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

Fragmented hospital silos and strict privacy rules often leave medical AI models staring at small, uneven, mismatched tabular cohorts, so models trained directly on those data tend to wobble when moved between sites. Here we sketch *PANDA* (Pretrained Adaptation Network with Domain Alignment)–a cross-hospital setup that leans on a pre-trained tabular foundation model, keeps the feature budget lean, and folds in unsupervised domain adaptation, even if calling it a framework is arguably generous. PANDA uses a TabPFN-style Transformer encoder meta-trained on millions of synthetic tables; that pretraining appears to capture higher-order interactions that tuned gradient-boosting ensembles often miss when samples are scarce. A cross-cohort RFE step uses the foundation model to identify eight biomarkers that stay predictive across both hospitals, cutting data-collection demands and stabilizing interpretation. To ease distribution gaps, we add TCA to the training loop so source and target cohorts land in a shared latent space. This mix–foundation-model representations, RFE-filtered features, and TCA–seems to reduce covariate shift and keep those eight variables useful even when each site ranks them differently. On two lung-nodule cohorts (295 training, 190 external), PANDA lifts AUC and sensitivity over supervised and non-adaptive baselines, hinting that pairing foundation-model priors with statistical alignment may improve generalization in small, cross-domain medical tasks.

INSERT_TOC_HERE

1 Introduction

Early and accurate prediction of pulmonary nodule malignancy remains central to lung cancer screening, yet decision support tools routinely fail once they leave the academic centers in which they were born. Classical risk scores such as the Mayo Clinic, Veterans Affairs, Brock (PanCan), PKUPH, and Li models reach internal AUCs in the 0.80–0.94 range by fitting logistic regressions to carefully curated cohorts, then collapse to 0.60–0.75 when applied to community-screening sites, Asian hospitals, or solitary-nodule subgroups [1, 2, 3, 4, 5, 6, 7]. Meta-analyses covering more than 80,000 nodules emphasize that prevalence changes, acquisition protocols, and different baseline diseases (e.g., tuberculosis versus granulomas) all distort the learned decision boundaries, making non-adaptive risk calculators a clinical liability in cross-hospital practice.

In response, the medical AI community has assembled an ecosystem of algorithms that mirror the broader evolution of structured-data learning. Gradient-boosted decision trees, led by XGBoost/LightGBM successors, dominate many tabular benchmarks because they remain robust to heterogeneous feature scales and missing values, and they continue to anchor registries such as NLST [8, 9]. Radiomics pipelines engineer thousands of texture descriptors from CT volumes to capture subtle morphologic cues, but their scanner sensitivity and need for harmonization often erase cross-site gains [5, 10]. Neural "deep tabular" architectures—TabNet, TabTransformer, SAINT, FT-Transformer, NODE, and a wave of attention-based variants—extend differentiability to structured data and enable multimodal fusion, yet they demand large, well-calibrated cohorts and frequently trail tuned tree ensembles on clinical tabular benchmarks [11, 12, 13, 14, 9]. Foundation-style approaches push further: TabPFN uses synthetic structural-causal priors to deliver hyperparameter-free, small-sample inference; TabPFN-2.5 and drift-resilient variants relax attention bottlenecks and introduce explicit temporal priors; Tabular LLMs serialize rows into prompts to borrow reasoning skills from generative models; and researchers now explore re-purposing tabular foundation models for graph reasoning and multimodal prompts [15, 16, 17, 18, 19, 20, 21]. Complementary efforts examine federated optimization or on-device continual learning so that models can absorb new hospital evidence without breaching privacy constraints [22, 23].

Despite this diversity of techniques, cross-hospital transfer remains fragile. Performance fails because the three dominant pathologies of medical tabular data co-occur: (i) sample scarcity—most nodular cohorts comprise a few hundred labeled patients, limiting the stability of purely supervised training; (ii) distribution shift—label prevalence, scanner kernels, and demographics change the marginal $P(X)$ and even the conditional $P(Y|X)$ between hospitals; and (iii) feature heterogeneity—sites log disjoint biomarker panels, measurement units, and coding policies that invalidate naive feature alignment [24, 25, 26]. Domain adaptation research in imaging and wearables demonstrates that adversarial training, optimal transport, or statistical moment matching can rescue some performance, but these methods are rarely tailored to structured clinical data, and benchmarks like TableShift show that off-the-shelf algorithms still suffer large out-of-distribution gaps even when in-distribution accuracy is high [22, 27, 28]. Large-scale regulators now treat shift detection and recalibration as core parts of post-market surveillance, underscoring that robustness cannot be an afterthought [24].

Tabular foundation models partially alleviate the sample-size constraint, yet they inherit a closed-world assumption: the context set used during in-context learning must reflect the same joint distribution as the query samples. When shifts in biomarkers, acquisition settings, or feature schemata emerge, even TabPFN variants can become overconfident because their attention weights are tied to the geometry of the source cohort [21, 18]. Emerging iterations like TabPFN-2.5 and drift-resilient TabPFN extend context length and bake simulated drifts into the prior, but they remain sensitive to mismatched feature spaces or unlabeled target domains without an explicit alignment step [16, 19]. Consequently, bridging the gulf 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 ideas because every stage of the pipeline can drift. Traditional clinical scores (Mayo, VA, Brock, PKUPH, Li) and their LASSO or gradient-boosted successors were derived from carefully curated cohorts with narrow demographic spreads and fixed scanner protocols, so their coefficients silently encode source-specific prevalence, upper-lobe priors, and calcification heuristics [1, 2, 3, 4, 6]. Meta-analyses across Asian screening programs and European cancer centers show that the same score threshold yields wildly different sensitivities (50–90%) once smoking histories, granulomatous disease burdens, or acquisi-

tion kernels change, even before considering that benign nodules dominate community-screening cohorts [5, 7]. Radiomics pipelines and 3D CNNs attain impressive internal AUCs on NLST and LIDC, yet external validations reveal double-digit drops when voxel spacing, reconstruction kernels, or ethnic mix shift, and shortcut learning can prompt models to key off hospital-specific artifacts rather than biology [29, 30, 10]. Domain adaptation techniques drawn from imaging—adversarial discriminators, cycle-consistent transfers, optimal transport—help when both domains share feature schemas, but they rarely address the missing-variable problem or the strict small-N regime of tabular nodular cohorts [22, 28]. Even TableShift, Wild-Time, and BRFSS benchmarks illustrate that strong in-distribution accuracy does not guarantee out-of-distribution reliability, and that label shift dominates error budgets unless prevalence-aware sampling or calibration is performed [27, 28].

Three recurring fault lines run through cross-hospital deployments. First, protocol heterogeneity induces covariate shift: radiology departments swap reconstruction kernels, slice thicknesses, and iterative denoisers across scanner upgrades, while laboratory information systems change assay vendors and reference ranges; even BRFSS survey wording drifts across years, warping marginal feature distributions. Second, label prevalence shifts with setting: tertiary oncology centers see far more malignant nodules than community-screening sites, and diabetes rates differ sharply across racial cohorts. Thresholds tuned to one prevalence produce over-biopsy or missed cancers elsewhere, creating safety and regulatory risk. Third, feature mismatches and missingness invalidate naive alignment: hospitals log different biomarker panels, use distinct coding for smoking status, or drop variables entirely when tests are not ordered. Without schema-aware pruning, models overfit site-specific artifacts or fail on missing columns. These shifts accumulate over time (concept drift), so one-off calibration cannot guarantee safe operation.

The algorithmic landscape mirrors these stresses. Gradient-boosted trees cope with messy scales and missingness, but they require abundant data to keep variance in check and cannot be fine-tuned across domains without rebuilding from scratch. Deep tabular models offer differentiable representations and multimodal fusion, yet they are data hungry, sensitive to hyperparameters, and often collapse when the effective sample size drops below a few thousand [9, 14]. Tabular foundation models such as TabPFN relax the data requirement through heavy pre-training and in-context learning, but they inherit closed-world assumptions: the context window expects a stable joint distribution and a consistent feature schema. When any of the three fault lines above appear, attention weights focus on non-comparable neighbors, inflating confidence while accuracy erodes [21, 18, 19].

Safety guidance now emphasizes designing for shift rather than reacting to it. Agencies and hospital governance boards increasingly demand evidence that models remain calibrated when equipment, demographics, or policies change [24]. In practice, relying on AUC alone hides threshold failures: a model can keep rank-ordering patients but still trigger excessive false positives after prevalence drifts. Cross-hospital nodule tools must therefore surface calibration behavior and maintain sensitivity where early intervention matters most, especially under privacy rules that preclude sharing target labels. These requirements push method design toward unsupervised alignment, feature budget discipline, and explicit handling of prevalence drift.

Taken together, the research gap is stark: tree ensembles and deep tabular nets struggle with small, heterogeneous cohorts; foundation models lift small-sample performance but assume matched domains; and generic domain adaptation rarely accounts for missing features or label drift in clinical tables. A credible solution must (i) retain sample efficiency via strong priors, (ii) discard site-specific signals that cannot transfer, and (iii) align source and target representations without target labels or schema changes.

Pulmonary nodule screening crystallizes these issues. Tuberculosis and pneumoconiosis inflate upper-lobe benign nodules in many Asian cohorts, confounding upper-lobe priors baked into Western-derived scores; smoking histories differ by region and era; and scanner upgrades alter texture features that radiomics and 3D CNNs depend upon [5, 10, 30]. Thresholds optimized on tertiary centers with high malignancy prevalence overcall cancer in community settings, triggering unnecessary biopsies. Conversely, down-tuned thresholds can miss aggressive lesions in high-risk clinics. Similar tensions surface in the BRFSS race-shift task: demographic composition, socioeconomic exposures, and survey-year wording alter the marginal distribution of risk factors, while diabetes prevalence rises from 12.5% (White) to 17.4% (non-White), so fixed operating points misfire. Any method that ignores these shifts risks brittle, non-actionable predictions.

Design constraints follow from these observations. Models must be frugal with labels, avoid depen-

dence on site-specific variables, expose calibration behavior across prevalences, and handle unlabeled target domains where privacy bars cross-site annotations. They must also degrade gracefully when partial feature overlap forces a reduced schema. These constraints shape our approach to pair pre-trained priors with statistical alignment and minimal, stable feature sets instead of relying on brute-force training.

The broader AI arc for tabular healthcare data provides both ingredients and warnings. Tree ensembles remain strong baselines because they tolerate mixed scales and missingness, yet their non-differentiable nature makes them hard to adapt across sites or fuse with other modalities [8, 14]. Deep tabular models (TabNet, TabTransformer, SAINT, FT-Transformer) import attention and gating to structured data, but they are data hungry, hyperparameter-sensitive, and vulnerable to batch-statistic or encoding drift when hospitals differ in coding or preprocessing [11, 12, 13, 9]. Tabular foundation models promise sample efficiency via massive synthetic pre-training and in-context learning, yet they still assume aligned schemas and stable covariates; when scanners, assay panels, or demographics shift, attention can anchor on non-comparable neighbors [21, 18, 19]. Tabular large language models serialize rows into prompts and leverage general-purpose reasoning but incur heavy latency and often struggle with precise numerical reasoning demanded by biomarkers [20]. Across these families, robustness hinges less on raw capacity and more on respecting feature overlap and shift.

Small-sample, high-dimensional, and imbalanced regimes further amplify brittleness. Pulmonary nodule cohorts rarely exceed a few hundred labeled patients, while radiomics or biomarker panels can exceed a hundred variables; naive inclusion of all features raises variance and encodes site-specific artifacts. Stability-driven feature selection (e.g., RFE on shared features) mitigates this variance and curbs schema mismatch, especially when positive classes are scarce [31]. Class imbalance and prevalence drift also distort thresholds: an operating point tuned on a tertiary center with 60–70% malignancy prevalence over-calls in community screening, while diabetes prevalence jumps between White and non-White cohorts in BRFSS, eroding precision and calibration.

Concrete cross-hospital failures underscore these themes. Meta-analyses of Mayo/VA/Brock successors show AUC drops of 0.1–0.3 when ported from U.S. academic centers to Asian screening programs, driven by different granuloma burdens, smoking histories, and scanner kernels [5, 7, 6]. Radiomics signatures tuned on sharp-kernel CTs lose discriminatory power on smooth-kernel images unless aggressively harmonized, and even then residual scanner bias can dominate texture features [10]. Cross-year BRFSS surveys alter wording and missingness patterns; features such as self-reported health or smoking show discrete shifts that break models calibrated on earlier years. These cases illustrate that without explicit feature pruning and alignment, both classic and modern models can become confidently wrong.

To demonstrate robustness beyond our private hospital cohorts, we additionally validate on a public cross-domain benchmark (TableShift BRFSS Diabetes) that introduces a race-driven shift (White $\rightarrow$ non-White) and survey-year drift. This setting mirrors the same covariate, label, and concept shift trio while operating at national scale, ensuring that the proposed approach addresses both clinical and population-level distribution shifts without changing the chapter structure or adding new modeling components.

Safety and regulation make these failures more than academic. Post-market surveillance guidelines now expect evidence of calibration and drift monitoring when models are deployed across equipment upgrades, demographic mixes, or policy changes [24]. AUC alone cannot certify safe decision support: over-diagnosis from optimistic thresholds causes unnecessary biopsies, while under-diagnosis from prevalence shifts can miss aggressive lesions. Privacy constraints often forbid labeled target data, ruling out supervised recalibration. Any deployable system must therefore assume unlabeled targets, partial feature overlap, and shifting priors, while still exposing confidence and calibration behavior to human overseers.

The shift landscape is multifaceted and worth making explicit. Covariate shift ($P_s(X) \neq P_t(X)$) emerges when scanner kernels, survey wording, or coding changes alter feature distributions; label shift ($P_s(Y) \neq P_t(Y)$) follows from different prevalences across centers, races, or years; concept shift ($P_s(Y \mid X) \neq P_t(Y \mid X)$) appears when new clinical guidelines, demographics, or comorbidities change the meaning of a feature vector [24]. Classical ERM optimizes source risk and leaves the divergence term uncontrolled, so even strong in-distribution accuracy fails to upper-bound target error. In practice, the three shifts co-occur: BRFSS race splits bundle covariate drift (lifestyle and socioeconomic factors), label shift (diabetes prevalence), and concept changes (different risk weight for identical behaviors).

Pulmonary nodules see the same mix: granulomatous disease confounds location priors, and protocol upgrades change radiomic textures. These conditions invalidate the implicit closed-world assumptions behind most off-the-shelf models.

Prior attempts to bridge domains reveal recurring limitations. Adversarial discriminators and style-transfer methods from imaging presume shared feature grids and plentiful target data; in tabular medicine, missing columns, mixed data types, and unlabeled targets induce instability or mode collapse [22, 28]. Statistical alignment such as MMD/CORAL is more stable but still assumes overlapping schemas and can degrade discriminative variance when applied naively. Invariant risk minimization and GroupDRO show promise in vision but routinely underperform tuned GBDTs on tabular benchmarks like TableShift and Wild-Time [27]. These results motivate combining alignment with strong priors rather than expecting any single robustness trick to suffice.

Meanwhile, feature engineering choices matter as much as model class. Clinical tables mix continuous labs, ordinal scores, sparse categorical codes, and structured missingness; simply one-hot encoding expands dimensionality and sparsity, hurting small-N generalization. Stability-driven feature pruning, hierarchical encoding of categorical variables, and unit-aware normalization reduce spurious site signatures and keep attention focused on shared, clinically interpretable signals [31]. Recursive feature elimination across domains further enforces schema overlap, trading a slight drop in ceiling accuracy for substantial gains in portability when hospitals differ.

The same caution extends to emerging tabular large-language-model approaches. Serializing rows into text prompts allows reuse of general reasoning, but tokenizing high-cardinality numerical columns bloats context windows, invites quantization error, and increases latency; moreover, LLM priors trained on web text do not encode clinical calibration by default [20]. Without explicit calibration or domain alignment, TabLLM-style systems risk confident misclassification when faced with out-of-template lab panels or race-specific prevalence changes.

Regulatory and clinical workflows impose further constraints on deployment. Hospitals require traceable decision rationales, audit logs of model updates, and clear operating thresholds tied to disease prevalence. When labels cannot be shared across sites, calibration transfer must rely on unsupervised statistics or prior knowledge; model updates must avoid catastrophic forgetting of earlier domains while accommodating drift. These practical requirements narrow the design space toward approaches that separate representation learning from alignment and that make minimal assumptions about target supervision or schema completeness.

This study therefore proceeds from a pragmatic stance: embrace tabular foundation models for their sample efficiency, but surround them with schema-aware feature pruning and unlabeled alignment so that attention operates on comparable examples even when hospitals, years, or races differ. The remainder of this manuscript formalizes the cross-domain problem, surveys prior art, and presents PANDA—a pipeline that chains cross-domain RFE, Transfer Component Analysis, and TabPFN inference—to restore calibration and discrimination under realistic deployment constraints.

Across these categories, shortcomings accumulate rather than cancel. Tree ensembles are non-differentiable and brittle in the small-sample regime; modest covariate shifts or low positive fractions push them toward overfitting and preclude gradient-based adaptation or calibration transfer across sites [8, 9]. Deep tabular models introduce differentiable representations but remain data hungry and tuning sensitive, and batch-statistic drift or coding changes can collapse learned embeddings when cohorts span hospitals or survey years [14]. Tabular foundation models lift sample efficiency but keep a closed-world view of the feature schema and marginal distributions: attention looks for nearest neighbors that may be non-comparable once scanners, biomarker panels, or demographic mixes shift, leading to confident but wrong matches [21, 18]. Finally, adaptation tricks borrowed from imaging—adversarial discriminators, cycle/style transfer, or optimal-transport aligners—assume shared feature grids and label access; they falter when target domains are unlabeled, miss variables entirely, or experience label drift, as documented on TableShift and medical DA surveys [28, 27, 24].

Despite rapid progress in deep tabular modeling, the combination of small-sample regimes, covariate drift, and feature-space mismatch remains largely unsolved in cross-hospital pulmonary nodule prediction. Existing AI methods—tree ensembles, deep tabular networks, or foundation-model variants—usually presume stable schemas or labeled targets, assumptions that rarely hold in real deployments. These gaps motivate a hybrid framework that integrates pre-trained tabular priors, schema-aware feature selection, and unsupervised domain alignment, which we develop in this study.

In cross-hospital pulmonary nodules or BRFSS race-shift diabetes prediction, these gaps become

acute: feature sets differ, prevalence drifts, and privacy blocks target labels, so neither trees, deep tabular models, foundation models alone, nor imaging-style DA offer a complete remedy. Any viable approach must combine strong priors, schema-aware feature pruning, and unlabeled distribution alignment to regain calibration and sensitivity under shift.

Our study therefore targets two representative settings: (i) cross-hospital pulmonary nodule prediction where Cohort A provides labels but Cohort B remains unlabeled, and (ii) the TableShift BRFSS diabetes race-split benchmark where White respondents form the source domain and non-White respondents form the target. Both settings reflect the same deployment realities: privacy constraints, schema mismatch, prevalence drift, and the need for sensitivity at clinically actionable thresholds. They also expose failure modes of purely supervised training and of foundation models without adaptation, offering a stress test for any proposed remedy.

Because HIPAA/GDPR rules forbid sharing labeled target data, supervised domain adaptation and threshold tuning on the target side are off the table. Methods that assume label access or perfect feature overlap therefore cannot be deployed in these scenarios. Any practical solution must work with source labels only, respect schema intersections, and deliver calibrated probabilities despite prevalence changes. This motivation drives the alignment-heavy, feature-prudent strategy developed here.

Existing AI toolkits each leave holes relative to these constraints. Tree ensembles cope with mixed scales and missingness but cannot be fine-tuned across domains and quickly overfit when positive cases are rare. Deep tabular models promise differentiable representations and multimodal fusion, yet they require large, clean cohorts and collapse when categorical codes or batch statistics shift. Tabular foundation models address the small-N barrier but assume matched schemas and stable covariates, so attention retrieves misleading neighbors when acquisition protocols change or when hospitals omit variables. Generic domain-adaptation tricks from imaging—adversarial discriminators, style transfer, or optimal transport—presume either shared feature grids or labeled targets; they seldom consider missingness shift, prevalence drift, or unlabeled target domains that dominate clinical deployments. Without explicit feature pruning and alignment, these methods can become overconfident while making non-comparable comparisons across sites.

We therefore introduce *PANDA* (Pretrained Adaptation Network with Domain Alignment), a pragmatic framework that chains three proven ideas. First, TabPFN supplies a strong inductive prior for small cohorts by meta-learning across millions of synthetic tabular tasks [15]. Second, Transfer Component Analysis (TCA) aligns source and target distributions in a shared reproducing-kernel subspace without labeled target data, minimizing divergence while preserving clinical variance [32]. Third, cross-domain Recursive Feature Elimination prunes to the biomarkers that are consistently available and stable, mitigating schema mismatch and noisy hospital artifacts [31]. PANDA targets the explicit goal of cross-hospital pulmonary nodule prediction with screening-level sensitivity by combining these components rather than relying on any single modeling breakthrough.

2 Related 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, 21]. 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, 14]. Benchmarking studies covering hundreds of OpenML tasks show that GBDTs still beat most neural base-

lines 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. On the small, heavily imbalanced cohorts typical of lung-screening registries ($N \approx 300$), the same capacity becomes a liability: a few malignant cases can be memorized by deep trees, calibration deteriorates in low-prevalence subgroups, and there is no clear way to "warm start" or fine-tune an existing model when a new hospital adds or removes variables. 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 DA objectives, motivating methods that transfer tree-like priors into differentiable architectures.

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 [11, 39, 40, 13]. 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: 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. 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, 25, 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 [15, 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 [16, 17, 18]. Drift-resilient variants model temporal shift with secondary structural-causal modules and record measurable gains when patient mixes evolve [48, 19]. 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, 20, 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, 20, 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 WildTab 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, Tab-Transformer, 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 $P(Y \mid X)$ [24, 25]. 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 [25, 58, 22]. 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, 27, 24]. 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 [22, 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 [28, 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 [27, 67, 68, 55, 28]. 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 [24]. 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, 22]. 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, 25]. 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. 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 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 [70]. 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 [71, 72, 73]. 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 [71, 72, 73]. 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 [11, 12, 13, 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 compact "best7" and "best8" subsets that are 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.

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, 74, 2, 3, 4, 5, 6, 7]. 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 [5, 6, 7, 75]. 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.

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, 74]. 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, 7]. 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 [75]. 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 [5, 6, 7].

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 2 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 2: 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 [75]
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, 7]

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 [10, 5]. 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 [29, 30, 24]. These failures illustrate that radiomics alone cannot guarantee transportability and that alignment plus feature vetting are required before cross-hospital use. 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 [10]. 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 [5, 10]. 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 [29]. Subsequent works combine deep features with handcrafted radiomics or clinical covariates, showing further gains on curated datasets [76, 77]. 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 [30, 10, 24]. 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, 5, 76, 77]. 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 [5, 10]. 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 [5, 10]. 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 [5, 10]. 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 [27, 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 [28]. 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 [27, 67]. Wild-Time isolates temporal drift, showing monotonic degradation without continual recalibration [28]; 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 [24, 25]. 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 [27, 28]. Finally, reproducibility crises in radiomics and deep imaging models show that aggressive priors or harmonization cannot replace explicit alignment and feature governance [10, 24].

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 shared feature sets ("best7", "best8") that remain predictive in both institutions and provide 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 [27, 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 medical classification mixes distribution shift, sample scarcity, and feature heterogeneity. We cast it as an unsupervised domain adaptation (UDA) problem on structured clinical data: the goal is reliable prediction in a target hospital without target labels. The framing mirrors common deployment constraints in medical AI.

Cross-Domain Learning Setup

Let the labeled source cohort be $\mathcal{D}_s = (X_s, Y_s) = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ and the unlabeled target cohort be $\mathcal{D}_t = (X_t, \varnothing) = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$. Each instance $\mathbf{x} \in \mathbb{R}^d$ collects structured clinical variables and $y \in \{0, 1\}$ encodes malignancy for nodules or diabetes status for TableShift. Domains expose imperfectly overlapped feature sets: $\mathcal{F}_s$ and $\mathcal{F}_t$ denote the recorded indices, $\mathcal{F}_\cap = \mathcal{F}_s \cap \mathcal{F}_t$ the shared subset used for modeling, and $\mathcal{F}_\setminus = \mathcal{F}_s \triangle \mathcal{F}_t$ the features seen in only one hospital or demographic group. We write $d_\cap = |\mathcal{F}_\cap|$ for the dimensionality after intersection.

Admissible models operate on $\mathcal{X}_\cap = \mathbb{R}^{d_\cap}$. The objective is to learn $f : \mathcal{X}_\cap \to \mathcal{Y}$ that minimizes the target risk

$$\mathcal{R}_t(f) = \mathbb{E}_{(\mathbf{x},y)\sim P_t}[\ell(f(\mathbf{x}), y)]$$

subject to the constraint that Y_t remains unobserved during training. Privacy policies (HIPAA, GDPR) render this unsupervised domain adaptation framing the only feasible option in many multi-institution collaborations.

The two tasks tackled in this dissertation differ strongly in their feature vocabularies yet share the notation above. Table 3 summarizes the recorded variables, the overlapping subsets, and the information discarded when aligning hospitals or demographic splits.

Table 3: Feature partitioning for the two cross-domain tasks. $\mathcal{F}_\cap$ contains the variables usable across domains; $\mathcal{F}_\setminus$ collects site- or demographic-specific factors excluded from modeling.

Task	Shared feature families ($\mathcal{F}_\cap$)	Site-/group-specific factors ($\mathcal{F}_\setminus$)
Cross-hospital pulmonary nodules	Age, sex, smoking indicators, upper-lobe flags, diameter, calcification, spiculation, tumor biomarkers (CEA, NSE, Cyfra21-1), ventilatory metrics (VC, DLCO)	CT reconstruction kernel IDs, segmentation quality scores, rare lab assays, hospital-specific comorbidities, imaging vendor tags
TableShift BRFSS race split	Demographics (age, sex, education), chronic-disease history, lifestyle (smoking, alcohol, physical activity), metabolic labs (BMI proxies, hypertension indicators), survey year	State-specific policy items, optional socioeconomic questions only asked in some years, race-restricted modules, state sampling weights

Challenges in Cross-Institutional Learning

Clinical tabular cohorts usually include only a few hundred labeled patients. For hypothesis classes on $d_\cap$ shared features, estimation error scales as $\widetilde{O}(\sqrt{d_\cap / n_s})$, making high-capacity models unreliable once $n_s \leq 500$. Many UDA techniques implicitly bank on larger sample sizes than most hospitals can release.

Distributional mismatch compounds the limits. Under the standard adaptation bound

$$\mathcal{R}_t(f) \leq \mathcal{R}_s(f) + \tfrac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(P_s, P_t) + \lambda,$$

the divergence term $d_{\mathcal{H}\Delta\mathcal{H}}$ dominates when variability is substantial—differences in CT scanners, assay calibrations, demographics, or survey wording. Partial feature overlap means source and target supports only partly coincide, straining assumptions behind kernel alignment and adversarial methods.

Shift types manifest differently across the two tasks but lead to the same failure mode of inflated λ:

- **Covariate shift:** $P_s(\mathbf{x}) \neq P_t(\mathbf{x})$ emerges when Hospital B observes higher upper-lobe prevalence or when BRFSS non-White respondents show distinct BMI and lifestyle distributions. Without alignment, the similarity kernel inside TabPFN attends to mismatched neighbors and the risk bound loosens.
- **Label shift:** $P_s(y) \neq P_t(y)$ appears in lung nodules when malignancy prevalence changes from tertiary centers to community hospitals, and in BRFSS when diabetes rates vary by race. Thresholds tuned on P_s miscalibrate decision curves once applied to P_t.
- **Concept shift:** $P_s(y|\mathbf{x}) \neq P_t(y|\mathbf{x})$ captures definition drift, such as tuberculosis confounding upper-lobe malignancy cues in some regions or policy changes altering how survey responses map to the DIABETES label.

The combination of $d_{\mathcal{H}\Delta\mathcal{H}}$ growth and concept shift means that even small empirical risk on Cohort A or the White BRFSS split does not guarantee acceptable target risk. Explicit alignment and feature harmonization are therefore prerequisites for any foundation-model method deployed in these settings.

4 Solution

PANDA targets the three core limitations identified in sample scarcity, distribution shift, and feature heterogeneity.

Compositional Architecture

PANDA consists of four sequential operators, each resolving a specific challenge in cross-hospital prediction, as depicted in Fig. 1.

(1) Cross-domain feature selection. The operator $\mathcal{T}_{\text{RFE}} : \mathbb{R}^d \to \mathbb{R}^{d'}$ selects a domain-stable subset of features via cross-domain recursive elimination:

$$\mathcal{T}_{\text{RFE}}(\mathbf{x}) = \mathbf{x}_{\mathcal{F}^*}, \qquad \mathcal{F}^* = \arg\min_{\mathcal{F}'} \sum_{j \in \mathcal{F}'} \text{Var}_{\text{domain}}(\mathbf{x}_j) + \lambda |\mathcal{F}'|.$$

This yields a compact and clinically consistent feature set shared across institutions.

(2) Foundation-model representation. The pretrained TabPFN encoder $\Phi_{\text{FM}} : \mathbb{R}^{d'} \to \mathbb{R}^h$ maps the reduced features into a smooth latent space:

$$\Phi_{\text{FM}}(\mathbf{x}) = \text{Transformer}_{\theta^*}(\text{Tokenize}(\mathbf{x})).$$

This step injects inductive priors learned from millions of synthetic tasks, yielding representations that generalize even when few labeled samples exist.

(3) Domain-invariant alignment via TCA. Transfer Component Analysis (TCA) learns a projection that reduces distribution discrepancies between hospitals:

$$\min_{W} \operatorname{tr}(W^\top KLK^\top W) + \mu \operatorname{tr}(W^\top KHK^\top W),$$

where L encodes maximum mean discrepancy (MMD), H is a centering matrix, and K is a kernel matrix (linear kernel in our implementation). The aligned representation is

$$\mathbf{z} = W^\top \phi(\mathbf{x}), \qquad \phi : \mathbb{R}^d \to \mathbb{R}^k,$$

with k chosen automatically to preserve information while enabling effective alignment.

(4) Classification head with ensemble aggregation. The final classifier $h : \mathbb{R}^k \to [0, 1]$ operates on aligned features and aggregates predictions across multiple preprocessing branches and random seeds:

$$f(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} h_b\Big(\mathcal{A}_{\mathrm{TCA}}(\Phi_{\mathrm{FM}}^{(b)}(\mathbf{x}))\Big).$$

Unified Objective

The complete PANDA mapping is:

$$f(\mathbf{x}) = h(\mathcal{A}_{\mathrm{TCA}}(\Phi_{\mathrm{FM}}(\mathcal{T}_{\mathrm{RFE}}(\mathbf{x})))).$$

The joint optimization objective minimizes source-domain classification loss while aligning source and target distributions:

$$\min_{W,h} \frac{1}{n_s} \sum_{i=1}^{n_s} \ell(h(\mathcal{A}_{\mathrm{TCA}}(\Phi_{\mathrm{FM}}(\mathbf{x}_i^s))), y_i^s) + \lambda_1 d_{\mathrm{MMD}}(\mathbf{Z}_s, \mathbf{Z}_t),$$

where $\mathbf{Z}_s = \mathcal{A}_{\mathrm{TCA}}(\Phi_{\mathrm{FM}}(\mathcal{D}_s))$ and $\mathbf{Z}_t = \mathcal{A}_{\mathrm{TCA}}(\Phi_{\mathrm{FM}}(\mathbf{X}_t))$.

5 Methods

5.1 Motivating Challenges and Methodological Response

Cross-hospital malignancy prediction and public-health surveillance generate intertwined constraints: tiny labeled cohorts, label imbalance, feature mismatch, and multiple forms of distribution shift. PANDA is organized around these constraints rather than around model novelty. Table 4 distills the major obstacles and the mechanisms assigned to them.

This architecture ensures that every component answers a crisp question: why do small-sample medical deployments fail, and what prior or alignment tool counters that failure?

5.2 Training and Inference Pipeline

The full workflow treats both pulmonary nodules and BRFSS as unlabeled-target problems. The steps below summarize the data dependencies and the order in which components are invoked.

PANDA cross-domain pipeline.

1. **Schema harmonization:** compute the shared index set $\mathcal{F}_\cap$ between $\mathcal{D}_s$ and $\mathcal{D}_t$; align categorical codes and clinical units, producing matrices $X_s^\cap, X_t^\cap$.
2. **Cross-domain RFE:** run recursive feature elimination with permutation-based importances on $\mathcal{D}_s$ while constraining candidates to $\mathcal{F}_\cap$. Record stable subsets (best7/best8) for downstream use.
3. **Multi-branch preprocessing:** for each retained subset, construct four preprocessing branches (raw order, rotated order, quantile transform, ordinal encoding). Each branch yields context matrices and associated metadata.

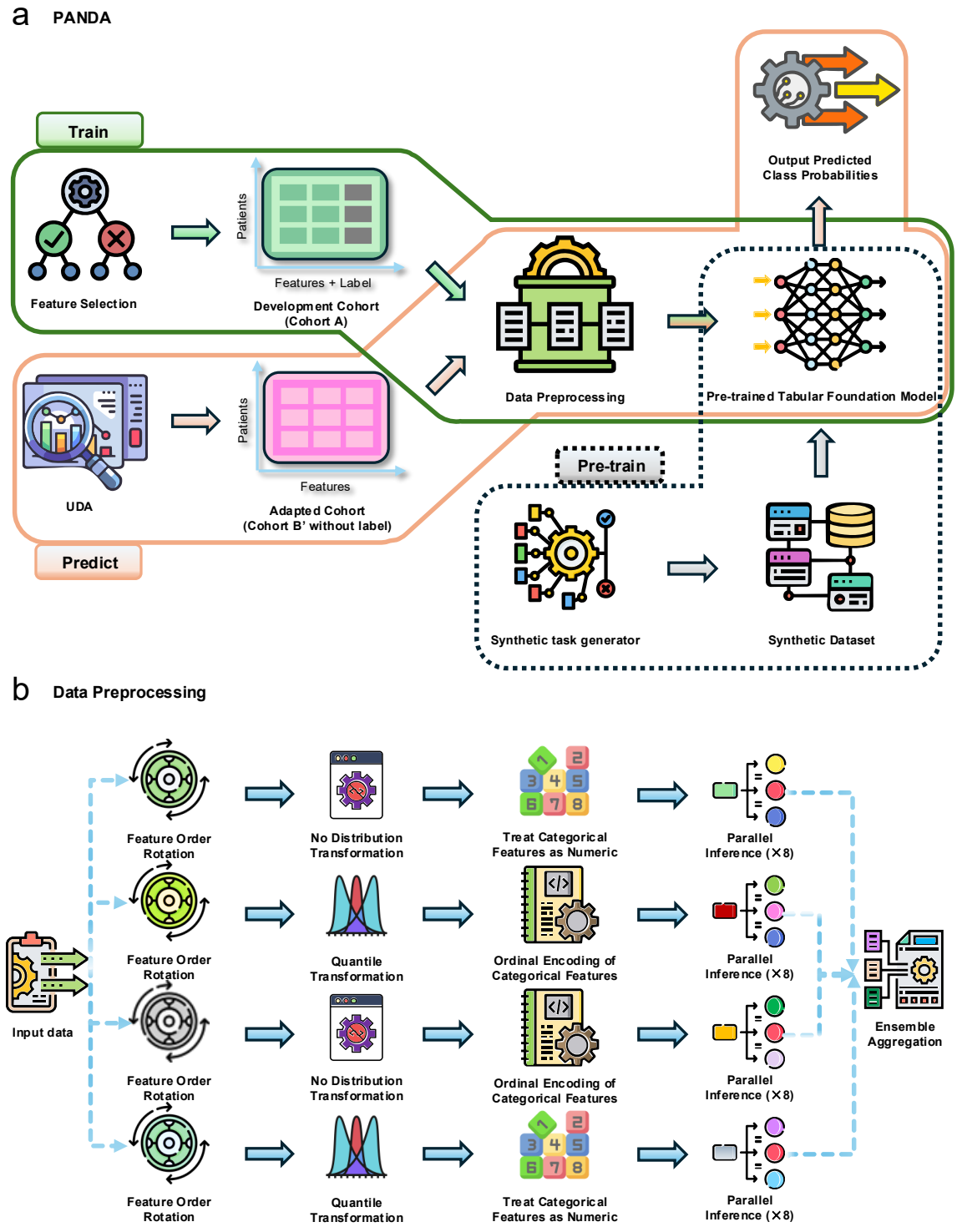

Figure 1: **The PANDA framework architecture.** (a) Compositional pipeline: from original tabular data through ensemble training, prediction aggregation, class imbalance adjustment, to final classification output. (b) Multi-branch ensemble with $B = 4$ preprocessing strategies, each generating $S = 8$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

Table 4: Challenge–mechanism mapping in PANDA. Each component targets a known failure mode, and the same design is reused for pulmonary nodules and the TableShift BRFSS race-shift task.

Challenge	Mechanism	Expected benefit
Small n with high-dimensional covariates	TabPFN prior-data fitted network performs in-context learning with frozen weights	Transfers structural priors from millions of synthetic tasks, reducing estimation variance without local fine-tuning
Feature heterogeneity across institutions/demographics	Cross-domain RFE surfaces stable subsets ("best7", "best8") definable in every site, plus schema alignment utilities	Removes site-specific artefacts before adaptation and guarantees that downstream models only consume shared attributes
Covariate shift and mixed acquisition protocols	TCA applied to TabPFN embeddings realigns marginal distributions before the classifier head	Shrinks 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	Injects diversity without retraining new weights and stabilizes predictions under minor data perturbations

4. **TabPFN context construction:** concatenate support (training) samples and target queries per branch, tokenize as sequences, and feed to the frozen TabPFN backbone to obtain contextual embeddings.

5. **Transfer Component Analysis:** solve the TCA objective on each branch's embeddings using both $X_s^{\cap}$ and $X_t^{\cap}$ to obtain projection $\mathbf{W}$; project embeddings into an aligned latent space.

6. **Prediction and aggregation:** pass aligned embeddings through the foundation-model classifier head to get probabilities; apply class-balanced calibration/temperature scaling; average across 32 branch-seed combinations to form the final risk score.

7. **Thresholding and reporting:** select operating points (e.g., 0.5 default or clinically mandated net-benefit thresholds) separately for each cohort, but keep the classifier weights fixed.

5.3 Foundation Model Architecture

5.3.1 TabPFN Backbone Details

TabPFN uses a 12-layer Transformer with four attention heads and 128-dimensional embeddings. Clinical samples are tokenized as $[\text{CLS}, \mathbf{x}_1, \ldots, \mathbf{x}_d, \text{SEP}]$ with positional encodings to preserve ordering. Training instances and test queries are processed jointly in one forward pass, enabling in-context learning without gradient updates.

5.3.2 Synthetic Task Generation

Pre-training draws diverse synthetic classification tasks from several function priors, including Gaussian processes, multilayer perceptrons, and ridge regression families. This variety teaches generalizable tabular reasoning patterns that appear to transfer to real-world medical classification tasks.

5.4 Feature Selection and Preprocessing

5.4.1 Cross-Domain RFE Algorithm

We recursively eliminate features based on domain-invariant importance scores:

$$\text{Importance}(\mathbf{x}_j) = \frac{1}{M}\sum_{m=1}^{M}\left|\mathcal{R}_s^{(m)}(\mathcal{F}\setminus\{\mathbf{x}_j\}) - \mathcal{R}_s^{(m)}(\mathcal{F})\right|$$

where $M = 5$ permutation repeats evaluate feature stability. At each iteration the least important element within $\mathcal{F}_\cap$ is dropped, producing nested subsets. Clinical input from collaborators constrains the search to candidates physically measurable at every institution, so the procedure naturally returns compact sets (best7, best8) that survive both availability and robustness checks.

5.4.2 Multi-Branch Preprocessing Pipeline

The 32-model ensemble comes from four simple branches: two keep the original or rotated feature order with plain numerical encodings, and two pair those orders with a quantile transform plus ordinal encoding. Each branch spits out eight runs with seeds 1–8, and a majority vote settles the label. Balanced-accuracy weights keep the malignant class from getting drowned out.

5.5 Domain Adaptation Implementation

5.5.1 TCA Optimization

Transfer Component Analysis learns domain-invariant representations by solving:

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

where $\mathbf{L}$ is the MMD kernel matrix with entries $L_{ij} = K_{ij}/(n_s^2) + K_{ij}/(n_t^2) - 2K_{ij}/(n_s n_t)$. The kernel matrix K adopts Gaussian RBF kernels with bandwidth σ set via the median heuristic.

The alignment step preserves discriminative information while reducing domain discrepancy:

$$\mathbf{z} = \mathbf{W}^\top \phi(\mathbf{x}), \quad \phi : \mathbb{R}^d \to \mathbb{R}^h$$

where latent dimensionality $h = 15$ balances information preservation with alignment effectiveness.

5.5.2 Label Prevalence Handling

Unsupervised deployment cannot rely on target labels to recalibrate prevalence. PANDA therefore performs class-balanced sampling on the source cohort so that minority cases remain visible during TabPFN context construction. During inference we maintain cohort-specific threshold tables: pulmonary nodule deployments retain clinically validated malignancy cutoffs, whereas BRFSS experiments default to 0.5 for AUC computation and adapt thresholds only when policy guidance requires specific sensitivity ranges. Temperature scaling is applied using a held-out source fold to avoid leaking target information.

5.6 Assumptions and Mitigation Strategies

The pipeline rests on several operational assumptions. Table 5 lists the most salient ones, describes the failure mode if they are violated, and notes the mitigation built into PANDA. This keeps the discussion in the methods section distinct from later empirical performance claims.

5.7 Ethics Statement and Data Collection

This study received Institutional Review Board approval from two participating hospitals in China and followed the Declaration of Helsinki. Patient data were retrospectively extracted from electronic medical records and fully de-identified before analysis. Written informed consent for research use of clinical information was obtained from all patients with solitary pulmonary nodules (SPNs) at admission, and no identifiable personal data were retained.

The training cohort (Cohort A, $n = 295$) originated from Hospital A between January 2011 and December 2016. The external test cohort (Cohort B, $n = 190$) was collected at Hospital B. All participants provided written informed consent for scientific use of their clinical data at the time of admission.

Table 5: Key assumptions behind PANDA and the mitigation strategies employed when deployments deviate from them.

Assumption	Failure mode if violated	Mitigation in PANDA
Shared schema retains predictive signals	Missing biomarkers or survey items make $\mathcal{F}_\cap$ too small, degrading TabPFN context quality	Cross-domain RFE enforces minimum-cardinality subsets (best7/best8) vetted by clinicians; fallback to manual feature engineering when subset shrinks below safety threshold
Covariate shift is smooth enough for kernel alignment	Abrupt scanner or questionnaire shifts push samples outside the kernel bandwidth, leading to large $d_{\mathcal{H}\Delta\mathcal{H}}$	Online statistics monitor MMD and trigger retraining/alignment updates; kernel bandwidth adapts via median heuristic recalculated per batch
Label prevalence drift is moderate	Thresholds calibrated on $P_s(y)$ misclassify minority cohorts	Maintain cohort-specific threshold dictionaries and class-balanced sampling; clinicians can override cutoffs per site
TabPFN context window covers source+target batches	Excessively large batches overflow the Transformer input, forcing truncation	Use stratified subsampling (1k context rows default) with deterministic seeds; for BRFSS we slice the unlabeled pool before feeding TabPFN
Independence of samples within context	Correlated patients (e.g., repeat visits) bias in-context learning	Deduplicate patient identifiers and enforce visit-level sampling prior to TabPFN ingestion

5.7.1 Data Variables and Measurements

Collected variables included demographics (age, sex, height, weight, body mass index), smoking history, family cancer history, and symptoms (fever, cough, hemoptysis, chest pain). Radiologic descriptors of SPNs covered anatomical location (lung side and lobe), nodule diameter and area, calcification, cavity, spiculation, pleural thickening, and adhesion. Laboratory data comprised hematologic and biochemical indices such as white blood cell count (WBC), neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), albumin/globulin ratio (AGR), liver and renal function markers, and tumor biomarkers including CEA, Cyfra21-1, and NSE.

5.8 Experimental Procedures

5.8.1 Cross-Validation Protocol

For internal validation, we applied 10-fold cross-validation on Cohort A. The dataset was randomly split into 10 equal parts with class balance preserved. Each fold served once as validation while the remaining nine folds trained the model. This cycle was repeated 10 times with different random seeds to strengthen robustness of performance estimates.

5.8.2 Baseline Methods

For comparison, we included a few familiar baselines:

- Decision Tree (CART) [78]
- Gradient Boosting Decision Tree [79]
- Random Forest [80]
- XGBoost [8]
- Support Vector Machine [81]
- LASSO Logistic Regression for nodule risk [4]
- Clinical scores (Mayo Clinic, PKUPH) [1, 82]

6 Analysis

We trace how PANDA deals with the main sources of failure in cross-site medical AI. Each component is tied to a specific hurdle rather than bolted on for convenience, and the mechanics show up in both the math and the observed gains.

6.1 In-Context Learning for Small-Sample Robustness

Deep models tend to overfit on small cohorts (e.g., $n_s = 295$) and swing wildly with minor perturbations. PANDA avoids heavy re-training by casting classification as in-context learning. The TabPFN backbone uses a *Per-Feature Transformer Architecture*, treating each input $\mathbf{x} = [x_1, x_2, \ldots, x_d] \in \mathbb{R}^d$ as a token sequence:

$$\mathbf{e}_i = \text{Embed}(x_i) + \mathbf{p}_i, \quad i = 1, \ldots, d$$

where $\text{Embed}(\cdot) : \mathbb{R} \to \mathbb{R}^{d_{\text{model}}}$ maps features to a d_{model}-dimensional space. This embedded sequence $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_d]$ is processed through a 12-layer Transformer encoder:

$$\mathbf{H}^{(\ell)} = \text{LayerNorm}(\text{MultiHead}(\mathbf{H}^{(\ell-1)}) + \mathbf{H}^{(\ell-1)})$$

$$\mathbf{H}^{(\ell+1)} = \text{LayerNorm}(\text{FFN}(\mathbf{H}^{(\ell)}) + \mathbf{H}^{(\ell)})$$

where $\mathbf{H}^{(0)} = \mathbf{E}$. To circumvent data scarcity, the model is pre-trained using a stochastic task generator that synthesizes classification problems from diverse function priors. For each batch, we sample a prior family and hyperparameters:

$$r \sim \text{Categorical}(\boldsymbol{\pi}), \qquad \boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid r),$$

where $r \in \{\text{gp}, \text{mlp}, \text{ridge}, \text{mix\_gp}\}$. Inputs are sampled independently from a factorized base distribution and optionally transformed:

$$\mathbf{x}_t \sim p_{\text{base}}(\mathbf{x}), \qquad \tilde{\mathbf{x}}_t = \psi_{\boldsymbol{\theta}}(\mathbf{x}_t)$$

During inference, the model performs in-context learning by processing the entire sequence of context examples $\mathcal{D}_{\text{ctx}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_{\text{ctx}}}$ and query inputs $\mathbf{x}_{\text{query}}$:

$$\mathbf{z} = [\mathbf{x}_1, y_1, \ldots, \mathbf{x}_{n_{\text{ctx}}}, y_{n_{\text{ctx}}}, \mathbf{x}_{\text{query}}]$$

The prediction minimizes the in-context loss over query positions (averaged over a batch of size B):

$$\mathcal{L}_{\text{ICL}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{t=n_{\text{ctx}}+1}^{T} \ell(f_\Theta(\mathbf{z}_{1:t-1}), y_t)$$

The pre-trained priors act as regularizers, helping the model interpolate in sparse regions where conventional models often fail to learn stable boundaries.

6.2 Mitigating Distributional Heterogeneity

Performance usually drops when moving across hospitals because of small shifts in encoding and feature distributions. A dual preprocessing strategy tackles positional bias and distribution mismatch. To reduce ordering bias in the Transformer input, each ensemble member applies a cyclical permutation to the features:

$$\mathbf{x}_{\text{rotated}}^{(m)} = \text{rotate}(\mathbf{x}, m) = [x_{(m) \bmod d}, x_{(m+1) \bmod d}, \ldots, x_{(m+d-1) \bmod d}]$$

with rotation offsets generated deterministically for each ensemble member $m \in [0, N-1]$. In parallel, we employ *Adaptive Feature Transformation* to bridge distributional gaps. The **Enhanced Feature Transformation** performs a quantile transform followed by dimensionality expansion:

$$\mathbf{x}_{\text{quantile}} = \text{QuantileTransformer}(\mathbf{x}, n_{\text{quantiles}} = \max(\lfloor n_{\text{samples}}/10 \rfloor, 2))$$

$$\mathbf{X}_{\text{expanded}} = \text{SVD}(\mathbf{X}_{\text{quantile}}, n_{\text{components}} = \min(4, d))$$

yielding a final representation $\mathbf{x}_{\text{final}} = [\mathbf{x}_{\text{original}}; \mathbf{x}_{\text{quantile}}; \mathbf{x}_{\text{SVD}}]$. A complementary **Preserved Feature Transformation** keeps the raw feature distribution:

$$\mathbf{x}_{\text{preserved}} = \mathbf{x}_{\text{original}}$$

Categorical variables are processed using *Intelligent Categorical Encoding*:

$$\text{encode}(x_{ij}) = \begin{cases} \phi_j(x_{ij}) & \text{if feature } j \text{ has frequently occurring categories} \\ x_{ij} & \text{otherwise} \end{cases}$$

where $\phi_j = \pi(\{0, 1, \ldots, |U_j|-1\})$ employs randomized integer assignment. Alternatively, the **Numeric Treatment Strategy** treats categorical features as continuous:

$$\text{encode}(x_{ij}) = \text{float}(x_{ij})$$

Providing multiple "views" of the data lets the model marginalize hospital-specific artifacts and focus on the clinical signal.

6.3 Addressing Feature Inconsistency

Noisy or missing variables across cohorts make careful selection essential, and RFE offers a fairly transparent way to handle it. The workflow is straightforward:

1. Train the Pre-trained Tabular Foundation Model $f_\Theta^{(t)}$ on the current feature subset $\mathcal{F}^{(t)}$.
2. Estimate importance scores $\mathbf{I}^{(t)} = [I_1^{(t)}, I_2^{(t)}, \ldots, I_{|\mathcal{F}^{(t)}|}^{(t)}]$ using permutation-based evaluation.
3. Remove the feature with the smallest score: $\mathcal{F}^{(t+1)} \leftarrow \mathcal{F}^{(t)} \setminus \{\arg\min_j I_j^{(t)}\}$.
4. Repeat until the subset reaches the target size $|\mathcal{F}^{(t+1)}| = k$.

Feature importance here is defined by how much performance drops when a variable is randomly shuffled:

$$I_j = \frac{1}{R}\sum_{r=1}^{R}\left[\mathrm{AUC}(f_\Theta, \mathcal{D}) - \mathrm{AUC}(f_\Theta, \mathcal{D}^{(r)}_{\mathrm{perm}(j)})\right].$$

To determine the optimal feature subset, we optimize a comprehensive cost-effectiveness index:

$$\mathrm{CostEffectiveness}(k) = w_1 \cdot S_{\mathrm{perf}}(k) + w_2 \cdot S_{\mathrm{eff}}(k) + w_3 \cdot S_{\mathrm{stab}}(k) + w_4 \cdot S_{\mathrm{simp}}(k)$$

where the component scores are normalized as follows:

- **Performance Score**:

$$S_{\mathrm{perf}}(k) = 0.5 \cdot \mathrm{AUC}(k) + 0.3 \cdot \mathrm{Accuracy}(k) + 0.2 \cdot \mathrm{F1}(k)$$

- **Efficiency Score**:

$$S_{\mathrm{eff}}(k) = 1 - \frac{T(k) - T_{\min}}{T_{\max} - T_{\min}}$$

- **Stability Score**:

$$S_{\mathrm{stab}}(k) = 1 - \frac{CV(k) - CV_{\min}}{CV_{\max} - CV_{\min}}$$

- **Simplicity Score**:

$$S_{\mathrm{simp}}(k) = \exp(-\alpha \cdot k)$$

The optimal subset is chosen as $k^* = \arg\max_k \mathrm{CostEffectiveness}(k)$, yielding a feature set that keeps strong discriminative value while still matching what hospitals can reliably collect.

6.4 Latent Space Alignment for Covariate Shift

A noticeable gap between internal and external validation often hints at covariate shift ($P_s(\mathbf{x}) \neq P_t(\mathbf{x})$). *Transfer Component Analysis (TCA)* addresses this by mapping both domains into a shared latent subspace where their distributions look closer. Let $X_s \in \mathbb{R}^{n_s \times d}$ and $X_t \in \mathbb{R}^{n_t \times d}$ be source and target feature matrices. A combined kernel matrix $K \in \mathbb{R}^{(n_s+n_t)\times(n_s+n_t)}$ with a linear kernel $K(x_i, x_j) = x_i^\top x_j$ is partitioned as:

$$K = \begin{bmatrix} K_{ss} & K_{st} \\ K_{ts} & K_{tt} \end{bmatrix}$$

A projection matrix $W \in \mathbb{R}^{(n_s+n_t)\times k}$ is learned by solving:

$$\min_W \ \mathrm{tr}(W^\top K L K^\top W) + \mu \cdot \mathrm{tr}(W^\top K H K^\top W),$$

where the alignment matrix L encourages domain alignment:

$$L = \begin{bmatrix} \frac{1}{n_s^2}\mathbf{1}_{n_s \times n_s} & -\frac{1}{n_s n_t}\mathbf{1}_{n_s \times n_t} \\ -\frac{1}{n_s n_t}\mathbf{1}_{n_t \times n_s} & \frac{1}{n_t^2}\mathbf{1}_{n_t \times n_t} \end{bmatrix}$$

and the centering matrix $H = I - \frac{1}{n_s+n_t}\mathbf{1}\mathbf{1}^\top$ ensures zero-centered features. The eigen-decomposition $(I + \mu KLK)S = KHKS$ yields W, and source and target samples project via $Z_s = K_s W$ and $Z_t = K_t W$. Distances are computed in the TCA space using pooled statistics $\hat{\mu}, \hat{\sigma}$ and standardized features $\mathbf{X}_s^{\text{norm}}, \mathbf{X}_t^{\text{norm}}$:

$$\mathbf{X}_s^{\text{norm}} = \frac{\mathbf{X}_s - \hat{\mu}}{\hat{\sigma}}, \quad \mathbf{X}_t^{\text{norm}} = \frac{\mathbf{X}_t - \hat{\mu}}{\hat{\sigma}}$$

These metrics include **Wasserstein Distance**:

$$W_{\text{norm}}(\mathbf{X}_s, \mathbf{X}_t) = \frac{1}{d}\sum_{i=1}^{d} W_1(X_{s,i}^{\text{norm}}, X_{t,i}^{\text{norm}})$$

Symmetric KL Divergence:

$$KL_{\text{norm}}(\mathbf{X}_s, \mathbf{X}_t) = \frac{1}{d}\sum_{i=1}^{d} \frac{KL(P_{s,i}^{\text{norm}} || P_{t,i}^{\text{norm}}) + KL(P_{t,i}^{\text{norm}} || P_{s,i}^{\text{norm}})}{2}$$

and **MMD with RBF Kernel**:

$$\text{MMD}^2(\mathbf{X}_s, \mathbf{X}_t) = \frac{1}{n_s(n_s-1)}\sum_{i\neq j} k(x_i^s, x_j^s) + \frac{1}{n_t(n_t-1)}\sum_{i\neq j} k(x_i^t, x_j^t) - \frac{2}{n_s n_t}\sum_{i,j} k(x_i^s, x_j^t)$$

where $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma||\mathbf{x} - \mathbf{y}||^2)$.

6.5 Stabilizing Predictions with Ensemble Aggregation

Single models often give poorly calibrated scores that drift toward the majority class. PANDA tempers this tendency with an ensemble setup, which aggregates multiple slightly varied representations to steady both calibration and overall stability. **Class imbalance handling** uses inverse-frequency reweighting:

$$\hat{p}_i^{\text{balanced}} = \frac{\hat{p}_i/\pi_i}{\sum_{j=1}^{C} \hat{p}_j/\pi_j}$$

where $\hat{p} = (p_1, \ldots, p_C)$ are predicted probabilities and π the empirical class distribution. **Ensemble aggregation** takes a simple but surprisingly steadying approach: it averages the temperature-scaled outputs from $N = 32$ members,

$$p(y = c \mid \mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \frac{\exp(z_i^c/T)}{\sum_{c'=1}^{C}\exp(z_i^{c'}/T)},$$

where z_i^c are the logits and $T = 0.9$ sets the softmax temperature. This kind of averaging tends to smooth out the quirks of any single model. It usually improves calibration and cuts down variance, giving risk scores that feel a bit more stable–something clinicians often care about more than a marginal bump in accuracy.

6.6 Why PANDA Outperforms Baselines

Before applying TCA, the PCA and t-SNE plots (Fig. 3a,c) show that the two hospitals' data don't quite land in the same neighborhood–there's some separation, though perhaps not as dramatic as one might expect from a textbook domain-shift example. Still, the shape of the clusters hints at meaningful differences in how the two cohorts distribute themselves in feature space. After alignment (Fig. 3b,d), those clouds pull a bit closer together. They don't collapse into a single blob, but the overlap becomes tighter in a way that feels more reassuring than the raw-input view.

When we looked at the numbers behind the scenes–the MMD, Wasserstein-1 distance, and symmetric KL divergence computed on the latent representations–they all moved in the direction we hoped for: smaller gaps, less tug-of-war between hospitals. These weren't included as explicit figures, but the calculations (following the definitions in Sec. 6) back up the visual impression. It's not perfect alignment, but it seems to argue that the method is at least nudging the domains toward the same latent "language."

Another piece that quietly helps is the cross-domain RFE step. By trimming the features down to the eight variables both hospitals actually measure–and that stay predictive across both–it strips away a lot of those site-specific quirks that often masquerade as signal. This makes the alignment problem less messy. There's even a theoretical hint supporting this: the covariance bound discussed in the Theoretical Foundation section on feature selection and domain adaptation suggests that selecting lower-variance shared features may shrink the alignment complexity. In practice, that seems to match what we observed: once the feature set stops dragging along hospital-specific noise, TCA has an easier time finding a common subspace that both cohorts can live with.

The same pattern showed up in the public TableShift BRFSS Diabetes benchmark, where a race shift creates measurable covariate and prevalence gaps. Pre-trained TabPFN representations kept the drop from source tuning to OOD small, suggesting that the learned priors already bridge some of the domain gap. Cross-domain RFE removed site- or survey-specific artifacts, stabilizing which features enter the Transformer context. TCA then shaved off the remaining discrepancy, yielding a small but repeatable gain over TabPFN without alignment. Tree-based models and TabPFN-no-TCA degraded more under the race shift, underscoring that pre-training alone is not sufficient when feature distributions move. Together, the trio–pre-trained representations, cross-domain RFE, and TCA–acts as a layered buffer against shift rather than a single brittle fix.

7 Evaluation

We assess PANDA across cross-institutional performance, domain adaptation, interpretability, and clinical utility, using a protocol meant to resemble what deployment would actually look like.

7.1 Evaluation Metrics and Statistical Analysis

7.1.1 Classification Performance Metrics

Results are averaged over 10-fold stratified cross-validation to temper label imbalance. The metrics are:

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

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the full dataset, and $\mathcal{D}_k$ be the k-th fold. For 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:** Plot $TPR(\tau)$ versus $FPR(\tau)$ for $\tau \in [0,1]$ to see the sensitivity-specificity trade-off.
- **Calibration Curves:** Check agreement between predicted probability $\hat{p}_i$ and observed frequency y_i. For K equal-width bins $B_k = [k/K, (k+1)/K)$:

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

- **Decision Curve Analysis (DCA):**

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

With benchmark strategies:

$$NB_{all}(p_t) = \text{Prevalence} - (1 - \text{Prevalence}) \cdot \frac{p_t}{1-p_t}, \quad NB_{none} = 0$$

where $\text{Prevalence} = \frac{1}{n}\sum_{i=1}^{n} y_i$

7.2 Experimental Setup and Results

Structured clinical data from two cancer centers in China provided a training cohort (Cohort A, $n_s = 295$) and an external test cohort (Cohort B, $n_t = 190$). Cohort A contained 63 structured features; Cohort B contained 58 (Table 6).

Table 6: The 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
CRE	73.41 ± 17.16	62.94 ± 13.64
NSE	13.07 ± 3.90	13.82 ± 4.36
Outcome (Malignant)		
Yes/Positive	189 (64.1%)	125 (65.8%)
No/Negative	106 (35.9%)	65 (34.2%)

In source-domain evaluation (10-fold cross-validation on Cohort A), PANDA led on all metrics (Fig. 2): AUC 0.829, accuracy 0.746, F1-score 0.810, precision 0.786, recall 0.846. The high recall is what screening workflows tend to care about. Classical machine learning methods were moderate (Random Forest AUC 0.752; XGBoost 0.742), and clinical scores fared poorly.

For external validation (train on Cohort A, test on Cohort B), the TCA-enhanced PANDA model again came out ahead (AUC 0.705, F1-score 0.808, recall 0.944), with the non-adaptive version slightly behind at AUC 0.698. Among baselines, LASSO LR reached AUC 0.668 with recall 0.943; Random Forest dropped to 0.632; SVM, GBDT, and XGBoost fell below 0.59, underscoring shift sensitivity.

Feature-space checks (Fig. 3) suggest TCA is doing its job: PCA and t-SNE views tighten the alignment between source and target after transformation, even if some scatter remains.

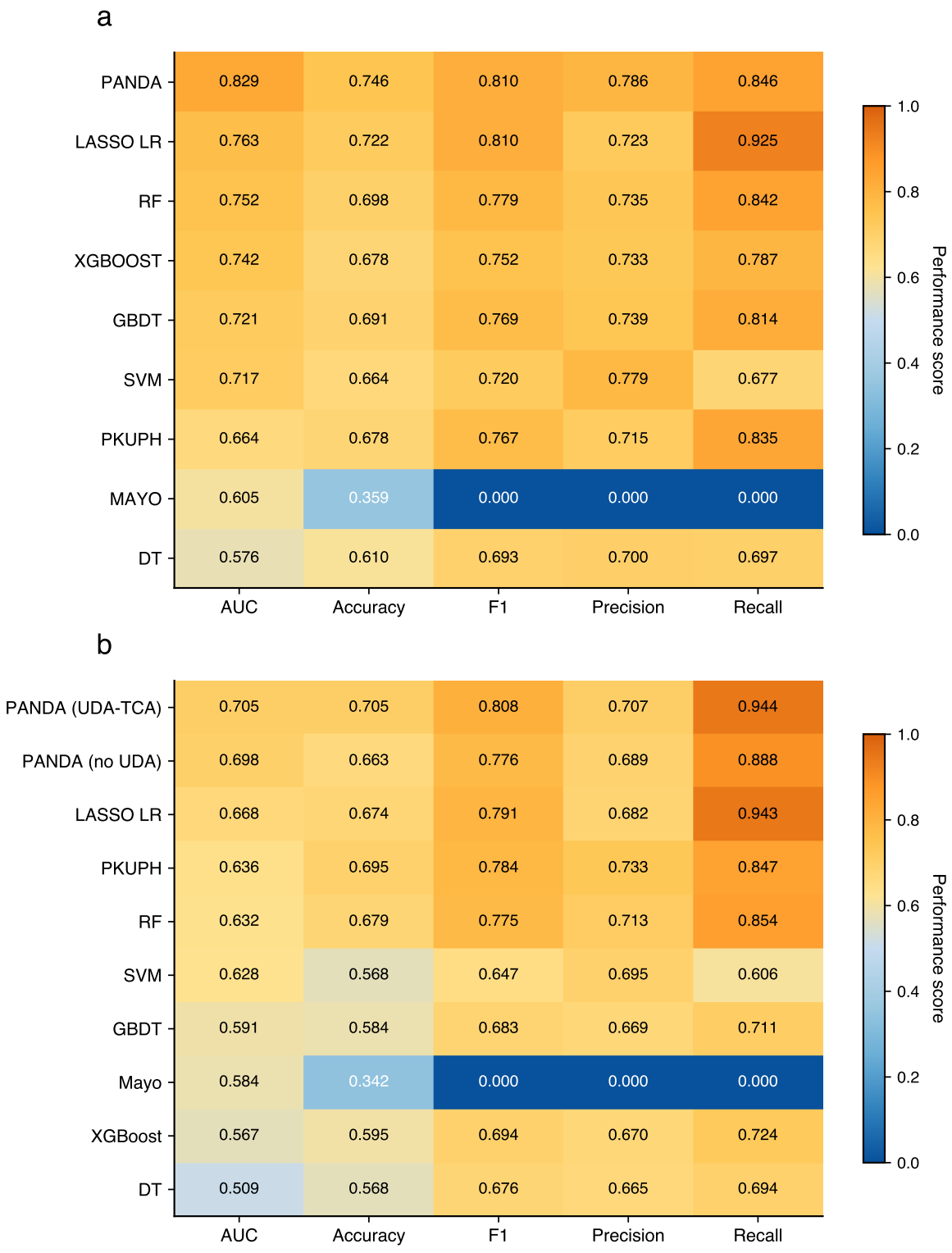

Figure 2: **Performance comparison across source and target domains.** **a** Source domain 10-fold cross-validation performance heatmap across five classification metrics. The PANDA framework achieves the best overall performance across all metrics. **b** Cross-domain performance heatmap on the external validation set. The TCA-enhanced PANDA model shows the highest AUC and recall, indicating improved generalization under domain shift.

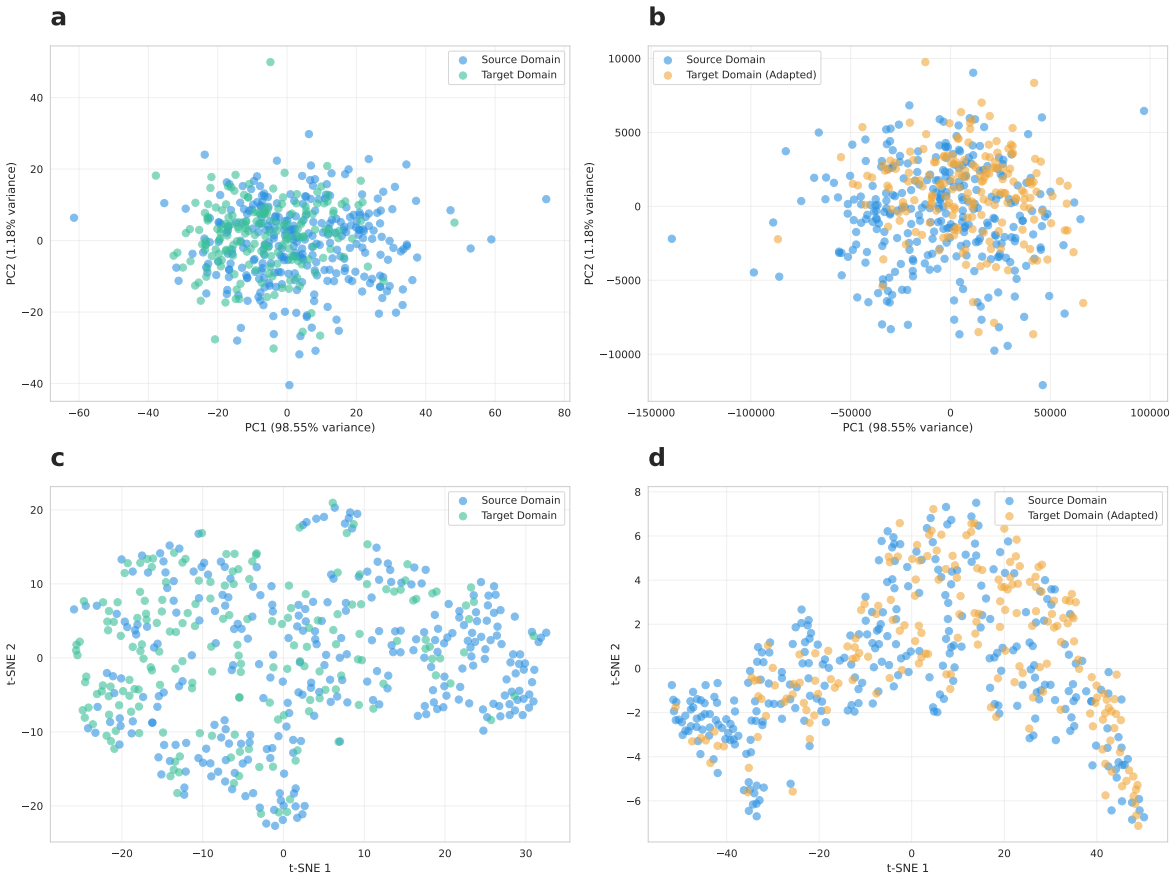

Figure 3: TCA-based domain adaptation visualization. **a,b** PCA visualization before and after TCA transformation, showing improved alignment of target samples with source samples. **c,d** t-SNE visualization before and after TCA transformation, demonstrating enhanced cluster center alignment and distribution consistency.

7.3 Additional Cross-Domain Validation on a Public Benchmark

Beyond the private cross-hospital evaluation, we stress-test PANDA on a public benchmark with an explicit domain shift: the TableShift BRFSS Diabetes task (race shift: White → non-White). All plots were re-rendered with high-DPI fonts for print clarity. Table 7 summarizes the full cohorts prior to subsampling, highlighting the large scale, the difference in diabetes prevalence (12.5% vs. 17.4%), and the definition of the shift variable (PRACE1). For a like-for-like comparison with our small-sample hospital setting, we sample $n_{\text{train}} = 1024$ and $n_{\text{test}} = 2048$ while maintaining the source/target label balance.

Table 7: In-distribution (ID) versus out-of-distribution (OOD) BRFSS Diabetes cohorts under race shift.

Characteristic	**Source / ID (`PRACE1` = 1)**	**Target / OOD (`PRACE1` $\in \{2,3,4,5,6\}$)**
Sample size	Train: 969,229 Val: 121,154 Test: 121,154	Val: 23,264 Test: 209,375
Diabetes positive rate	12.5% (train)	17.4% (OOD test)
Survey years	2015: 245,675 2016: 5,789 2017: 244,996 2018: 6,403 2019: 221,847 2020: 9,630 2021: 223,088 2022: 11,801	2015: 49,216 2016: 1,507 2017: 52,150 2018: 1,424 2019: 48,012 2020: 3,147 2021: 50,595 2022: 3,324
Domain shift variable	`PRACE1`=1 (non-Hispanic White)	`PRACE1`$\in \{2,3,4,5,6\}$ (other races)
Label definition	`DIABETES`=1 (Yes)	`DIABETES`=0 (No / Prediabetes / Borderline)
Modelling sample	1,024 training samples	2,048 evaluation samples
Sampled diabetes rate	13.2% (train subset)	17.3% (OOD subset)
Diabetes outcome		
Positive	135 (13.2%)	355 (17.3%)
Negative	889 (86.8%)	1,693 (82.7%)
Sampled years	2015: 245 2016: 8 2017: 278 2018: 2 2019: 232 2020: 7 2021: 241 2022: 11	2015: 497 2016: 17 2017: 488 2018: 17 2019: 445 2020: 30 2021: 525 2022: 29

The dataset contains 142 numerical features with cross-year alignment. Feature names have underscores removed for consistency, while the `IYEAR` variable is retained to preserve temporal information. Preprocessing steps include removing NA rows and outliers (`DRNK_PER_WEEK`=99900), recoding healthday responses from 88 to 0, mapping `SEX` to binary 0/1 values, and imputing missing values for `TOLDHI` and `SMOKDAY2` as `NOTASKED_MISSING`.

Because we keep the native 0.5 threshold and do not apply class weighting on a 17% positive cohort, classifiers lean toward the majority (negative) class. The calibration and decision-curve panels in Fig. 4 therefore pair high accuracy with low recall and F1: the default threshold misses many positives, while precision is moderate but not enough to lift the harmonic mean. We retain these raw numbers to reflect the default-deployment setting; threshold tuning or balanced losses could raise recall if desired. Fig. 5 contrasts the source and race-shift OOD metrics, showing how PANDA_TCA holds the smallest

deltas while tree models drop sharply under the shift.

Across seven baselines and variants, PANDA with TCA attained the highest OOD AUC (0.804) and accuracy (0.848), edging the non-adaptive TabPFN variant (AUC 0.796, accuracy 0.847) and tree ensembles (best GBDT AUC 0.783). Using the source cross-validation tuning run (AUC 0.809, accuracy 0.848) as a reference, the race-shift drop is mild (ΔAUC ≈ 0.005; ΔAccuracy < 0.001), while adaptation provides a small but consistent gain over PANDA_NoUDA (+0.008 AUC). Classic SVM (AUC 0.646) and decision trees (AUC 0.566) degrade far more under the race shift, underscoring the value of explicit alignment even on large public data.

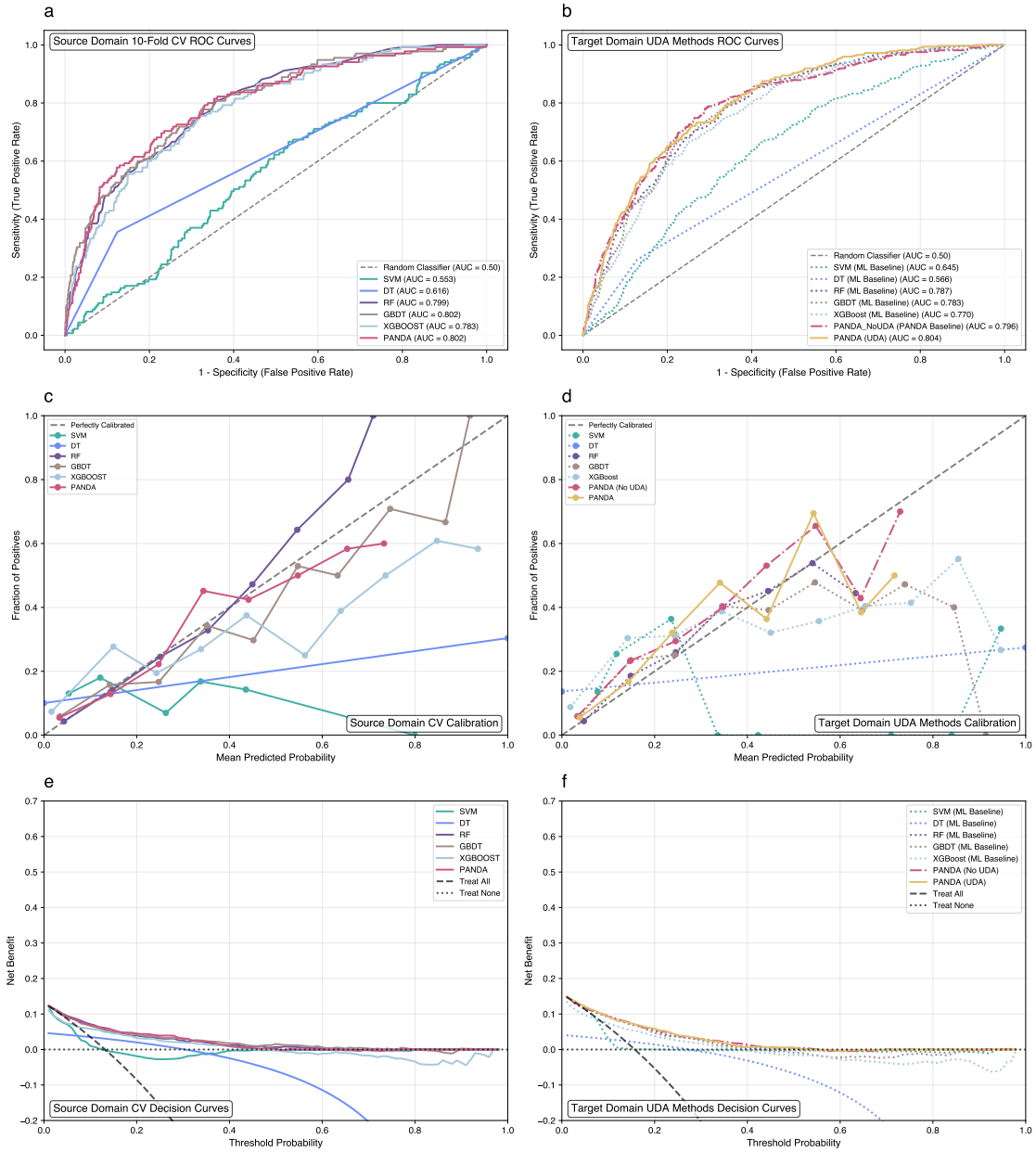

Figure 4: **TableShift BRFSS Diabetes (race shift: White → non-White). a,b** ROC curves for baselines and PANDA variants on the public benchmark. **c,d** Calibration curves highlighting probability alignment post-adaptation. **e,f** Decision curves illustrating net benefit across threshold ranges.

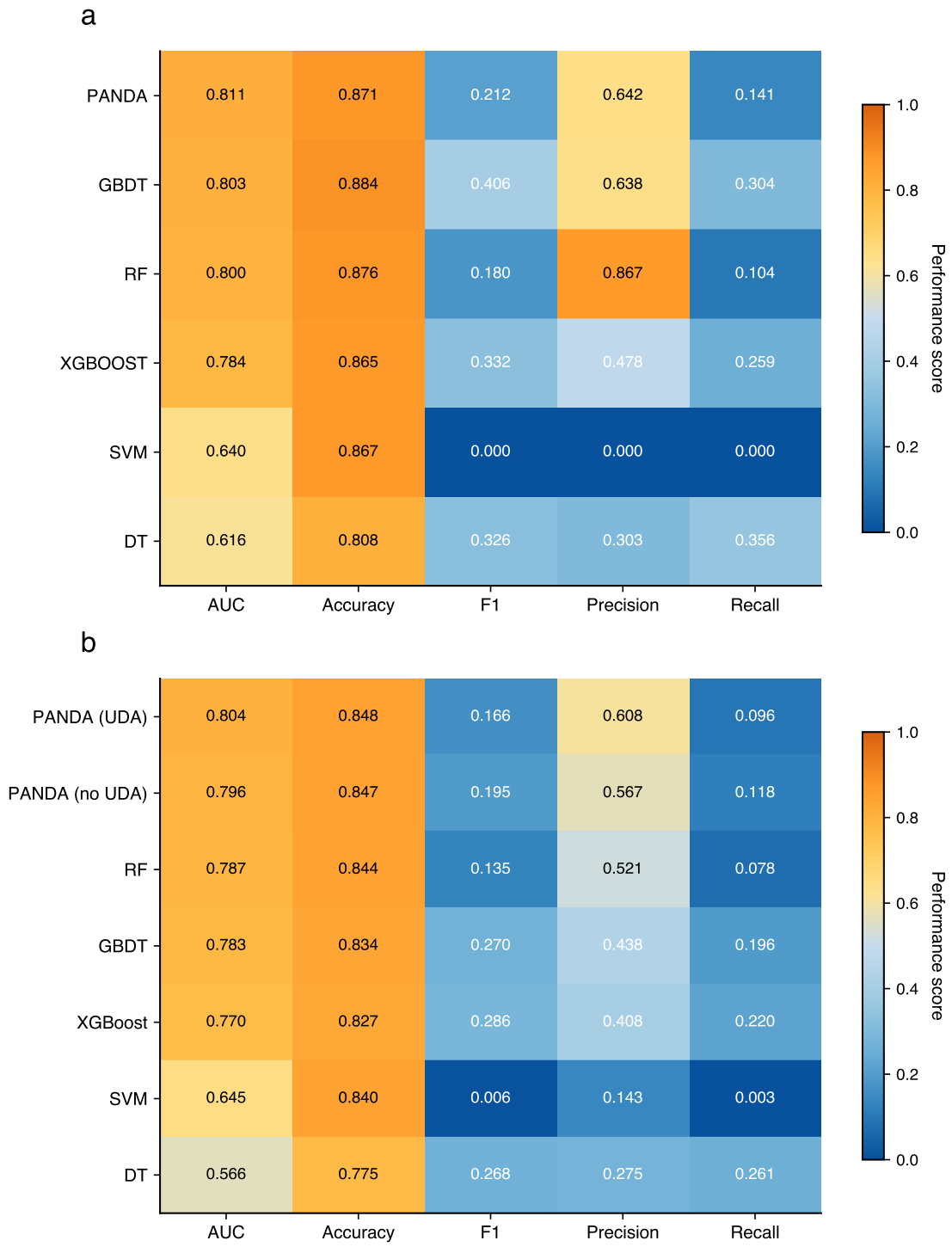

Figure 5: **Performance heatmaps for the TableShift BRFSS Diabetes task.** **a** Source cross-validation metrics. **b** Race-shift (White → non-White) OOD metrics.

7.4 Model Explainability, Reliability, and Clinical Utility

RFE with the pre-trained model kept interpretation manageable, and performance across subset sizes leveled off around 9–13 features (Fig. 6). In terms of reliability, the ROC curves give PANDA a clear edge–AUC 0.829 on the source cohort and 0.705 for the TCA-augmented model on the external one. Calibration plots also place PANDA closer to the diagonal, with TCA nudging the target-side curve a bit nearer to what we would hope for. Decision curves, which weigh net clinical benefit across thresholds, tilt in PANDA's favor as well, and the TCA variant adds a small but noticeable gain on the external cohort.

8 Conclusion

This work links pre-trained tabular foundation models with domain adaptation to address long-standing issues in tabular learning under distribution shift. PANDA suggests that foundation-model priors and statistical alignment can reinforce one another, helping models generalize from scarce, heterogeneous samples where standard supervised approaches often stumble. The evidence is not sweeping, but it does point toward a practical recipe rather than a one-off trick.

Several methodological themes stand out. Pre-trained representations reduce the effective sample burden, letting high-capacity models behave sensibly in low-data regimes. Cross-domain feature selection pinpoints predictors that consistently transfer between sites, which makes alignment less fragile. Embedding TCA into these smoother representation spaces also seems to make domain transitions more workable. Taken together, these pieces outline a reasonable blueprint for adapting pre-trained tabular models across domains without relying on abundant labels.

Beyond pulmonary nodules, the same ingredients likely extend to other structured settings with small samples and noticeable shift–financial risk scores that change across branches, industrial monitoring when sensors drift, or hospital-adjacent analytics where coding practices evolve. PANDA is meant as a reusable template that treats pre-trained representations as portable priors rather than site-specific quirks.

The claims about smoother representations, feature-selection interactions, and reduced sample complexity align with the observed reduction in discrepancy and the improved external performance, hinting that pre-trained tabular models may broaden what is feasible in domain adaptation.

Open questions remain: scaling to larger tabular foundation models, moving toward multimodal pre-training, tightening feature selection for distributional robustness, and handling continual shift. As tabular models mature, pairing them with principled alignment may redefine how we handle shift.

In sum, PANDA frames tabular domain adaptation around pre-trained representations that support cross-domain generalization, aiming for deployments where shift is the rule rather than the exception. The same recipe now holds across private clinical cohorts and a public TableShift benchmark, hinting at dataset-agnostic, shift-resilient generalization beyond pulmonary nodules.

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.

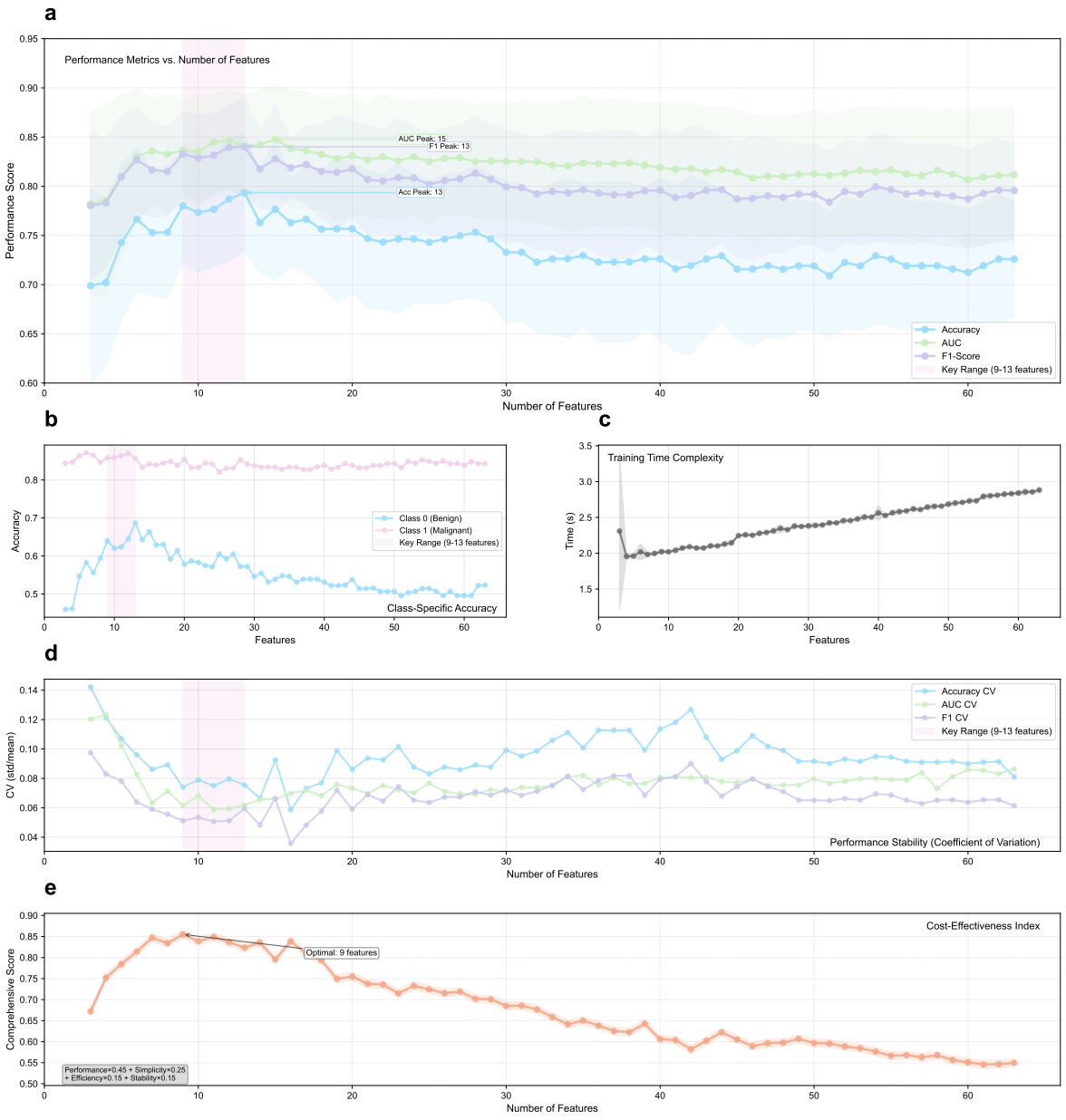

Figure 6: Comprehensive feature selection and performance analysis using recursive feature elimination (RFE). **a** AUC, accuracy, and F1 curves as functions of the number of selected features. Performance plateaus around 9–13 features, aligning with the preference for simpler models. Shaded regions show variance across 10-fold cross-validation. **b** Class-specific accuracy for malignant and benign cases across feature 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.

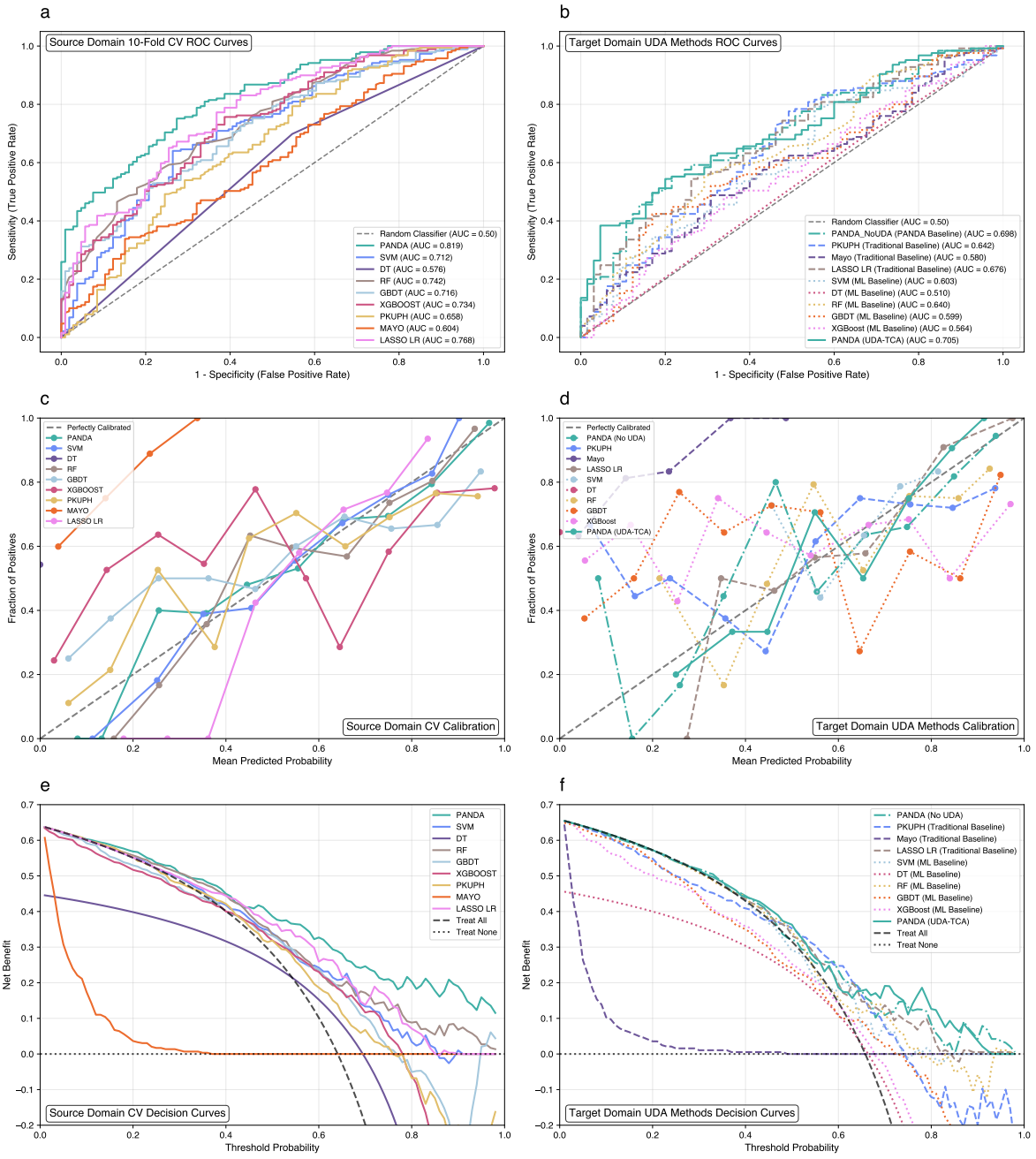

Figure 7: **Cross-hospital pulmonary nodule task (Cohort A → Cohort B). a,b** ROC curves on source and target cohorts. **c,d** Calibration plots showing probability reliability after TCA. **e,f** Decision curves quantifying net benefit for internal versus external deployment.

[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] 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.

[6] 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.

[7] 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.

[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] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Prior labs.

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

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

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

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

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

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

[23] 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.

[24] 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.

[25] 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.

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

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

[28] 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.

[29] 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.

[30] 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.

[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 multitask 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] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

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

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

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

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

[82] 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.