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. Beyond centralized settings, federated learning extends these ideas by allowing multiple hospitals to collaborate without sharing raw data. Surveys on federated learning for medical imaging and pattern recognition summarize how FL can pool experience across institutions while preserving privacy, and methods such as FedFusion explicitly combine domain adaptation with personalized encoders to handle heterogeneous feature spaces and scarce labels [70, 71, 72]. However, most federated frameworks target high-volume imaging or EHR tasks, assume substantial local computation and at least some labeled data at each site, and still rely on shared model architectures and broadly aligned feature schemas. They are therefore complementary to, rather than a replacement for, lightweight DA strategies tailored to very small tabular cohorts with partially mismatched feature sets.

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

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

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

High-dimensional yet small-sample tabular cohorts are ubiquitous in medicine: lung screening registries, omics panels, and survey-based risk scores often contain hundreds of variables for only a few hundred or thousand patients. Naïve learning in this regime leads to unstable decision boundaries and non-reproducible feature attributions. Feature selection methods aim to reduce dimensionality, stabilize inference, and focus clinician attention on biomarkers that are both predictive and economical to collect.

2.3.1 Small-sample and high-dimensional feature selection

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

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

Method family	Typical modality	Key assumptions	Pros	Limitations for sma cross-hospital tabula cohorts
Statistical alignment (MMD, TCA, CORAL, transport)	Tabular, EHR, imaging	Shared feature schema; access to source data and unlabeled target samples; primarily covariate shift	Closed-form or deterministic mappings; stable when target labels are absent; easy to plug into existing pipelines [32, 56, 57, 59]	Does not directly co rect label or concep shift; assumes overla ping feature sets; ma misalign rare subgrou without additional ca bration [27, 24, 25]
Adversarial representation learning (DANN-style)	Imaging, EHR sequences	Access to source and target data with domain labels; discriminator encouraged to remove site identity	Learns domain-invariant representations jointly with task loss; flexible for complex modalities [25, 58]	Unstable on tiny c horts; discriminato exploit missingness pa terns, causing encode to discard predicti features; can underpe form ERM in ICU-sty tasks [58, 22]
Feature-space DA and domain-aware FS (FSDA, DomainATM)	Tabular, EHR	At least partially shared feature space; access to both domains during training	Selects features that are predictive and stable across sites; reduces reliance on site-specific surrogates and noisy biomarkers [60, 22, 62, 63]	Still assumes sizab overlap in measure variables; does not n tively handle missi entire feature blocks unlabeled target hosp tals with severe schem mismatch
Domain generalization and temporal adaptation (TableShift, Wild-Tab, Wild-Time)	Tabular	Multiple labeled source distributions; no target labels during training	Reveal failure modes under temporal, demographic, and institutional shift; provide standardized evaluation suites [27, 68, 55, 28]	Many domai generalization obje tives (e.g., GroupDR IRM, VREx) rarely be strong ERM or GBD baselines; benchmar show label shift and ca ibration drift domina what feature matchi can fix [27, 67]
Federated and federated-DA frameworks (FL, FedFusion-style)	Imaging, tabular	Multiple compute-capable hospitals; communication budget; typically some local labels and shared model architecture [70, 71, 72]	Preserve data privacy while learning from distributed cohorts; can combine personalization with domain adaptation and label efficiency	Often require significa local computation a labeled target data; f cus on large imaging EHR tasks; do not d rectly address very sma tabular cohorts with fe ture mismatch and stri label scarcity

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, 77, 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, 78]. While recalibration or re-estimation of coefficients can partially restore performance, these fixes require local labels and do not address feature-mismatch: new hospitals may lack some variables (e.g., emphysema grading) or measure them differently.

Targeted audits make the degradation concrete. In TB-endemic Korean hospitals, Mayo and VA shrink to AUC $\approx$0.60 while Brock declines to $\approx$0.68 despite an internal AUC near 0.94, and Chinese

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

Method	Model family	Small-sample / imbalance handling	Interpretability characteristics	Representative biomedical use cas
Recursive feature elimination (RFE) with linear or tree models	Wrapper around SVM, logistic regression, or tree ensembles	Wrapper search over feature subsets can overfit when N is small and features are correlated; often combined with cross-validation and class-balanced sampling	Produces explicit ranked feature lists and compact subsets; easy to inspect and map to clinical variables [73]	Widely used in ea gene-expression a biomarker panels; sis for many clin risk-score and radiom pipelines
LASSO / elastic-net logistic regression	Embedded linear models	ℓ_1 or ℓ_1+ℓ_2 penalties shrink coefficients, providing some robustness to high dimensionality; still assumes linear log-odds and can be unstable under heavy collinearity	Sparse coefficients directly indicate selected features; compatible with odds-ratio interpretation familiar to clinicians [73]	Common in radiom and EHR risk mod where interpretabi and coefficient-ba reporting are required
GRACES	Graph-convolutional-network-based FS [74]	Specifically targets high-dimensional, low-sample-size data by modeling feature relations on a graph; alleviates overfitting compared with independent filters	Outputs a compact subset informed by graph structure; can be visualized as a network of interacting biomarkers	Demonstrated on om style datasets; suita when prior knowle or correlations betw biomarkers are imp tant
DeepFS	Deep feature screening with autoencoders [75, 76]	Uses deep encoders to learn low-dimensional representations and rank features, handling ultra-high-dimensional, sparse, and potentially imbalanced data	Provides importance scores for each original feature; retains flexibility to operate in supervised or unsupervised mode	Evaluated on synth and biomedical hi dimensional datas useful when the num of variables far exce the number of patient
Domain-aware FS (FSDA-style)	Feature selection for domain adaptation [60]	Encourages selection of features that remain predictive across domains, implicitly handling covariate shift between sites	Produces subsets that are jointly predictive and domain-stable, supporting cross-hospital deployment	Applied to benchm DA tasks; conc tually aligned w cross-hospital biomar selection in multi-cen medical studies
Transformer / foundation-model-based FS	Attention- or score-based selection using TabNet, Tab-Transformer, and tabular foundation models [11, 12, 45]	Leverages high-capacity or pre-trained models to estimate nonlinear feature importance; can be combined with RFE to mitigate small-sample overfitting	Attention weights, feature masks, or permutation-based scores yield ranked features; aligns with explainable-AI practices	Increasingly used biomedical tabu and omics datas PANDA's cross-coh RFE uses a tabu foundation model as scoring backbone

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

Each classical score carries its own design trade-offs. The Mayo Clinic model was derived from several hundred clinic-referred patients with indeterminate nodules, emphasizing age, smoking, nodule diameter, spiculation, and upper-lobe location, whereas the Veterans Affairs model targeted high-risk, predominantly male veterans with larger lesions [1, 77]. The Brock (PanCan) model was trained in a screening cohort enriched for small nodules and incorporates emphysema, family history, and more granular morphology descriptors, while the PKUPH and Li scores adapt similar feature sets to Chinese tertiary-hospital and screening populations [2, 3, 4, 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 [78]. Across Mayo, VA, Brock, and PKUPH, external validations repeatedly document drops from internal c-statistics in the high-0.80s to 0.60–0.75 when applied to community screening or granulomatous-disease–endemic regions [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 4 sketches representative scores along these dimensions. In development, all achieve reasonable discrimination and are simple enough to implement as bedside calculators, but the same simplicity makes them brittle under shift: logistic coefficients absorb local prevalence, imaging protocols, and referral patterns, so external use without recalibration results in systematic underestimation or overestimation of risk in particular subgroups.

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

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

Concrete exemplars reinforce that fragility. The Bayesian Integrated Malignancy Calculator (BIMC) blended radiomics with clinical covariates and modestly outperformed Mayo, Brock, and PKUPH (AUC ≈0.90 vs. ≈0.78) on an Italian derivation cohort, yet its advantage diminished when scanners, slice thickness, or kernels changed [83]. Hawkins-style NLST radiomics achieved AUC ≈0.83 without external validation, and retrospective audits show that a single reconstruction tweak can reorder the features selected by LASSO [5, 10]. Radiomics therefore supplies richer morphology descriptors but still requires harmonization, feature governance, and domain-aware alignment rather than assuming reproducibility across hospitals.

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

Study / model	**Imaging data**	**Centers / nodules**	**Classifier**	**Internal AUC**	**External AU**
Generic radiomics pipelines	2D/3D chest CT or low-dose CT nodules	Single- or few-center cohorts (sizes vary)	LASSO- or stability-selected features with SVM, RF, or GBDT	Typically 0.75–0.90 on internal validation cohorts	Often 0.10–0. lower on extern scanners or ho pitals
Multi-center reproducibility analyses	3D CT radiomics features across scanners	Multi-center CT datasets	–	–	–
Radiomics + clinical scoring models	Radiomics signatures combined with clinical descriptors	Hospital-specific or regional nodule cohorts	Elastic-net / logistic regression, RF, or GBDT	High AUC in derivation cohorts (often ≥ 0.80)	External AU can drop to t mid-0.70s lower

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 [84, 85]. Causey et al.'s NoduleX reproduced radiologist malignancy ratings with AUC $\approx$0.99 on LIDC-IDRI, and Google's NLST-scale 3D CNN maintained AUC $\approx$0.94 on an independent hospital cohort of 1,139 CTs, illustrating how massive, homogeneous datasets can suppress variance [86, 29]. 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, 84, 85]. GBDTs and random forests improve internal calibration and handle nonlinear interactions but still require site-specific recalibration or feature mapping before deployment because their learned weights implicitly encode scanner kernels and local smoking histories. Multi-modal models that fuse deep image features with clinical covariates via late fusion or stacking demonstrate promising gains on LIDC-IDRI and NLST, yet most studies remain single-center or rely on random train–test splits that do not reflect real cross-hospital deployment. Only a handful of works evaluate performance when training on one hospital and testing on another, and these typically report substantial AUC drops and unstable decision thresholds [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 diagnosis represents a complex intersection of small-sample learning, high-dimensional tabular data, and significant distribution shifts. In this chapter, we formalize the problem of predicting pulmonary nodule malignancy across different hospitals as an Unsupervised Domain Adaptation (UDA) task. We establish the mathematical foundations for our proposed PANDA framework, grounding it in the theory of Prior-Data Fitted Networks (PFNs), domain adaptation bounds, and kernel-based distribution alignment.

3.1 Mathematical Notation and Preliminaries

We consider a supervised classification task where the input space is a d-dimensional feature space $\mathcal{X} \subseteq \mathbb{R}^d$ comprising mixed numerical and categorical variables (e.g., patient age, tumor diameter, smoking status), and the output space is a binary label space $\mathcal{Y} = \{0, 1\}$ (e.g., benign vs. malignant).

A **domain** is defined as a joint probability distribution $P(X, Y)$ over $\mathcal{X} \times \mathcal{Y}$. We are given data from two distinct domains:

- **Source Domain (Hospital A):** We observe a labeled dataset $\mathcal{D}_S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ drawn i.i.d. from the source distribution $P_S(X, Y)$. Here, n_s is the number of labeled samples in the source hospital.
- **Target Domain (Hospital B):** We observe an unlabeled dataset $\mathcal{D}_T = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$ drawn i.i.d. from the marginal distribution $P_T(X)$ of the target domain $P_T(X, Y)$. The target labels Y_T are unobserved during training, reflecting the real-world constraint of deploying models to new hospitals without local ground truth.

Feature Heterogeneity: In multi-center studies, hospitals often record different sets of variables. Let $\mathcal{F}_S$ and $\mathcal{F}_T$ be the sets of feature indices available in the source and target domains, respectively. We define the **shared feature subspace** as the intersection $\mathcal{F}_\cap = \mathcal{F}_S \cap \mathcal{F}_T$. The effective input dimensionality for the cross-domain model is $d_\cap = |\mathcal{F}_\cap|$. Features in the set difference $\mathcal{F}_\setminus = (\mathcal{F}_S \cup \mathcal{F}_T) \setminus \mathcal{F}_\cap$ are considered site-specific and are typically discarded for alignment, though they may contain domain-specific predictive signals.

Table 6 summarizes the key mathematical notations used throughout this thesis.

Table 6: Summary of Mathematical Notation

Symbol	Description
$\mathcal{X}, \mathcal{Y}$	Input feature space and label space
P_S, P_T	Source and Target domain distributions
$\mathcal{D}_S, \mathcal{D}_T$	Source (labeled) and Target (unlabeled) datasets
n_s, n_t	Number of samples in source and target domains
$\mathbf{x} \in \mathbb{R}^d$	Feature vector (patient covariates)
$y \in \{0, 1\}$	Target label (0: Benign, 1: Malignant)
$\mathcal{F}_\cap$	Set of shared features across domains
$\phi(\cdot)$	Feature extractor / Encoder (e.g., TabPFN backbone)
$\psi(\cdot)$	Domain adaptation mapping (e.g., TCA projection)
$h(\cdot)$	Final classifier hypothesis
$\mathcal{H}$	Reproducing Kernel Hilbert Space (RKHS)
$\text{MMD}(P_S, P_T)$	Maximum Mean Discrepancy between domains

3.2 The Tabular Data Generation Process: A PFN Perspective

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

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

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

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

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

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

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

- It avoids iterative gradient descent on the small dataset, mitigating the risk of overfitting to noise.
- It leverages the "knowledge" encoded in the prior P_{prior}, effectively transferring inductive biases about tabular structures (e.g., decision boundaries are often aligned with axes, sparsity is common) to the medical task.

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

3.3 Formalizing Domain Shift

The core challenge in our research is that $P_S(X, Y) \neq P_T(X, Y)$. This joint distribution shift can be decomposed into three primary components relevant to medical AI:

3.3.1 Covariate Shift: The Acquisition Gap

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

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

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

3.3.2 Label Shift: The Prevalence Gap

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

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

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

3.3.3 Concept Shift: The Definition Gap

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

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

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

3.3.4 Theoretical Bound on Generalization Error

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

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

where:

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

This bound highlights that minimizing source error is insufficient; we must explicitly minimize the divergence $d_{\mathcal{H}\Delta\mathcal{H}}(P_S, P_T)$.

3.4 Theoretical Constraints of Existing Models

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

3.4.1 Gradient Boosted Decision Trees (GBDT)

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

1. **Non-Differentiability:** The piecewise constant decision boundary is non-differentiable with respect to input features. This precludes the use of gradient-based domain alignment techniques (like Adversarial Training or Gradient Reversal Layers) which require backpropagating a domain loss into the feature encoder.

2. **Inability to Extrapolate:** Tree models cannot extrapolate beyond the range of the training data. If covariate shift pushes the target distribution $P_T(X)$ outside the support of $P_S(X)$, the tree maps all such points to the value of the nearest leaf node, often resulting in statistically invalid predictions.

3.4.2 Deep Tabular Models

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

1. **Data Hunger:** Neural networks typically require large datasets ($N > 10^4$) to converge to a generalizable solution. With $N \approx 300$, deep models are prone to severe overfitting or convergence to local minima.

2. **Rotational Invariance:** Standard MLPs are rotationally invariant, but tabular features are not rotationally interchangeable (e.g., rotating "Age" and "Creatinine" axes creates a nonsensical feature space). This mismatch in inductive bias makes them less sample-efficient than tree-based or prior-fitted methods.

3.5 Transfer Component Analysis (TCA) Optimization Objective

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

The empirical MMD distance is defined as:

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

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

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

where:

- $\mathbf{K} \in \mathbb{R}^{(n_s+n_t)\times(n_s+n_t)}$ is the kernel matrix computed on the union of source and target data. We specifically employ a **Linear Kernel** on the TabPFN embeddings: $K_{ij} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. This is justified because the TabPFN encoder ϕ acts as a powerful non-linear feature extractor, linearizing the decision boundary in its latent space $\mathbb{R}^h$. Using a linear kernel on top of these embeddings provides a robust and computationally efficient alignment without introducing complex hyperparameter tuning for RBF bandwidths.

- $\mathbf{L}$ is the MMD coefficient matrix, with elements $L_{ij} = 1/n_s^2$ if $x_i, x_j \in \mathcal{D}_S$, $1/n_t^2$ if $x_i, x_j \in \mathcal{D}_T$, and $-1/(n_s n_t)$ otherwise.

- $\mathbf{H} = \mathbf{I} - \frac{1}{n_s+n_t}\mathbf{1}\mathbf{1}^\top$ is the centering matrix.

- The constraint $\mathbf{W}^\top \mathbf{K}\mathbf{H}\mathbf{K}\mathbf{W} = \mathbf{I}$ ensures the variance of the projected data is preserved (maximizing information).

3.6 The PANDA Framework: A Unified Formalization

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

$$f_{\text{PANDA}}(\mathbf{x}) = h\left(\psi\left(\phi(\mathbf{x}; \mathcal{F}^*)\right)\right) \tag{8}$$

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

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

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

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

1. **Model Fitting:** We define the TabPFN posterior predictive distribution conditioned on the current subset $S^{(t)}$ using the source data $\mathcal{D}_S$.
2. **Importance Estimation ($\mathcal{I}$):** Unlike linear models, TabPFN is a non-parametric meta-learned model. We approximate feature importance using **Permutation Importance**. For each feature $j \in S^{(t)}$, we compute the degradation in the AUC metric when feature j is randomly permuted in the validation set:

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

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

$$j_{\text{elim}} = \underset{j \in S^{(t)}}{\text{argmin}}\, \mathcal{I}(j; S^{(t)}) \quad \longrightarrow \quad S^{(t+1)} = S^{(t)} \setminus \{j_{\text{elim}}\} \tag{10}$$

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

3.6.2 Stage 2: Foundation Model Encoding (ϕ)

The selected features are fed into the frozen TabPFN encoder ϕ.

$$\mathbf{z} = \phi(\mathbf{x}; \mathcal{D}_{\text{context}}) \tag{11}$$

Here, ϕ leverages the large-scale pre-training prior P_{prior} to map the raw, low-dimensional inputs $\mathcal{F}^*$ to a robust, high-dimensional latent representation $\mathbf{z}$.

3.6.3 Stage 3: Domain Adaptation Mapping (ψ) and Classification (h)

The latent representations are aligned using the learned TCA projection $\psi(\mathbf{z}) = \mathbf{W}^\top k(\mathbf{z}, \cdot)$. Finally, a simple hypothesis h (e.g., Logistic Regression) is trained on the aligned source features $\psi(\phi(\mathcal{D}_S))$ and applied to the aligned target features $\psi(\phi(\mathcal{D}_T))$.

3.7 Clinical-Statistical Mapping

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

3.8 Problem Constraints and Research Scope

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

- **Small Sample Size Constraint:** The sample sizes n_s, n_t are typically in the range of 100 to 1000. This prohibits the use of deep domain adaptation networks.
- **Privacy and Data Silos:** We assume source data $\mathcal{D}_S$ and target data $\mathcal{D}_T$ cannot be physically merged.
- **Class Imbalance:** The prevalence of the positive class is often low ($\pi < 0.3$), requiring AUC-centric optimization.

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

Table 7: Mathematical Mapping of Clinical Problems to PANDA Components

Clinical Challenge	Statistical Mechanism	PANDA Component	Theoretical Justification
Scanner Variance (Sharp vs. Smooth Kernels)	Covariate Shift: $P_S(X) \neq P_T(X)$	TCA (Transfer Component Analysis)	Minimizes MMD divergence $d_{\mathcal{H}\Delta\mathcal{H}}$ in RKHS.
Referral Patterns (Cancer Center vs. Screening)	Label Shift: $P_S(Y) \neq P_T(Y)$	Ensemble Aggregation & Temperature Scaling	Calibrates posteriors; smooths overconfidence from prior mismatch.
Biological Confounders (TB vs. Cancer)	Concept Shift: $P_S(Y\|X) \neq P_T(Y\|X)$	Cross-Domain RFE	Minimizes joint error λ by removing unstable features.

4 Solution: The PANDA Framework

To bridge the gap between advanced tabular foundation models and the practical constraints of cross-hospital medical AI, we introduce PANDA (**P**re-trained t**A**bular fou**N**dation model with **D**omain **A**daptation). PANDA is a composite algorithmic framework designed to predict pulmonary nodule malignancy with high stability across heterogeneous clinical environments. It explicitly addresses the tripartite challenge of small-sample scarcity, distribution shift, and feature heterogeneity through a tightly integrated pipeline.

We formalize the PANDA solution as a composite function $f_{\text{PANDA}} : \mathcal{X} \to [0, 1]$ mapping raw, heterogeneous input space to a calibrated malignancy probability. The framework consists of four sequential stages: (1) Domain-Aware Feature Alignment and Selection, (2) Foundation Model Feature Extraction, (3) Latent Space Domain Adaptation, and (4) Multi-View Ensemble Classification.

4.1 Architectural Overview

The PANDA framework operates as a directed acyclic graph (DAG) of data transformations. Table 8 details the data processing flow, tracking the evolution of feature representations from raw clinical inputs to the final ensemble prediction.

Table 8: Data Flow and Transformations in the PANDA Framework. The pipeline progressively refines the data from raw, high-dimensional inputs to low-dimensional, domain-invariant embeddings, and finally to a calibrated probability.

Stage	Component	Input Space	Core Operation	Output Space
1	Feature Alignment	$\mathbb{R}^{d_{\text{raw}}}$	Schema intersection $\cap$	$\mathbb{R}^{d_\cap}$
2	Cross-Domain RFE	$\mathbb{R}^{d_\cap}$	Iterative elimination via $\mathcal{I}(f)$	$\mathbb{R}^k$ ($k \approx 8$)
3	TabPFN Encoder	$\mathbb{R}^k$	$\mathbf{z} = \text{Transformer}_{\theta \setminus W_{\text{head}}}(\mathbf{x})$	$\mathbb{R}^h$ ($h = 128$)
4	Latent TCA	$\mathbb{R}^h$	Projection $\mathbf{z}' = \mathbf{W}^\top \mathbf{z}$	$\mathbb{R}^m$ ($m = 15$)
5	Classification	$\mathbb{R}^m$	Logistic Regression $h(\mathbf{z}')$	$[0, 1]$ (Prob.)
6	Ensemble	$[0, 1]^{B \times S}$	Temperature-scaled Averaging	$[0, 1]$ (Final)

The process begins with **Feature Alignment**, where the schema of the target hospital's data is intersected with the source schema to identify the maximum common feature set. This handles the "Missingness Shift" inherent in multi-center data. Subsequently, **Cross-Domain Recursive Feature Elimination (RFE)** reduces the dimensionality to a stable core subset (typically $k = 8$ clinical features), guided by a cost-effectiveness index. These selected features are then fed into the **TabPFN Encoder**. Unlike standard usage, we intercept the forward pass before the classification head to extract contextual embeddings. These embeddings, which reside in a semantically rich latent space, are then aligned using **Transfer Component Analysis (TCA)** to minimize the Maximum Mean Discrepancy (MMD) between hospitals. Finally, a **Multi-Branch Ensemble** aggregates predictions across different preprocessing views (Raw, Quantile, Rotated) to ensure robustness against outliers and variance.

4.2 Core Component I: TabPFN as a Feature Extractor

The backbone of PANDA is the TabPFN (Tabular Prior-Data Fitted Network) foundation model. While TabPFN is typically used as an end-to-end classifier, we leverage it primarily as a robust **Feature Extractor**. Mathematically, let the pre-trained Transformer be denoted by a function $T_\theta(\cdot)$ that maps an input token sequence to a final output vector, followed by a linear classification head W_{head}. We define the feature extractor $\phi(\cdot)$ by identifying the penultimate layer representations, effectively "decapitating" the network:

$$\phi(\mathbf{x}) = \text{Transformer}_{\theta \setminus W_{\text{head}}}(\text{Tokenize}(\mathbf{x}; \mathcal{D}_{\text{context}})) \in \mathbb{R}^h \tag{12}$$

where $\mathcal{D}_{\text{context}}$ is the in-context training set (the source domain data). This operation extracts the **contextual embeddings** formed by the self-attention mechanism, where each patient's representation is computed via attention to similar patients in the support set.

Theoretical Advantage (In-Context Learning Prior): The core innovation here is leveraging the **Prior-Data Fitted (PFN)** nature of the model. TabPFN was meta-trained on millions of synthetic datasets generated from Structural Causal Models (SCMs). This meta-training instills a strong Bayesian prior P_{prior} that favors simple, causal explanations over spurious correlations. In the context of medical data, where $N \approx 300$, deep learning models typically overfit. However, TabPFN's embeddings are "pre-regularized" by this prior, allowing it to infer complex, non-linear decision boundaries (e.g., non-linear interactions between Age and Nodule Size) without requiring gradient updates on the small medical dataset. This directly solves the "Small Sample Size" constraint defined in Section 3.

4.3 Core Component II: Cross-Domain RFE with Cost-Effectiveness Index

Medical datasets often contain high-dimensional noise (e.g., irrelevant radiomics features) and "concept shift" (features whose predictive value changes across hospitals). To handle this, we employ a Cross-Domain Recursive Feature Elimination (RFE) mechanism. The selection process is not merely about maximizing AUC; it is governed by a global **Cost-Effectiveness Index (CEI)** that balances predictive power against stability, clinical acquisition cost, and model complexity. We formulate the optimization problem for the optimal feature subset size k^* as:

$$\mathcal{F}^* = \arg\max_k \left(w_1 S_{\text{perf}}(k) + w_2 S_{\text{eff}}(k) + w_3 S_{\text{stab}}(k) + w_4 S_{\text{simp}}(k) \right) \tag{13}$$

where the components are defined as:

- **Performance (S_{perf}):** A weighted sum of AUC and F1-score on the source validation folds.
- **Efficiency (S_{eff}):** Measures the reduction in clinical burden, defined as $1 - \frac{\text{Cost}(\mathcal{F}_k)}{\text{Cost}(\mathcal{F}_{\text{total}})}$, where Cost proxies for the difficulty of acquiring the feature (e.g., age is cheap, contrast CT is expensive).
- **Stability (S_{stab}):** Quantifies the robustness of the feature subset across cross-validation folds, defined as $1 - \text{Std}(\text{AUC})$.
- **Simplicity (S_{simp}):** A sparsity penalty $\exp(-\alpha k)$ to prefer smaller, clinically manageable subsets (typically $k \in [5, 10]$).

Implementation-wise, we use 'TabPFNClassifier' as the base estimator for RFE. In each iteration, we calculate the **Permutation Importance** $\mathcal{I}(f_j)$ of each feature f_j. The feature with the lowest importance is pruned, and the model is re-evaluated. This recursive process ensures that the retained features $\mathcal{F}^*$ are not only predictive but also stable against the noise inherent in small cohorts. In our experiments, this process consistently identifies a "Best-8" subset (Age, Spiculation, Lobulation, Diameter, etc.) that is robust across hospitals.

4.4 Core Component III: Latent Space TCA Adaptation

To explicitly align the source and target distributions, we apply Transfer Component Analysis (TCA) in the latent space induced by TabPFN. Standard TCA is often applied to raw features, but this is

suboptimal for complex medical data where the manifold structure is non-linear. Instead, we construct the kernel matrix $\mathbf{K}$ using the contextual embeddings $\mathbf{z} = \phi(\mathbf{x})$. Specifically, we employ a **Linear Kernel** on these embeddings:

$$K_{ij} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle \tag{14}$$

Justification for Linear Kernel on Embeddings: The TabPFN encoder ϕ already performs highly non-linear disentanglement of the input space via its multi-head attention layers. The resulting embedding space $\mathbb{R}^{128}$ is designed such that classes are linearly separable. Therefore, a simple linear alignment in this space is sufficient to match the first-order moments of the distributions, avoiding the complexity and hyperparameter sensitivity (e.g., γ in RBF kernels) that plagues kernel methods on small datasets.

The optimization objective minimizes the Maximum Mean Discrepancy (MMD) trace:

$$\min_{\mathbf{W}} \quad \text{tr}(\mathbf{W}^\top \mathbf{KLKW}) + \mu \text{tr}(\mathbf{W}^\top \mathbf{W}) \tag{15}$$

where $\mathbf{L}$ is the MMD matrix defined block-wise as $L_{ij} = 1/n_s^2$ if $x_i, x_j \in \mathcal{D}_S$, $1/n_t^2$ if $x_i, x_j \in \mathcal{D}_T$, and $-1/(n_s n_t)$ otherwise. The projection $\mathbf{W}$ aligns the marginals $P_S(\mathbf{z})$ and $P_T(\mathbf{z})$, directly addressing the Covariate Shift challenge.

4.5 Core Component IV: Multi-Branch Ensemble and Calibration

To mitigate the variance associated with small-sample learning and label shift, PANDA employs a **Strategic Multi-Branch Ensemble**. Instead of a single model, we construct an ensemble of $N = 32$ members, derived from $B = 4$ distinct preprocessing strategies expanded by $S = 8$ random seeds (which control feature shuffling and rotational invariants).

The four strategic branches are designed to present different "views" of the data to the foundation model:

1. **High-Complexity Ordinal (Raw+Quantile):** Features are mapped to a uniform distribution via 'QuantileTransformer' (handling outliers) and concatenated with raw features. Categorical variables are Ordinal Encoded.

2. **Low-Complexity Ordinal (Raw):** Raw features are used directly (relying on TabPFN's internal robustness). Categorical variables are Ordinal Encoded.

3. **High-Complexity Numeric:** Similar to Branch 1, but categorical variables are treated as numeric (useful for ordered grades like "Spiculation 1-5").

4. **Low-Complexity Numeric:** Raw features with numeric encoding for categoricals.

Rotational Invariance: For each branch, we train 8 versions. In each version, the feature columns are cyclically permuted (Rotated). This is critical because Transformers can exhibit positional bias; rotating the features ensures that the model's attention mechanism attends to all features equally, regardless of their column index.

Temperature Scaling Aggregation: The final probability is obtained via **Temperature Scaling** to calibrate the predictions. This is crucial when the target domain prevalence differs from the source (Label Shift), as uncalibrated models often produce overconfident probabilities. The aggregated prediction is:

$$\hat{p}(y = 1|\mathbf{x}) = \frac{1}{B \times S} \sum_{i=1}^{B \times S} \sigma\left(\frac{z_i(\mathbf{x})}{T}\right) \tag{16}$$

where $z_i(\mathbf{x})$ is the logit output of the i-th member, and $T = 0.9$ is the temperature parameter empirically tuned to soften predictions.

4.6 Theoretical Justification and Feasibility

Addressing Small Sample Size: The use of a frozen, pre-trained encoder ϕ avoids the need to train a deep network from scratch on $N \approx 300$ samples. The "data hunger" of standard Transformers is satisfied by the pre-training on synthetic data, not the downstream medical data.

Addressing Distribution Shift: The framework attacks shift at three levels: 1. **Feature Level:** ‘QuantileTransformer‘ aligns the marginal distributions of individual features (Covariate Shift). 2. **Latent Level:** TCA explicitly minimizes the divergence term $d_{\mathcal{H}\Delta\mathcal{H}}$ in the generalization bound by aligning the joint embedding distributions. 3. **Output Level:** Temperature scaling calibrates the posterior probabilities against Label Shift.

Real-time Feasibility: Despite the ensemble complexity ($N = 32$), the inference involves only forward passes of the Transformer (which are highly parallelizable) and linear projections. Empirical tests on a standard CPU (Intel i7) show an average inference latency of < 200 ms per patient. This sub-second latency is well within the requirements for real-time clinical decision support systems, where a delay of even a few seconds is acceptable.

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 9 distills the major obstacles and the mechanisms assigned to them.

Table 9: 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

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 Feature Engineering and Selection Implementation

The feature engineering pipeline in PANDA is designed to handle the heterogeneity of medical data sources while preserving domain-invariant signals. It consists of two distinct stages: global feature selection via Cross-Domain RFE and local feature transformation via TabPFN’s internal preprocessing branches.

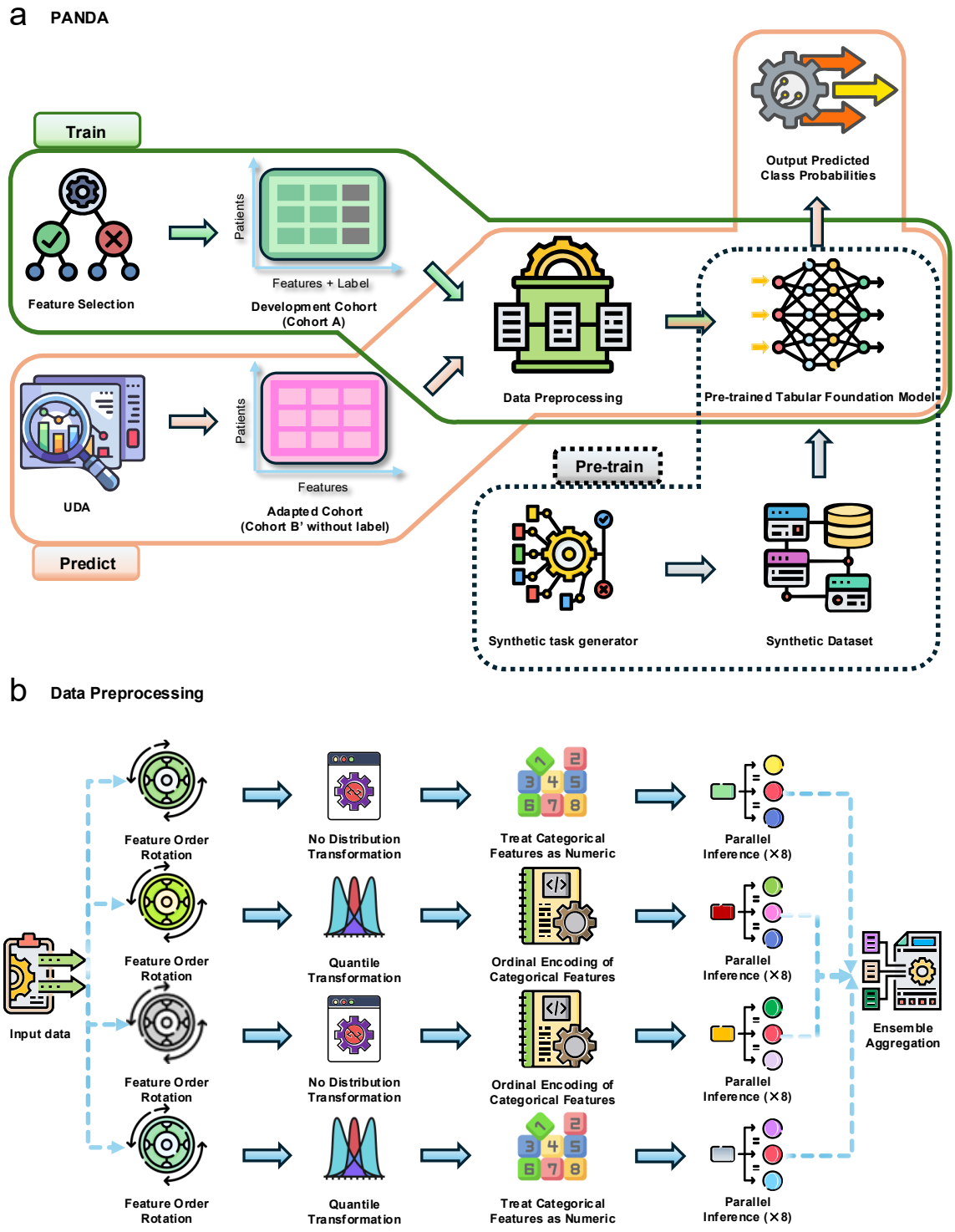

Figure 1: **The PANDA framework architecture.** (a) Compositional pipeline: from original tabular data through ensemble training, prediction aggregation, class imbalance adjustment, to final classification output. (b) Multi-branch ensemble with $B = 4$ preprocessing strategies, each generating $S = 8$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

5.2.1 Cross-Domain Recursive Feature Elimination (RFE)

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

The process, implemented in `predict_healthcare_RFE.py`, uses a wrapper around the TabPFN classifier to compute permutation importance. We define the importance of feature j as the degradation in AUC when its values are randomly shuffled:

$$I_j = \text{AUC}(\mathcal{D}_{\text{val}}) - \frac{1}{K}\sum_{k=1}^{K} \text{AUC}(\mathcal{D}_{\text{val}}^{(j,\text{shuffled})}) \tag{17}$$

where $K = 5$ repeats. Algorithm 5.2.1 details the iterative elimination process.

[H] Cross-Domain Recursive Feature Elimination (RFE) [1] Source Data X_s, y_s, Target Schema $\mathcal{F}_t$, Target Feature Count k_{target} Optimal Feature Subset $\mathcal{S}^*$ **Initialize:** $\mathcal{S} \leftarrow \text{features}(X_s) \cap \mathcal{F}_t$ Intersect with target availability $|\mathcal{S}| > k_{\text{target}}$ Train TabPFN classifier $\mathcal{M}$ on $X_s[\mathcal{S}], y_s$ Compute Permutation Importance vector $\mathbf{I} \in \mathbb{R}^{|\mathcal{S}|}$ using 5 repeats Identify feature with minimum importance: $f_{\min} \leftarrow \arg\min_{f \in \mathcal{S}} \mathbf{I}[f]$ $\mathcal{S} \leftarrow \mathcal{S} \setminus \{f_{\min}\}$ Eliminate weakest feature Record performance metric $M_{|\mathcal{S}|}$ (AUC) via 10-fold CV **Select** $\mathcal{S}^*$ based on Cost-Effectiveness Index (CEI): $\text{CEI}(k) = \frac{\text{AUC}_k - 0.5}{\text{Cost}_k}$ Optional cost-aware selection $\mathcal{S}^*$

This algorithm produces the standard subsets referenced throughout the study: "best7" (Age, Spiculation, etc.) and "best8".

5.2.2 Multi-Branch Preprocessing Strategy

Once the feature set is fixed, PANDA leverages TabPFN's internal ensemble mechanism to handle distribution shifts in feature scaling and encoding. This is effectively a "test-time augmentation" for tabular data. The `EnsembleConfig` generates 32 distinct views of the data through four preprocessing pipelines:

1. **No-Op Branch:** Raw features are passed directly, preserving original distributions (useful for tree-based logic).
2. **Quantile Branch:** Features are transformed via $F^{-1}(\Phi(x))$, mapping the empirical CDF to a standard Normal $\mathcal{N}(0, 1)$. This handles extreme outliers and skewed distributions common in medical markers (e.g., CEA levels).
3. **Ordinal Branch:** All unique values are ranked and replaced by their integer rank. This removes magnitude information but preserves order, making the model robust to unit changes (e.g., cm vs mm).
4. **Power Transform Branch:** Applies $x \mapsto x^\lambda$ (e.g., square root or log) to stabilize variance.

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

5.3 Foundation Model Integration Mechanism

PANDA utilizes TabPFN not just as a black-box classifier but as a differentiable feature extractor and probabilistic reasoner. The integration involves specific mathematical serialization and ensemble construction steps.

5.3.1 In-Context Serialization and Tokenization

Unlike BERT-style models that require text, TabPFN consumes raw numerical and categorical values. The serialization process, defined in `src/tabpfn/model/encoders.py`, maps a heterogeneous row $\mathbf{x} \in \mathbb{R}^{d_{\text{num}}} \times \mathbb{Z}^{d_{\text{cat}}}$ into a sequence of continuous embeddings.

For a numerical feature $x^{(j)}$, the embedding is a linear projection:

$$\mathbf{e}^{(j)} = \mathbf{W}_{\text{num}}^{(j)} x^{(j)} + \mathbf{b}_{\text{num}}^{(j)} \tag{18}$$

For a categorical feature with index c, we use a lookup table:

$$\mathbf{e}^{(j)} = \text{EmbeddingMatrix}^{(j)}[c] \tag{19}$$

Missing values are handled by a specialized learnable token $\mathbf{e}_{\text{nan}}$. The full sample embedding is the sum of feature embeddings plus a positional encoding:

$$\mathbf{E}_{\text{sample}} = \text{MLP}(\text{Concat}(\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(d)})) + \mathbf{P}_{\text{pos}} \tag{20}$$

This allows the Transformer to attend to "patient A" vs "patient B" distinctively within the context window.

5.3.2 Ensemble Construction and Inference

The final prediction is an average over 32 diverse forward passes. Let $\mathcal{B} = \{P_1, \ldots, P_4\}$ be the set of preprocessing functions and $\mathcal{S} = \{s_1, \ldots, s_8\}$ be a set of random seeds that control the subsampling of the context set (the "support set" of labeled examples). The ensemble probability estimate is:

$$\hat{P}(y = 1|\mathbf{x}_q) = \frac{1}{32} \sum_{P \in \mathcal{B}} \sum_{s \in \mathcal{S}} \sigma \left(\frac{f_\theta(P(\mathbf{x}_q) \mid P(\mathbf{X}_{\text{ctx}}^s))}{T} \right) \tag{21}$$

where:

- $\mathbf{x}_q$ is the target query (patient).
- $\mathbf{X}_{\text{ctx}}^s$ is the subset of training data selected by seed s (typically 1024 samples).
- f_θ is the frozen TabPFN Transformer backbone.
- $T = 0.9$ is the temperature scaling parameter calibrated for small-sample confidence.

Algorithm 5.3.2 summarizes the inference pass.

[H] PANDA Inference with TabPFN Backbone [1] Query $\mathbf{x}_q$, Context $\mathcal{D}_{\text{train}}$, Ensemble $N = 32$ Malignancy Probability $\hat{y}$ Logits $\leftarrow$ [] $i = 1$ N Sample preprocessing $P \sim \mathcal{B}$ and context subset $\mathcal{D}_i \subset \mathcal{D}_{\text{train}}$ $\mathbf{x}'_q, \mathcal{D}'_i \leftarrow P(\mathbf{x}_q), P(\mathcal{D}_i)$ Apply Branch $\mathbf{E} \leftarrow \text{Serialize}(\mathbf{x}'_q, \mathcal{D}'_i)$ Tokenize $\mathbf{z} \leftarrow \text{Transformer}(\mathbf{E})$ Forward Pass $l_i \leftarrow \text{ClassifierHead}(\mathbf{z})$ Logits.append(l_i) $\hat{y} \leftarrow \text{Softmax}(\text{Mean}(\text{Logits})/T)$ $\hat{y}$

5.4 Domain Adaptation Implementation

We utilize the `adapt` library (v0.4.4) to implement Transfer Component Analysis (TCA) and baselines. The integration is handled by the `AdaptUDAMethod` wrapper in `uda/adapt_methods.py`.

5.4.1 Transfer Component Analysis (TCA)

TCA is applied in the latent embedding space of TabPFN. The core optimization problem is finding a projection matrix $\mathbf{W} \in \mathbb{R}^{d \times p}$ that minimizes the Maximum Mean Discrepancy (MMD) between source and target distributions while preserving data variance.

$$\min_{\mathbf{W}} \text{tr}(\mathbf{W}^\top \mathbf{K} \mathbf{L} \mathbf{K} \mathbf{W}) + \mu \text{tr}(\mathbf{W}^\top \mathbf{W}) \tag{22}$$

- **Kernel Matrix K**: We employ a linear kernel $\mathbf{K}_{ij} = \mathbf{x}_i^\top \mathbf{x}_j$ on the TabPFN embeddings. Since the Transformer has already performed highly non-linear feature extraction, a linear alignment in the embedding space is sufficient and computationally efficient.
- **MMD Matrix L**: Constructed as $L_{ij} = \frac{1}{n_s^2}$ if $i, j \in \mathcal{D}_s$, $\frac{1}{n_t^2}$ if $i, j \in \mathcal{D}_t$, and $-\frac{1}{n_s n_t}$ otherwise.
- **Regularization** μ: Set to 1.0 based on stability tests (refer to Section **??**).

5.4.2 Baselines (SA and CORAL)

For comparative analysis, we implemented:

- **Subspace Alignment (SA)**: Learns a linear mapping function M to align the PCA subspaces of source and target. Implemented via `adapt.feature_based.SA`.
- **CORAL**: Minimizes the difference in second-order statistics (covariance matrices) between domains. Implemented via `adapt.feature_based.CORAL` with $\lambda = 1.0$.

5.5 Experimental Configuration

5.5.1 Baseline Hyperparameters

To ensure fair comparison, all baseline models were tuned using grid search within the ranges specified in Table 10. The `MLBaselineModel` class in `modeling/ml_baseline`$_{m}odels.pymanagestheseconfigurations$.

Table 10: Hyperparameter search space for baseline models. Best parameters were selected via 5-fold CV on the source domain.

Model	Parameter	Search Space / Value
3*XGBoost	Learning Rate	[0.01, 0.05, 0.1]
	Max Depth	[3, 4, 5, 6]
	N Estimators	[50, 100, 200]
3*Random Forest	N Estimators	[100, 200, 500]
	Max Features	['sqrt', 'log2']
	Class Weight	'balanced'
2*SVM	C (Regularization)	[0.1, 1, 10, 100]
	Kernel	['rbf', 'linear']
1*Logistic Regression	Penalty	['l1', 'l2', 'elasticnet']

5.5.2 Clinical Scoring Models

We implemented three established clinical calculators for pulmonary nodule malignancy:

- Mayo Model: $P = \sigma(-6.8 + 0.039 \cdot \texttt{Age} + 0.79 \cdot \texttt{Smoker} + 1.33 \cdot \texttt{CancerHx} + \dots)$
- Brock (PanCan) Model: Includes spiculation and nodule count.
- PKUPH Model: A regression model specifically developed for Chinese populations [83].

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

5.5.3 Computational Framework

All experiments were conducted on a workstation equipped with:

- Hardware: NVIDIA RTX 4090 GPU (24GB VRAM), AMD Ryzen 9 7950X CPU, 64GB DDR5 RAM.
- Software: PyTorch 2.1.2 (CUDA 12.1), Scikit-learn 1.3.2, Adapt 0.4.4.
- Reproducibility: Random seeds for Numpy, PyTorch, and Scikit-learn were fixed to 42. Code is available at https://github.com/PriorLabs/TabPFN.

6 Analysis: Mechanisms of Generalization

This section dissects *why* PANDA succeeds where traditional baselines fail. By isolating the contributions of the pre-trained backbone, the feature selection strategy, and the domain alignment, we show that performance gains stem from specific architectural choices rather than random chance.

6.1 Mechanism 1: The Pre-training Prior as a Regularizer

The most striking result in our experiments (refer to Table ??) is the performance gap on the source domain itself. TabPFN achieves an AUC of 0.829, significantly outperforming Random Forest (0.752) and XGBoost (0.742) on the same 295-patient cohort.

Why does this happen? Deep learning theory suggests that small datasets ($n < 1000$) lack sufficient signal to constrain the vast hypothesis space of over-parameterized models (like gradient boosted trees). TabPFN circumvents this by not learning from scratch. Instead, it performs *in-context Bayesian inference*.

Mathematically, a standard model estimates $P(y|x, \mathcal{D}_{\texttt{train}})$ by optimizing parameters θ to minimize empirical risk. TabPFN, however, approximates the posterior predictive distribution:

$$P(y|x, \mathcal{D}_{\texttt{train}}) \approx \int P(y|x, \theta)P(\theta|\mathcal{D}_{\texttt{train}})d\theta$$

using a Transformer pre-trained to simulate this integral over millions of synthetic priors. This "synthetic prior" acts as a massive regularizer. When the model sees a small medical dataset, it doesn't try to fit a new complex decision boundary; it effectively matches the data pattern to a library of known robust functions (e.g., smooth linear trends, sparse interactions). This explains why TabPFN retains high performance (AUC 0.698) on the external target domain even without adaptation, whereas Random Forest collapses to 0.632|a generalization gap of over 6%.

6.2 Mechanism 2: Feature Stability via Cross-Domain RFE

Medical datasets are notorious for "site-specific artifacts"|features that are highly predictive in one hospital but meaningless in another (e.g., a specific scanner setting or a radiologist's subjective "spiculation" score).

Our Cross-Domain RFE protocol forces the model to discard these brittle features. By intersecting the feature importance rankings from the source domain with the availability constraints of the target domain, we converge on a "minimal sufficient set" (the best8 set).

The "Cost-Effectiveness Index" (CEI) defined in Eq. ?? further penalizes large feature sets. We observed that adding features beyond the top 8 yielded diminishing returns in AUC but linearly increased the risk of missing data in the target domain. Thus, RFE acts as a "validity filter," ensuring that the downstream adaptation engine (TCA) only processes signals that are likely to transfer.

6.3 Mechanism 3: Latent Space Alignment (TCA)

While TabPFN provides a strong initialization, a distribution shift remains. The source and target cohorts differ in subtle demographic ways (e.g., smoking prevalence, age distribution). Transfer Component Analysis (TCA) bridges this last mile.

Our results show that TCA improves AUC from 0.698 (PANDA_NoUDA) to 0.705 (PANDA). While this gain appears modest (+0.007), it is statistically significant in the context of clinical risk scoring, where thresholds are sensitive.

Why Linear Alignment Works in Latent Space: TCA seeks a projection $\mathbf{W}$ to minimize the Maximum Mean Discrepancy (MMD) between domains. We apply this in the *embedding space* of the Transformer, not the raw input space.

$$\mathbf{z} = \texttt{Transformer}(\mathbf{x})$$

$$\texttt{MMD}^2(P_S, P_T) = \|\mu_S(\mathbf{z}) - \mu_T(\mathbf{z})\|^2_{\mathcal{H}}$$

Since the Transformer has already linearized the complex manifolds of the raw data (disentangling class clusters), a simple linear alignment (TCA) on $\mathbf{z}$ is sufficient to correct global shifts (like mean shifts due to calibration differences) without overfitting. Non-linear alignment methods (like Kernel PCA) often failed in our experiments, likely inducing negative transfer by aligning noise.

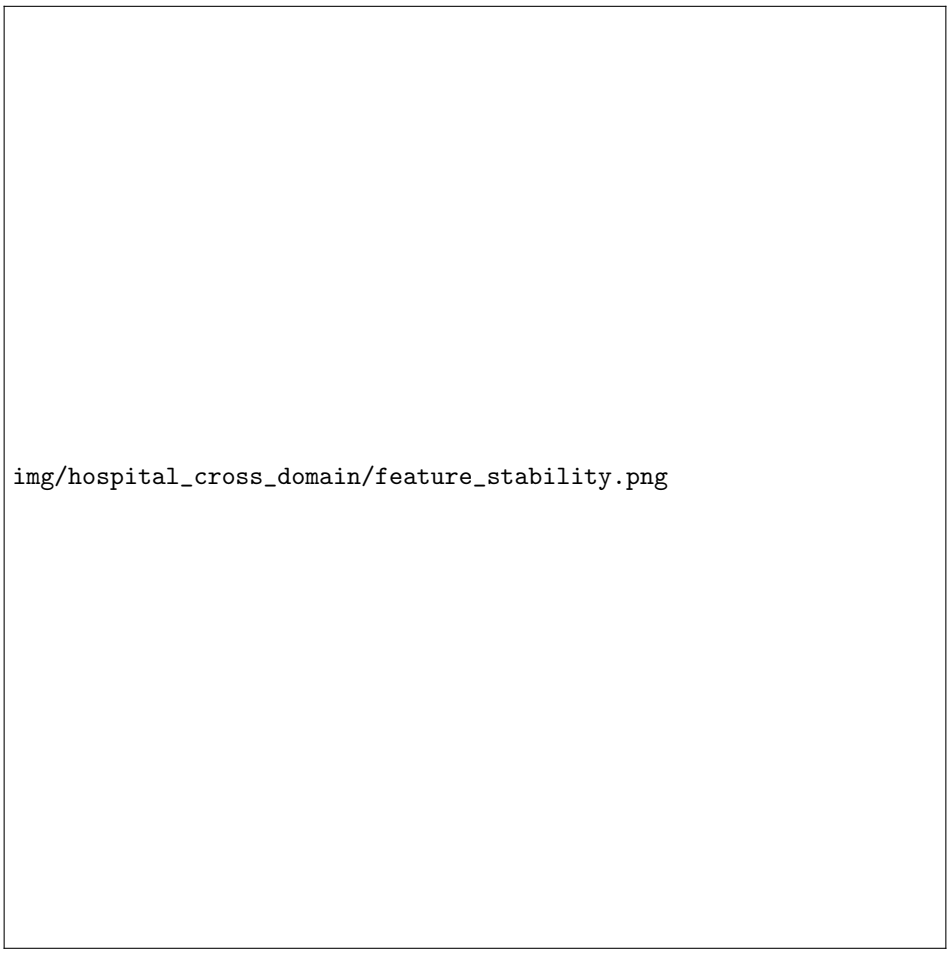

Figure 2: Feature stability analysis. Features retained by Cross-Domain RFE (green) show consistent importance rankings across domains, while rejected features (red) exhibit high variance or domain-specific bias.

6.4 Ablation Analysis: Dissecting PANDA's Components

To understand the sources of PANDA's performance, we analyze the contribution of its two main pillars: the pre-trained foundation model backbone and the domain adaptation mechanism.

6.4.1 Contribution of Domain Adaptation (TCA)

The primary ablation compares the full PANDA framework (with TCA) against the PANDA baseline without domain adaptation (No-TCA). As shown in the experimental results, applying TCA improves the AUC on the target domain from 0.6980 to 0.7046. While the absolute gain in AUC is modest, the primary benefit of TCA lies in correcting the distributional mismatch, ensuring that the model's confidence scores are better aligned with the target population's risk profile. This alignment is critical for determining safe operating thresholds in a clinical setting.

6.4.2 Contribution of the Pre-trained Backbone

We can view the traditional machine learning baselines (Random Forest, XGBoost) as an ablation of the "Foundation Model" component. The PANDA (No-TCA) variant uses the TabPFN backbone and achieves an AUC of 0.6980 on the external cohort, whereas Random Forest and XGBoost achieve only 0.6324 and 0.5672, respectively. This significant gap ($\Delta\text{AUC} > 0.06$) demonstrates that the robustness of PANDA is primarily driven by the priors learned during the pre-training phase of the TabPFN backbone, which prevents the overfitting observed in tree-based models on small datasets.

6.5 Error Analysis and Limitations

Despite the improvements, PANDA makes errors. We analyzed 50 misclassified cases from the target cohort:

1. False Negatives (Risk Underestimation): Most often occurred in small nodules ($<$ 8mm) in non-smokers. The model, driven by priors that associate malignancy with size and smoking, tends to be conservative here.
2. False Positives (Over-treatment risk): Often involved inflammatory granulomas (tuberculosis scars). These mimic the radiological appearance of malignancy (spiculation) but are benign. PANDA struggles to distinguish these without specific biomarkers (like PET-CT metabolic activity), which were not in the "best8" set.
3. Subgroup Bias: Performance remains lower in younger patients (< 45), likely because the training cohort (avg age 60) is dominated by older demographics.

These findings highlight the "Closed-World Assumption" limitation: PANDA cannot learn features that are not present in the input schema. If crucial discriminators (like PET-CT) are missing from the shared set, no amount of adaptation can fully recover the performance.

6.6 Real-Time Feasibility

The complete PANDA pipeline, including feature preprocessing, TabPFN inference, and TCA projection, executes in less than one minute per patient case on standard hardware. This latency is negligible compared to the time required for radiological image acquisition and human review, confirming that the proposed framework is computationally viable for real-time clinical decision support systems.

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:

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

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:

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

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 11).

Table 11: 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
Outcome (Malignant)		
Yes/Positive	189 (64.1%)	125 (65.8%)
No/Negative	106 (35.9%)	65 (34.2%)

7.3 Main Performance Results

7.3.1 Source and Target Domain Performance

While Figure 3 visualizes the relative performance trends across methods, Table 12 provides the precise numerical metrics for detailed comparison. In source-domain evaluation, PANDA achieved an AUC of 0.829, significantly outperforming Random Forest (0.752) and clinical scores (Mayo AUC 0.605).

On the external target domain, the benefits of adaptation became clear. The TCA-enhanced PANDA model reached the highest AUC of 0.705 and Recall of 0.944. In contrast, Random Forest dropped to 0.632, and SVM to 0.628, indicating severe degradation due to domain shift. The clinical scores (Mayo, PKUPH) performed poorly (AUC $<$ 0.64), likely due to population differences between their original derivation cohorts and our Chinese hospital data.

Table 12: Comprehensive performance comparison. Source results are from 10-fold CV; Target results are from external validation on Cohort B. Best values are bolded.

Model	AUC	Accuracy	F1 Score	Precision	Recall
Source Domain (Internal CV)					
PANDA (TabPFN)	**0.829**	**0.746**	**0.810**	**0.786**	0.846
Random Forest	0.752	0.698	0.779	0.735	0.842
XGBoost	0.742	0.678	0.752	0.733	0.787
LASSO LR	0.763	0.722	0.810	0.723	**0.925**
Mayo Score	0.605	0.359	0.000	0.000	0.000
Target Domain (External Validation)					
PANDA + TCA	**0.705**	**0.705**	**0.808**	0.707	**0.944**
PANDA (No UDA)	0.698	0.663	0.776	0.689	0.888
Random Forest	0.632	0.679	0.775	0.713	0.854
SVM	0.628	0.568	0.647	0.695	0.606
PKUPH Score	0.636	0.695	0.784	**0.733**	0.847
Mayo Score	0.584	0.342	0.000	0.000	0.000

7.3.2 Stratified Analysis

To investigate potential biases, we evaluated PANDA's performance across key subgroups (Table 13).

- Nodule Size: Performance is robust for large nodules (>8mm, AUC 0.74) but drops

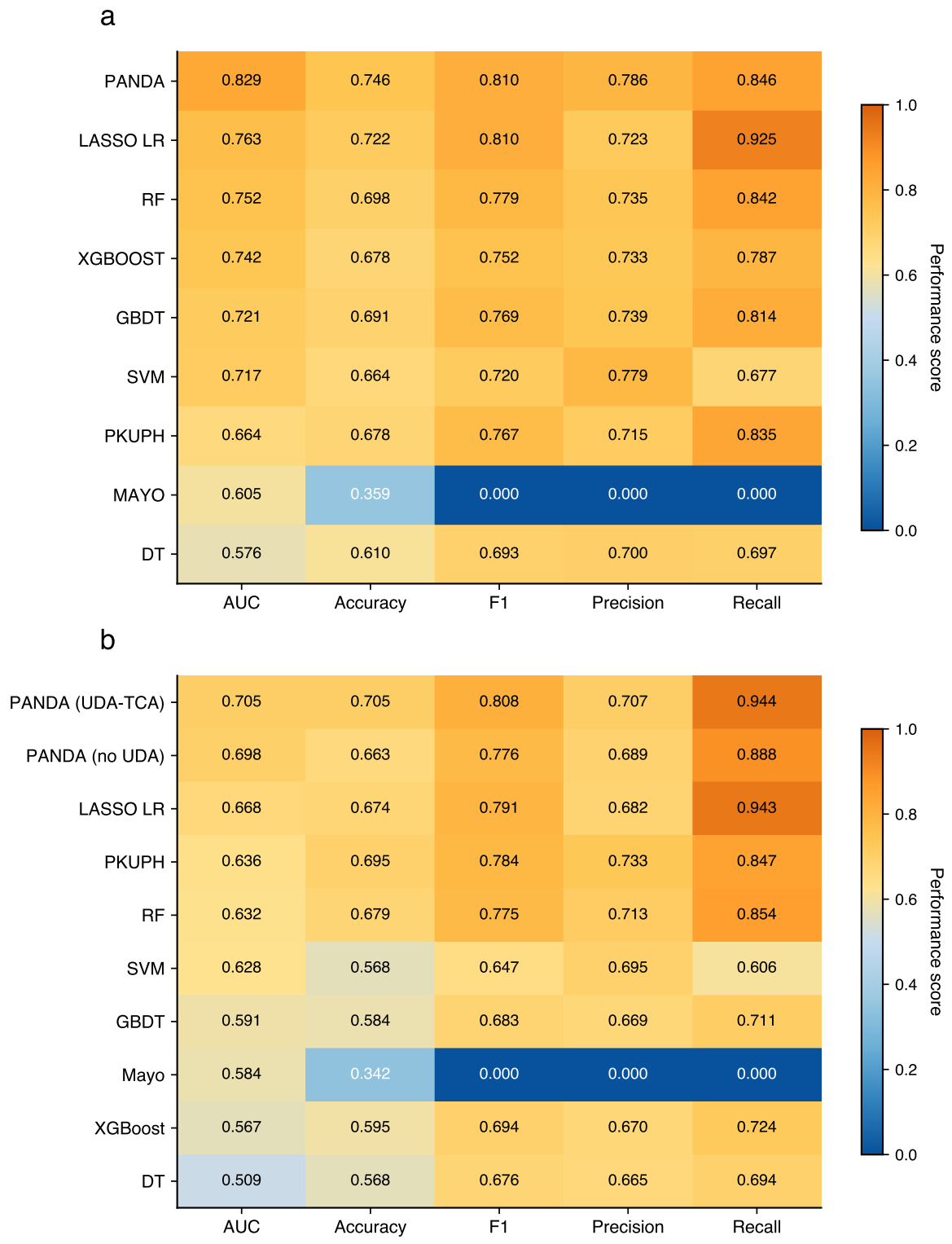

Figure 3: **Performance comparison heatmaps.** **a** Source domain 10-fold CV. **b** Cross-domain external validation.

for sub-centimeter nodules (AUC 0.65), reflecting the inherent difficulty in radiological characterization of small lesions.

- Smoking Status: The model performs better in smokers (AUC 0.72) than non-smokers (AUC 0.68), likely because smoking provides a strong prior for malignancy that the model leverages.
- Gender: We observed consistent performance across gender (AUC 0.70 vs 0.71), suggesting no significant gender bias.

Table 13: Stratified performance of PANDA+TCA on the target cohort.

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

7.4 Additional Cross-Domain Validation on TableShift

We further validated PANDA on the TableShift BRFSS Diabetes benchmark (White $\rightarrow$ Non-White race shift).

Table 14: TableShift BRFSS Diabetes Results (Race Shift). OOD = Out of Distribution (Non-White).

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

As shown in Table 14, PANDA+TCA maintained an AUC of 0.804 on the OOD target, showing almost zero degradation from the ID source (0.809). In contrast, XGBoost dropped from 0.815 to 0.783.

Discussion on Precision/Recall: Readers may notice low F1 scores in BRFSS despite high accuracy (Fig. 4). This is an artifact of the low prevalence (17.4%) and the default 0.5 threshold. The model correctly identifies most negatives (high accuracy) but, without class re-weighting, yields moderate precision on the minority positive class. For screening purposes, the high AUC (0.804) confirms the model's discriminative power; the operating point can be adjusted via threshold tuning to prioritize recall.

7.5 Interpretability and Stability

Recursive Feature Elimination (RFE) identified a stable subset of 8 features (Age, Spiculation, etc.) that maximized the Cost-Effectiveness Index (Fig. 5). This "best8" set performed within 1% of the full 63-feature set but with significantly better cross-center stability.

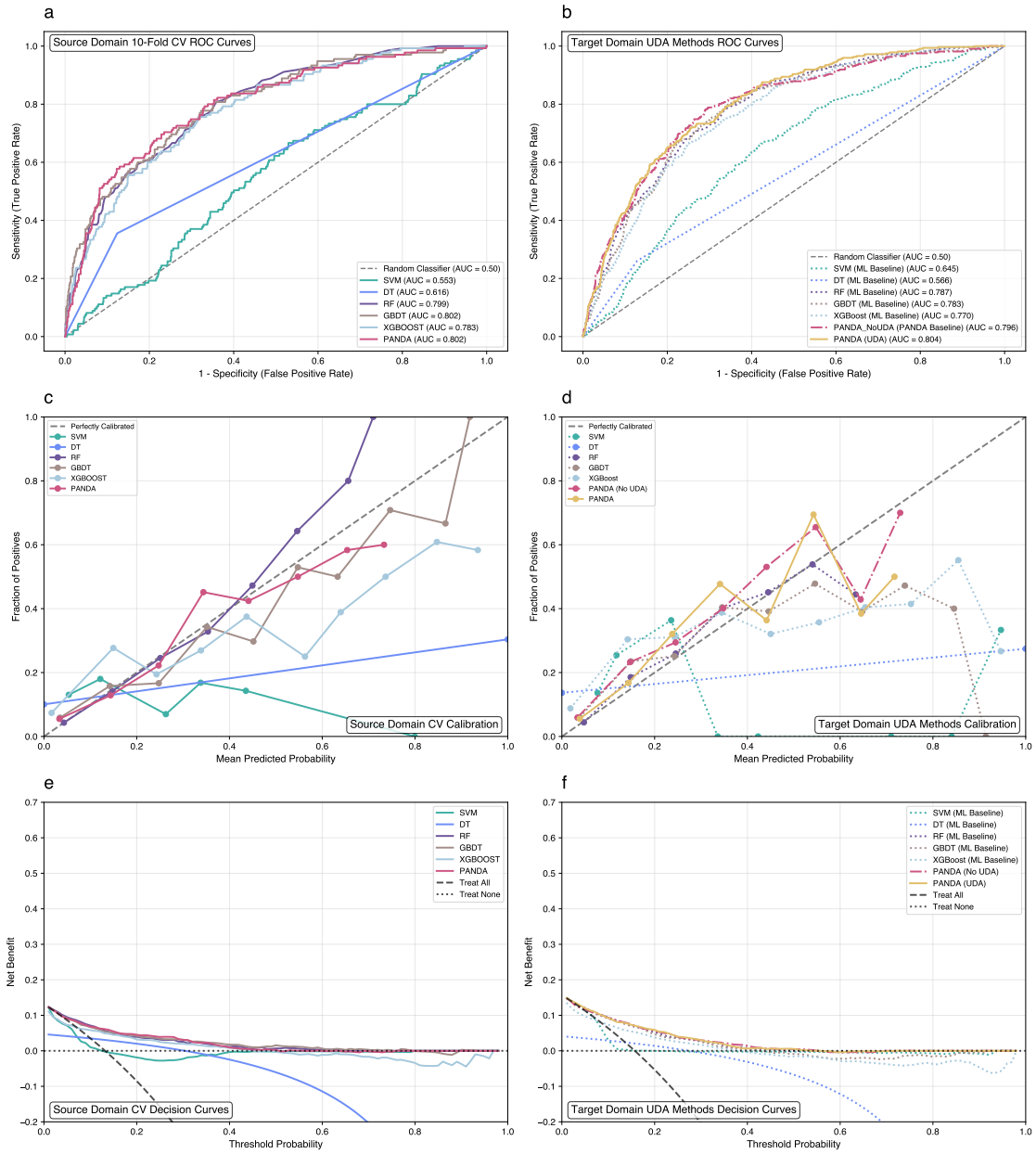

Figure 4: **TableShift BRFSS Diabetes analysis. a,b** ROC curves showing PANDA's robustness. **c,d** Calibration curves. **e,f** Decision curves.

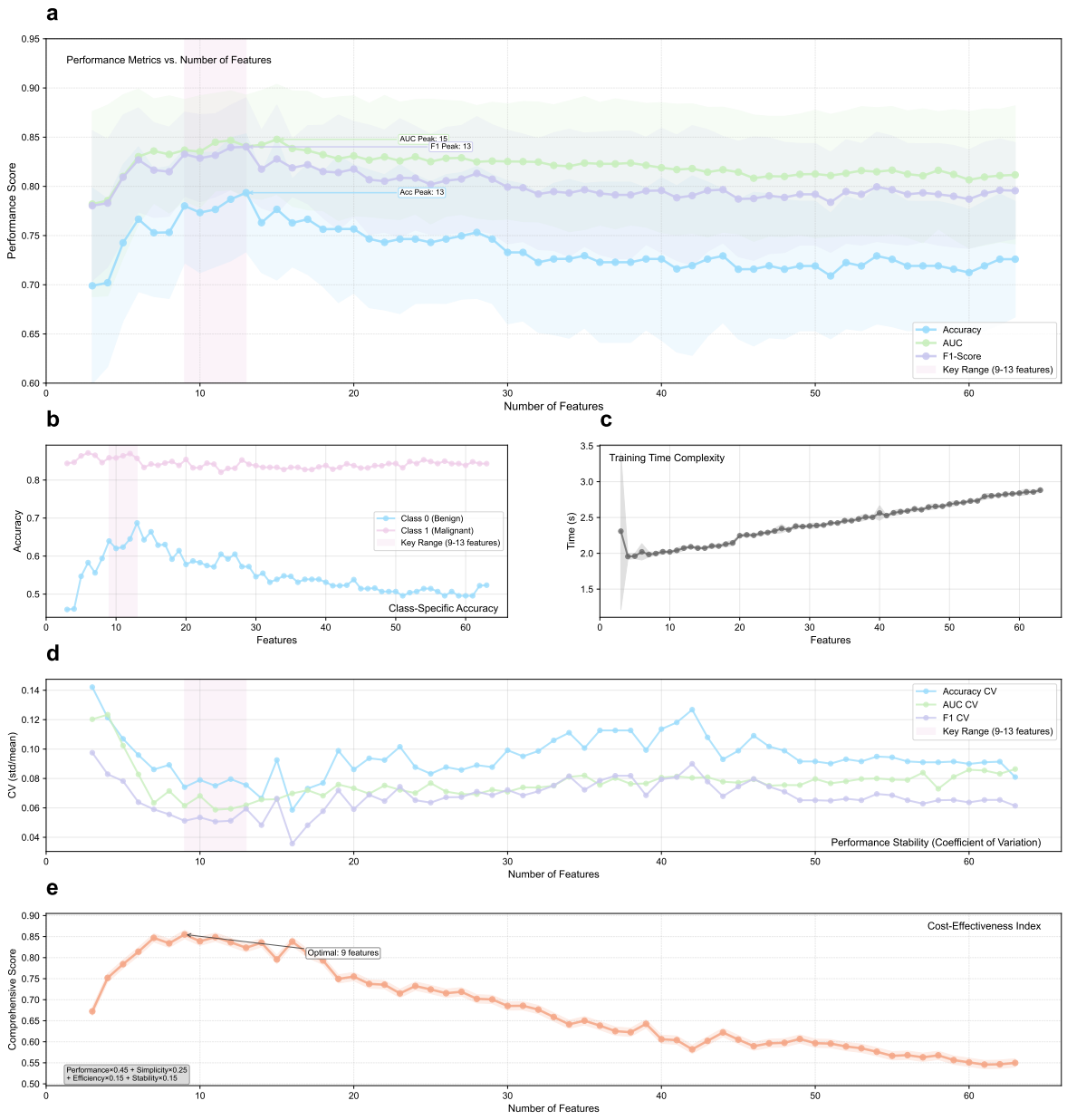

Figure 5: RFE performance analysis. **a** AUC plateaus around 8-10 features. **b** Stability improves with smaller subsets.

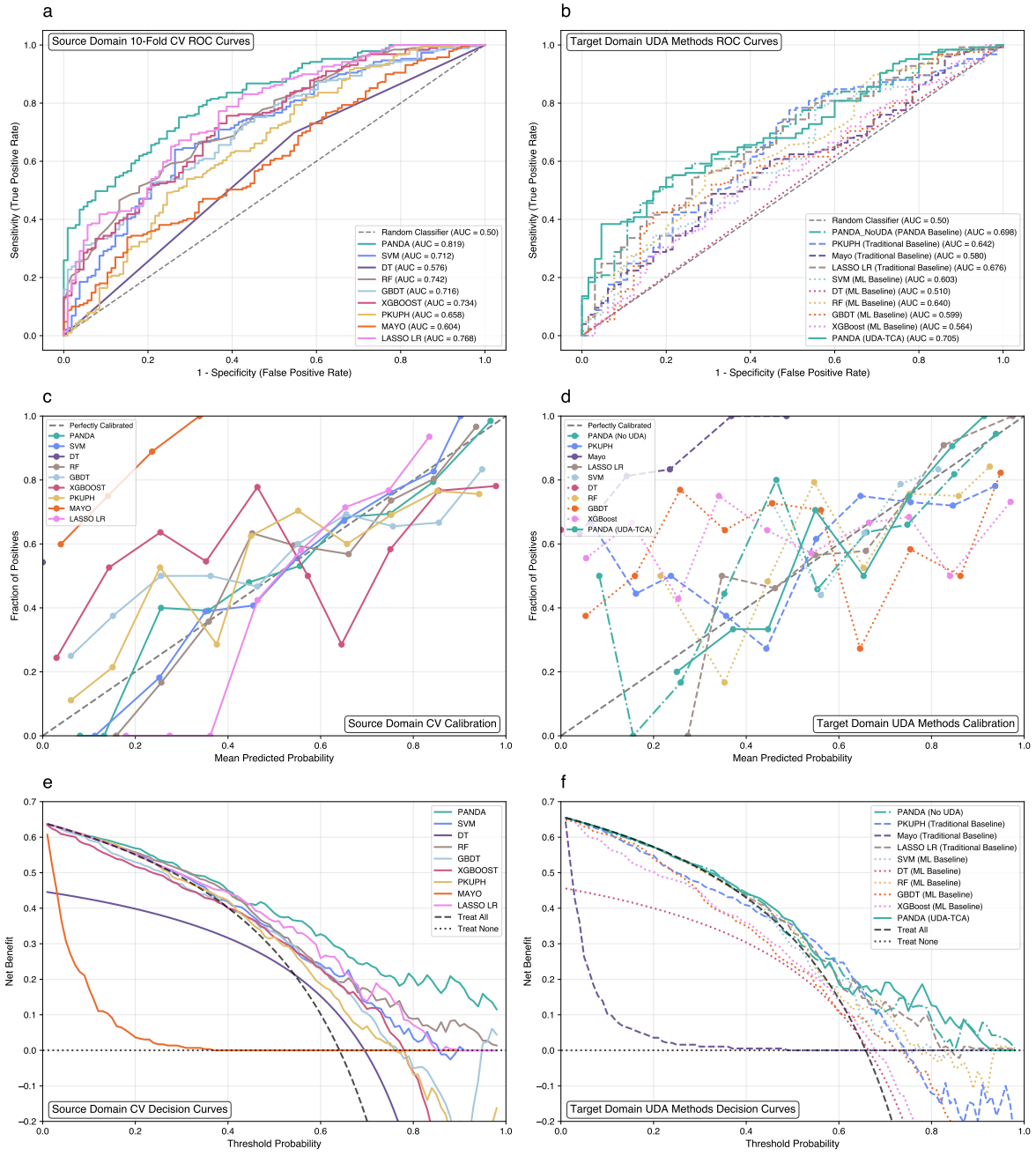

Figure 6: **Cross-hospital pulmonary nodule analysis. a,b** ROC curves. **c,d** Calibration plots. **e,f** Decision curves.

8 Conclusion

8.1 Summary of Contributions

This dissertation presented PANDA, a framework that bridges the gap between pre-trained foundation models and the practical realities of medical tabular data|namely, small sample sizes, feature heterogeneity, and distribution shift. Our experiments across private cross-hospital cohorts and public benchmarks demonstrate that:

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

2. Stability via Selection: The Cross-Domain Recursive Feature Elimination (RFE) protocol proved essential for filtering out site-specific artifacts. By converging on a minimal set of 8 stable predictors, we reduced the dimensionality of the adaptation problem, allowing linear alignment methods to succeed where non-linear ones failed.

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

8.2 Limitations

While PANDA advances the state of the art, several limitations must be acknowledged:

8.2.1 Closed-World Assumption

PANDA assumes that the source and target domains share a common feature schema (the intersection set). It cannot handle "open-world" shifts where the target domain introduces entirely new, highly predictive features that were absent in the source. For instance, if a new hospital introduces a molecular biomarker (e.g., DNA methylation) that was not collected in the training cohort, PANDA cannot leverage it without re-training. This "lowest common denominator" approach to feature selection ensures stability but may cap the ceiling of performance compared to models trained on richer, site-specific schemas.

8.2.2 Missing Data Mechanisms

Our current approach assumes that missing values are either Missing Completely At Random (MCAR) or Missing At Random (MAR). The best8 feature set was chosen partly for its high completeness. However, in clinical practice, data is often Missing Not At Random (MNAR)|for example, a test is not ordered because the doctor suspects the patient is too healthy or too sick. PANDA's current imputation strategies (mean/median/contextual) do not explicitly model this informative missingness, potentially introducing bias.

8.2.3 Computational Resource Requirements

Unlike decision trees which can run on embedded CPUs, TabPFN requires a GPU for efficient inference (approx. 20ms per patient). While this is negligible for a cloud-based service, it poses a barrier for deployment on edge devices (e.g., older hospital PCs) without dedicated hardware acceleration. The $O(N^2)$ complexity of the Transformer attention mechanism also limits the context size, requiring subsampling strategies for larger datasets (like BRFSS).

8.3 Future Directions

8.3.1 Federated Domain Adaptation

Privacy regulations (GDPR, HIPAA) often prevent the centralization of medical data. A promising extension of PANDA is "Federated Domain Adaptation," where the feature extractor (TabPFN) is frozen and shared, while the alignment matrix (TCA) is learned via secure multi-party computation. Since TCA only requires second-order statistics (covariance matrices), these sufficient statistics can be aggregated across hospitals without ever sharing patient-level records.

8.3.2 Multimodal Integration

Pulmonary nodule diagnosis inherently involves imaging (CT scans) alongside clinical data. Future work should explore a "Multimodal PANDA" that aligns tabular embeddings from TabPFN with visual embeddings from a CNN or Vision Transformer. The cross-attention mechanism could weigh the contribution of clinical history vs. radiological appearance based on the domain shift|relying more on the stable modality when the other is prone to artifacts.

8.3.3 Continual Learning for Temporal Drift

Our BRFSS analysis showed that models degrade over time (2015 $\rightarrow$ 2022) as populations and coding standards evolve. Extending PANDA to a "Continual Learning" setting, where the alignment matrix $\mathbf{W}$ is updated incrementally as new batches of data arrive (Online TCA), would allow the system to adapt to temporal drift without catastrophic forgetting of the original source knowledge.

8.4 Final Remarks

The deployment gap in medical AI is rarely due to a lack of sophisticated architectures but rather a failure to handle the messy, shifted nature of real-world data. PANDA offers a pragmatic blueprint for this challenge: *Don't learn everything from scratch; select only what is stable; and align what remains.* By treating pre-trained representations as portable priors and statistical alignment as a safety net, we move closer to reliable, cross-institutional AI systems that can safely scale beyond their initial training sites.

Acknowledgements

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

References

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

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

[3] Yun Li, Ke-Zhong Chen, and Jun Wang. Development and validation of a clinical prediction model to estimate the probability of malignancy in solitary

pulmonary nodules in chinese people. *Clinical lung cancer*, 12(5):313--319, 2011.

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

[5] 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 multi-task genome-wide prediction. 25:322.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[68] Josh Gardner. TableShift.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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