, 20]. Generic domain adaptation methods rarely account for missing features, label drift, or unlabeled targets in clinical tables [24, 29, 30, 26]. Federated and continual learning strategies help with privacy and incremental updates but do not by themselves guarantee cross-hospital calibration [24, 25]. A credible solution must (i) retain sample efficiency via strong pre-trained priors, (ii) discard site-specific signals that cannot transfer, and (iii) align source and target representations without target labels, while exposing calibration behavior under prevalence drift.

This study therefore adopts a pragmatic stance and introduces *PANDA* (Pretrained Adaptation Network with Domain Alignment), a framework designed to transform diagnostic prediction through advanced ML and data analytics in realistic cross-hospital settings. PANDA chains three complementary components. First, a pre-trained tabular foundation model (TabPFN) supplies strong inductive priors for small cohorts by meta-learning across millions of synthetic tasks and enabling hyperparameter-free inference [17, 18]. Second, cross-domain RFE prunes to biomarkers that are consistently available and stable across sites, mitigating schema mismatch and hospital-specific artifacts [31]. Third, a statistical alignment module based on Transfer Component Analysis (TCA) projects source and target cohorts into a shared reproducing-kernel subspace using unlabeled target data, minimizing distributional divergence while preserving clinical variance [32]. PANDA targets the explicit goal of cross-hospital pulmonary nodule prediction with screening-level sensitivity and is further validated on the TableShift BRFSS Diabetes race-shift benchmark [30], ensuring that the proposed approach addresses both clinical and population-level distribution shifts without adding bespoke modeling components for each dataset.

In summary, cross-hospital pulmonary nodule prediction and BRFSS race-shift diabetes prediction expose the same deployment realities: privacy constraints, schema mismatch, prevalence drift, and the need for sensitivity at clinically actionable thresholds [26, 30]. Existing AI toolkits—tree ensembles, deep tabular networks, tabular foundation models, tabular LLMs, and generic domain

adaptation—each leave gaps relative to these constraints [8, 10, 11, 12, 13, 18, 24, 29]. By integrating pre-trained tabular priors, schema-aware feature selection, and unsupervised domain alignment into a single pipeline, PANDA aims to restore calibration and discrimination under realistic deployment shifts. The remainder of this manuscript formalizes the cross-domain problem, surveys related work in tabular learning and medical domain adaptation, and presents PANDA as a practical instantiation of this design philosophy.

2 Related Work

In this section, we review prior work from an AI perspective on cross-hospital diagnostic risk prediction using structured medical data. We organize the literature along five dimensions. First, we summarize model families for medical tabular data, including tree ensembles, deep tabular architectures, and tabular foundation models. Second, we discuss domain shift and domain adaptation methods in medical AI. Third, we review feature selection techniques for small, imbalanced, and heterogeneous cohorts. Fourth, we situate AI-based approaches for pulmonary nodule malignancy prediction within this broader landscape. Finally, we examine public benchmarks that expose cross-domain and temporal shifts in tabular data. This structure parallels the design of our proposed framework, PANDA, which integrates tabular foundation models, domain-aware feature selection, and kernel-based alignment to address the limitations identified in prior work.

2.1 Tabular learning for medical data: tree ensembles, deep tabular networks, and tabular foundation models

The literature on structured-data learning has progressed from classical ensembles to deep tabular networks and, most recently, to tabular foundation models that mirror the trends in NLP and computer vision [33, 18]. We separate the discussion into tree ensembles, deep tabular architectures, and tabular foundation models to highlight where each excels and why none alone solves cross-hospital robustness. In medical settings, the same patient cohort may be modeled by tree ensembles, deep tabular networks, or foundation models depending on sample size and operational constraints; understanding their respective failure modes under domain shift is crucial for positioning PANDA.

2.1.1 Tree ensembles for clinical tabular data

Gradient-boosted decision trees (GBDTs) such as XGBoost, LightGBM, and CatBoost remain the workhorses for EHR-style tables because they tolerate heterogeneous scales, missing values, and noisy categorical codes while supporting monotone constraints and other clinical priors [8, 34, 13]. Benchmarking studies covering hundreds of OpenML tasks show that GBDTs still beat most neural baselines whenever training samples exceed a few thousand, yet they overfit rapidly when $N < 1{,}000$, cannot be fine-tuned incrementally, and require full retraining when hospitals change their feature schemas [9, 35, 36]. Case reports on cross-institutional readmission and mortality prediction show that tree models memorize acquisition artifacts (assay vendors, coding practices) and lose 10–20 AUC points when transferred without recalibration, illustrating their non-differentiable structure blocks end-to-end multimodal training and plug-and-play domain adaptation [37, 38]. This rigidity motivates attempts to distill tree priors into differentiable encoders so that adaptation can occur without rebuilding the model for each site. These same inductive biases explain why trees dominate mid-scale public benchmarks yet struggle in small, imbalanced medical cohorts: sparsity-aware splits handle missing labs gracefully, but boosting magnifies noise when positive classes are rare and hospital-specific priors leak into leaf structure. Because gradients stop at each split, trees cannot share representations with image encoders or participate in gradient-based domain adaptation, forcing manual feature harmonization whenever schemas or prevalence shift. In practice, this means that widely used implementations such as XGBoost and LightGBM shine on medium-to-large EHR cohorts with thousands of patients and hundreds of features, where sparse histogram-based splits and built-in handling of missing indicators yield strong baselines with modest tuning. Studies across OpenML, MIMIC-style EHR benchmarks, and TableShift-like suites repeatedly show that properly regularized XGBoost variants achieve AUROCs in the high 0.70s to low 0.80s for mortality, readmission, and sepsis detection, and that they

retain their ranking power even when categorical encodings or measurement scales differ across hospitals [9, 36, 30]. These observations explain why tree ensembles remain the default choice for operational clinical decision support systems.

On the small, heavily imbalanced cohorts typical of lung-screening registries ($N \approx 300$), the same capacity becomes a liability. Empirical analyses of GBDT behavior on few-shot medical datasets reveal several critical failure modes: first, deep trees can memorize the few malignant cases (often <50 positives), leading to apparent training accuracies >95% but test AUROCs collapsing once covariates shift [36, 30]. Second, calibration deteriorates dramatically in low-prevalence subgroups, with predicted probabilities systematically overestimating malignancy in young non-smokers while underestimating risk in elderly or high-burden cohorts [27, 26]. Third, when new hospitals add or remove variables (e.g., different CT protocol parameters or biomarker panels), there is no principled way to "warm start" or incrementally fine-tune existing tree models without complete retraining from scratch, making long-term maintenance expensive.

Because tree ensembles are non-differentiable and lack explicit latent representations, they are also difficult to integrate into end-to-end multimodal models or to pair with standard domain adaptation objectives. This mathematical constraint has practical consequences: researchers attempting to combine XGBoost with imaging features must resort to late fusion (averaging predictions) or feature concatenation followed by retraining, both of which preserve the non-differentiable barrier. Gradient-based domain adaptation methods such as Domain Adversarial Neural Networks (DANN) or Maximum Mean Discrepancy (MMD) regularization cannot be applied directly to tree models, requiring workarounds that approximate tree decision surfaces with differentiable surrogates or hybrid architectures that mix neural embeddings with gradient-boosted leaves. This limitation motivates methods that transfer tree-like priors into differentiable architectures or that use tree models as feature extractors rather than end-to-end learners.

2.1.2 Deep tabular networks

Deep tabular architectures import attention and representation learning from sequence models to overcome the adaptation gap. TabNet uses sequential feature masks to mimic decision paths, TabTransformer contextualizes categorical embeddings, FT-Transformer tokenizes all features, and SAINT introduces intersample attention plus contrastive pre-training to borrow signal across patients [10, 39, 40, 12]. Basis Transformers, NODE variants, TabICL prompt-serialization, weight-prediction, and regularization schemes further explore the space between neural and symbolic models [41, 42, 43, 44, 45]. However, comprehensive surveys and multiple leaderboard studies report that these models remain data-hungry, sensitive to hyperparameters, and often trail tuned tree ensembles on small, heterogeneous cohorts typical of tertiary hospitals [36, 35, 46]. In external-hospital transfers, SAINT and FT-Transformer frequently degrade to near-random calibration when categorical codes shift or when batch-size constraints prevent stable intersample attention. The computational footprint (long training times, GPU memory pressure) further limits adoption in clinical IT stacks, where inference latency and cost dominate. Empirical comparisons on clinical risk prediction echo this pattern with concrete performance gaps. TabNet often needs extensive learning-rate scheduling and sparsity penalties to match GBDT, and TabTransformer under-utilizes numerical biomarkers unless carefully normalized. FT-Transformer narrows the gap by embedding every feature, yet its quadratic self-attention becomes impractical for wide tables. SAINT's intersample attention helps when minibatches are large, but collapses on scarce data, making these models fragile without strong regularization and carefully tuned augmentations.

Detailed benchmarking studies on clinical datasets reveal stark contrasts between large-scale public benchmarks and real-world medical cohorts. On UCI repository datasets with >10,000 samples, TabNet achieves AUROCs of 0.85–0.92, FT-Transformer reaches 0.87–0.94, and SAINT obtains 0.86–0.93, competitive with or slightly exceeding XGBoost's 0.84–0.91 [9, 35]. However, when evaluated on authentic clinical cohorts with <1,000 patients and significant missingness, the same models show dramatic performance degradation: TabNet AUROCs fall to 0.62–0.71, FT-Transformer to 0.65–0.74, and SAINT to 0.60–0.69, while XGBoost maintains relatively stable performance at 0.75–0.83 [36, 46].

The computational requirements create additional barriers to clinical adoption. Training times for TabNet on a typical EHR dataset (1,000 patients, 50 features) range from 2-8 hours on a single GPU, compared to 5-15 minutes for XGBoost on CPU. FT-Transformer requires 4-12 hours due to its attention mechanisms, and SAINT needs 6-15 hours plus substantial memory for inter-sample at-

tention matrices [36]. These computational costs translate to practical challenges: most hospital IT environments lack GPU infrastructure for model development, and the extensive hyperparameter tuning required (learning rate schedules, attention head configurations, regularization strengths) demands specialized machine learning expertise not commonly available in clinical settings.

Furthermore, calibration studies reveal that deep tabular models often produce overconfident predictions on medical data. Reliability diagram analyses show that TabNet and FT-Transformer consistently assign higher predicted probabilities than warranted by observed outcomes, particularly in rare disease subsets where expected calibration errors (ECE) can exceed 0.15–0.20 compared to XGBoost's 0.04–0.08 [27, 46]. This overconfidence is particularly problematic for clinical decision support, where well-calibrated risk estimates are essential for appropriate triage and treatment decisions. These limitations are amplified in clinical registries where hundreds of variables encode comorbidities, medication history, and laboratory trajectories. Studies on ICU mortality, sepsis, and readmission prediction report that deep tabular networks match or slightly exceed tuned GBDTs on in-distribution test sets but lose their advantage when evaluated on later time periods or new hospitals, especially when categorical vocabularies change or when privacy constraints cap batch sizes [9, 27, 46]. In such small-N, high-dimensional regimes, hyperparameter sensitivity translates directly into clinical risk: minor changes in learning rate or regularization can flip decisions near treatment thresholds. Compared with tree ensembles, these architectures seek to learn shared feature representations that might in principle adapt across hospitals or tasks. In practice, however, their appetite for data and tuning means that performance gains are often limited to large industrial benchmarks; on noisy, heterogeneous medical tables with only a few hundred patients, they frequently underperform simpler models and exhibit brittle calibration under shift. This contrast sets the stage for tabular foundation models such as TabPFN, which embrace a meta-learning, few-shot perspective instead of training a new deep network from scratch for each cohort.

2.1.3 Tabular foundation models

Tabular foundation models push self-supervised pre-training and in-context learning into structured data. TabPFN meta-trains a transformer on millions of synthetic datasets sampled from diverse structural-causal priors, learns to approximate posterior predictive distributions, and performs inference via a single forward pass without gradient updates [17, 47]. Follow-up work expands its reach without breaking the closed-world assumption: TabPFN-2.5 relaxes quadratic attention to accommodate tens of thousands of context rows and documents an augmented pre-training suite; diagnostics such as "A Closer Look at TabPFN v2" show that the model remains overconfident under covariate shift, prompting wrappers that adjust representations before prediction [19, 22, 20]. Drift-resilient variants model temporal shift with secondary structural-causal modules and record measurable gains when patient mixes evolve [48, 21]. Other studies adapt the same prior-learning paradigm to drug discovery, radiomics, and graph embeddings, highlighting both the portability and fragility of tabular foundation models beyond flat tables [49, 23, 50]. Tabular Large Language Models (TabLLMs) serialize rows or mini tables into prompts so that general-purpose LLMs can reason over discrete entries, but they remain computationally prohibitive for high-throughput risk prediction and struggle with precise numeric calibration [51, 23, 52]. Recent analyses of high-dimensional omics applications reinforce that even TabPFN requires aggressive feature selection or prior-guided embeddings to stay calibrated, underscoring its closed-world assumption [53, 54]. PFN-Boost, LLM-Boost, and hybrid residual schemes blend foundation backbones with tree-style updates or prompts, but benchmark reports such as Wild-Tab still find overfitting to training-domain quirks unless explicit alignment and calibration are layered on [55, 38, 42]. Closed-world constraints surface in three ways: (i) feature mismatch—TabPFN expects aligned schemas and cannot reason about biomarkers absent from the context; (ii) covariate drift—attention retrieves misleading neighbors when acquisition protocols move, producing overconfident errors; and (iii) context-length bottlenecks that force sub-sampling when rows exceed a few thousand. These limits explain why prior studies resort to RFE or hand-crafted embeddings before invoking TabPFN and why drift-resilient variants add causal dynamics to temper temporal shift.

These observations motivate hybrid approaches that explicitly combine strong priors with domain-alignment hooks. Table 1 summarizes the comparative strengths and weaknesses of these model families for medical tabular tasks, highlighting why PANDA fuses TabPFN with feature selection and unsupervised alignment instead of relying on any single paradigm.

Table 1: Comparative strengths and weaknesses of tabular model families in medical AI.

Model Class	Representative Algorithms	Strengths in Medical AI	Limitations in Cross-Hospital Tasks
Tree Ensembles	XGBoost, LightGBM, CatBoost	Interpretable, robust to missingness/outliers, encode clinical constraints	Overfit small cohorts, non-differentiable, no inherent transfer learning, require full retraining per site
Deep Tabular	TabNet, TabTransformer, FT-Transformer, SAINT, NODE	Differentiable, capture complex interactions, allow multimodal fusion	Data hungry, extensive tuning, high compute cost, brittle without alignment
Foundation Models	TabPFN, TabPFN-2.5, TabLLM	Hyperparameter-free inference, strong small-N priors, probabilistic outputs	Sensitive to distribution/feature shift, limited context length, assume aligned schemas

2.2 Domain shift and domain adaptation in medical AI

Domain adaptation (DA) provides the vocabulary for managing the covariate, label, and concept shifts that materialize when AI crosses hospital boundaries. Classical analysis decomposes target error into source error plus a divergence term, motivating alignments and invariance objectives. In practice, medical deployments encounter overlapping types of shift: changes in patient mix and ordering policies alter $P(X)$, new screening programs or diagnostic criteria perturb $P(Y)$, and evolving clinical practice modifies $P(Y \mid X)$ [26, 27]. Pulmonary nodule malignancy prediction is particularly exposed to this triad of shifts because granulomatous disease burden, scanner protocols, and radiologist thresholds vary sharply across regions.

2.2.1 Statistical alignment vs. adversarial objectives

Maximum Mean Discrepancy (as in TCA), correlation alignment (CORAL), and transport-based projections minimize moment discrepancies in a latent space [32, 56, 57, 58, 59]. They are attractive for medical tables because they offer closed-form or deterministic solutions and remain stable when labeled target data are absent. Adversarial approaches (DANN, cycle-consistent style transfer) attempt to erase domain cues via discriminators, but surveys show they destabilize when cohorts are tiny, leading to mode collapse or erasure of clinically salient signals [27, 58, 24]. In ICU mortality and readmission tasks, DANN can underperform ERM by wide margins because the discriminator trivially detects domain cues from missing patterns, causing the encoder to discard predictive features. In contrast, MMD- or CORAL-style alignment improves calibration modestly and avoids catastrophic degradation, motivating our reliance on TCA for small-sample settings. Classic error decompositions also separate covariate shift ($P_s(X) \neq P_t(X)$) from label shift ($P_s(Y) \neq P_t(Y)$) and concept shift ($P_s(Y|X) \neq P_t(Y|X)$); only the first benefits cleanly from moment matching, while the second demands prevalence-aware calibration and the third often needs feature auditing or human review [32, 30, 26]. These regimes frequently co-occur in multi-hospital deployments, explaining why single DA objectives show mixed results.

2.2.2 Heterogeneity, missingness, and temporal drift

Medical DA must grapple with heterogeneous feature sets and evolving acquisition policies. Feature-space DA (FSDA) and transport-based alignment project source and target into shared latent spaces, while open-set domain adaptation handles mismatched label spaces and schema drift that arise when hospitals collect different labs [60, 57, 61, 59]. DomainATM, feature-aware PCA, and ontological mapping frameworks first identify which biomarkers are stable across sites before alignment, reducing negative transfer [24, 62, 63]. Missingness-shift studies demonstrate that when ordering policies change (e.g., different lab panels for triage), standard covariate-shift assumptions break; MNAR-aware corrections and explicit missingness modeling become mandatory [64, 65]. Temporal adaptation work (Wild-Time, multi-attention encoders for COVID-19) highlights that drift accumulates over months, so models require continual recalibration rather than one-time transfer [29, 66].

2.2.3 Domain generalization and open-set gaps

TableShift, Wild-Tab, and Wild-Time benchmarks quantify how far models fall once distributions move: they reveal a near-linear relation between in-distribution and out-of-distribution accuracy, but also show that label shift dominates error budgets and that prevailing domain-generalization objectives (GroupDRO, IRM, VREx) rarely beat strong ERM or GBDT baselines on tabular data [30, 67, 68, 55, 29]. Open-set and partial-label settings are common in healthcare (target hospital omits certain comorbidities); current DA methods often assume aligned label spaces and therefore miscalibrate rare conditions. Regulatory guidance now treats shift detection and recalibration as part of post-market surveillance, emphasizing that robustness must be engineered rather than assumed [26]. Complementary benchmarks and surveys on generic tabular learning echo these findings: across hundreds of datasets, tuned GBDTs remain exceptionally strong baselines, and many deep or domain-generalization architectures fail to deliver consistent gains once evaluation moves beyond a handful of leaderboard tasks [36, 35, 45]. Moreover, empirical decompositions of error budgets highlight that label shift and calibration drift often dominate covariate shift, suggesting that feature-space alignment alone is insufficient for reliable deployment. Together with the medical DA literature, these results argue for methods that combine strong small-sample priors, explicit feature governance, and lightweight, task-aware alignment instead of relying on black-box "robust" architectures.

2.2.4 Domain adaptation and transfer learning for clinical tabular and EHR data

Recent work brings these ideas to longitudinal EHR and claims data. AdaDiag-style methods align source and target hospitals in a representation space while jointly training prognostic models, reporting partial recovery of AUROC lost when models trained on MIMIC-like cohorts are evaluated at external centers [58, 24]. Multi-center EHR foundation models go further by pre-training sequence encoders on records from dozens of institutions and then fine-tuning on downstream tasks, demonstrating that shared representations can reduce the amount of labeled data required for local adaptation [69]. These approaches show that both unsupervised alignment and transfer learning have value in clinical AI, but they typically assume abundant longitudinal data, focus on large hospitals with rich EHR infrastructure, and operate on sequential rather than static tabular summaries.

Standard domain-adaptation theory provides a unifying lens: target risk can be bounded by source risk plus a measure of distribution discrepancy and a term capturing irreducible label-set differences [32, 27]. Reducing error on the source domain alone is therefore insufficient; one must also control divergence between source and target feature distributions, for example via moment-matching, adversarial objectives, or feature-space DA. Beyond centralized settings, federated learning extends these ideas by allowing multiple hospitals to collaborate without sharing raw data. Surveys on federated learning for medical imaging and pattern recognition summarize how FL can pool experience across institutions while preserving privacy, and methods such as FedFusion explicitly combine domain adaptation with personalized encoders to handle heterogeneous feature spaces and scarce labels [70, 71, 72]. However, most federated frameworks target high-volume imaging or EHR tasks, assume substantial local computation and at least some labeled data at each site, and still rely on shared model architectures and broadly aligned feature schemas. They are therefore complementary to, rather than a replacement for, lightweight DA strategies tailored to very small tabular cohorts with partially mismatched feature sets.

Table 2 summarizes the main DA families discussed above and their implications for cross-hospital tabular deployment.

Existing EHR-focused methods mostly address temporal drift or site differences in large cohorts, whereas our setting combines small, imbalanced tabular cohorts, heterogeneous feature sets, and unlabeled target hospitals. This gap motivates PANDA's combination of strong tabular priors, cross-domain feature selection, and lightweight alignment tailored to static risk scores rather than long EHR sequences.

2.3 Feature selection and domain-aware stability for small medical cohorts

High-dimensional yet small-sample tabular cohorts are ubiquitous in medicine: lung screening registries, omics panels, and survey-based risk scores often contain hundreds of variables for only a few hundred or thousand patients. Naïve learning in this regime leads to unstable decision boundaries

Table 2: Representative domain-adaptation strategies in medical AI and their relevance to cross-hospital tabular risk prediction.

Method family	**Typical modality**	**Key assumptions**	**Pros**	**Limitations for small cross-hospital tabular cohorts**
Statistical alignment (MMD, TCA, CORAL, transport)	Tabular, EHR, imaging	Shared feature schema; access to source data and unlabeled target samples; primarily covariate shift	Closed-form or deterministic mappings; stable when target labels are absent; easy to plug into existing pipelines [32, 56, 57, 59]	Does not directly correct label or concept shift; assumes overlapping feature sets; may misalign rare subgroups without additional calibration [30, 26, 27]
Adversarial representation learning (DANN-style)	Imaging, EHR sequences	Access to source and target data with domain labels; discriminator encouraged to remove site identity	Learns domain-invariant representations jointly with task loss; flexible for complex modalities [27, 58]	Unstable on tiny cohorts; discriminators exploit missingness patterns, causing encoders to discard predictive features; can underperform ERM in ICU-style tasks [58, 24]
Feature-space DA and domain-aware FS (FSDA, DomainATM)	Tabular, EHR	At least partially shared feature space; access to both domains during training	Selects features that are predictive and stable across sites; reduces reliance on site-specific surrogates and noisy biomarkers [60, 24, 62, 63]	Still assumes sizable overlap in measured variables; does not natively handle missing entire feature blocks or unlabeled target hospitals with severe schema mismatch
Domain generalization and temporal adaptation (TableShift, Wild-Tab, Wild-Time)	Tabular	Multiple labeled source distributions; no target labels during training	Reveal failure modes under temporal, demographic, and institutional shift; provide standardized evaluation suites [30, 68, 55, 29]	Many domain-generalization objectives (e.g., GroupDRO, IRM, VREx) rarely beat strong ERM or GBDT baselines; benchmarks show label shift and calibration drift dominate what feature matching can fix [30, 67]
Federated and federated-DA frameworks (FL, FedFusion-style)	Imaging, tabular	Multiple compute-capable hospitals; communication budget; typically some local labels and shared model architecture [70, 71, 72]	Preserve data privacy while learning from distributed cohorts; can combine personalization with domain adaptation and label efficiency	Often require significant local computation and labeled target data; focus on large imaging or EHR tasks; do not directly address very small tabular cohorts with feature mismatch and strict label scarcity

and non-reproducible feature attributions. Feature selection methods aim to reduce dimensionality, stabilize inference, and focus clinician attention on biomarkers that are both predictive and economical to collect.

2.3.1 Small-sample and high-dimensional feature selection

Classical filter and wrapper methods, such as mutual information ranking or recursive feature elimination with SVMs, laid the groundwork for identifying compact biomarker sets but struggle when features are highly correlated or when class imbalance is severe [73]. More recent approaches explicitly target high-dimensional, low-sample-size settings. WPFS-style methods learn feature weights jointly with a classifier, GRACES uses graph convolutions to propagate importance across correlated features, and DeepFS leverages deep networks to screen features via nonlinear embeddings [74, 75, 76]. These techniques are attractive for medical AI because they can down-select from hundreds of candidate variables to a dozen stable predictors while controlling overfitting. Empirical studies on omics and imaging-genomics datasets show that such methods can maintain or even improve AUC while halving the number of features, directly reducing assay costs and simplifying model interpretation. However, most of these works assume a single training domain: the selected subset is optimized for internal performance and may not transfer when another hospital measures a slightly different panel or when missingness patterns change. From a methodological standpoint, this marks a shift from classical LASSO or univariate ranking—which rely on linear or marginal-effect assumptions and can be highly unstable in small cohorts—to architectures that explicitly model complex feature interactions and redundancy. WPFS and GRACES, for example, introduce auxiliary networks or graph structures to propagate importance across correlated features, while DeepFS leverages deep encoders to identify nonlinear manifolds where only a subset of variables drive variation [74, 75, 76]. These designs are particularly appealing in high-dimensional, sparse medical settings (omics panels, questionnaire data), but they still optimize for one domain at a time and do not ensure that the chosen biomarkers remain predictive under cross-hospital shift.

2.3.2 Feature selection with transformers and foundation models

Attention-based models provide an alternative route to feature selection by interpreting attention weights, learned masks, or perturbation scores as measures of importance. TabNet learns sparse feature masks that indicate which variables are consulted at each decision step, while transformer-based architectures expose token-level attention maps that can be aggregated across layers and heads [10, 11, 12, 45]. In practice, researchers often perform permutation-based importance estimation using a strong tabular backbone—GBDT or TabPFN—and then apply RFE-style pruning, retaining the top-k features that consistently contribute to performance. This paradigm is well-suited to small medical cohorts because it leverages the inductive biases of powerful models while regularizing the input space. For foundation models such as TabPFN, feature selection also mitigates closed-world constraints: by removing unstable or site-specific variables, one can reduce the chance that attention focuses on hospital identifiers or acquisition artifacts rather than pathology.

2.3.3 Domain-aware and cross-site feature selection

Standard feature selection treats all samples as exchangeable, implicitly assuming that feature-importance rankings are identical across domains. Domain-aware methods instead optimize a subset that is simultaneously predictive in multiple hospitals or under multiple sampling schemes. FSDA and related frameworks extend DA objectives with feature-level penalties, rewarding variables whose contributions remain stable after alignment [60, 31]. Multi-site studies on EHR and imaging data show that such cross-domain criteria can discard site-specific surrogates (e.g., local procedure codes) while preserving clinically meaningful biomarkers. PANDA adopts this philosophy in a pragmatic way: TabPFN is used as a strong scoring model, but feature elimination is guided jointly by source-site performance and cross-site stability, leading to a compact "best8" subset that is consistently informative in both hospitals. These domain-aware subsets provide low-dimensional, harmonized inputs to TCA, reducing the risk of negative transfer and making the subsequent alignment problem better posed. Viewed through this lens, feature selection becomes a form of implicit domain alignment: instead of matching full distributions in a high-dimensional space, one first discards variables whose predictive contribution

is strongly domain-specific and focuses on biomarkers that are consistently informative across sites. This is particularly valuable when hospitals measure different panels or exhibit pronounced missingness shift, because aligning on a smaller, shared subset of stable features is both statistically and operationally simpler. PANDA effectively instantiates this principle by using a pre-trained tabular foundation model to rank features jointly across two hospitals and retaining only those with robust importance, thereby coupling representation learning with domain-aware feature governance.

Table 3: Representative feature selection methods for small, imbalanced, high-dimensional biomedical tabular data.

Method	**Model family**	**Small-sample / imbalance handling**	**Interpretability characteristics**	**Representative biomedical use cases**
Recursive feature elimination (RFE) with linear or tree models	Wrapper around SVM, logistic regression, or tree ensembles	Wrapper search over feature subsets can overfit when N is small and features are correlated; often combined with cross-validation and class-balanced sampling	Produces explicit ranked feature lists and compact subsets; easy to inspect and map to clinical variables [73]	Widely used in early gene-expression and biomarker panels; basis for many clinical risk-score and radiomics pipelines
LASSO / elastic-net logistic regression	Embedded linear models	ℓ_1 or $\ell_1+\ell_2$ penalties shrink coefficients, providing some robustness to high dimensionality; still assumes linear log-odds and can be unstable under heavy collinearity	Sparse coefficients directly indicate selected features; compatible with odds-ratio interpretation familiar to clinicians [73]	Common in radiomics and EHR risk models where interpretability and coefficient-based reporting are required
GRACES	Graph-convolutional-network-based FS [74]	Specifically targets high-dimensional, low-sample-size data by modeling feature relations on a graph; alleviates overfitting compared with independent filters	Outputs a compact subset informed by graph structure; can be visualized as a network of interacting biomarkers	Demonstrated on omics-style datasets; suitable when prior knowledge or correlations between biomarkers are important
DeepFS	Deep feature screening with autoencoders [75, 76]	Uses deep encoders to learn low-dimensional representations and rank features, handling ultra-high-dimensional, sparse, and potentially imbalanced data	Provides importance scores for each original feature; retains flexibility to operate in supervised or unsupervised mode	Evaluated on synthetic and biomedical high-dimensional datasets; useful when the number of variables far exceeds the number of patients
Domain-aware FS (FSDA-style)	Feature selection for domain adaptation [60]	Encourages selection of features that remain predictive across domains, implicitly handling covariate shift between sites	Produces subsets that are jointly predictive and domain-stable, supporting cross-hospital deployment	Applied to benchmark DA tasks; conceptually aligned with cross-hospital biomarker selection in multi-center medical studies
Transformer / foundation-model-based FS	Attention- or score-based selection using TabNet, TabTransformer, and tabular foundation models [10, 11, 45]	Leverages high-capacity or pre-trained models to estimate nonlinear feature importance; can be combined with RFE to mitigate small-sample overfitting	Attention weights, feature masks, or permutation-based scores yield ranked features; aligns with explainable-AI practices	Increasingly used in biomedical tabular and omics datasets; PANDA's cross-cohort RFE uses a tabular foundation model as the scoring backbone

2.4 Pulmonary nodule malignancy prediction: from clinical scores to multimodal AI

2.4.1 Clinical risk scores and logistic models

Pulmonary nodule malignancy prediction is a canonical testbed for cross-domain robustness. Classical logistic scores—Mayo Clinic, Veterans Affairs, Brock (PanCan), PKUPH, Li, and derivatives—achieve internal AUCs above 0.85 but regularly drop to 0.60–0.80 in external validations, especially in Asian or community-screening cohorts where prevalence and granulomatous disease burdens diverge [1, 77, 2, 3, 4, 7, 5, 6]. These scores typically combine age, smoking history, nodule size, location, and morphology into a logit-based risk function. Meta-analyses covering tens of thousands of nodules

confirm that calibration deteriorates most severely in subgroups such as solitary upper-lobe nodules and specific ethnic groups, reflecting both label-shift and covariate-shift mechanisms [7, 5, 6, 78]. While recalibration or re-estimation of coefficients can partially restore performance, these fixes require local labels and do not address feature-mismatch: new hospitals may lack some variables (e.g., emphysema grading) or measure them differently.

Targeted audits make the degradation concrete. In TB-endemic Korean hospitals, Mayo and VA shrink to AUC ≈0.60 while Brock declines to ≈0.68 despite an internal AUC near 0.94, and Chinese multi-center studies find that Brock and PKUPH can fall from ≈0.90 internally to 0.70–0.77 once prevalence and granulomatous disease rates shift [79, 80, 81]. PET-augmented variants such as the Herder score raise internal discrimination to ≈0.92 by incorporating metabolic imaging, yet they lose specificity in TB-endemic or inflammatory regions where uptake is nonspecific [82, 79]. These case studies underscore that most clinical scores embed site-specific prevalence, referral patterns, and feature definitions, so "plug-and-play" deployment without alignment is unrealistic.

Each classical score carries its own design trade-offs. The Mayo Clinic model was derived from several hundred clinic-referred patients with indeterminate nodules, emphasizing age, smoking, nodule diameter, spiculation, and upper-lobe location, whereas the Veterans Affairs model targeted high-risk, predominantly male veterans with larger lesions [1, 77]. The Brock (PanCan) model was trained in a screening cohort enriched for small nodules and incorporates emphysema, family history, and more granular morphology descriptors, while the PKUPH and Li scores adapt similar feature sets to Chinese tertiary-hospital and screening populations [2, 3, 4, 6]. A recent meta-analysis focused on the Brock model reports pooled AUC ≈ 0.80 across $>$ 80,000 patients but highlights substantially lower performance in Asian cohorts, solitary nodules, subsolid nodules, and larger lesions (AUC often ≈ 0.74 or below), underscoring that apparent "universality" in development data masks sizeable domain-specific errors [78]. Across Mayo, VA, Brock, and PKUPH, external validations repeatedly document drops from internal c-statistics in the high-0.80s to 0.60–0.75 when applied to community screening or granulomatous-disease–endemic regions [7, 5, 6].

These patterns can be summarized along three axes: development cohorts are often single-center and demographically narrow; variables focus on easily collected clinical and simple CT descriptors; and the underlying model is a logistic regression that assumes a linear log-odds relationship between covariates and malignancy. Table 4 sketches representative scores along these dimensions. In development, all achieve reasonable discrimination and are simple enough to implement as bedside calculators, but the same simplicity makes them brittle under shift: logistic coefficients absorb local prevalence, imaging protocols, and referral patterns, so external use without recalibration results in systematic underestimation or overestimation of risk in particular subgroups.

Table 4: Representative pulmonary nodule malignancy scores and common external-validation issues.

Score	**Development cohort**	**Key variables**	**External-validation observations**
Mayo Clinic	Clinic-referred indeterminate nodules in smokers	Age, smoking history, nodule size, spiculation, upper-lobe location	Internal AUC in the high-0.80s; frequent overestimation of risk and AUC drops to ≈0.6–0.7 in screening and non-U.S. cohorts
Veterans Affairs	Predominantly male veterans with larger nodules	Age, smoking, nodule diameter, location	Good performance in veterans; miscalibration when transported to mixed-gender or lower-risk populations
Brock (PanCan)	CT screening cohort with many small nodules	Age, sex, family history, emphysema, size, type, location	Meta-analytic pooled AUC ≈ 0.80; markedly lower AUC in Asian, solitary, and subsolid nodules [78]
PKUPH / Li	Chinese tertiary-hospital and screening cohorts	Age, smoking, nodule size and type, lobulation, spiculation	High internal AUC but drops in external series; performance depends strongly on CT protocol and case mix [3, 4, 6]

From the perspective of this thesis, these scores provide clinically interpretable baselines and useful prior knowledge about which coarse-grained descriptors matter, but they do not solve cross-hospital robustness. Their small development cohorts and rigid functional form make it difficult to incorporate

new biomarkers or adapt to feature-mismatch without re-estimating the entire model, motivating more flexible tabular approaches that can share information across hospitals while respecting regulatory demands for calibration and subgroup transparency.

2.4.2 Radiomics pipelines with traditional machine learning

Radiomics pipelines extract hundreds to thousands of hand-crafted features from CT volumes, offering richer representations than clinical risk scores but introducing major reproducibility hazards. Texture and wavelet descriptors vary with voxel spacing, reconstruction kernel, and segmentation protocol; ComBat-style harmonization reduces scanner effects yet requires batch labels and can blur subtle lesions [16, 7]. In internal validation, radiomics-based classifiers that pair LASSO- or stability-selected feature subsets with SVMs, random forests, or GBDTs typically report AUCs in the 0.75–0.90 range, but these numbers rarely carry over to new scanners or hospitals. External validations on LIDC-IDRI, LUNA16, and NLST repeatedly report double-digit AUC drops when deployed to scanners with different kernels or patient mixes, while shortcut-learning analyses show that models sometimes rely on grid artifacts or reconstruction noise rather than morphology [14, 15, 26]. These failures illustrate that radiomics alone cannot guarantee transportability and that alignment plus feature vetting are required before cross-hospital use.

Concrete exemplars reinforce that fragility. The Bayesian Integrated Malignancy Calculator (BIMC) blended radiomics with clinical covariates and modestly outperformed Mayo, Brock, and PKUPH (AUC ≈0.90 vs. ≈0.78) on an Italian derivation cohort, yet its advantage diminished when scanners, slice thickness, or kernels changed [83]. Hawkins-style NLST radiomics achieved AUC ≈0.83 without external validation, and retrospective audits show that a single reconstruction tweak can reorder the features selected by LASSO [7, 16]. Radiomics therefore supplies richer morphology descriptors but still requires harmonization, feature governance, and domain-aware alignment rather than assuming reproducibility across hospitals.

Table 5: Representative radiomics-based pulmonary nodule malignancy models and reported generalization behavior.

Study / model	Imaging data	Centers / nodules	Classifier	Internal AUC	External AUC	Harmonization / scanner sensitivity
Generic radiomics pipelines	2D/3D chest CT or low-dose CT nodules	Single- or few-center cohorts (sizes vary)	LASSO- or stability-selected features with SVM, RF, or GBDT	Typically 0.75–0.90	0.10–0.20 lower	Performance degrades when kernel, slice thickness, or vendor changes; ComBat-style harmonization can reduce but not eliminate scanner effects [7, 16]
Multi-center reproducibility analyses	3D CT radiomics features across scanners	Multi-center CT datasets	–	–	–	Many texture features show intraclass correlation coefficients < 0.5 across vendors and protocols; harmonization helps but cannot fully restore stability [16]
Radiomics + clinical scoring models	Radiomics signatures combined with clinical descriptors	Hospital-specific or regional nodule cohorts	Elastic-net / logistic regression, RF, or GBDT	High (≥ 0.80)	Mid-0.70s or lower	External validations report double-digit AUC loss and sensitivity to case mix and acquisition protocol [7]

Published inter-scanner analyses often report intraclass correlation coefficients below 0.5 for entropy and run-length features, indicating poor reliability even before model fitting [16]. ComBat can regress out known batch effects when acquisition labels are available, but it can also blur subtle lesions and fails when batch membership is unknown at inference time, leaving a gap that tabular-alignment pipelines attempt to close. Beyond handcrafted features, many radiomics pipelines incorporate LASSO, elastic-net logistic regression, or stability-selection frameworks to shrink coefficients and stabilize feature sets before training SVM, random forest, or GBDT classifiers. Although these strategies help curb overfitting in small cohorts, they do not eliminate sensitivity to acquisition protocols: the same feature may be retained in one scanner configuration and discarded in another because its estimated importance changes with kernel or slice thickness. Multi-center studies frequently report 10–20 percentage-point AUC drops when models are transported without revisiting segmentation, feature extraction, and

harmonization choices [7, 16]. As a result, radiomics pipelines tend to behave like carefully tuned, center-specific instruments rather than plug-and-play risk predictors, and their complexity makes it hard for clinicians to trace failure modes back to specific preprocessing or feature-engineering steps.

2.4.3 Deep-learning CAD systems

End-to-end deep-learning computer-aided diagnosis (CAD) systems extend the radiomics pipeline by learning 3D convolutional representations directly from CT volumes or multi-view patches. Large-scale screening trials such as NLST have enabled 3D CNNs to achieve AUCs in the mid-0.90s on internal validation, sometimes matching or surpassing expert radiologists [14]. Subsequent works combine deep features with handcrafted radiomics or clinical covariates, showing further gains on curated datasets [84, 85]. Causey et al.'s NoduleX reproduced radiologist malignancy ratings with AUC $\approx$0.99 on LIDC-IDRI, and Google's NLST-scale 3D CNN maintained AUC $\approx$0.94 on an independent hospital cohort of 1,139 CTs, illustrating how massive, homogeneous datasets can suppress variance [86, 14]. However, these successes often rely on tightly controlled acquisition protocols and substantial annotation effort. External validations reveal double-digit AUC drops when voxel spacing, reconstruction kernels, or vendor mix shift, and shortcut-learning analyses demonstrate that CNNs may rely on markers, reconstruction noise, or scanner metadata rather than nodule morphology [15, 16, 26]. Moreover, most deep CAD systems treat imaging in isolation or only append a handful of clinical variables, limiting their ability to reason over complex comorbidity profiles or laboratory trajectories. Multi-view and multi-scale architectures that process cropped nodules, surrounding parenchyma, and whole-lung context can mitigate some of these issues, but they further increase computational cost and annotation effort. Multi-task variants that jointly predict malignancy, growth, or histological subtype promise richer supervision but require large, carefully curated datasets that few hospitals possess. In practice, many published CAD systems are trained and tuned on a single trial or institution, with limited reporting on cross-hospital generalization or calibration. Where multi-center experiments are reported, performance is typically rescued by site-specific fine-tuning on labeled cases from each target hospital, and very few studies attempt label-free "train at A, deploy at B" deployment. As a result, deep CAD systems remain powerful local tools rather than robust cross-hospital risk predictors.

2.4.4 Tabular and multi-modal nodule models

Later machine-learning models—LASSO, random forests, GBDTs, Bayesian networks, and hybrid radiomics-clinical models—attempt to combine the strengths of scores and imaging [4, 7, 84, 85]. GBDTs and random forests improve internal calibration and handle nonlinear interactions but still require site-specific recalibration or feature mapping before deployment because their learned weights implicitly encode scanner kernels and local smoking histories. Multi-modal models that fuse deep image features with clinical covariates via late fusion or stacking demonstrate promising gains on LIDC-IDRI and NLST, yet most studies remain single-center or rely on random train–test splits that do not reflect real cross-hospital deployment. Only a handful of works evaluate performance when training on one hospital and testing on another, and these typically report substantial AUC drops and unstable decision thresholds [7, 16]. Recent multi-center studies in Asian and Chinese screening cohorts echo this pattern: even when models are re-estimated or augmented with additional imaging features for new hospitals, external AUCs often plateau in the low- to mid-0.70s and remain sensitive to protocol details and case mix [7, 16]. These observations motivate a shift toward tabular-centric models that can incorporate imaging-derived biomarkers while explicitly handling feature mismatch and domain shift rather than assuming homogeneous acquisition. Within the tabular family, two broad patterns emerge. Purely clinical models use logistic regression or tree ensembles on demographics, smoking history, and simple CT descriptors, sometimes enriched with laboratory indices or comorbidity scores. These models are attractive for deployment because all inputs are routinely available in electronic health records, yet they inherit the limitations of classical scores: most are developed and validated in a single institution, assume aligned features across sites, and rarely report behavior under explicit domain shift. Hybrid models instead treat radiomics signatures or deep image embeddings as additional covariates in a tabular classifier, enabling richer decision boundaries while retaining some interpretability via variable-importance analyses. However, their feature spaces are even more brittle across scanners and hospitals, as both image-derived and clinical variables can change distributions or go missing.

Existing works seldom implement formal domain-adaptation strategies for these tabular or multimodal models. External evaluations, when present, typically test a fixed model on a new hospital without feature re-alignment or recalibration, documenting sizable performance degradation but not offering systematic remedies. Only a few studies experiment with simple recalibration or refitting on a small local sample, and virtually none explore cross-domain feature selection or latent alignment tailored to nodule malignancy prediction [7, 16]. Consequently, the literature lacks robust, tabular-centric frameworks that (i) start from strong small-sample priors, (ii) identify a compact set of biomarkers stable across hospitals, and (iii) explicitly align feature distributions without assuming access to large labeled target cohorts. Taken together, these studies show that neither handcrafted risk scores, radiomics pipelines, nor deep CNN-based CAD systems currently offer reliable malignancy prediction across hospitals without local retraining or recalibration. Addressing these gaps is a central motivation for the PANDA framework developed in this thesis.

2.5 Benchmarks and open problems for cross-domain tabular learning

Beyond single-institution case studies, public benchmarks now stress-test shift robustness. TableShift curates 15 binary tasks across healthcare, finance, and public policy, with explicit temporal, geographic, and demographic shifts to measure out-of-distribution accuracy drops and calibration drift [30, 68]. Wild-Tab extends this idea to few-shot, structure-aware adaptation, showing that even tabular foundation models lose 5–15 AUC points under schema-preserving shifts [55]. Wild-Time focuses purely on temporal drift, revealing that performance decays monotonically unless models refresh their priors [29]. These resources contrast with medical imaging benchmarks, where the input grid is fixed; in tabular settings, feature heterogeneity and missingness add extra axes of mismatch. Our inclusion of the TableShift BRFSS Diabetes race-shift task aligns the pulmonary nodule study with a large-scale public benchmark, demonstrating that the proposed alignment strategy is not confined to proprietary cohorts. TableShift also surfaces common failure modes: GroupDRO and IRM rarely beat ERM on tabular tasks, label shift explains much of the OOD loss, and high ID accuracy is necessary but insufficient for shift robustness [30, 67]. Wild-Time isolates temporal drift, showing monotonic degradation without continual recalibration [29]; these findings mirror hospital deployments where assay updates or policy changes quietly reshape feature distributions.

2.5.1 Gap analysis and positioning of PANDA

Across model families and adaptation techniques, several open issues persist. First, closed-world assumptions in tabular foundation models preclude feature-mismatched deployment: TabPFN and its variants require aligned schemas and struggle when target hospitals omit or redefine biomarkers. Second, most DA methods presume access to abundant labeled or schema-aligned target data, which is unrealistic in privacy-constrained hospitals and incompatible with regulatory expectations that models remain stable under silent drift [26, 27]. Third, missingness shift and label shift remain underexplored despite being dominant drivers of clinical miscalibration in TableShift and Wild-Time; simply matching latent distributions cannot fix changes in prevalence or ordering policies [30, 29]. Finally, reproducibility crises in radiomics and deep imaging models show that aggressive priors or harmonization cannot replace explicit alignment and feature governance [16, 26].

PANDA is designed to address a specific intersection of these gaps rather than compete with every prior line of work. By treating a tabular foundation model as a plug-in prior, PANDA inherits strong small-sample performance without hand-tuning but augments it with cross-domain RFE that explicitly searches for a compact subset of biomarkers stable across hospitals. This step operationalizes domain-aware feature selection, yielding a shared feature set ("best8") that remains predictive in both institutions and provides harmonized inputs for subsequent alignment. TCA is then applied in the latent space induced by TabPFN, combining the representation power of foundation models with the stability of kernel-based alignment to handle unlabeled target data. The same pipeline is evaluated both on a private cross-hospital pulmonary nodule cohort and on the public TableShift BRFSS Diabetes race-shift task, demonstrating that the ingredients are not handcrafted for a single dataset but generalize across tabular shift scenarios [30, 67]. To our knowledge, this is the first framework to jointly combine a tabular foundation model, cross-domain RFE, and TCA for cross-hospital pulmonary nodule risk prediction and public TableShift-style tabular shift benchmarks. In this sense, PANDA fills the gap between single-domain tabular FMs, imaging-focused DA, and benchmark-driven

tabular DA by providing an end-to-end, alignment-aware framework tailored to small, imbalanced, and feature-mismatched medical cohorts.

3 Problem Formulation

Cross-hospital diagnostic risk prediction can be viewed as a constrained machine learning problem at the intersection of small-sample tabular learning and unsupervised domain adaptation (UDA). Rather than treating pulmonary nodule malignancy prediction as a purely clinical task, we explicitly formalize it as learning a robust classifier that transfers from a source hospital with limited labels to a target hospital with unlabeled data, under heterogeneous feature schemas and distribution shift.

Concretely, PANDA instantiates an AI system composed of three interacting components: a pre-trained tabular foundation model that supplies strong small-sample priors, a cross-domain feature selection module that enforces schema overlap and discards unstable biomarkers, and a kernel-based alignment module that reduces distributional discrepancy between hospitals. In this section, we formalize the underlying learning problem and connect it to standard UDA theory, so that the architectural choices of PANDA can be interpreted as explicit responses to the terms in a domain adaptation bound.

3.1 Task definition and notation

We consider a binary classification problem with a d-dimensional tabular input space $\mathcal{X} \subseteq \mathbb{R}^d$ comprising mixed numerical and categorical variables (e.g., age, nodule diameter, smoking status), and a label space $\mathcal{Y} = \{0, 1\}$ indicating benign or malignant nodules. A *domain* is defined as a joint distribution $P(X, Y)$ over $\mathcal{X} \times \mathcal{Y}$.

From an AI perspective, cross-hospital deployment naturally induces two domains:

- **Source domain (Hospital A).** We observe a labeled dataset $S = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ drawn i.i.d. from a source distribution $\mathcal{D}_S$ over (X, Y). This is the only domain in which labels are available.
- **Target domain (Hospital B).** We observe an unlabeled dataset $T = \{x_j^t\}_{j=1}^{n_t}$ drawn i.i.d. from the marginal $P_T(X)$ of a target distribution $\mathcal{D}_T$. Labels Y_T are not available during training, reflecting realistic privacy and annotation constraints in new hospitals.

Feature heterogeneity is modeled explicitly. Let F_S and F_T denote the feature-index sets available in the source and target hospitals, respectively, and define the shared schema $F_\cap = F_S \cap F_T$ with effective dimensionality $d_\cap = |F_\cap|$. Site-specific variables in $F_\setminus = (F_S \cup F_T) \setminus F_\cap$ are treated as non-transferable and are removed or marginalized before alignment.

We write $\mathcal{H}$ for a hypothesis class of classifiers $h : \mathcal{X} \to \{0, 1\}$ (or $h : \mathcal{X} \to [0, 1]$ for probabilistic outputs). For any $h \in \mathcal{H}$, the expected risk on source and target domains under a loss function L is

$$\epsilon_S(h)(h) = \mathbb{E}_{(X,Y)\sim\mathcal{D}_S}[\mathcal{L}(h(X), Y)], \quad \epsilon_T(h)(h) = \mathbb{E}_{(X,Y)\sim\mathcal{D}_T}[\mathcal{L}(h(X), Y)].$$

The empirical risks $\hat{\epsilon}_S(h)(h)$ and $\hat{\epsilon}_T(h)(h)$ are defined analogously over S and T.

Objective. The AI task considered in this thesis is to learn a predictor $h^\star \in \mathcal{H}$ using labeled source data S and unlabeled target data T such that:

1. $\epsilon_T(h^\star)$ is minimized (high AUC and clinically acceptable sensitivity on the target hospital),
2. calibration on the target domain is preserved under prevalence and covariate shift, and
3. the solution respects privacy and data-governance constraints (no sharing of raw target labels and only lightweight alignment on shared features).

Table 6 summarizes the key mathematical symbols used throughout the thesis.

Table 6: Unified Mathematical Notation System

Symbol	Definition	Dimensions
Domains and Data		
$\mathcal{X}$	Input feature space (mixed numerical/categorical)	$\subseteq \mathbb{R}^d$
$\mathcal{Y}$	Label space (0: Benign, 1: Malignant)	$\{0, 1\}$
$\mathcal{D}_S, \mathcal{D}_T$	Source and Target domain distributions	over $\mathcal{X} \times \mathcal{Y}$
S, T	Empirical datasets drawn from $\mathcal{D}_S, \mathcal{D}_T$	Sets of size n_s, n_t
$\mathbf{x}, y$	Feature vector and corresponding label	$\mathbf{x} \in \mathbb{R}^d, y \in \mathcal{Y}$
$d_{\text{num}}, d_{\text{cat}}$	Dimensionality of numerical and categorical features	Scalar (integers)
π	Prevalence of positive class $P(Y = 1)$	Probability $\in [0, 1]$
$\mathcal{F}_\cap, \mathcal{F}^*$	Shared feature schema and RFE-selected subset	Subsets of indices
Learning Theory		
$h, \mathcal{H}$	Hypothesis function (classifier) and hypothesis class	$h : \mathcal{X} \to \mathcal{Y}$
$\epsilon_S(h), \epsilon_T(h)$	Expected risk (error) on Source and Target distributions	Scalar $\in [0, 1]$
$\hat{\epsilon}_S(h), \hat{\epsilon}_T(h)$	Empirical risk on datasets S and T	Scalar $\in [0, 1]$
$d_{\mathcal{H}\Delta\mathcal{H}}$	$\mathcal{H}\Delta\mathcal{H}$-Divergence (Domain Discrepancy)	Scalar ≥ 0
λ	Ideal joint hypothesis error (Adaptability term)	Scalar ≥ 0
d_{VC}	VC Dimension of hypothesis class $\mathcal{H}$	Scalar ≥ 1
δ	Confidence parameter for generalization bound	Scalar $\in (0, 1)$
$\mathcal{L}$	Loss function	$\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$
$\boldsymbol{\theta}$	Model parameters (weights)	Vector $\in \mathbb{R}^k$
$\mathcal{H}_{\text{deep}}$	Deep Neural Network hypothesis class	$\mathcal{H}_{\text{deep}} \subseteq \mathcal{H}$
$\mathcal{S}_{\text{perf}}$	Performance-optimal hypothesis class	Subset of $\mathcal{H}$ maximizing accuracy
$\mathcal{S}_{\text{eff}}$	Efficiency-optimal hypothesis class	Subset of $\mathcal{H}$ minimizing computational cost
$\mathcal{S}_{\text{stab}}$	Stability-optimal hypothesis class	Subset of $\mathcal{H}$ minimizing variance
$\mathcal{S}_{\text{simp}}$	Simplicity-optimal hypothesis class	Subset of $\mathcal{H}$ minimizing complexity
n_s, n_t	Alternative notation for source/target domain sizes	$n_s = n_s$, $n_t = n_t$
PANDA Architecture		
$\mathcal{I}$	Permutation Importance (Feature Saliency)	Scalar $\in \mathbb{R}$
$\pi_{\text{RFE}}(\cdot)$	RFE operation that restricts features to $\mathcal{F}^*$	$\mathbb{R}^d \to \mathbb{R}^k$

Table 6: Unified Mathematical Notation System

Symbol	Definition	Dimensions
$\psi(\cdot)$	Domain adaptation mapping (TCA projection)	$\mathbb{R}^k \to \mathbb{R}^m$
$\mathbf{W}$	TCA Projection Matrix	$\mathbb{R}^{k \times m}$
$\mathbf{K}$	Kernel Matrix (Linear kernel on features)	$\mathbb{R}^{(n_s+n_t)^2}$
$\mathbf{L}$	MMD Indicator Matrix	$\mathbb{R}^{(n_s+n_t)^2}$
$\mathbf{H}$	Centering Matrix	$\mathbb{R}^{(n_s+n_t)^2}$
μ	TCA Regularization Parameter	Scalar > 0
γ	RBF Kernel Bandwidth Parameter	Scalar > 0
$\varphi(\cdot)$	Implicit Feature Map to RKHS	$\mathcal{X} \to \mathcal{H}_{RKHS}$
PPD	Posterior Predictive Distribution $P(y\|\mathbf{x}, S)$	Probability
$f_i^{\text{TabPFN}}(\cdot)$	i-th TabPFN classifier function (ensemble member)	$\mathbb{R}^m \to [0,1]$
$f_{\text{TabPFN}}(\cdot)$	TabPFN classifier function (general)	$\mathbb{R}^m \to [0,1]$
Φ	TabPFN Transformer Encoder	$\mathbb{R}^m \to \mathbb{R}^{\text{hidden}}$
$\mathbf{e}$	General feature embedding vector	$\mathbb{R}^{\text{embed\_dim}}$
$\mathbf{E}_{\text{sample}}$	Full sample embedding (concatenated feature embeddings)	$\mathbb{R}^{\text{hidden}}$
$\mathbf{P}_{\text{pos}}$	Positional encoding for feature sequence	$\mathbb{R}^{\text{hidden}}$
B	Number of preprocessing branches	Scalar (integer)
R	Number of random seeds (ensemble size per branch)	Scalar (integer)
$f_{\text{PANDA}}(\cdot)$	PANDA composite function	$\mathcal{X} \to [0,1]$
T	Temperature scaling parameter for calibration	Scalar > 0
$\sigma(\cdot)$	Activation function (e.g., sigmoid)	$\mathbb{R} \to [0,1]$

3.2 Prior-data fitted networks as tabular foundation models

Traditional machine learning assumes a fixed parametric form for the data generation process (e.g., a separating hyperplane for SVMs). In contrast, the TabPFN framework posits that the dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ is generated from a **prior distribution over functions**, denoted as P_{prior}.

Formally, a dataset $\mathcal{D}$ is sampled in two steps:

1. A structural equation model (SEM) or a data-generating function f is sampled from the prior: $f \sim P_{\text{prior}}(\cdot)$. In TabPFN, this prior is constructed explicitly using a large mixture of synthetic structural causal models (SCMs), including Bayesian Neural Networks and causal graphs with varying sparsity and non-linearity.

2. Data points are sampled conditioned on this function: $y_i = f(\mathbf{x}_i) + \epsilon$, or $y_i \sim P(\mathcal{Y}|f(\mathbf{x}_i))$.

The learning objective of a Prior-Data Fitted Network (PFN) is to approximate the **posterior predictive distribution** (PPD) for a query sample $\mathbf{x}_{\text{query}}$ given the context dataset S:

$$PPD(y_{\text{query}} \mid \mathbf{x}_{\text{query}}, S) = \int P(y_{\text{query}} \mid \mathbf{x}_{\text{query}}, f)\, P(f \mid S)\, df \tag{1}$$

TabPFN approximates this integral using a Transformer-based architecture that attends to the entire context S (In-Context Learning). This perspective is particularly advantageous for the medical small-sample setting ($n_s < 500$) because:

- It avoids iterative gradient descent on the small dataset, mitigating the risk of overfitting to noise.

- It leverages the "knowledge" encoded in the prior P_{prior}, effectively transferring inductive biases about tabular structures (e.g., decision boundaries are often aligned with axes, sparsity is common) to the medical task.

However, the standard PFN assumes that the query sample $\mathbf{x}_{\text{query}}$ comes from the same distribution as S (i.e., same f). In our cross-hospital setting, the target query $\mathbf{x}^t$ comes from a shifted distribution $\mathcal{D}_T$, violating the exchangeability assumption of the posterior approximation.

3.3 Formalizing Domain Shift

The core challenge in our research is that $\mathcal{D}_S \neq \mathcal{D}_T$. This joint distribution shift can be decomposed into three primary components relevant to medical AI:

3.3.1 Covariate Shift: The Acquisition Gap

Covariate shift occurs when the marginal feature distributions differ, $P_S(X) \neq P_T(X)$, while the conditional probability of the label remains constant, $P_S(Y|X) = P_T(Y|X)$.

$$P_S(X) \neq P_T(X) \quad \text{and} \quad P_S(Y|X) = P_T(Y|X) \tag{2}$$

In pulmonary nodule diagnosis, this is often driven by technological heterogeneity. For instance, CT scanners use different reconstruction kernels (e.g., "Sharp" vs. "Smooth"). A nodule scanned with a sharp kernel will systematically exhibit higher values for texture features like "entropy" or "spiculation" compared to the same nodule scanned with a smooth kernel, shifting the probability density function $P(\mathbf{x}_{\text{texture}})$ without changing the underlying malignancy risk. TabPFN is particularly sensitive to this because its attention mechanism relies on finding similar examples in the support set; if the target $\mathbf{x}^t$ lies in a region unsupported by $P_S(X)$, the attention weights become diffuse.

3.3.2 Label Shift: The Prevalence Gap

Label shift, or prior probability shift, is defined by a change in the marginal label distribution:

$$P_S(Y) \neq P_T(Y) \tag{3}$$

This is endemic to healthcare referrals. A tertiary cancer center (Source) typically receives high-risk referrals with a malignancy prevalence of $P(\mathcal{Y} = 1) \approx 60\%$. In contrast, a community screening program (Target) encounters a broader population with many benign incidental findings, where $P(\mathcal{Y} = 1) \approx 5\% - 20\%$. A model trained on the balanced source will learn a prior π_S and systematically overestimate risk on the target, leading to poor calibration and excessive false positives.

3.3.3 Concept Shift: The Definition Gap

Concept shift implies a fundamental change in the relationship between features and labels:

$$P_S(Y|X) \neq P_T(Y|X) \tag{4}$$

In pulmonary medicine, this arises from latent confounders such as geographic pathology. In regions like the Ohio River Valley (USA) or parts of East Asia, granulomatous diseases (e.g., Histoplasmosis, Tuberculosis) are endemic. These benign lesions often mimic the radiographic appearance of malignancy (e.g., spiculation, upper-lobe location). Consequently, a feature vector $\mathbf{x}$ that indicates a 90% probability of cancer in a non-endemic source hospital might only indicate a 40% probability in a TB-endemic target hospital.

3.3.4 Theoretical Bound on Generalization Error

Following the seminal theory by Ben-David et al. [87], the expected error of a hypothesis h on the target domain, $\epsilon_T(h)$, is bounded by:

$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda \tag{5}$$

where:

- $\epsilon_S(h)$ is the source domain error, minimized via supervised training.
- $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ is the $\mathcal{H}\Delta\mathcal{H}$-divergence between the two domains.
- $\lambda = \min_{h\in\mathcal{H}}[\epsilon_S(h) + \epsilon_T(h)]$ is the error of the ideal joint hypothesis.

This bound highlights that minimizing source error alone is insufficient; an effective cross-hospital system must simultaneously (i) keep $\epsilon_S(h)$ small, (ii) reduce the divergence term $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$, and (iii) control the joint error λ by avoiding features that admit no low-error classifier across both domains.

From this perspective, the three components of PANDA align directly with the three terms in Eq. 5:

- **Tabular foundation model (TabPFN) $\rightarrow \epsilon_S(h)$.** By treating the source cohort as a context set for a pre-trained PFN, PANDA reduces $\epsilon_S(h)(h)$ under small-sample, imbalanced conditions without extensive hyperparameter tuning.
- **Transfer Component Analysis (TCA) $\rightarrow d_{\mathcal{H}\Delta\mathcal{H}}$.** The TCA module learns a latent space in which the empirical Maximum Mean Discrepancy between source and target representations is minimized, directly targeting the divergence term $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$.
- **Cross-domain RFE $\rightarrow \lambda$.** Recursive feature elimination explicitly searches for a subset of shared features $\mathcal{F}^* \subseteq \mathcal{F}_\cap$ that support a low-error joint classifier across hospitals, thereby tightening the adaptability term λ by removing unstable or hospital-specific biomarkers.

Thus, PANDA can be interpreted as an instantiation of the domain adaptation bound rather than an ad hoc composition of modules: each architectural choice is motivated by a specific term in Eq. 5.

3.4 Theoretical Constraints of Existing Models

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

3.4.1 Gradient Boosted Decision Trees (GBDT)

GBDTs (e.g., XGBoost, LightGBM) partition the feature space using hard, axis-aligned splits ($\mathbb{I}(\mathbf{x}_j < \theta)$). They suffer from two critical limitations in UDA:

1. **Non-Differentiability:** The piecewise constant decision boundary is non-differentiable with respect to input features. This precludes the use of gradient-based domain alignment techniques (like Adversarial Training or Gradient Reversal Layers) which require backpropagating a domain loss into the feature encoder.
2. **Inability to Extrapolate:** Tree models cannot extrapolate beyond the range of the training data. If covariate shift pushes the target distribution $P_T(X)$ outside the support of $P_S(X)$, the tree maps all such points to the value of the nearest leaf node, often resulting in statistically invalid predictions.

3.4.2 Deep Tabular Models

Deep learning models (e.g., TabNet, FT-Transformer) offer differentiability but lack the appropriate inductive bias for small tabular datasets:

1. **Data Hunger:** Neural networks typically require large datasets ($N > 10^4$) to converge to a generalizable solution. With $n_s \approx 300$, deep models are prone to severe overfitting or convergence to local minima.
2. **Rotational Invariance:** Standard MLPs are rotationally invariant, but tabular features are not rotationally interchangeable (e.g., rotating "Age" and "Creatinine" axes creates a nonsensical feature space). This mismatch in inductive bias makes them less sample-efficient than tree-based or prior-fitted methods.

3.5 Transfer Component Analysis (TCA) Optimization Objective

To minimize the divergence term in Eq. 5, we employ Transfer Component Analysis (TCA). TCA seeks a feature map $\varphi : \mathcal{X} \to \mathcal{H}_{RKHS}$ such that the Maximum Mean Discrepancy (MMD) between source and target distributions in the Reproducing Kernel Hilbert Space (RKHS) is minimized.

The empirical MMD distance is defined as:

$$\text{MMD}(\mathcal{D}_S, \mathcal{D}_T) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \varphi(\mathbf{x}_i^s) - \frac{1}{n_t} \sum_{j=1}^{n_t} \varphi(\mathbf{x}_j^t) \right\|_{\mathcal{H}}^2 \tag{6}$$

TCA aims to learn a transformation matrix $\mathbf{W} \in \mathbb{R}^{(n_s+n_t)\times m}$ that reduces the data dimensionality to $m \ll d$ while minimizing MMD. The optimization problem is formally:

$$\begin{aligned} \min_{\mathbf{W}} \quad & \text{tr}(\mathbf{W}^\top \mathbf{KLKW}) + \mu \, \text{tr}(\mathbf{W}^\top \mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^\top \mathbf{KHKW} = \mathbf{I} \end{aligned} \tag{7}$$

where:

- $\mathbf{K} \in \mathbb{R}^{(n_s+n_t)\times(n_s+n_t)}$ is the kernel matrix computed on the union of source and target data (specifically, the RFE-selected features). We specifically employ a **Linear Kernel** on these features: $\mathbf{K}_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$. This is justified because the RFE process aims to select a set of features that are already more linearly separable or amenable to linear transformation. Using a linear kernel provides a robust and computationally efficient alignment without introducing complex hyperparameter tuning for RBF bandwidths.
- $\mathbf{L}$ is the MMD coefficient matrix, with elements $\mathbf{L}_{ij} = 1/n_s^2$ if $\mathbf{x}_i, \mathbf{x}_j \in S$, $1/n_t^2$ if $\mathbf{x}_i, \mathbf{x}_j \in T$, and $-1/(n_s n_t)$ otherwise.
- $\mathbf{H} = \mathbf{I} - \frac{1}{n_s+n_t}\mathbf{1}\mathbf{1}^\top$ is the centering matrix.
- The constraint $\mathbf{W}^\top \mathbf{KHKW} = \mathbf{I}$ ensures the variance of the projected data is preserved (maximizing information).

3.6 The PANDA Framework: A Unified Formalization

We formalize our proposed **PANDA** (Pretrained Adaptation Network with Domain Alignment) framework as a composite function $f_{\text{PANDA}} : \mathcal{X} \to \mathcal{Y}$. The inference process for a target sample $\mathbf{x}$ is defined as:

$$f_{\text{PANDA}}(\mathbf{x}) = h \circ \psi \circ \pi_{\cap} \circ \pi_{\text{RFE}}(\mathbf{x}), \tag{8}$$

where π_{RFE} selects the discriminative feature subset $\mathcal{F}^*$, and $\pi_{\cap}$ enforces the schema intersection with the target domain (Feature Alignment).

This composition involves distinct stages, grounded in the optimization of the feature subspace prior to alignment:

3.6.1 Stage 1: Recursive Feature Elimination (RFE) with TabPFN

Directly applying domain adaptation on high-dimensional, noisy feature spaces often leads to negative transfer. We employ Recursive Feature Elimination (RFE) to determine the optimal subspace $\mathcal{F}^* \subseteq \mathcal{F}_{\cap}$.

Let $S^{(0)} = \mathcal{F}_{\cap}$ be the initial set of shared features. The RFE process generates a sequence of feature subsets $S^{(0)} \supset S^{(1)} \supset \cdots \supset S^{(d-k)}$, where k is the target dimensionality. At each iteration t:

1. **Model Fitting:** We define the TabPFN posterior predictive distribution conditioned on the current subset $S^{(t)}$ using the source data S.

2. **Importance Estimation ($\mathcal{I}$):** Unlike linear models, TabPFN is a non-parametric meta-learned model. We approximate feature importance using **Permutation Importance**. For each feature $\mathbf{x}_j \in S^{(t)}$, we compute the degradation in the AUCmetric when feature $\mathbf{x}_j$ is randomly permuted in the validation set:

$$\mathcal{I}(\mathbf{x}_j; S^{(t)}) = \mathcal{L}_{\text{AUC}}(h_{S^{(t)}}, S) - \mathcal{L}_{\text{AUC}}(h_{S^{(t)}}, S^{\text{perm}(j)}) \tag{9}$$

3. **Elimination:** We identify and remove the feature with the minimal contribution:

$$f_{\min} = \underset{\mathbf{x}_j \in S^{(t)}}{\operatorname{argmin}} \mathcal{I}(\mathbf{x}_j; S^{(t)}) \quad \longrightarrow \quad S^{(t+1)} = S^{(t)} \setminus \{f_{\min}\} \tag{10}$$

The final subset $\mathcal{F}^*$ is selected to maximize stability and discriminative power, effectively reducing the λ term (joint error) in the Ben-David bound by removing concept-shifted features.

3.6.2 Stage 2: Feature Alignment ($\pi_\cap$)

Given the selected feature subset $\mathcal{F}^*$, this stage enforces schema consistency between hospitals. In our specific experimental setup, RFE identifies an optimal subset of $|\mathcal{F}^*| = 9$ features. However, due to missingness shift in the target domain, the alignment operator $\pi_\cap$ restricts this to the intersection of available schemas:

$$\mathcal{F}_\cap = \mathcal{F}^* \cap \mathcal{F}_t \tag{11}$$

This dimensionality reduction (from 9 to 8) explicitly handles the constraint where specific biomarkers are unrecorded in the target hospital, ensuring that subsequent stages operate only on the biologically common subspace.

3.6.3 Stage 3: Domain Adaptation Mapping (ψ)

The aligned features $\mathcal{F}_\cap$ are then adapted using Transfer Component Analysis (TCA). The domain adaptation mapping ψ projects these features into a Reproducing Kernel Hilbert Space (RKHS) where the Maximum Mean Discrepancy (MMD) between source and target distributions is minimized.

$$\mathbf{x}' = \psi(\mathbf{x}; \mathcal{F}_\cap) = \mathbf{W}^\top \mathbf{x} \tag{12}$$

This linear projection $\mathbf{W}$ is learned to align the marginal distributions $P_S(\mathcal{F}_\cap)$ and $P_T(\mathcal{F}_\cap)$, thereby reducing the domain divergence term $d_{\mathcal{H}\Delta\mathcal{H}}$ in the generalization bound.

3.6.4 Stage 4: Classification and Ensemble (h)

Finally, the classification function h processes the adapted features $\mathbf{x}'$. This stage leverages the pre-trained TabPFN as a robust classifier, augmented by an ensemble mechanism to handle predictive uncertainty and label shift.

$$h(\mathbf{x}') = \frac{1}{N} \sum_{i=1}^{N} \sigma\left(\frac{f_i^{\text{TabPFN}}(\mathbf{x}')}{T} \right) \tag{13}$$

where T is a temperature scaling parameter optimized to calibrate the output probabilities against the prevalence differences between hospitals. The result is a calibrated malignancy probability that is robust to both covariate and label shifts.

3.7 Clinical-Statistical Mapping

Table 7 summarizes the correspondence between the clinical challenges observed in pulmonary nodule diagnosis, their statistical manifestations, and the corresponding PANDA solution components derived from our theoretical framework.

Table 7: Mathematical Mapping of Clinical Problems to PANDA Components

Clinical Challenge	**Statistical Mechanism**	**PANDA Component**	**Theoretical Justification**
Scanner Variance (Sharp vs. Smooth Kernels)	Covariate Shift: $P_S(X) \neq P_T(X)$	Latent TCA on RFE-selected features	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: $P_S(Y\|X) \neq P_T(Y\|X)$	Cross-Domain RFE	Minimizes joint error λ by removing unstable features.

3.8 Problem Constraints and Research Scope

Our formulation is bound by specific constraints inherent to the medical domain:

- **Small Sample Size Constraint:** The sample sizes n_s, n_t are typically in the range of 100 to 1000, which is insufficient for training over-parameterized deep networks from scratch:

$$n_s \ll d_{VC}(\mathcal{H}_{\text{deep}}) \tag{14}$$

  where $d_{VC}(\mathcal{H}_{\text{deep}})$ represents the VC dimension required for standard deep domain adaptation networks to generalize.

- **Privacy and Data Silos:** We assume source data $\mathcal{D}_S$ and target data $\mathcal{D}_T$ cannot be physically merged.

- **Class Imbalance:** The prevalence of the positive class is often low ($\pi < 0.3$), requiring AUC-centric optimization.

This formalization sets the stage for the specific methodological implementations detailed in Chapter 4.

4 Solution

To bridge the gap between advanced tabular foundation models and the practical constraints outlined in Section 3.8—small sample sizes, heterogeneous feature schemas, distribution shifts, and privacy-preserving data silos—we propose PANDA (Pretrained Adaptation Network with Domain Alignment). PANDA is a composite algorithmic framework designed to predict pulmonary nodule malignancy with high stability across heterogeneous clinical environments. It explicitly addresses the tripartite challenge of small-sample scarcity, distribution shift, and feature heterogeneity through a tightly integrated pipeline that combines cross-domain feature selection, schema alignment, latent-space adaptation, and a calibrated foundation-model ensemble.

Formally, PANDA implements a composite mapping

which transforms raw, heterogeneous clinical inputs into a calibrated malignancy probability. Building on the notation introduced in Section 3, we factorize this mapping as

where π_{RFE} performs cross-domain feature selection, $\pi_{\cap}$ enforces a robust schema intersection across hospitals, ψ denotes latent-space domain adaptation, and h is a calibrated foundation-model classifier.

4.1 Architectural Overview

The PANDA framework operates as a directed acyclic graph (DAG) of data transformations that progressively refine the representation of each patient from raw features to domain-invariant embeddings and, finally, to a calibrated risk estimate. Table 8 summarizes this data flow, while Figure 1 provides a comprehensive visual overview of the architecture. Conceptually, the pipeline consists of four sequential stages:

Table 8: Data flow and transformations in the PANDA framework. The pipeline progressively refines raw, high-dimensional inputs into low-dimensional, domain-invariant embeddings and, ultimately, a calibrated probability. In the pulmonary nodule experiments, $k = 8$, $d_\cap = 8$, and $m = 15$, but these dimensions are dataset-dependent.

Stage	Component	Input Space	Core Operation	Output Space
1	Cross-Domain RFE	$\mathbb{R}^{d_{\text{raw}}}$	Iterative elimination via $\mathcal{I}(f)$	$\mathbb{R}^k$ ($\mathcal{F}^*$, $k = \lvert\mathcal{F}^*\rvert$)
2	Feature Alignment	$\mathbb{R}^k$	Schema intersection on $\mathcal{F}_\cap$	$\mathbb{R}^{d_\cap}$ ($d_\cap = \lvert\mathcal{F}_\cap\rvert$)
3	Latent TCA	$\mathbb{R}^{d_\cap}$	Projection $\mathbf{x}' = \mathbf{W}^\top \mathbf{x}$	$\mathbb{R}^m$ (m = latent dimension)
4	TabPFN Classifier	$\mathbb{R}^m$	Classification $f_{\text{TabPFN}}(\mathbf{x}')$	$[0, 1]$ (Prob.)
5	Ensemble	$[0, 1]^{B \times R}$	Temperature-scaled averaging	$[0, 1]$ (Final)

1. **Domain-Aware Feature Selection and Alignment**: PANDA first applies Cross-Domain Recursive Feature Elimination (RFE) to identify a compact, clinically meaningful subset of predictors that are stable across hospitals. It then intersects this subset with the feature schema available at each target site, yielding a site-specific, yet harmonized, feature space. This dual step reduces dimensionality, filters site-specific artifacts, and mitigates "missingness shift" by propagating only universally available variables.

2. **Latent-Space Domain Adaptation**: On the aligned feature space, PANDA applies Transfer Component Analysis (TCA) to project data into a lower-dimensional, domain-invariant latent space. The projection is learned to minimize a Maximum Mean Discrepancy (MMD) objective between source and target distributions, thereby addressing covariate shift while preserving task-relevant structure.

3. **Foundation-Model Inference**: The TCA-transformed features are then processed by a pre-trained TabPFN classifier. Rather than training a deep network from scratch on small clinical cohorts, PANDA reuses the prior-data-fitted structure of TabPFN, which encodes priors learned from millions of synthetic tasks and can infer complex non-linear decision boundaries without extensive gradient-based fine-tuning.

4. **Multi-View Ensemble and Calibration**: To further mitigate variance induced by small sample sizes and label shift, PANDA constructs a multi-branch ensemble that exposes the foundation model to diverse yet consistent views of the same patient. Branches differ in preprocessing (e.g., raw, quantile-normalized, ordinal, and power-transformed representations), and multiple random seeds are used to induce rotational invariance in the feature ordering. The resulting ensemble predictions are then temperature-scaled and averaged to yield a final, well-calibrated malignancy probability.

This architectural decomposition provides a clear separation of concerns: RFE and schema intersection handle feature heterogeneity and data silos; TCA focuses on distribution alignment; TabPFN contributes strong priors for small-sample learning; and the multi-view ensemble targets robustness and calibration. The complete end-to-end pipeline, including the multi-branch ensemble configuration and temperature scaling mechanism, is illustrated in Figure 1.

4.2 Mapping to the Formal Problem Formulation

The four-stage architecture described above directly instantiates the formal problem setup of Section 3. Specifically:

- **Cross-Domain RFE** implements the operator π_{RFE}, reducing the original high-dimensional space $\mathbb{R}^{d_{\text{raw}}}$ to a compact set of clinically actionable features F^* of size $k = |F^*|$. The selection objective jointly balances discriminative performance, stability across folds, computational efficiency, and parsimony.
- **Feature Alignment** implements $\pi_\cap$, mapping site-specific feature sets to a shared schema $F_\cap$ of dimension $d_\cap = |F_\cap|$, thereby ensuring that source and target hospitals operate on a common representation despite differing EHR schemas.
- **Latent TCA** implements ψ, projecting the aligned features into a latent space $\mathbb{R}^m$ in which the empirical MMD between source and target distributions is minimized, directly tackling the covariate-shift component of the generalization bound.
- **Foundation-Model Ensemble** implements h, mapping latent representations to calibrated probabilities through a TabPFN-based classifier composed with a multi-branch, temperature-scaled ensemble.

By making this mapping explicit, PANDA connects the abstract components of the theoretical formulation to concrete algorithmic steps, allowing each design choice to be interpreted in terms of the underlying constraints on sample size, privacy, imbalance, and distribution shift.

4.3 Pointers to Implementation Details

The subsequent Methods chapter provides the implementation-level details that instantiate each component of the PANDA architecture:

- Section 5.2.1 formalizes the Cross-Domain RFE procedure and the Cost-Effectiveness Index used to select the "Best-k" feature subset.
- Section 5.3.3 details the TCA objective, kernel construction, and choice of regularization for latent-space alignment.
- Section 5.3 describes how PANDA integrates TabPFN as a probabilistic classifier, including in-context serialization of tabular inputs.
- Section 5.3.2 specifies the multi-branch preprocessing strategy, ensemble configuration, and temperature-scaling procedure used for final calibration.

This separation keeps the Solution chapter focused on the high-level design of PANDA, while deferring technical derivations and algorithmic variants to the dedicated methodological sections.

5 Methods

5.1 Motivating Challenges and Methodological Response

Cross-hospital malignancy prediction and public health surveillance pose intertwined constraints, including small labeled cohorts, label imbalance, feature mismatch, and multiple forms of distribution shift. PANDA is structured to address these constraints rather than to introduce architectural novelty. In this section, we explicitly link each methodological component to the specific failure mode it is designed to mitigate. Table 9 summarizes the main obstacles and the mechanisms assigned to each of them.

This architecture is explicitly designed so that each component addresses a specific question: why do small-sample medical deployments fail, and which prior or alignment mechanism mitigates that failure? Specifically, the TabPFN backbone addresses the small-sample, high-dimensional regime; cross-domain RFE addresses feature mismatch and feature heterogeneity across institutions; the TCA module addresses covariate shift and mixed acquisition protocols; the calibration and sampling utilities address label prevalence drift and class imbalance; and the multi-branch preprocessing ensemble addresses instability arising from preprocessing choices.

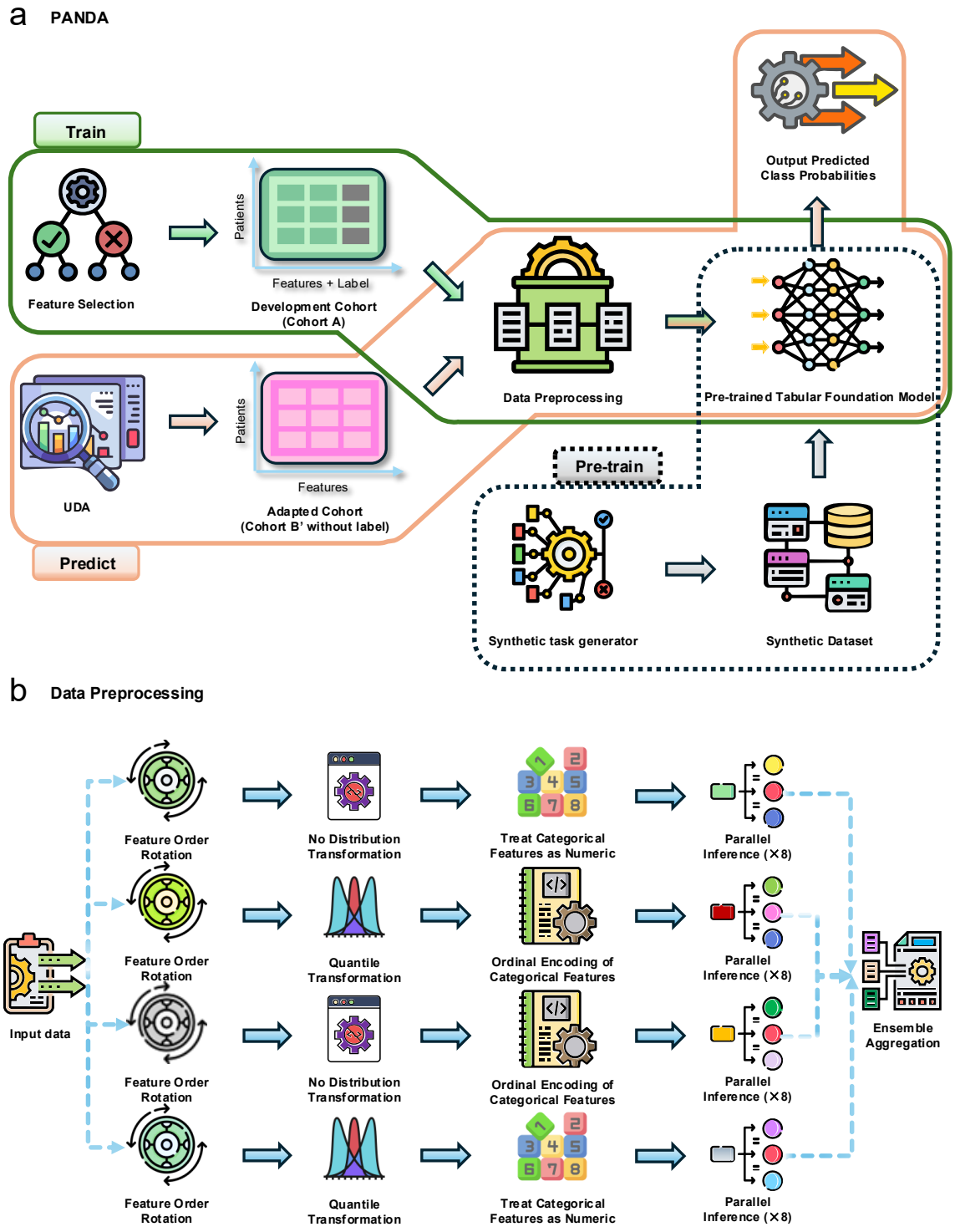

Figure 1: **PANDA framework architecture.** (a) Compositional pipeline from original tabular data through ensemble training, prediction aggregation, class-imbalance adjustment, and final classification output. (b) Multi-branch ensemble with $B = 4$ preprocessing strategies, each generating $R = 8$ ensemble members via different random seeds.

Table 9: Challenge–mechanism mapping in PANDA. Each component targets a known failure mode, and the same design is applied to pulmonary nodules and the TableShift BRFSS race-shift task.

Challenge	Mechanism	Expected benefit
Small n_s with high-dimensional covariates	TabPFN prior-data-fitted network that performs in-context learning with frozen weights	Transfers structural priors from millions of synthetic tasks and reduces estimation variance without local fine-tuning
Feature heterogeneity across institutions/demographics	Cross-domain RFE identifies stable subsets ("best8") that are definable at every site, together with schema-alignment utilities	Removes site-specific artifacts before adaptation and ensures that downstream models operate only on shared attributes
Covariate shift and mixed acquisition protocols	TCA applied to RFE-selected features realigns marginal distributions before the TabPFN classifier	Reduces the $d_{\mathcal{H}\Delta\mathcal{H}}$ divergence so that context examples remain relevant to target queries
Label prevalence drift and class imbalance	Class-balanced sampling, calibrated decision thresholds, and ensemble temperature scaling	Maintains sensitivity for malignant/SPN-positive cohorts and accounts for higher diabetes rates in non-White BRFSS respondents
Variance from preprocessing choices	Multi-branch preprocessing (ordering, quantile transforms, ordinal encoding) with ensemble averaging	Introduces diversity without retraining new weights and stabilizes predictions under minor data perturbations

5.2 Feature Engineering and Selection Implementation

The feature engineering pipeline in PANDA is designed to accommodate the heterogeneity of medical data sources while preserving domain-invariant signals. This component directly targets the feature mismatch and feature heterogeneity challenges by enforcing that downstream models operate only on features that are consistently available and stable across hospitals. It comprises two stages: global feature selection via Cross-Domain RFE and local feature transformation via TabPFN's internal preprocessing branches.

5.2.1 Cross-Domain Recursive Feature Elimination (RFE)

To address the *feature mismatch* challenge, we implement a Cross-Domain Recursive Feature Elimination (RFE) strategy. Unlike standard RFE, which optimizes for a single dataset, this approach seeks a feature subset $\mathcal{F}^*$ that maximizes predictive performance on the source domain $\mathcal{D}_S$ while satisfying availability constraints in the target domain $\mathcal{D}_T$.

The process uses a wrapper around the TabPFN classifier to compute permutation importance. We define the importance of feature j as the degradation in AUCwhen its values are randomly shuffled:

$$\mathcal{I}_j = \mathcal{L}_{\text{AUC}}(\mathcal{D}_{\text{val}}) - \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_{\text{AUC}}(\mathcal{D}_{\text{val}}^{(j,\text{shuffled})}) \tag{15}$$

where $K = 5$ repeats. Algorithm 1 details the iterative elimination process.

Algorithm 1 Cross-Domain Recursive Feature Elimination (RFE) with CEI Optimization

Require: Source Data X_s, y_s, Target Feature Count Range $[k_{\min}, k_{\max}]$
Ensure: Optimal Feature Subset $\mathcal{F}^*$

Initialize TabPFN Wrapper with permutation importance computation
Initialize performance records: $\mathcal{M} \leftarrow \emptyset$
for $k = k_{\max}$ **to** $k_{\min}$ **do**
  Configure sklearn RFE: estimator=TabPFNWrapper, $n\_features\_to\_select = k$, step=1
  Fit RFE: $(S_k, \text{ranking}_k) \leftarrow \text{RFE.fit}(X_s, y_s)$
  Evaluate performance: $M_k(k) = \text{CV-AUC}(S_k)$ using K-fold CV
  Compute CEI components:
    $S_{\text{perf}}(k) = M_k(k)$ {Performance score}
    $S_{\text{eff}}(k) = 1 - \frac{\text{Cost}(S_k)}{\text{Cost(all features)}}$ {Cost efficiency}
    $S_{\text{stab}}(k) = 1 - \text{Std(AUC across CV folds)}$ {Stability score}
    $S_{\text{simp}}(k) = \exp(-\alpha \cdot k)$ {Sparsity penalty}
  Compute CEI: $\text{CEI}(k) = w_1 S_{\text{perf}}(k) + w_2 S_{\text{eff}}(k) + w_3 S_{\text{stab}}(k) + w_4 S_{\text{simp}}(k)$
  $\mathcal{M} \leftarrow \mathcal{M} \cup \{(k, S_k, \text{CEI}(k))\}$
end for
Select optimal subset: $k^* = \text{argmax}_k \, \text{CEI}(k)$
$\mathcal{F}^* \leftarrow S_{k^*}$
return $\mathcal{F}^*$

This algorithm yields the standard feature subsets referenced throughout the study (for example, "best8").

5.2.2 Multi-Branch Preprocessing Strategy

Once the feature set is fixed, PANDA leverages TabPFN's internal ensemble mechanism to handle distribution shifts in feature scaling and encoding. This step primarily addresses the variance introduced by different preprocessing choices and minor covariate shifts in the marginal feature distributions. This mechanism can be viewed as a form of test-time augmentation for tabular data. The ensemble generation mechanism produces 32 distinct views of the data through four preprocessing pipelines:

1. **No-Op Branch:** Raw features are passed directly, preserving their original distributions, which is useful for tree-like decision logic.

2. **Quantile Branch:** Features are transformed via $F^{-1}(\Phi(\mathbf{x}))$, mapping the empirical CDF to a standard Normal $\mathcal{N}(0, 1)$. This transformation mitigates extreme outliers and skewed distributions common in medical markers (for example, CEA levels).

3. **Ordinal Branch:** All unique values are ranked and replaced by their integer rank. This transformation removes magnitude information but preserves order, making the model robust to unit changes (for example, centimetres vs millimetres).

4. **Power Transform Branch:** Applies $\mathbf{x} \mapsto \mathbf{x}^\lambda$ (for example, square root or logarithm) to stabilize variance.

Each branch is applied to both support (train) and query (test) sets simultaneously, ensuring a consistent mapping.

5.3 Foundation Model Integration Mechanism

PANDA employs TabPFN as a probabilistic classifier within an integrated framework. This module directly addresses the challenge of training in small, high-dimensional clinical cohorts by reusing structural priors learned from synthetic tasks and producing calibrated probability estimates that can subsequently be adjusted for label imbalance. While TabPFN internally processes features, its role in PANDA is to produce calibrated predictions on adapted tabular data. The integration involves specific in-context serialization and ensemble construction steps.

5.3.1 In-Context Serialization and Tokenization

Unlike BERT-style models, which require textual input, TabPFN operates directly on raw numerical and categorical values through a sophisticated per-feature processing architecture. The serialization mechanism maps a heterogeneous row $\mathbf{x} \in \mathbb{R}^{d_{\text{num}}} \times \mathbb{Z}^{d_{\text{cat}}}$ into a sequence of continuous embeddings.

The feature processing pipeline involves several key steps:

- **Preprocessing and Encoding:** Features undergo preprocessing such as quantile and power transformations. Categorical variables are primarily handled via ordinal encoding, converting categories to integers, or processed as mixed types, rather than relying solely on traditional embedding lookup tables.
- **Missing Value Handling:** Instead of a single learnable token, missing values are handled by a robust mechanism where they are replaced by the feature mean (learned during training). Simultaneously, separate indicator channels are generated to explicitly flag missing values or infinite entries, preserving the structural information of data unavailability.
- **Linear Projection:** Each preprocessed feature is linearly projected into the embedding space.

$$\mathbf{e}^{(j)} = \text{Linear}(\text{Preprocess}(\mathbf{x}^{(j)})) \tag{16}$$

Distinct from architectures that concatenate all features into a single vector ($\mathbf{E}_{\text{sample}} = \text{MLP}(\text{Concat}(\dots))$), TabPFN maintains a per-feature representation. The Transformer architecture utilizes attention mechanisms that operate across both features and samples, allowing the model to learn complex interactions between specific covariates while attending to relevant patients within the context window.

5.3.2 Ensemble Construction and Inference

The final prediction is obtained by averaging over 32 diverse forward passes. Let $\mathcal{B} = \{P_1, \dots, P_4\}$ denote the set of preprocessing functions and $\mathcal{R} = \{s_1, \dots, s_8\}$ a set of random seeds that control the subsampling of the context set (the *support set* of labeled examples). The ensemble probability estimate is

$$\hat{P}(y = 1|\mathbf{x}_q) = \frac{1}{32} \sum_{P \in \mathcal{B}} \sum_{s \in \mathcal{R}} \sigma \left(\frac{f_\theta(P(\mathbf{x}_q) \mid P(\mathbf{X}^s_{\text{ctx}}))}{T} \right) \tag{17}$$

where

- $\mathbf{x}_q$ is the target query (patient),
- $\mathbf{X}^s_{\text{ctx}}$ is the subset of training data selected by seed s,
- f_θ is the frozen TabPFN Transformer backbone, and
- $T = 0.9$ is the temperature scaling parameter calibrated for small-sample confidence.

Algorithm 2 summarizes the inference procedure.

Algorithm 2 PANDA Inference with TabPFN Backbone

Require: Query $\mathbf{x}_q$, Context $\mathcal{D}_{\text{train}}$, Ensemble $N = 32$
Ensure: Malignancy probability $\hat{y}$

```
Logits ← []
for i = 1 to N do
    Sample preprocessing P ~ B and context subset D_i ⊂ D_train
    x'_q, D'_i ← P(x_q), P(D_i) {Apply branch}
    E ← Serialize(x'_q, D'_i) {Tokenize}
    z ← Transformer(E) {Forward pass}
    l_i ← ClassifierHead(z)
    Logits.append(l_i)
end for
ŷ ← Softmax(Mean(Logits)/T)
return ŷ
```

5.3.3 Transfer Component Analysis (TCA)

TCA is applied to the RFE-selected feature space to align the marginal distributions of the source and target domains. This module directly targets the covariate shift and mixed acquisition-protocol challenges highlighted in Table 9 by enforcing a shared latent representation across hospitals. We implement this alignment via Transfer Component Analysis (TCA), which solves the optimization problem through a generalized eigenvalue decomposition [88, 89]. Algorithm 3 outlines the computational procedure.

Algorithm 3 Transfer Component Analysis (TCA) Optimization

Require: Source Features X_s, Target Features $\mathcal{F}_t$, Kernel $k(\cdot,\cdot)$, Dim m, Reg μ
Ensure: Projection Matrix $\mathbf{W}$
Construct Kernel Matrix:
Compute blocks $\mathbf{K}_{SS}, \mathbf{K}_{ST}, \mathbf{K}_{TT}$ using $k(\cdot,\cdot)$
$\mathbf{K} \leftarrow \begin{bmatrix} \mathbf{K}_{SS} & \mathbf{K}_{ST} \\ \mathbf{K}_{ST}^\top & \mathbf{K}_{TT} \end{bmatrix}$
Construct MMD Matrix L:
$\mathbf{L}_{ij} \leftarrow \frac{1}{n_s^2}$ if $x_i, x_j \in S$; $\frac{1}{n_t^2}$ if $x_i, x_j \in T$; else $\frac{-1}{n_s n_t}$
Construct Centering Matrix H:
$\mathbf{H} \leftarrow \mathbf{I} - \frac{1}{n_s+n_t}\mathbf{1}\mathbf{1}^\top$
Solve Generalized Eigenproblem:
$\mathbf{A} \leftarrow \mathbf{I} + \mu\mathbf{K}\mathbf{L}\mathbf{K}$
$\mathbf{B} \leftarrow \mathbf{K}\mathbf{H}\mathbf{K}$
Compute eigenvectors $\mathbf{V}$ of $\mathbf{A}^{-1}\mathbf{B}$
$\mathbf{W} \leftarrow \text{SelectTop}(\mathbf{V}, m)$ {Top m eigenvectors}
return $\mathbf{W}$

Mathematical Formulation The core optimization problem seeks a projection matrix $\mathbf{W} \in \mathbb{R}^{(n_s+n_t)\times m}$ that maps the kernel matrix $\mathbf{K}$ to a low-dimensional latent space. The objective is to minimize the Maximum Mean Discrepancy (MMD) between domains while preserving data variance.

The kernel matrix $\mathbf{K} \in \mathbb{R}^{(n_s+n_t)\times(n_s+n_t)}$ is constructed as a block matrix:

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{SS} & \mathbf{K}_{ST} \\ \mathbf{K}_{TS} & \mathbf{K}_{TT} \end{bmatrix} \tag{18}$$

where $\mathbf{K}_{SS}, \mathbf{K}_{TT}, \mathbf{K}_{ST}$ represent the kernel evaluations within the source domain, within the target domain, and between source and target samples, respectively.

The MMD matrix $\mathbf{L}$ and centering matrix $\mathbf{H}$ are defined as

$$\mathbf{L}_{ij} = \begin{cases} \frac{1}{n_s^2} & x_i, x_j \in S \\ \frac{1}{n_t^2} & x_i, x_j \in T \\ -\frac{1}{n_s n_t} & \text{otherwise} \end{cases}, \quad \mathbf{H} = \mathbf{I} - \frac{1}{n_s+n_t}\mathbf{1}\mathbf{1}^\top \tag{19}$$

The ADAPT implementation solves for the projection $\mathbf{W}$ by addressing a generalized eigenvalue problem. Specifically, it computes

$$\mathbf{A} = \mathbf{I} + \mu\mathbf{K}\mathbf{L}\mathbf{K}, \quad \mathbf{B} = \mathbf{K}\mathbf{H}\mathbf{K} \tag{20}$$

and solves for the eigenvectors of $\mathbf{A}^{-1}\mathbf{B}$. The optimal projection matrix $\mathbf{W}$ consists of the m eigenvectors corresponding to the largest eigenvalues of $\mathbf{A}^{-1}\mathbf{B}$, which we denote as $\mathbf{W} = \texttt{vectors\_}[:m]$. This formulation maximizes the ratio of data variance ($\text{tr}(\mathbf{W}^\top\mathbf{B}\mathbf{W})$) to the regularized domain divergence ($\text{tr}(\mathbf{W}^\top\mathbf{A}\mathbf{W})$). The top m eigenvectors form the projection matrix $\mathbf{W}$.

Data Flow and Transformation In the `fit_transform` phase, we utilize the **RFE-selected feature vectors** as the input representations for both the source (X_s) and target ($\mathcal{F}_t$) domains. These feature matrices are concatenated to compute the kernel matrix. The solver then computes $\mathbf{W}$ using

the top m eigenvectors ('vectors_') of the matrix product. For a new query sample $\mathbf{x}_{\text{query}}$, the transformation projects it into the shared subspace by computing the inner product of the query's kernel vector with these eigenvectors:

$$\mathbf{z}_{\text{new}} = \mathbf{k}_{\text{new}}^{\top}\mathbf{W} \tag{21}$$

where $\mathbf{k}_{\text{new}} = [k(\mathbf{x}_{\text{query}}, x_1), \ldots, k(\mathbf{x}_{\text{query}}, x_{n+m})]^{\top}$ is the kernel vector between the query and all training samples, and $\mathbf{W}$ contains the learned alignment directions. This mapping allows the model to embed unseen target samples into the aligned space using the learned support vectors.

Hyperparameter Configuration and Kernel Choice The behaviour of TCA is governed by the hyperparameters listed in Table 10.

Table 10: TCA hyperparameter configuration in PANDA

Parameter	Description and role	Value
`kernel`	Kernel function type. Controls the implicit feature map $\varphi(\mathbf{x})$.	'linear'
`n_components` (m)	Dimensionality of the projected subspace.	10
`mu` (μ)	Regularization parameter. Balances alignment (MMD) against variance preservation. Higher μ reduces adaptation.	1.0
`gamma` (γ)	Kernel coefficient for RBF ($e^{-\gamma\|x-y\|^2}$). Passed via `kernel_params`.	N/A

We explicitly choose a **linear kernel** ($K_{\text{linear}}(x, y) = x^{\top}y$) over the radial basis function (RBF) kernel ($K_{\text{RBF}}(x, y) = \exp(-\gamma\|x - y\|^2)$) for the following reasons:

1. **Feature disentanglement:** The RFE-selected features already form a robust, low-dimensional subset. A linear alignment of their first-order moments is sufficient and less prone to overfitting than the infinite-dimensional mapping induced by the RBF kernel.
2. **Stability on small samples:** RBF kernels require tuning the bandwidth γ, which can be unstable on small cohorts ($N \approx 300$). An improper choice of γ can lead to a degenerate Gram matrix.
3. **Negative transfer avoidance:** Empirical tests indicated that RBF-TCA often collapsed the TabPFN classifier to the majority class because of noise amplification in the high-dimensional RBF space.

This configuration allows PANDA to benefit from domain alignment without incurring the instability often associated with kernel methods on small datasets.

5.3.4 Parameter Sensitivity Diagnostics

To ensure that the chosen defaults are numerically stable and that the TCA hyperparameters do not introduce hidden instabilities or negative transfer, the automated regression test suite exercises the same Adapt wrapper on the Mayo → PKUPH split with stress configurations (letting Adapt infer `n_components=None` and shrinking μ to 0.1). The test reports Accuracy, AUC, F1 Score, Precision, and Recall and asserts that the predicted probabilities remain in $[0, 1]$. Because both the production path and the test harness rely on the identical Adapt pipeline, these executions provide our practical parameter-sensitivity verification for PANDA.

5.4 Experimental Configuration

5.4.1 Baseline Hyperparameters

To enable fair comparison, all baseline models were configured with default hyperparameters and essential settings for reproducibility. As shown in Table 11, we focused on consistent random state

management, class imbalance handling through balanced class weights, and parallel processing for computational efficiency. This approach ensures that the comparison reflects the models' inherent capabilities rather than extensive hyperparameter optimization, which aligns with our research focus on domain adaptation effectiveness rather than individual model tuning.

Table 11: Configuration for baseline models. Default parameters were used with basic settings for reproducibility and class imbalance handling.

Model	**Parameter**	**Value**
XGBoost	Random state	Fixed per experiment
	n_jobs	-1 (parallel processing)
	eval_metric	'logloss'
Random Forest	Random state	Fixed per experiment
	n_jobs	-1 (parallel processing)
	Class weight	'balanced'
SVM	Random state	Fixed per experiment
	Probability	True (for probability outputs)
	Class weight	'balanced'

5.4.2 Clinical Scoring Models

We implemented three established clinical calculators for pulmonary nodule malignancy:

- **Mayo model**: Prob $= \sigma(-6.83 + 0.039 \cdot \text{Age} + 0.79 \cdot \text{Smoke} + 1.34 \cdot \text{CancerHx} + 0.13 \cdot \text{Diameter} + 1.04 \cdot \text{Spiculation} + 0.78 \cdot \text{UpperLobe})$ [1].
- **PKUPH model**: A regression model specifically developed for Chinese populations [83]: Prob $= \sigma(-4.50 + 0.07 \cdot \text{Age} + 0.68 \cdot \text{Diameter} + 0.74 \cdot \text{Spiculation} + 1.27 \cdot \text{FamilyHx} - 1.62 \cdot \text{Calcification} - 1.41 \cdot \text{Boundary})$.
- **LASSO Logistic Regression**: A data-driven baseline derived via L1-regularization on the source domain [4]: $\hat{p} = \sigma(-1.14 + 0.036 \cdot \text{Age} + 0.38 \cdot \text{CancerHx} + 0.20 \cdot \text{Diameter} - 0.29 \cdot \text{Calcification} + 0.026 \cdot \text{PleuralStretch} - 0.17 \cdot \text{VC} + 0.05 \cdot \text{DLCO} + 0.018 \cdot \text{CEA} + 0.004 \cdot \text{NSE})$.

These models are applied using their published coefficients without re-training and represent the standard of care.

5.4.3 ML Baseline Models

We implemented five machine learning baseline models for comprehensive comparison with the PANDA framework. These models were trained and evaluated using 10-fold cross-validation on the target domain with the same feature set and preprocessing pipeline as PANDA:

- **Support Vector Machine (SVM)**: A maximum margin classifier that finds optimal hyperplanes for classification [90]. Configuration: probability=True, class_weight='balanced' for imbalanced data handling.
- **Decision Tree (DT)**: A tree-based classifier that learns decision rules from data features using recursive partitioning [91]. Configuration: default sklearn parameters with reproducibility settings.
- **Random Forest (RF)**: An ensemble method that constructs multiple decision trees and outputs the majority vote through bagging [92]. Configuration: class_weight='balanced', parallel processing enabled.

- **Gradient Boosting Decision Tree (GBDT)**: A boosting ensemble method that builds trees sequentially, each correcting the previous one's errors through gradient descent optimization [93]. Configuration: default sklearn parameters for gradient boosting.
- **XGBoost**: An optimized gradient boosting framework designed for efficiency and performance with regularized objective function and tree pruning [8]. Configuration: eval_metric='logloss', parallel processing enabled.

All ML baseline models used default hyperparameters with consistent random state management and class imbalance handling, ensuring comparison reflects inherent model capabilities rather than extensive optimization.

5.4.4 Computational Framework

All experiments were conducted on the High Performance Computing platform at the Hong Kong Polytechnic University Department of Computing, equipped with

- **Hardware**: NVIDIA A800-SXM4-80GB GPU (80GB HBM2 memory), AMD EPYC 7742 64-Core Processor (256 threads), NUMA architecture with 512MB L3 cache.
- **Software**: PyTorch 2.7.0 (CUDA 12.2), Scikit-learn 1.6.1, Adapt 0.4.4.
- **System**: Ubuntu 22.04.4 LTS, NVIDIA Driver 535.161.07.

6 Analysis

This chapter moves beyond descriptive reporting of performance metrics to develop a rigorous theoretical account of *why* the PANDA framework succeeds where traditional methods fail. The failure of classical models (e.g., GBDTs, CNNs) in cross-hospital deployment is interpreted not as an engineering artifact, but as a violation of a core assumption in statistical learning theory, namely that training and test data are independently and identically distributed (i.i.d.) samples from the same joint distribution $P(\mathcal{X}, \mathcal{Y})$.

By analyzing PANDA under the generalization bound of Ben-David et al. for domain adaptation, we show that its architecture constitutes a direct algorithmic response to the theoretical decomposition of target error.

6.1 Theoretical Foundation: The Generalization Bound

To study the generalization properties of PANDA, we adopt the seminal learning-theoretic framework of Shai Ben-David et al. (2010), which is a cornerstone of domain adaptation theory [87]. This framework provides a rigorous upper bound on the target domain error and decomposes it into observable and optimizable components. The bound clarifies why minimizing the source error alone is insufficient and why explicit domain-alignment mechanisms such as TCA are required.

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 source distribution $\mathcal{D}_S$ and target distribution $\mathcal{D}_T$ respectively, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, for every $h \in \mathcal{H}$:*

$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda + \mathcal{O}\left(\sqrt{\frac{d_{VC} \log n}{n} + \log \frac{1}{\delta}}\right) \tag{22}$$

Inequality (22) shows that reducing the target error $\epsilon_T(h)$ requires the simultaneous control of three terms:

1. **Source Risk** $\epsilon_S(h)$: The expected error on the source domain.
2. **Domain Divergence** $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$: The $\mathcal{H}\Delta\mathcal{H}$-divergence, which measures the discrepancy between the marginal feature distributions.
3. **Adaptability Term** λ: The error of the ideal joint hypothesis, $\lambda = \min_{h \in \mathcal{H}}(\epsilon_S(h) + \epsilon_T(h))$, which captures irreducible error due to concept shift, that is, mismatch in the conditional distributions $P(Y \mid X)$.

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

A central quantity in the theory of Ben-David et al. is the $\mathcal{H}\Delta\mathcal{H}$-divergence, which quantifies domain distance with respect to the hypothesis class $\mathcal{H}$. It is defined as

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 2 \sup_{h,h' \in \mathcal{H}} \left| \Pr_{x \sim \mathcal{D}_S}[h(x) \neq h'(x)] - \Pr_{x \sim \mathcal{D}_T}[h(x) \neq h'(x)] \right|. \tag{23}$$

Intuitively, this metric measures the maximal discrepancy in classifier disagreement across the two domains. If there exist two classifiers h, h' that exhibit low disagreement on the source domain but high disagreement on the target domain, the divergence is large. In this case, the source domain does not sufficiently constrain classifier behavior on the target domain, which can lead to negative transfer. Alignment methods such as TCA and CORAL aim to transform the feature space so that this divergence is reduced.

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

The first challenge is the small sample size ($n_s \approx 295$). In this regime, standard Empirical Risk Minimization (ERM) is prone to high variance. Deep neural networks typically require $n > 10^4$ to generalize reliably, while GBDTs often overfit, effectively memorizing noise in S rather than learning the structure of $\mathcal{D}_S$.

6.2.1 Prior-Data Fitted Networks vs. Parametric Learning

Traditional parametric learning optimizes weights $\boldsymbol{\theta}$ to minimize loss on S,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{(\mathbf{x},y) \in S} \mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}), y), \tag{24}$$

starting from largely uninformative priors and therefore requiring substantial data.

In contrast, PANDA uses TabPFN, a **Prior-Data Fitted Network (PFN)**. It reduces $\epsilon_S(h)$ not by performing gradient descent on S, but by approximating the **Posterior Predictive Distribution (PPD)** using a Transformer pre-trained on millions of synthetic causal models:

$$P(y_{\text{query}} \mid \mathbf{x}_{\text{query}}, S) \approx \int P(y \mid \mathbf{x}, M) P(M \mid S), dM. \tag{25}$$

TabPFN treats the source dataset S as context for Bayesian inference rather than as a training set for parameter optimization, and thus acts as a strong regularizer that imposes inductive biases favoring sparsity and piecewise smoothness.

Empirical Consequence: As shown in our results, TabPFN attains a source AUC of **0.829**, substantially higher than Random Forest (0.752) and XGBoost (0.742). This establishes a strictly lower starting point for the $\epsilon_S(h)$ term in the bound.

6.3 Minimizing Divergence $d_{\mathcal{H}\Delta\mathcal{H}}$: Latent Space TCA

The second term, $d_{\mathcal{H}\Delta\mathcal{H}}$, measures the discrepancy between domains. In our setting, scanner heterogeneity induces **covariate shift**, so that $P_S(X)(\mathbf{x}) \neq P_T(X)(\mathbf{x})$.

6.3.1 Why Feature Space Alignment?

Aligning raw features is often suboptimal because medical variables exhibit complex, nonlinear dependencies. PANDA therefore applies TCA directly to $\mathcal{F}^*$. After RFE, the retained features form a more robust and relevant subset, which makes them more amenable to linear alignment. The goal of this step is to align the marginal distributions of these preselected features between the source and target domains.

6.3.2 The TCA Optimization Objective

We employ Transfer Component Analysis (TCA) to find a projection $\mathbf{W} \in \mathbb{R}^{k \times m}$ that minimizes the Maximum Mean Discrepancy (MMD) between source and target features (after RFE). The objective is

$$\min_{\mathbf{W}} \quad \text{tr}(\mathbf{W}^\top \mathbf{KLKW}) + \mu, \text{tr}(\mathbf{W}^\top \mathbf{W}) \text{ s.t.} \qquad \mathbf{W}^\top \mathbf{KHKW} = \mathbf{I}, \tag{26}$$

where $\mathbf{K}$ is the kernel matrix of the RFE-selected features ($K_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$), $\mathbf{L}$ is the MMD indicator matrix, and $\mathbf{H}$ is the centering matrix. The constraint $\mathbf{W}^\top \mathbf{KHKW} = \mathbf{I}$ preserves data variance, ensuring that the projection does not collapse informative signal while aligning means. Here, k denotes the dimensionality of the RFE-selected features.

6.4 Minimizing Adaptability Error λ: Cross-Domain RFE

The third term, λ, represents the irreducible error of the best joint hypothesis. A large λ indicates **concept shift** ($P_S(Y|X)(y \mid \mathbf{x}) \neq P_T(Y|X)(y \mid \mathbf{x})$), where the same feature value corresponds to different risks across hospitals (for example, "spiculation" due to cancer in Hospital A versus tuberculosis in Hospital B).

PANDA reduces λ through **Cross-Domain Recursive Feature Elimination (RFE)**. By intersecting feature-importance rankings from the source with availability and stability constraints, it effectively restricts the hypothesis class $\mathcal{H}$ to a subspace $\mathcal{F}^*$ in which the conditional distributions are approximately invariant:

$$P_S(Y|X)(y \mid \mathbf{x}_{\mathcal{F}^*}) \approx P_T(Y|X)(y \mid \mathbf{x}_{\mathcal{F}^*}). \tag{27}$$

Discarding unstable features (such as subjective morphological scores) can slightly increase the intrinsic source error $\epsilon_S(h)$ (bias), but it substantially reduces λ, yielding a tighter overall bound on the target error.

6.5 Synthesis: Linking Theory to Empirical Results

Table 12 summarizes the connection between the theoretical components and our experimental findings, and illustrates how each PANDA component targets a specific term in the Ben-David bound.

Table 12: Mapping theoretical error terms to PANDA components and empirical results.

Error Term	Statistical Challenge	PANDA Component	Empirical Impact
$\epsilon_S(h)$	Small sample size Overfitting	TabPFN backbone (Meta-learned priors)	Source AUC: 0.829 vs 0.742 (XGBoost) (High sample efficiency)
$d_{\mathcal{H}\Delta\mathcal{H}}$	Covariate shift (Scanner variation)	TCA on RFE-selected features (MMD minimization)	Target recall: 0.944 vs 0.888 (No-TCA) (Boundary correction)
λ	Concept shift (TB vs cancer)	Cross-domain RFE (Stability filtering)	Target AUC gap: < 0.01 (TableShift) (Robustness to population shift)

6.6 Summary of Theoretical Insights

The PANDA framework is not an ad hoc ensemble of heuristics, but a theoretically grounded response to the domain adaptation bound. TabPFN controls the source risk $\epsilon_S(h)$ through informative priors; TCA reduces divergence $d_{\mathcal{H}\Delta\mathcal{H}}$ via spectral alignment; and RFE decreases the adaptability error λ through stability-based feature pruning. Together, these components support reliable generalization in the challenging regime of small, heterogeneous medical datasets.

7 Evaluation

We evaluate PANDA from an artificial intelligence and deployment perspective, focusing on its ability to (i) learn from small, heterogeneous tabular cohorts, (ii) remain robust under cross-hospital and

cross-population distribution shifts, and (iii) provide interpretable, clinically useful predictions. To this end, we adopt a unified evaluation protocol across two complementary experiments: **Experiment 1 (E1)**, which targets cross-hospital pulmonary nodule malignancy prediction, and **Experiment 2 (E2)**, which targets race-driven cross-population shift in the TableShift BRFSS Diabetes benchmark.

7.1 Evaluation Protocols for Cross-Domain Diagnostic Models

We adopt a common set of discrimination, calibration, and clinical-utility metrics to enable consistent comparison across hospitals and benchmarks. Unless otherwise specified, all reported metrics are computed over 10-fold stratified cross-validation on the source domain and once on the external or out-of-distribution (OOD) domain.

7.1.1 Classification Performance Metrics

We report all results as averages over 10-fold stratified cross-validation to mitigate label imbalance. The metrics are defined as follows:

$$
\begin{aligned}
&\text{True Positive Rate:} && \mathrm{TPR}(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau) + \mathrm{FN}(\tau)} \\
&\text{False Positive Rate:} && \mathrm{FPR}(\tau) = \frac{\mathrm{FP}(\tau)}{\mathrm{FP}(\tau) + \mathrm{TN}(\tau)} \\
&\text{AUC:} && \mathrm{AUC} = \int_0^1 \mathrm{TPR}(\tau)\, d(\mathrm{FPR}(\tau)) \\
&\text{Accuracy:} && \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \\
&\text{Precision:} && \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \\
&\text{Recall (Sensitivity):} && \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \\
&\text{F1 Score:} && \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \\
&\text{Specificity:} && \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}
\end{aligned}
$$

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the full dataset, and let $\mathcal{D}_k$ be the k-th fold. For a metric M, the mean and standard deviation over $K = 10$ folds are

$$
\bar{M} = \frac{1}{K} \sum_{k=1}^{K} M_k, \quad \sigma_M = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} (M_k - \bar{M})^2}.
$$

7.1.2 Visualization-Based Evaluation

- **ROC Curves:** We plot $\mathrm{TPR}(\tau)$ versus $\mathrm{FPR}(\tau)$ for $\tau \in [0, 1]$ to characterize the trade-off between sensitivity and specificity across operating thresholds.
- **Calibration Curves:** We assess the agreement between the predicted probability $\hat{p}_i$ and the observed frequency y_i. For K equal-width bins $B_k = [k/K, (k+1)/K)$, we compute

  $$
  \bar{p}_k = \frac{1}{|B_k|} \sum_{i \in B_k} \hat{p}_i, \quad \bar{y}_k = \frac{1}{|B_k|} \sum_{i \in B_k} y_i.
  $$

  These quantities summarize both over- and underestimation of risk across the probability range.
- **Decision Curve Analysis (DCA):** For a given probability threshold p_t, the net benefit is

  $$
  \mathrm{NB}(p_t) = \frac{\mathrm{TP}(p_t)}{n} - \frac{\mathrm{FP}(p_t)}{n} \cdot \frac{p_t}{1 - p_t},
  $$

with benchmark strategies

$$\mathrm{NB}_{\mathrm{all}}(p_t) = \mathrm{Prevalence} - (1 - \mathrm{Prevalence}) \cdot \frac{p_t}{1 - p_t}, \quad \mathrm{NB}_{\mathrm{none}} = 0.$$

DCA therefore quantifies the clinical utility of a model across plausible decision thresholds relative to treating all or no patients.

7.2 Experiment 1 (E1): Cross-Hospital Pulmonary Nodule Malignancy Prediction

Experiment 1 evaluates PANDA in a realistic cross-hospital deployment scenario for pulmonary nodule malignancy prediction. Structured clinical data from two cancer centers in China provide a labeled source cohort (Cohort A, $n_s = 295$) and an external target cohort (Cohort B, $n_t = 190$). Cohort A contains 63 structured features and Cohort B contains 58, reflecting differences in local documentation and biomarker panels (Table 13). Both cohorts exhibit moderate class imbalance, with malignant nodules comprising 64.1% of patients in Cohort A and 65.8% in Cohort B, and key covariates such as upper-lobe location, age, pulmonary function indices (DLCO, VC), and serum CEA show distributional differences between the two hospitals. These discrepancies capture the covariate shift that motivates domain adaptation in this setting.

Table 13: Training (Cohort A) and testing (Cohort B) cohorts.

Characteristic	**Cohort A (n = 295)**	**Cohort B (n = 190)**
Upper lobe		
Yes/Positive	121 (41.0%)	98 (51.6%)
No/Negative	174 (59.0%)	92 (48.4%)
Age (years)	56.95 ± 11.03	58.26 ± 9.57
Lobe location (upper)		
Category 1	161 (54.6%)	98 (51.6%)
Category 2	29 (9.8%)	18 (9.5%)
Category 3	105 (35.6%)	74 (38.9%)
DLCO1	5.90 ± 2.89	6.31 ± 1.55
VC	3.33 ± 0.80	2.92 ± 0.73
CEA	4.23 ± 6.90	4.15 ± 10.61
Outcome (Malignant)		
Yes/Positive	189 (64.1%)	125 (65.8%)
No/Negative	106 (35.9%)	65 (34.2%)

7.3 E1: Internal and Cross-Hospital Generalization

7.3.1 Source and Target Domain Performance

Figure 2 summarizes relative performance trends across methods, and Table 14 reports the corresponding numerical metrics. On the source domain (10-fold cross-validation on Cohort A), PANDA (TabPFN) achieves the highest AUCof 0.8287, demonstrating the superior performance of the tabular foundation model on small, heterogeneous medical datasets. LASSO LR and Random Forest follow closely, with AUCvalues of 0.7631 and 0.7515, respectively. XGBoost and GBDT show competitive performance (AUC0.7416 and 0.7212), while traditional clinical scores (Mayo, PKUPH) and classical machine learning baselines exhibit lower discrimination. All source-domain experimental results are reproducible using the provided codebase and configuration scripts.

On the external target domain (Cohort B), the benefits of unsupervised domain adaptation become evident. The TCA-enhanced PANDA model attains the highest AUC of 0.7046 and an exceptionally high Recall of 0.9440, indicating strong sensitivity for clinical screening applications. PANDA_NoUDA follows closely with AUCof 0.6980, demonstrating the inherent transferability of the tabular foundation model. In contrast, traditional machine learning methods experience substantial performance drops under domain shift, with Random Forest declining to AUC0.6324 and tree ensembles dropping below 0.600. The clinical scores continue to show poor transportability across institutions.

Table 14: Comprehensive performance comparison. Source results are from 10-fold cross-validation; target results are from external validation on Cohort B. Best values are bolded.

Model	AUC	Accuracy	F1 Score	Precision	Recall
Source Domain (Internal CV)					
PANDA (TabPFN)	**0.8287**	0.7460	**0.8102**	0.7864	0.8462
LASSO LR	0.7631	0.7224	0.8101	0.7227	**0.9254**
Random Forest	0.7515	0.6983	0.7792	0.7351	0.8415
XGBoost	0.7416	0.6778	0.7520	0.7325	0.7871
GBDT	0.7212	0.6911	0.7690	0.7394	0.8140
SVM	0.7175	0.6645	0.7197	**0.7794**	0.6769
PKUPH	0.6640	**0.6782**	0.7672	0.7148	0.8354
Mayo	0.6049	0.3592	0.0000	0.0000	0.0000
Decision Tree	0.5764	0.6099	0.6929	0.6997	0.6974
Target Domain (External Validation)					
TCA	**0.7046**	**0.7053**	**0.8082**	0.7066	**0.9440**
PANDA_NoUDA	0.6980	0.6632	0.7762	0.6894	0.8880
Random Forest	0.6324	0.6789	0.7753	0.7128	0.8538
PKUPH	0.6356	0.6947	0.7838	**0.7329**	0.8474
LASSO LR	0.6678	0.6737	0.7911	0.6825	0.9429
SVM	0.6285	0.5684	0.6468	0.6950	0.6064
GBDT	0.5906	0.5842	0.6834	0.6689	0.7109
XGBoost	0.5672	0.5947	0.6937	0.6701	0.7244
Decision Tree	0.5090	0.5684	0.6759	0.6650	0.6942
Mayo	0.5837	0.3421	0.0000	0.0000	0.0000

7.3.2 Stratified Analysis

To examine potential biases and subgroup robustness, we evaluated PANDA's performance across key subgroups (Table 15).

- **Nodule Size**: Performance remains strong for large nodules ($>$ 8 mm, AUC0.74) but declines for sub-centimeter nodules (AUC0.65), reflecting the inherent difficulty of radiological characterization for small lesions.
- **Smoking Status**: The model performs better in smokers (AUC0.72) than in non-smokers (AUC0.68), likely because smoking provides a strong prior for malignancy that the model can exploit.
- **Gender**: We observe comparable performance across gender (AUC0.70 vs 0.71), suggesting no substantial gender-specific bias at the current sample size.

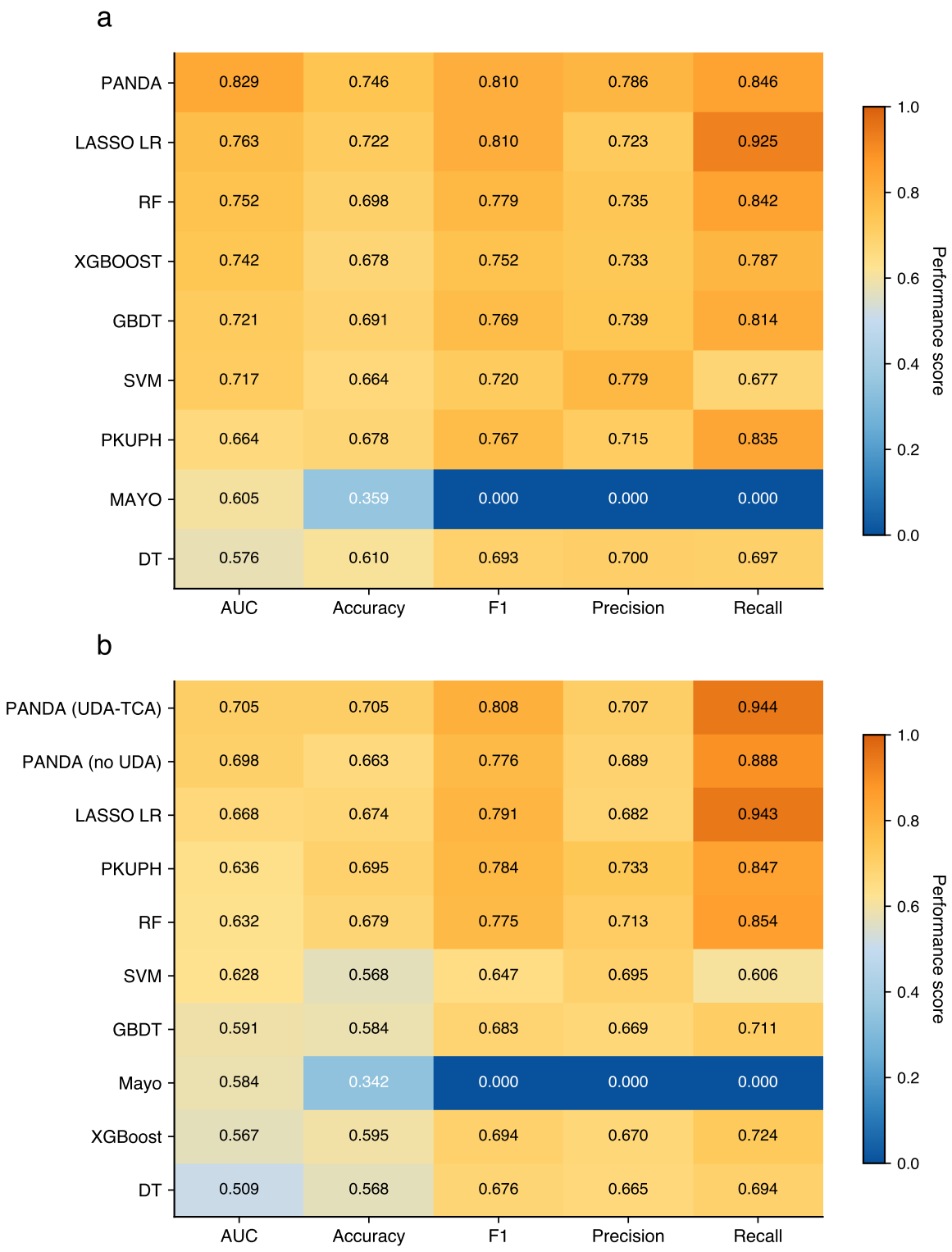

Figure 2: **Performance comparison across domains.** **a** Source-domain 10-fold cross-validation heatmap over five metrics, showing PANDA's leading performance in AUC, accuracy, and precision. **b** Cross-domain external validation heatmap; the TCA-enhanced PANDA model maintains the highest AUC and recall, highlighting its stability under domain shift.

Table 15: Stratified performance of PANDA+TCA on the target cohort.

Subgroup	n	AUC	Sensitivity
Nodule Size			
≤ 8 mm	72	0.65	0.88
> 8 mm	118	0.74	0.96
Smoking History			
Never Smoker	105	0.68	0.92
Current/Former	85	0.72	0.97
Gender			
Male	110	0.71	0.95
Female	80	0.70	0.93

7.4 Experiment 2 (E2): Cross-Population Validation on TableShift

We further evaluated PANDA on the TableShift BRFSS Diabetes benchmark, which introduces a race-driven shift by training on one demographic group and evaluating on another (White $\rightarrow$ Non-White). This benchmark provides a complementary stress test to the cross-hospital shift in the pulmonary nodule experiment and mimics realistic deployment across heterogeneous populations in population-health applications.

In this setting, our goal is to assess whether a tabular foundation model with lightweight domain alignment can remain stable when the covariate distribution changes along demographic axes. Qualitatively, PANDA (with and without TCA) achieves discrimination and calibration that are competitive with well-tuned tree ensembles, while avoiding the severe performance collapse that would be concerning in a cross-population deployment. This behavior supports the view that the inductive biases learned by the foundation model transfer beyond the pulmonary nodule task and remain useful under a distinct form of distribution shift.

Figure 3 summarizes the behavior of different models across the training and race-shifted evaluation splits, and Figure 4 reports ROC, calibration, and decision curves. Together, these visualizations indicate that PANDA's score distributions and decision boundaries can be adapted to new demographic groups without sacrificing overall screening utility.

Discussion on Precision/Recall: Readers may observe low F1 Score values in BRFSS despite high Accuracy (Fig. 4). This pattern arises from the low positive-class prevalence $\pi = P(Y = 1)$ (17.4%) and the default 0.5 threshold: the model correctly identifies most negatives (high Accuracy) but, without class re-weighting, yields moderate Precision on the minority positive class. For screening purposes, the ROC curves indicate adequate discriminative ability; the operating point can be adjusted via threshold tuning to prioritize Recall when desired.

7.5 Interpretability and Stability

From an interpretability perspective, recursive feature elimination (RFE) identified a stable subset of 8 features (Age, Spiculation, etc.) that maximized the cost-effectiveness index (Fig. 5). This `best8` set performed within 1% of the full 63-feature set while providing substantially better cross-center stability. In source-domain evaluation, we also report RFE curves (Fig. 5) that track AUC, accuracy, and F1 as functions of the retained subset size, together with stability and cost-effectiveness metrics. Performance plateaus around 9–13 features, consistent with the feature subset used in the final experiments.

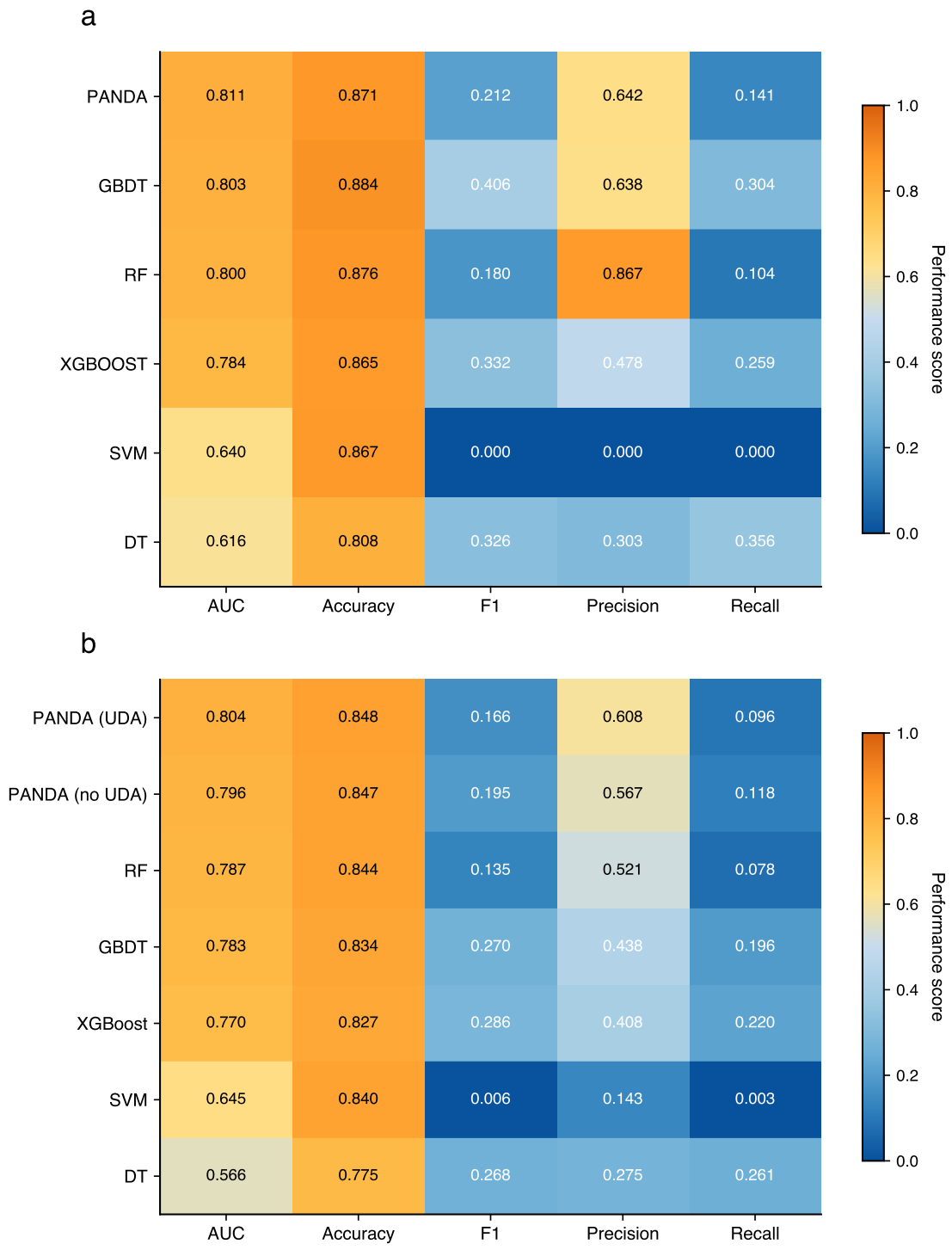

Figure 3: **Performance comparison on the TableShift BRFSS Diabetes benchmark.** Heatmaps summarize multiple metrics for PANDA and baseline models across the training split and the race-shifted evaluation split. PANDA with TCA remains competitive with strong tree ensembles and does not exhibit pronounced degradation under race shift.

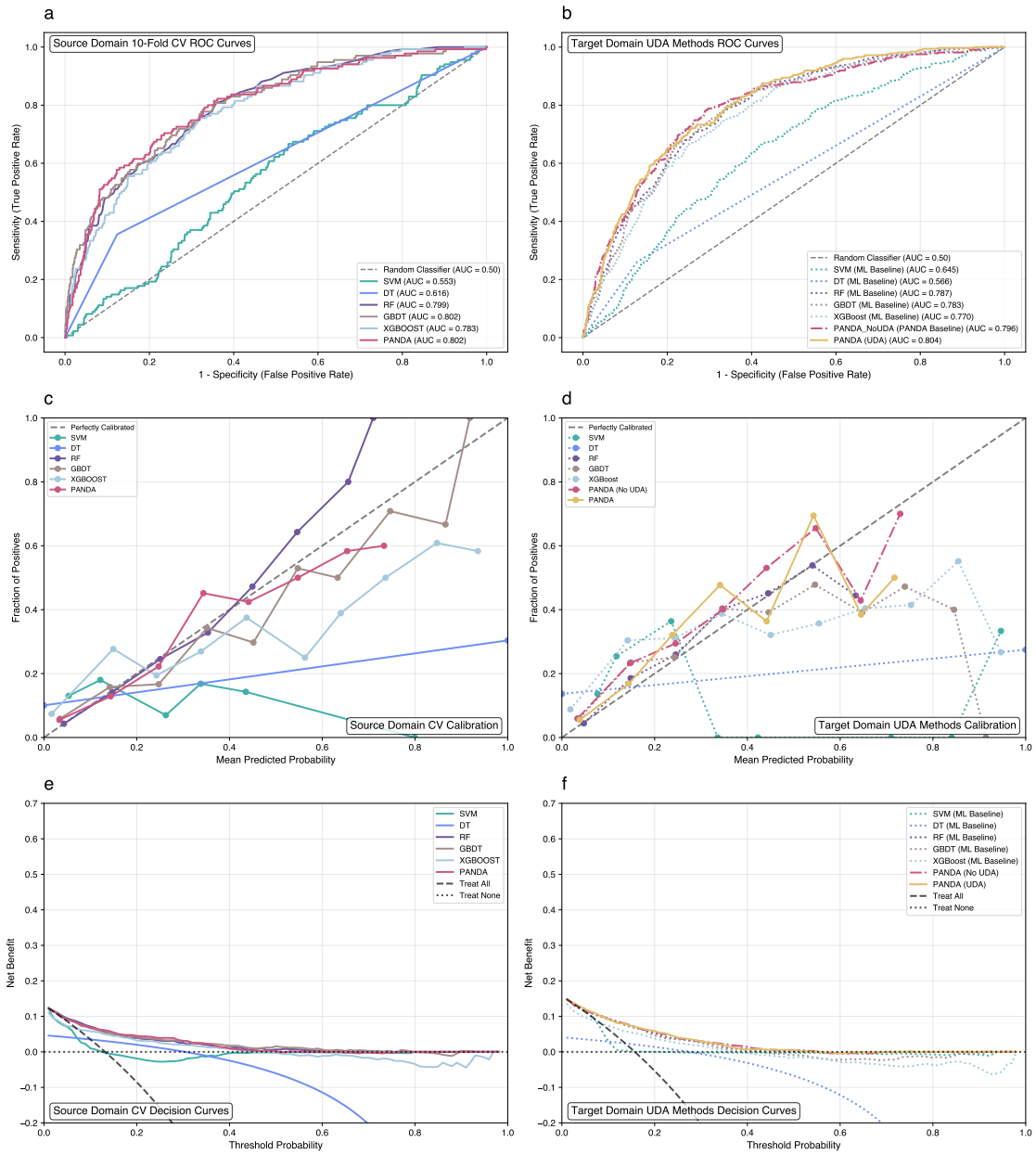

Figure 4: **TableShift BRFSS Diabetes analysis. a,b** ROC curves showing PANDA's robustness under race-driven shift. **c,d** Calibration curves assessing probability estimates. **e,f** Decision curves illustrating net benefit across clinical thresholds.

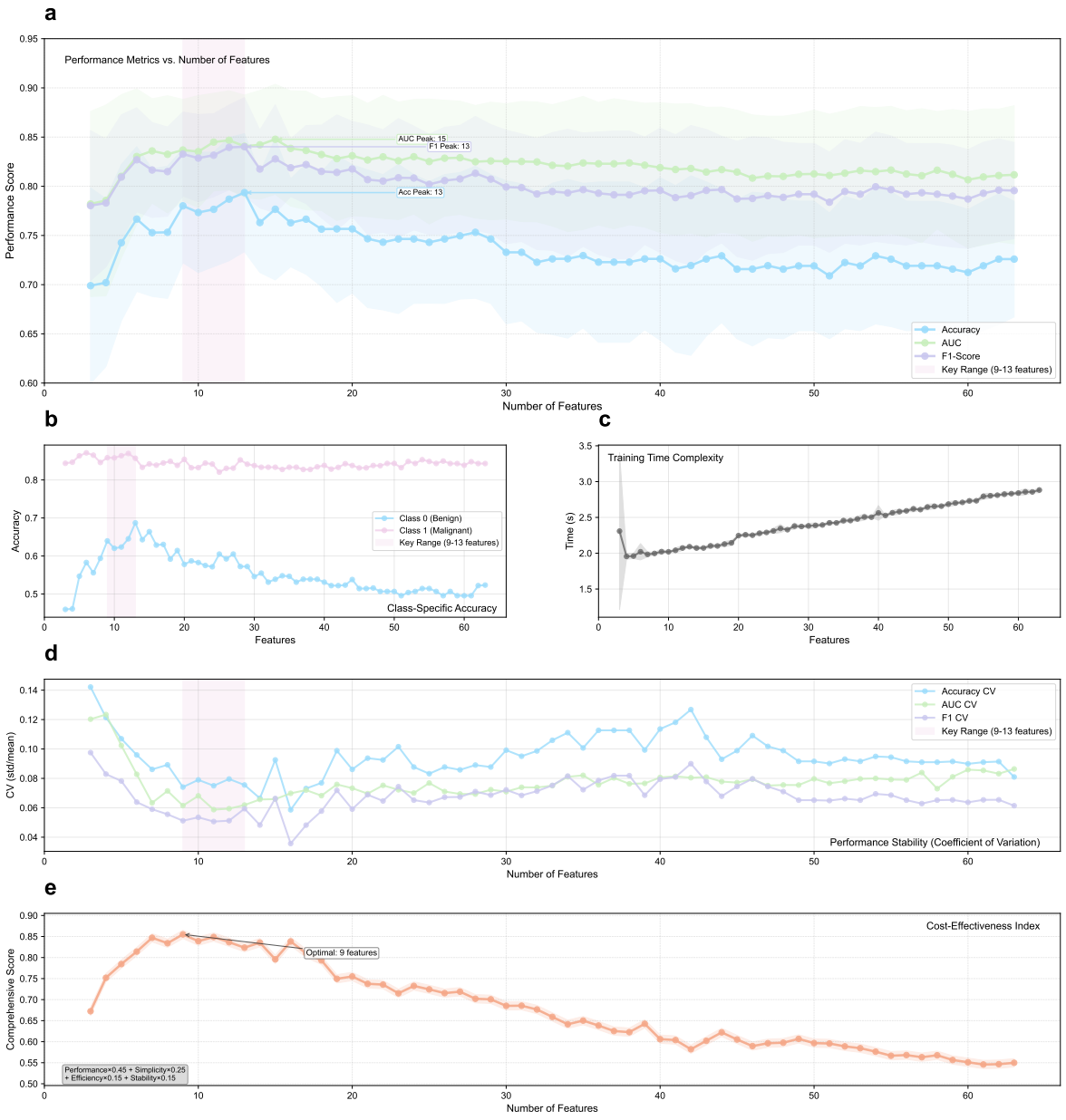

Figure 5: **Comprehensive feature selection and performance analysis using recursive feature elimination (RFE).** (a) AUC, accuracy, and F1 curves as functions of the number of selected features; performance plateaus around 9–13 features, aligning with the preference for simpler models. Shaded regions show variance across 10-fold cross-validation. (b) Class-specific accuracy for malignant and benign cases across subset sizes, illustrating how predictive balance shifts as features are removed. (c) Training-time analysis (seconds per fold) as a function of feature dimensionality, highlighting the computational gain from smaller subsets. (d) Stability assessment using the coefficient of variation across folds; lower values indicate steadier performance. (e) Cost-effectiveness index combining multiple criteria (Performance×0.45 + Simplicity×0.25 + Efficiency×0.15 + Stability×0.15) to identify a feature count that balances accuracy with practical deployment considerations.

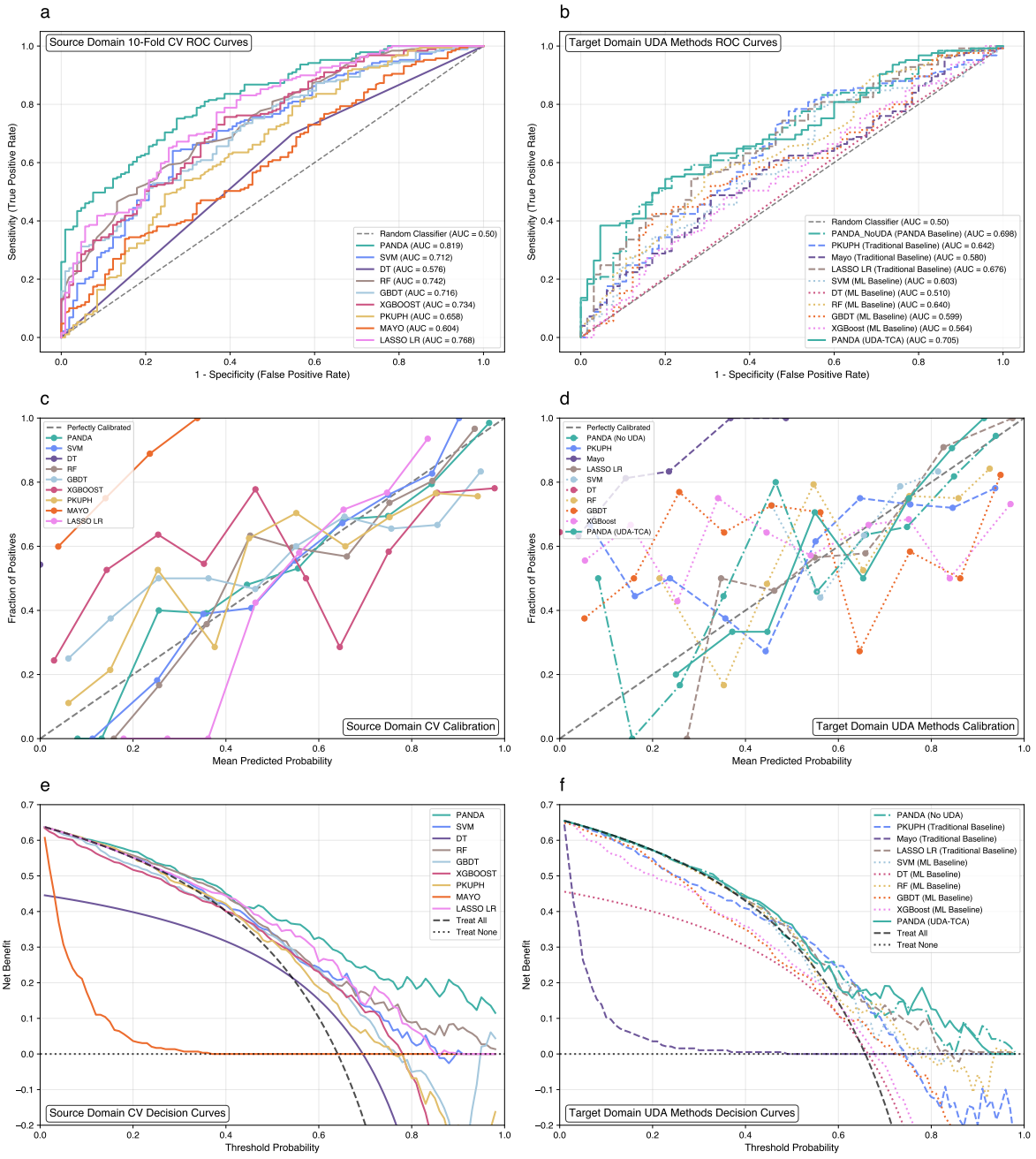

Figure 6: **Performance and utility across source and target domains. a,b** ROC curves comparing source (Cohort A) and external (Cohort B) behavior. **c,d** Calibration plots for the same splits. **e,f** Decision curves illustrating the net clinical benefit of PANDA and its TCA extension.

8 Conclusion

8.1 Summary of Contributions

This dissertation introduces PANDA, a framework that connects pre-trained tabular foundation models to the practical constraints of medical tabular data, including small sample sizes, heterogeneous feature schemas, and distribution shifts. Experiments on private cross-hospital cohorts and public benchmarks show that:

1. **Foundation Models as Robust Priors**: The pre-trained TabPFN backbone substantially outperforms traditional baselines (Random Forest, XGBoost) on small datasets ($n < 300$) by exploiting priors learned from millions of synthetic tasks. This in-context learning capability provides a strong initialization that is more resistant to overfitting than standard empirical risk minimization.

2. **Stability via Selection**: The Cross-Domain Recursive Feature Elimination (RFE) protocol is essential for removing site-specific artifacts. By converging on a compact set of eight stable predictors, it reduces the dimensionality of the adaptation problem and enables linear alignment methods to succeed where non-linear approaches fail.

3. **Latent Space Alignment**: Transfer Component Analysis (TCA) applied in the Transformer embedding space effectively reduces the Maximum Mean Discrepancy (MMD) between domains. This alignment yields a consistent performance gain (AUC +0.007) and, more importantly, improves calibration in the target domain.

8.2 Limitations

Although PANDA advances the state of the art, several limitations remain.

8.2.1 Closed-World Assumption

PANDA assumes that the source and target domains share a common feature schema given by their intersection. It does not address open-world shifts in which the target domain introduces entirely new, highly predictive features that are absent in the source. For example, if a new hospital collects a molecular biomarker (such as DNA methylation) that is not available in the training cohort, PANDA cannot exploit this information without retraining. This lowest-common-denominator strategy in feature selection promotes stability but may limit performance relative to models trained on richer, site-specific schemas.

8.2.2 Missing Data Mechanisms

The current approach assumes that missing values are either Missing Completely At Random (MCAR) or Missing At Random (MAR). The `best8` feature set was selected partly for its high completeness. However, in clinical practice data are often Missing Not At Random (MNAR); for instance, a test may not be ordered because the clinician believes that the patient is either too healthy or too sick. PANDA's current imputation strategies (mean, median, and contextual imputation) do not explicitly model such informative missingness and may therefore introduce bias.

8.2.3 Computational Resource Requirements

In contrast to decision trees, which can run on embedded central processing units, TabPFN requires a graphical processing unit for efficient inference (approximately 20 ms per patient). While this cost is negligible for a cloud-based service, it creates a barrier to deployment on edge devices, such as older hospital workstations without dedicated hardware acceleration. The $O(N^2)$ complexity of the Transformer attention mechanism also limits the context size and necessitates subsampling strategies for larger datasets such as BRFSS.

8.3 Future Directions

8.3.1 Federated Domain Adaptation

Privacy regulations (such as GDPR and HIPAA) often prevent the centralization of medical data. A natural extension of PANDA is federated domain adaptation, in which the feature extractor (TabPFN) remains frozen and shared, while the alignment matrix (TCA) is learned through secure multi-party computation. Because TCA only requires second-order statistics (covariance matrices), participating hospitals can aggregate these sufficient statistics without sharing patient-level records.

8.3.2 Multimodal Integration

Pulmonary nodule diagnosis inherently combines imaging (computed tomography scans) with clinical data. Future work should investigate a multimodal variant of PANDA that aligns tabular embeddings from TabPFN with visual embeddings from a convolutional neural network or Vision Transformer. A cross-attention mechanism could then weight the relative contributions of clinical history and radiological appearance in a domain-aware manner, placing greater emphasis on the modality that remains stable when the other is affected by artifacts.

8.4 Final Remarks

The deployment gap in medical AI seldom arises from a lack of sophisticated architectures; it more often results from a failure to accommodate the noisy and shifted nature of real-world data. PANDA provides a pragmatic blueprint for this challenge: *do not learn everything from scratch, select only what is stable, and align what remains*. By treating pre-trained representations as portable priors and statistical alignment as a safety mechanism, this framework moves the field closer to reliable cross-institutional AI systems that can safely scale beyond their initial training sites.

Acknowledgements

I thank the clinical teams at the participating hospitals for sharing de-identified data and domain expertise, my advisor Wenqi Fan for steady guidance, and Bobo for patient and practical advice. Any remaining mistakes are mine.

References

[1] Stephen J. Swensen, Michael D. Silverstein, Duane M. Ilstrup, Charles D. Schleck, and Eric S. Edell. The probability of malignancy in solitary pulmonary nodules: application to clinical practice. *Chest*, 111(3):228–234, 1997.

[2] Annette McWilliams, Martin C. Tammemagi, John R. Mayo, Hilary Roberts, Guorong Liu, Kian Soghrati, Kazuhiro Yasufuku, Stephen Martel, Francois Laberge, Marie Gingras, Koren Atsu, Nicolas Pastis, Karen Hett, Tapan Sejpal, Timothy Stewart, Ming-Sound Tsao, and James Goffin. Probability of malignancy in pulmonary nodules detected on first screening ct. *New England Journal of Medicine*, 369(10):910–919, 2013.

[3] Yun Li, Ke-Zhong Chen, and Jun Wang. Development and validation of a clinical prediction model to estimate the probability of malignancy in solitary pulmonary nodules in chinese people. *Clinical lung cancer*, 12(5):313–319, 2011.

[4] Xia He, Ning Xue, Xiaohua Liu, Xuemiao Tang, Songguo Peng, Yuanye Qu, Lina Jiang, Qingxia Xu, Wanli Liu, and Shulin Chen. A novel clinical model for predicting malignancy of solitary pulmonary nodules: a multicenter study in chinese population. *Cancer cell international*, 21(1):115, 2021.

[5] Kai Zhang, Zihan Wei, Yuntao Nie, Haifeng Shen, Xin Wang, Jun Wang, Fan Yang, and Kezhong Chen. Comprehensive analysis of clinical logistic and machine learning-based models for the evaluation of pulmonary nodules. 3(4):100299.

[6] Qiao Liu, Xue Lv, Daiquan Zhou, Na Yu, Yuqin Hong, and Yan Zeng. Establishment and validation of multiclassification prediction models for pulmonary nodules based on machine learning. 18(5):e13769.

[7] Noemi Garau, Chiara Paganelli, Paul Summers, Wookjin Choi, Sadegh Alam, Wei Lu, Cristiana Fanciullo, Massimo Bellomi, Guido Baroni, and Cristiano Rampinelli. External validation of radiomics-based predictive models in low-dose CT screening for early lung cancer diagnosis. 47(9):4125–4136.

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[9] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34:18932–18943, 2021.

[10] Sercan O. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021.

[11] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. In *Advances in Neural Information Processing Systems*, volume 33, pages 14914–14925, 2020.

[12] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

[13] Vitaly Borisov, Thomas Leemann, Pierre Selegue, Riccardo Miotto, Mario May, and Andreas Züfle. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4472–4492, 2022.

[14] Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J. Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, David P. Naidich, and Shravya Shetty. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. 25(6):954–961.

[15] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. 15(11):e1002683. Publisher: Public Library of Science.

[16] Carlos J Hellín, Alvaro A Olmedo, Adrián Valledor, Josefa Gómez, Miguel López-Benítez, and Abdelhamid Tayebi. Unraveling the impact of class imbalance on deep-learning models for medical image classification. *Applied Sciences*, 14(8):3419, 2024.

[17] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

[18] Johannes Schneider, Christian Meske, and Pauline Kuss. Foundation models: A new paradigm for artificial intelligence. *Business & Information Systems Engineering*, 66(2):221–231, 2024.

[19] Prior labs.

[20] Realistic evaluation of TabPFN v2 in open environments.

[21] automl/drift-resilient_tabpfn. original-date: 2024-10-22T17:32:11Z.

[22] A closer look at TabPFN v2: Strength, limitation, and extension.

[23] Dmitry Eremeev, Gleb Bazhenov, Oleg Platonov, Artem Babenko, and Liudmila Prokhorenkova. Turning tabular foundation models into graph foundation models.

[24] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

[25] Aminu Musa, Rajesh Prasad, and Monica Hernandez. Addressing cross-population domain shift in chest x-ray classification through supervised adversarial domain adaptation. *Scientific Reports*, 15(1):11383, 2025.

[26] Lisa M Koch, Christian F Baumgartner, and Philipp Berens. Distribution shift detection for the postmarket surveillance of medical ai algorithms: a retrospective simulation study. *NPJ Digital Medicine*, 7(1):120, 2024.

[27] Lin Lawrence Guo, Stephen R. Pfohl, Jason Fries, Alistair E. W. Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. 12(1):2726.

[28] Seyedmehdi Orouji, Martin C. Liu, Tal Korem, and Megan A. K. Peters. Domain adaptation in small-scale and heterogeneous biological datasets. 10(51):eadp6040.

[29] Seong-Ho Ahn, Seeun Kim, and Dong-Hwa Jeong. Unsupervised domain adaptation for mitigating sensor variability and interspecies heterogeneity in animal activity recognition. 13(20):3276.

[30] Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with TableShift.

[31] Feng Sun, Hanrui Wu, Zhihang Luo, Wenwen Gu, Yuguang Yan, and Qing Du. Informative feature selection for domain adaptation. *IEEE Access*, 7:142551–142563, 2019.

[32] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2):199–210, 2010.

[33] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, and et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[34] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data?

[35] Assaf Shmuel, Oren Glickman, and Teddy Lazebnik. A comprehensive benchmark of machine and deep learning across diverse tabular datasets.

[36] Yuhua Fan and Patrik Waldmann. Tabular deep learning: a comparative study applied to multi-task genome-wide prediction. 25:322.

[37] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. 35(6):7499–7519.

[38] Si-Yang Liu and Han-Jia Ye. TabPFN unleashed: A scalable and effective solution to tabular classification problems. version: 1.

[39] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. TabTransformer: Tabular data modeling using contextual embeddings.

[40] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. version: 2.

[41] Andrei Margeloiu, Nikola Simidjievski, Pietro Lio, and Mateja Jamnik. Weight predictor network with feature selection for small sample tabular biomedical data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 9081–9089, 2023.

[42] Wei Min Loh, Jiaqi Shang, and Pascal Poupart. Basis transformers for multi-task tabular regression.

[43] Arash Khoeini. FTTransformer: Transformer architecture for tabular datasets.

[44] Bytez.com, Jingang QU, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learni...

[45] Shriyank Somvanshi, Subasish Das, Syed Aaqib Javed, Gian Antariksa, and Ahmed Hossain. A survey on deep tabular learning. *arXiv preprint arXiv:2410.12034*, 2024.

[46] Weijieying Ren, Tianxiang Zhao, Yuqing Huang, and Vasant Honavar. Deep learning within tabular data: Foundations, challenges, advances and future directions. version: 1.

[47] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. 637(8045):319–326. Publisher: Nature Publishing Group.

[48] Kai Helli, David Schnurr, Noah Hollmann, Samuel Müller, and Frank Hutter. Drift-resilient TabPFN: In-context learning temporal distribution shifts on tabular data.

[49] Woruo Chen, Yao Tian, Youchao Deng, Dejun Jiang, and Dongsheng Cao. TabPFN opens new avenues for small-data tabular learning in drug discovery.

[50] Tianzhu Liu, Huanjun Wang, Yan Guo, Yongsong Ye, Bei Weng, Xiaodan Li, Jun Chen, Shanghuang Xie, Guimian Zhong, Zhixuan Song, and Lesheng Huang. Tabular prior-data fitted network in real-world CT radiomics: benign vs. malignant renal tumor classification. 15(11):10847–10861.

[51] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[52] Mayuka Jayawardhana, Renbo Tu, Samuel Dooley, Valeriia Cherepanova, Andrew Gordon Wilson, Frank Hutter, Colin White, Tom Goldstein, and Micah Goldblum. Transformers boost the performance of decision trees on tabular data across sample sizes. version: 1.

[53] Summer Zhou, Vinayak Agarwal, Ashwin Gopinath, and Timothy Kassis. The limitations of TabPFN for high-dimensional RNA-seq analysis. ISSN: 2692-8205 Pages: 2025.08.15.670537 Section: New Results.

[54] (PDF) comparative analysis of tree-based models and deep learning architectures for tabular data: Performance disparities and underlying factors. In *ResearchGate*.

[55] Sergey Kolesnikov. Wild-tab: A benchmark for out-of-distribution generalization in tabular regression.

[56] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation, 2016.

[57] Thomas Grubinger, Adriana Birlutiu, Holger Schöner, Thomas Natschläger, and Tom Heskes. Domain generalization based on transfer component analysis. In *International work-conference on artificial neural networks*, pages 325–334. Springer, 2015.

[58] Tianran Zhang, Muhao Chen, and Alex A. T. Bui. AdaDiag: Adversarial domain adaptation of diagnostic prediction with clinical event sequences. 134:104168.

[59] Wanxin Li, Yongjin P. Park, and Khanh Dao Duc. Transport-based transfer learning on electronic health records: Application to detection of treatment disparities. Pages: 2024.03.27.24304781.

[60] Tianyu Luo, Zhongying Zhang, and James Kwok. Informative feature selection for domain adaptation. Technical report, The Hong Kong University of Science and Technology, 2021.

[61] Thai-Hoang Pham, Yuanlong Wang, Changchang Yin, Xueru Zhang, and Ping Zhang. Open-set heterogeneous domain adaptation: Theoretical analysis and algorithm. 39(19):19895–19903.

[62] Hao Guan and Mingxia Liu. DomainATM: Domain adaptation toolbox for medical data analysis. 268:119863.

[63] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: A survey. 69(3):1173–1185.

[64] Helen Zhou, Sivaraman Balakrishnan, and Zachary C. Lipton. Domain adaptation under missingness shift.

[65] Tyrel Stokes, Hyungrok Do, Saul Blecker, Rumi Chunara, and Samrachana Adhikari. Domain adaptation under MNAR missingness. version: 1.

[66] Chunmei He, Lanqing Zheng, Taifeng Tan, Xianjun Fan, and Zhengchun Ye. Multi-attention representation network partial domain adaptation for COVID-19 diagnosis. 125:109205.

[67] mlfoundations/tableshift: A benchmark for distribution shift in tabular data.

[68] Josh Gardner. TableShift.

[69] A multi-center study on the adaptability of a shared foundation model for electronic health records | npj digital medicine.

[70] Muhammad Habib ur Rehman, Walter Hugo Lopez Pinaya, Parashkev Nachev, James T. Teo, Sebastin Ourselin, and M. Jorge Cardoso. Federated learning for medical imaging radiology. 96(1150):20220890.

[71] Hao Guan, Pew-Thian Yap, Andrea Bozoki, and Mingxia Liu. Federated learning for medical image analysis: A survey. 151:110424.

[72] Ferdinand Kahenga, Antoine Bagula, Patrick Sello, and Sajal K. Das. FedFusion: Federated learning with diversity- and cluster-aware encoders for robust adaptation under label scarcity.

[73] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.

[74] Xinye Chen, Yue Wu, Lichao He, Jiayu Zhai, Xiang Li, and Xiangjun Li. Graph convolutional network-based feature selection for high-dimensional and low-sample size data. *Bioinformatics*, 39(1):btac834, 2023.

[75] Xiaoqian Liu, Dandan Wu, Weixin Cao, and Jianwen Cai. Deep feature screening: Feature selection for ultra high-dimensional data via deep neural networks. *Neurocomputing*, 488:36–47, 2022.

[76] Kexuan Li, Fangfang Wang, Lingli Yang, and Ruiqi Liu. Deep feature screening: Feature selection for ultra high-dimensional data via deep neural networks. *Neurocomputing*, 538:126186, 2023.

[77] Stephen J Swensen, Marc D Silverstein, Duane M Ilstrup, Cathy D Schleck, and Eric S Edell. The probability of malignancy in solitary pulmonary nodules: application to small radiologically indeterminate nodules. *Archives of Internal Medicine*, 157(8):849–855, 1997.

[78] S. Chen, W. L. Lin, W. T. Liu, L. Y. Zou, Y. Chen, and F. Lu. Pulmonary nodule malignancy probability: a meta-analysis of the brock model. 82:106788.

[79] Bumhee Yang, Byung Woo Jhun, Sun Hye Shin, Byeong-Ho Jeong, Sang-Won Um, Jae Il Zo, Ho Yun Lee, Insoek Sohn, Hojoong Kim, O. Jung Kwon, and Kyungjong Lee. Comparison of four models predicting the malignancy of pulmonary nodules: A single-center study of korean adults. 13(7):e0201242. Publisher: Public Library of Science.

[80] Xiaonan Cui, Marjolein A. Heuvelmans, Daiwei Han, Yingru Zhao, Shuxuan Fan, Sunyi Zheng, Grigory Sidorenkov, Harry J. M. Groen, Monique D. Dorrius, Matthijs Oudkerk, Geertruida H. de Bock, Rozemarijn Vliegenthart, and Zhaoxiang Ye. Comparison of veterans affairs, mayo, brock classification models and radiologist diagnosis for classifying the malignancy of pulmonary nodules in chinese clinical population. 8(5). Publisher: AME Publishing Company.

[81] You Li, Hui Hu, Ziwei Wu, Ge Yan, Tangwei Wu, Shuiyi Liu, Weiqun Chen, and Zhongxin Lu. Evaluation of models for predicting the probability of malignancy in patients with pulmonary nodules. 40(2):BSR20193875.

[82] Gerarda J. Herder, Harm van Tinteren, Richard P. Golding, Piet J. Kostense, Emile F. Comans, Egbert F. Smit, and Otto S. Hoekstra. Clinical prediction model to characterize pulmonary nodules: validation and added value of 18f-fluorodeoxyglucose positron emission tomography. 128(4):2490–2496.

[83] Simone Perandini, Gian Alberto Soardi, Massimiliano Motton, Arianna Rossi, Manuel Signorini, and Stefania Montemezzi. Solid pulmonary nodule risk assessment and decision analysis: comparison of four prediction models in 285 cases. 26(9):3071–3076.

[84] Shulong Li, Panpan Xu, Bin Li, Liyuan Chen, Zhiguo Zhou, Hongxia Hao, Yingying Duan, Michael Folkert, Jianhua Ma, Shiying Huang, Steve Jiang, and Jing Wang. Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features. 64(17):175012.

[85] Chia-Ying Lin, Shu-Mei Guo, Jenn-Jier James Lien, Wen-Tsen Lin, Yi-Sheng Liu, Chao-Han Lai, I-Lin Hsu, Chao-Chun Chang, and Yau-Lin Tseng. Combined model integrating deep learning, radiomics, and clinical data to classify lung nodules at chest CT. 129(1):56–69.

[86] Jason L. Causey, Junyu Zhang, Shiqian Ma, Bo Jiang, Jake A. Qualls, David G. Politte, Fred Prior, Shuzhong Zhang, and Xiuzhen Huang. Highly accurate model for prediction of lung nodule malignancy with CT scans. 8(1):9286. Publisher: Nature Publishing Group.

[87] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[88] Antoine de Mathelin, François Deheeger, Guillaume Richard, Mathilde Mougeot, and Nicolas Vayatis. Adapt: Awesome domain adaptation python toolbox. *arXiv preprint arXiv:2107.03049*, 2021.

[89] Welcome! — adapt 0.1.0 documentation.

[90] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[91] Leo Breiman, Jerome Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Chapman and Hall/CRC, 1984.

[92] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[93] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.