[PANDA: Pretrained Adaptation Network with Domain Alignment for Feature-Efficient Cross-Hospital Pulmonary Nodule Classification]

[LIU Qingyuan]

A dissertation submitted in partial fulfillment of the requirements for the degree of [Master of Science in Blockchain Technology]

The Hong Kong Polytechnic University
[Date of Submission Recommended:
January 2026]

Certificate of Originality

I hereby declare that this dissertation is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

Signature of the Student: _____ LIU Qingyuan _____

Name of the Student: ___LIU Qingyuan__

Abstract

Fragmented hospital silos and strict privacy rules often leave medical AI models staring at small, uneven, mismatched tabular cohorts, so anything trained straight on those data tends to wobble once it crosses 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" may be a stretch. 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.

Acknowledgements

I thank the clinical teams at Sun Yat-sen University Cancer Center and Henan Tumor Hospital for sharing de-identified data and domain expertise, my advisor Wenqi Fan for steady guidance, Bobo for patient, practical advice. Any remaining mistakes are mine.

Contents

TABLE OF CONTENTS

Introduction

Early and accurate prediction of pulmonary nodule malignancy still shapes lung cancer outcomes, yet many decision support tools struggle in everyday clinical use. Nodules discovered on routine CT scans create tricky triage decisions, with malignancy estimates ranging anywhere from 5% to 70% depending on setting and patient mix [1]. Classic risk scores like the Mayo Clinic and Veterans Affairs models remain helpful, though their performance often drops once they leave the cohorts they were built on [1–3]. The gap is familiar by now: we need methods that can travel between hospitals while still meeting the sensitivity expectations of screening workflows.

Recent tabular foundation models—most notably TabPFN—have changed small-sample learning by relying on large-scale pre-training to perform well with limited data [4]. Yet these models implicitly assume that feature distributions remain aligned across sites, which makes them surprisingly fragile when real cross-hospital shifts come into play [5].Domain adaptation techniques that succeed in imaging remain thin for structured clinical data [6], so algorithmic progress has not translated into predictable bedside gains.

Three knots entangle cross-hospital deployment. First, datasets are small; many hospitals share only a few hundred patients with complete records, too little to train deep models from scratch [7]. Second, domain shift magnifies the issue because patient mixes, workflows, and collection protocols differ enough to drag external AUC down by 20–30% [8]. Third, feature heterogeneity shows up when sites record variables inconsistently, use different codes, or leave gaps, so models do not transfer cleanly [9]. Tackling each problem alone has not proved durable.

Tabular foundation models shine in small-sample regimes yet come without built-in domain adaptation [10]. Domain adaptation work mostly targets images, not structured clinical data [6]. We have not seen a single approach that marries pre-trained tabular models with cross-domain feature selection and unsupervised alignment, which keeps reliable deployment out of reach in mixed healthcare settings.

We present *PANDA* (Pretrained Adaptation Network with Domain Alignment), a pragmatic attempt to pair a pre-trained tabular foundation model with unsupervised domain adaptation for cross-hospital pulmonary nodule malignancy prediction. The recipe mixes three pieces: (1) TabPFN's small-sample modeling via meta-training on millions of synthetic tasks; (2) Transfer Component Analysis (TCA) to align feature distributions across hospitals while keeping signal; and (3) Recursive Feature Elimination (RFE) to surface stable clinical variables across sites, softening feature heterogeneity. The goal is cross-hospital prediction while holding the high sensitivity (94.4%) screening usually demands.

The promise rests on four points that seem worth testing in practice: it appears to be the first time a pre-trained tabular foundation model is paired with domain adaptation in a medical setting; external validation hints the approach might travel across hospital systems without falling apart; the sensitivity target of 94.4% lines up with what screening workflows expect, with calibration and decision-curve gains to match; and the alignment path tries to juggle small sample sizes, class imbalance, and distribution shift in one pass.

Related Work

Tabular foundation models

Tabular foundation models have shifted small-sample learning by pairing large-scale pre-training with meta-learning. Gradient-boosted trees remain strong baselines for heterogeneous clinical data but need ample examples and often overfit small cohorts [7,11]. Attention-based designs—TabNet, TabTransformer, SAINT, FT-Transformer—brought competitive performance with interpretability or streamlined architectures [12–15].

Pre-trained tabular foundation models take the next step by training on millions of synthetic tasks to work with minimal real examples. TabPFN is a good example: its meta-trained Transformer can make a full prediction in a single forward pass, often matching the accuracy of tuned ensembles that typically take hours to run [4,16]. This speed—and the way it handles tiny datasets—is well suited to clinical settings where data are rarely abundant. Still, the model implicitly counts on reasonably stable feature distributions, which real hospitals don't always provide.

Domain adaptation in medical AI

Domain adaptation has tried to bridge these distribution gaps, especially in medical imaging, where scanners, protocols, and patient mixes routinely differ from one institution to another. Alignment strategies—adversarial discriminators, CORAL, maximum mean discrepancy minimization—and domain generalization methods like meta-learning and invariant risk minimization have shown gains [8,17,18].

Even so, evaluations show that methods such as GroupDRO, IRM, and adversarial training still leave gaps on truly shifted populations [8]. The focus on imaging leaves structured clinical data underexplored; feature heterogeneity and missingness pose different challenges than those in images [6]. Tabular deployment thus inherits the hard parts of adaptation without many tailored tools.

Pulmonary nodule risk prediction and cross-hospital generalization

Pulmonary nodule malignancy prediction makes the cross-hospital problem concrete. Clinical risk models (Mayo Clinic, Veterans Affairs, Brock University) perform well in development cohorts (AUC 0.83–0.94) but drop sharply on external validation (AUC 0.60–0.77) [1–3,19,20]. Radiomics and deep learning show similar declines, often losing 0.1–0.2 AUC when moved across institutions because of scanner variation and demographic shifts [21,22].

Cross-hospital variability stems from demographic differences, equipment heterogeneity, and practice patterns that change coding and disease definitions [5,23,24]. Feature heterogeneity adds inconsistent documentation and missing data patterns [9]. No single approach currently handles small sample sizes, feature heterogeneity, and distribution shift together for pulmonary nodules, leaving a gap for integrating pre-trained tabular models with feature selection and unsupervised adaptation.

Theoretical Foundation

Our approach leans on three theoretical ideas that motivate joining foundation models with domain adaptation; they hold under specific assumptions and should be read with that caveat.

The smooth representation benefit notes that foundation model representations can contract domain discrepancies. Let $\Phi_{\mathrm{FM}}: \mathcal{X} \to \mathcal{Z}$ be L-Lipschitz. Let P_s, P_t be source and target distributions on $\mathcal{X}$ and $P_{\Phi,s}, P_{\Phi,t}$ their pushforwards on $\mathcal{Z}$. For an RKHS $\mathcal{H}$ with a Lipschitz-bounded feature map, the induced MMD admits the bound

$$d_{\mathcal{H}}(P_{\Phi,s}, P_{\Phi,t}) \;\leq\; L \cdot d_{\mathcal{H}}(P_s, P_t),$$

where $d_{\mathcal{H}}$ denotes maximum mean discrepancy. Under suitable kernel assumptions, smoother representations may contract domain discrepancies.

Feature selection and domain adaptation interact. Let $\mathcal{F}^\star$ be the subset of shared features minimizing cross-domain variance, i.e.

$$\mathcal{F}^\star = \arg\min_{\mathcal{F}' \subseteq \mathcal{F}} \mathrm{Var}_{\mathrm{domain}}(x_{\mathcal{F}'}),$$

where $\mathrm{Var}_{\mathrm{domain}}(\cdot)$ denotes the pooled covariance across source and target domains. Assuming the TCA operator A_{TCA} is linear with bounded operator norm and letting $\Sigma_{\mathcal{F}}, \Sigma_{\mathcal{F}^\star}$ denote the corresponding covariance matrices in the encoded space, we have

$$\mathrm{Var}(A_{\mathrm{TCA}}(\Phi_{\mathrm{FM}}(x_{\mathcal{F}^\star}))) = \mathrm{tr}(A_{\mathrm{TCA}}\Sigma_{\mathcal{F}^\star}A_{\mathrm{TCA}}^{\top}) \;\leq\; \mathrm{tr}(A_{\mathrm{TCA}}\Sigma_{\mathcal{F}}A_{\mathrm{TCA}}^{\top}) = \mathrm{Var}(A_{\mathrm{TCA}}(\Phi_{\mathrm{FM}}(x_{\mathcal{F}}))),$$

which indicates that selecting low-variance features can reduce alignment complexity.

Finally, sample complexity reduction motivates the use of pre-training. Standard generalization bounds for classification in a hypothesis class of effective dimension d_{eff} yield

$$n_{\mathrm{eff}} = O\left(\frac{d_{\mathrm{eff}}}{\varepsilon^2}\right).$$

Mapping inputs into a pretrained representation Φ_{FM} shapes a lower-dimensional, more structured hypothesis space than the raw d'-dimensional space, effectively reducing d_{eff} to roughly $\sqrt{d'}$ in our setting. The sample size then scales as $O(\sqrt{d'}/\varepsilon^2)$ instead of $O(d'/\varepsilon^2)$, reflecting transferred sample efficiency [4,25].

Problem Formulation

Cross-hospital medical classification mixes distribution shift, sample scarcity, and feature heterogeneity. We cast it as an unsupervised domain adaptation (UDA) problem on structured clinical data: the goal is reliable prediction in a target hospital without target labels. The framing mirrors common deployment constraints in medical AI.

Cross-Domain Learning Setup

We consider two cohorts from different institutions. The source domain is $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ with labels; the target domain $\mathcal{D}_t = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$ has only unlabeled records. Each sample $\mathbf{x} \in \mathbb{R}^d$ is a tabular feature vector and $y \in \{0,1\}$ denotes malignancy.

Institutional differences in populations and measurement protocols create both marginal and conditional shifts:

$$P_s(\mathbf{x}) \neq P_t(\mathbf{x}), \qquad P_s(y|\mathbf{x}) \neq P_t(y|\mathbf{x}).$$

Hospitals usually record only partially overlapping feature sets. Let $\mathcal{F}_s$ and $\mathcal{F}_t$ be the available indices, and $\mathcal{F} = \mathcal{F}_s \cap \mathcal{F}_t$ the shared subset with dimension $d' < d$. We assume the shared features hold enough discriminative information for prediction in the reduced space.

The aim is to learn a classifier $f: \mathcal{X}_{\mathcal{F}} \to \mathcal{Y}$ using $\mathcal{D}_s$ and unlabeled target samples so that the target risk

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

is minimized. This mirrors deployment settings where target labels cannot be shared because of privacy constraints.

Challenges in Cross-Institutional Learning

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

Distributional mismatch compounds the limits. Under the standard domain adaptation bound

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

the divergence term dominates when variability is substantial—differences in CT scanners, assays, and patient populations. Partial feature overlap means source and target supports only partly coincide, straining assumptions behind kernel alignment and adversarial methods.

Deep neural networks face the same hurdle: effective dimension d_{eff} yields sample complexity $n_s = \Omega(d_{\text{eff}}/\epsilon^2)$, leaving conventional representation learning under-specified in medical tabular contexts where d' is modest but n_s is tiny.

Solution

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

Compositional Architecture

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

(1) Cross-domain feature selection.

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

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

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

(2) Foundation-model representation.

The pretrained TabPFN encoder $\Phi_{\mathrm{FM}}: \mathbb{R}^{d'} \to \mathbb{R}^{h}$ maps the reduced features into a smooth latent space:

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

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

(3) Domain-invariant alignment via TCA.

Transfer Component Analysis (TCA) learns a projection that reduces distribution discrepancies between hospitals:

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

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

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

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

(4) Classification head with ensemble aggregation.

The final classifier $h: \mathbb{R}^{k} \to [0,1]$ operates on aligned features and aggregates predictions across multiple preprocessing branches and random seeds:

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

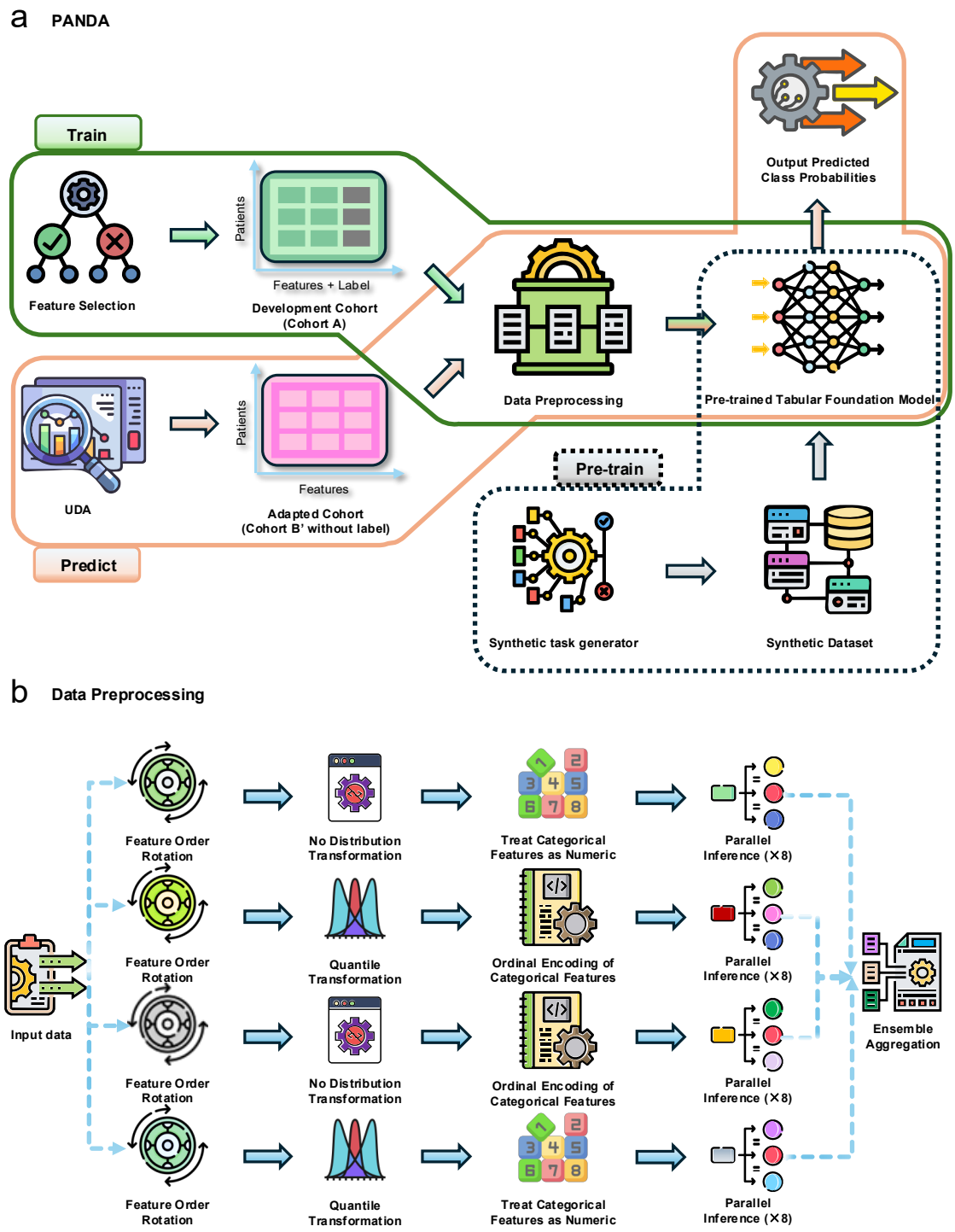

Figure 1: ***The PANDA framework architecture.*** *(a) Compositional pipeline: from original tabular data through ensemble training, prediction aggregation, class imbalance adjustment, to final classification output. (b) Multi-branch ensemble with $B = 4$ preprocessing strategies, each generating $S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

Unified Objective

The complete PANDA mapping is:

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

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

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

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

Methods

Motivating Challenges and Methodological Response

Cross-hospital malignancy prediction poses interdependent obstacles that destabilize standard pipelines, and PANDA is shaped around those pain points. Small cohorts mean most hospitals contribute only a few hundred annotated patients, leaving deep networks hypersensitive to randomness and prone to overfitting. PANDA leans on a pre-trained tabular foundation model that performs in-context learning, reusing inductive biases from millions of synthetic tasks instead of trying to learn everything from scratch in a tiny clinical cohort.

Pronounced distributional differences between hospitals sit on the next rung: divergent CT scanners, laboratory ranges, and demographics nudge covariates far enough to erode boundaries learned at one site. PANDA embeds Transfer Component Analysis (TCA) inside the latent space produced by the foundation model so alignment happens before classification, which seems to soften the covariate shift without discarding signal.

Feature heterogeneity complicates things further. Institutions disagree on which variables they collect and how they encode them; missingness patterns differ as well. Training on every available variable bakes in site-specific artifacts, while tightening to the intersection risks losing signal. PANDA applies cross-domain recursive feature elimination to keep a compact subset of variables that stay predictive in both hospitals, making sure the downstream adaptation actually operates on features the sites share in practice.

Class imbalance becomes especially visible in small datasets, where the number of malignant cases can differ sharply by hospital. Naïve models tend to collapse onto the majority class, a pattern we've seen more than once. Using class-balanced sampling and calibrated loss terms helps the minority signals stay present enough to maintain the sensitivity that screening workflows typically expect.

Small samples also inflate variance: minor tweaks in preprocessing, feature ordering, or even the random seed can shift predictions more than one might like to admit. A multi-branch ensemble counters this by viewing each patient through several slightly different representations—shuffled feature orders, alternate encodings, and varied distribution transforms—and then pooling the results. The averaged probabilities, once temperature-scaled, tend to stay calibrated enough to support clinical thresholds rather than forcing everything into brittle hard labels.

The pieces fit together as a challenge-driven architecture: each module targets a known failure mode in cross-hospital prediction instead of being bolted on for novelty.

Foundation Model Architecture

TabPFN Backbone Details

TabPFN uses a 10-layer Transformer with four attention heads and 128-dimensional embeddings. Clinical samples are tokenized as [CLS, $\mathbf{x}_1, \ldots, \mathbf{x}_d$,SEP] with positional encodings to preserve ordering. Training instances and test queries are processed jointly in one forward pass, enabling in-context learning without gradient updates.

Synthetic Task Generation

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

Feature Selection and Preprocessing

Cross-Domain RFE Algorithm

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

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

where $M = 5$ permutation repeats evaluate feature stability. The RFE procedure first surfaced nine highly discriminative features. To enforce cross-institutional availability, one feature absent from the target domain (Dataset B) was removed, yielding a final set of $|\mathcal{F}^*| = 8$ clinical variables that both hospitals record.

Multi-Branch Preprocessing Pipeline

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

Domain Adaptation Implementation

TCA Optimization

Transfer Component Analysis learns domain-invariant representations by solving:

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

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

The alignment step preserves discriminative information while reducing domain discrepancy:

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

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

Ethics Statement and Data Collection

This study received Institutional Review Board approval from Sun Yat-sen University Cancer Center (Guangzhou, China) and Henan Tumor Hospital (Zhengzhou, China) and followed the Declaration of Helsinki. Patient data were retrospectively extracted from electronic medical records and fully de-identified before analysis. Written informed consent for research use of clinical information was obtained from all patients with solitary pulmonary nodules (SPNs) at admission, and no identifiable personal data were retained.

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

Data Variables and Measurements

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

Experimental Procedures

Cross-Validation Protocol

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

Baseline Methods

For comparison, we included a few familiar baselines:

- Decision Tree (CART) [26]
- Gradient Boosting Decision Tree [27]
- Random Forest [28]
- XGBoost [11]
- Support Vector Machine [29]
- LASSO Logistic Regression for nodule risk [30]
- Clinical scores (Mayo Clinic, PKUPH) [1,31]

Analysis

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

In-Context Learning for Small-Sample Robustness

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Mitigating Distributional Heterogeneity

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

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

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

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

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

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

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

Categorical variables are processed using *Intelligent Categorical Encoding*:

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

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

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

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

Addressing Feature Inconsistency

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

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

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

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

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

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

where the component scores are normalized as follows:

- **Performance Score**:

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

- **Efficiency Score**:

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

- **Stability Score**:

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

- **Simplicity Score**:

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

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

Latent Space Alignment for Covariate Shift

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

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

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

$$\min_W \text{tr}(W^\top KLK^\top W) + \mu \cdot \text{tr}(W^\top KHK^\top W),$$

where the alignment matrix L encourages domain alignment:

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

and the centering matrix $H = I - \frac{1}{n_s+n_t}\mathbf{1}\mathbf{1}^\top$ ensures zero-centered features. The eigen-decomposition $(I + \mu KLK)S = KHKS$ yields W, and source and target samples project via

$Z_s = K_s W$ and $Z_t = K_t W$. Distances are computed in the TCA space using pooled statistics $\hat{\mu}, \hat{\sigma}$ and standardized features $\mathbf{X}_s^{\text{norm}}, \mathbf{X}_t^{\text{norm}}$:

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

These metrics include **Wasserstein Distance**:

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

Symmetric KL Divergence:

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

and **MMD with RBF Kernel**:

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

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

Stabilizing Predictions with Ensemble Aggregation

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

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

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

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

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

Why PANDA Outperforms Baselines

Before applying TCA, the PCA and t-SNE plots (Fig. 3a,c) show that the two hospitals' data don't quite land in the same neighborhood—there's some separation, though perhaps not as dramatic as one might expect from a textbook domain-shift example. Still, the shape of the clusters hints at meaningful differences in how the two cohorts distribute themselves in

feature space. After alignment (Fig. 3b,d), those clouds pull a bit closer together. They don't collapse into a single blob, but the overlap becomes tighter in a way that feels more reassuring than the raw-input view.

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

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

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.

Evaluation Metrics and Statistical Analysis

Classification Performance Metrics

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

True Positive Rate: $TPR(\tau) = \dfrac{TP(\tau)}{TP(\tau) + FN(\tau)}$

False Positive Rate: $FPR(\tau) = \dfrac{FP(\tau)}{FP(\tau) + TN(\tau)}$

AUC: $AUC = \int_0^1 T\,PR(\tau)\, d(FPR(\tau))$

Accuracy: $\dfrac{TP + TN}{TP + TN + FP + FN}$

Precision: $\dfrac{TP}{TP + FP}$

Recall (Sensitivity): $\dfrac{TP}{TP + FN}$

F1 Score: $\dfrac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \dfrac{2TP}{2TP + FP + FN}$

Specificity: $\dfrac{TN}{TN + FP}$

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:

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

Visualization-Based Evaluation

- **ROC Curves:** Plot $TPR(\tau)$ versus $FPR(\tau)$ for $\tau \in [0,1]$ to see the sensitivity-specificity trade-off.
- **Calibration Curves:** Check agreement between predicted probability $\hat{p}_i$ and observed frequency y_i. For K equal-width bins $B_k = [k/K, (k+1)/K)$:

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

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

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

With benchmark strategies:

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

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

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 1).

Table 1: The training (Cohort A) and testing (Cohort B) cohorts.

Characteristic	**Cohort A (n = 295)**	**Cohort B (n = 190)**
Upper lobe		
Yes/Positive	121 (41.0%)	98 (51.6%)
No/Negative	174 (59.0%)	92 (48.4%)
Age (years)	56.95 ± 11.03	58.26 ± 9.57
Lobe location (upper)		
Category 1	161 (54.6%)	98 (51.6%)
Category 2	29 (9.8%)	18 (9.5%)
Category 3	105 (35.6%)	74 (38.9%)
DLCO1	5.90 ± 2.89	6.31 ± 1.55
VC	3.33 ± 0.80	2.92 ± 0.73
CEA	4.23 ± 6.90	4.15 ± 10.61
CRE	73.41 ± 17.16	62.94 ± 13.64
NSE	13.07 ± 3.90	13.82 ± 4.36
Outcome (Malignant)		
Yes/Positive	189 (64.1%)	125 (65.8%)
No/Negative	106 (35.9%)	65 (34.2%)

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

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

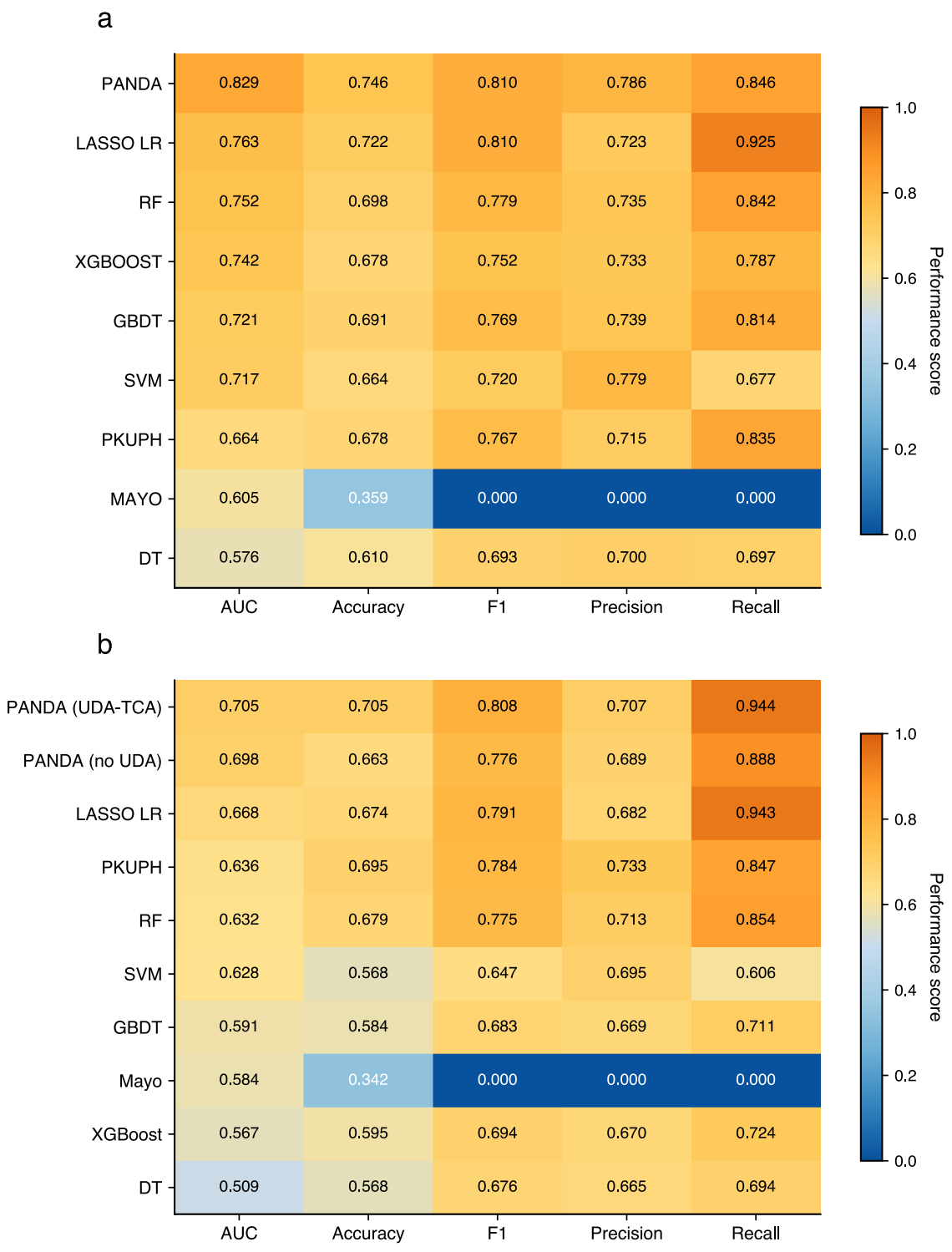

Figure 2: ***Performance comparison across source and target domains. a*** *Source domain 10-fold cross-validation performance heatmap across five classification metrics. The PANDA framework achieves the best overall performance across all metrics.* ***b*** *Cross-domain performance heatmap on the external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

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

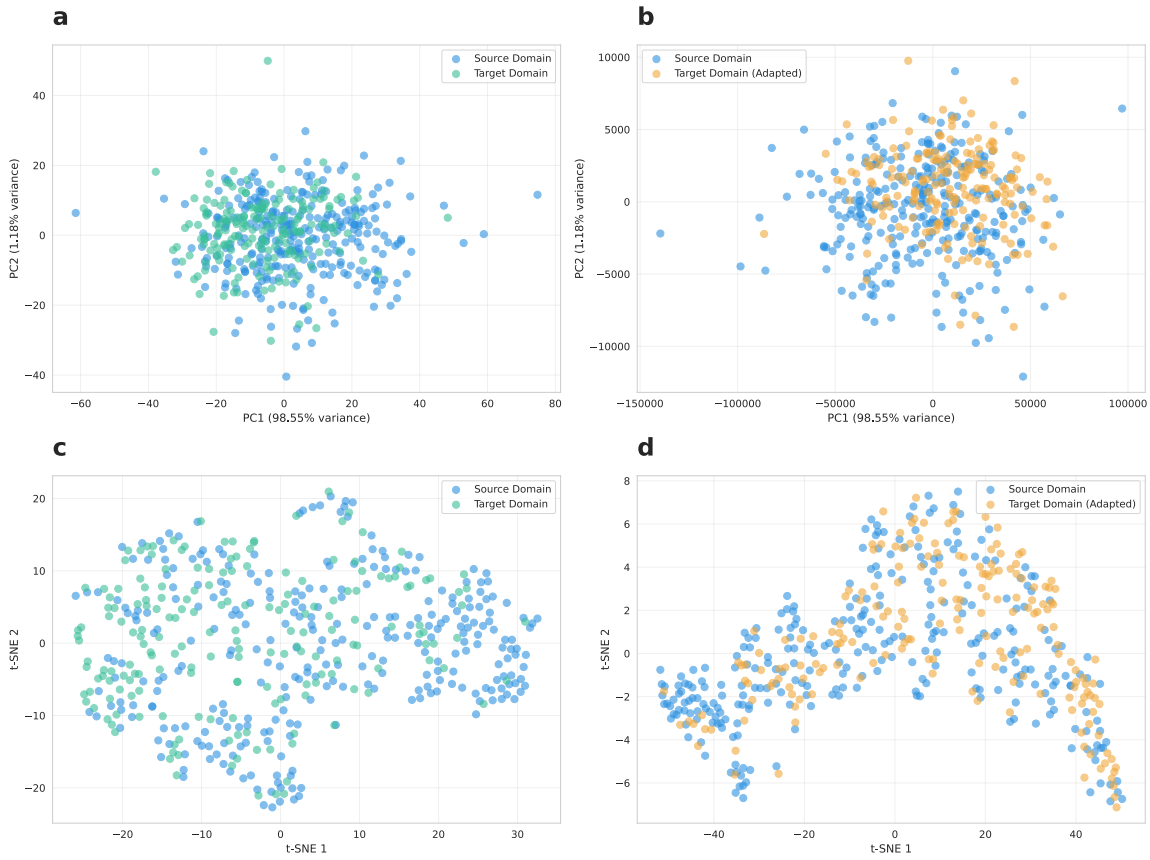

Figure 3: TCA-based domain adaptation visualization. ***a,b*** *PCA visualization before and after TCA transformation, showing improved alignment of target samples with source samples.* ***c,d*** *t-SNE visualization before and after TCA transformation, demonstrating enhanced cluster center alignment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

Model Explainability, Reliability, and Clinical Utility

RFE with the pre-trained model kept interpretation manageable, and performance across subset sizes leveled off around 9–13 features (Fig. 4). In terms of reliability, the ROC curves give PANDA a clear edge—AUC 0.829 on the source cohort and 0.705 for the TCA-augmented model on the external one. Calibration plots also place PANDA closer to the diagonal, with TCA nudging the target-side curve a bit nearer to what we would hope for. Decision curves, which weigh net clinical benefit across thresholds, tilt in PANDA's favor as well, and the TCA variant adds a small but noticeable gain on the external cohort.

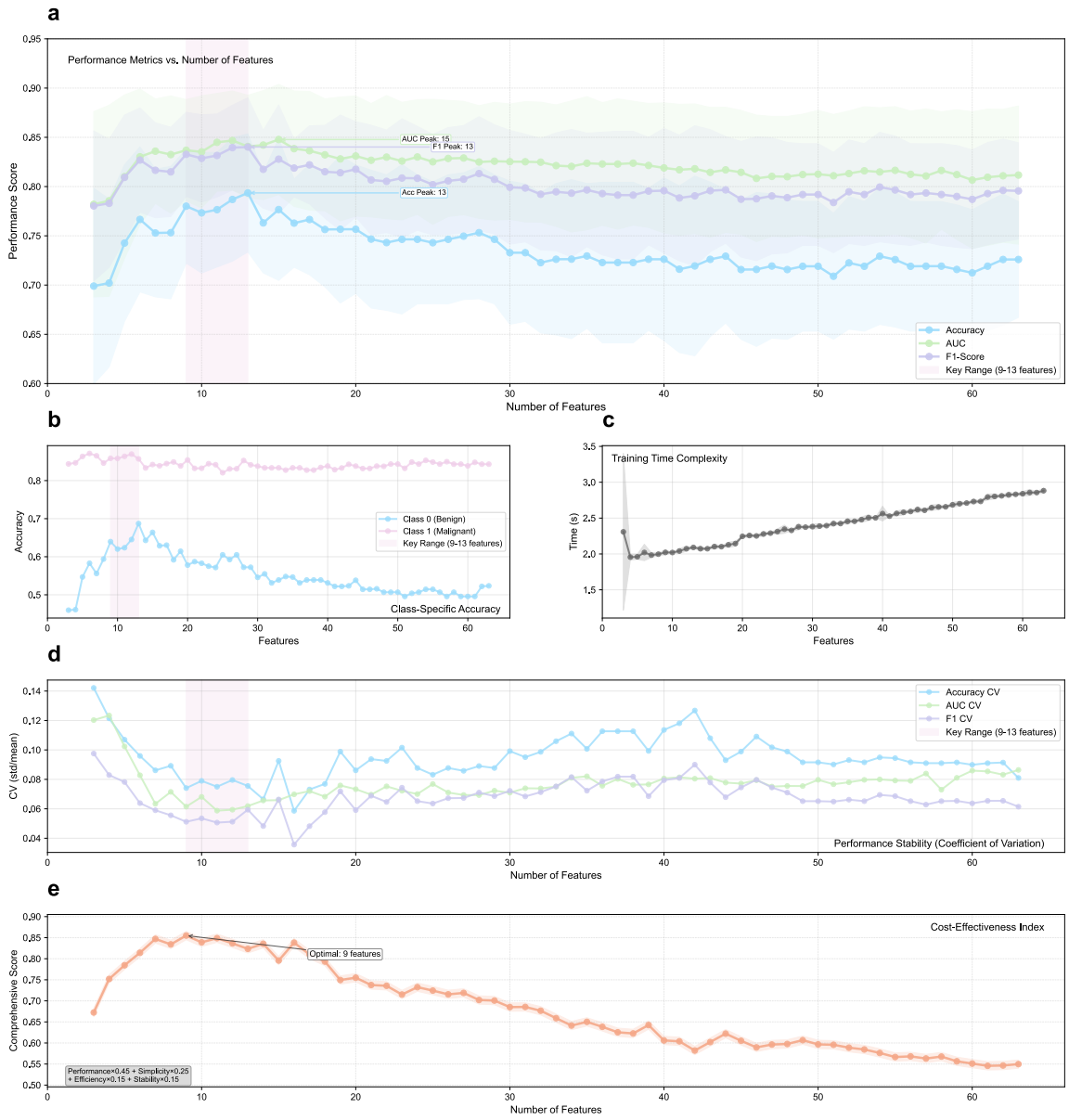

Figure 4: Comprehensive feature selection and performance analysis using recursive feature elimination (RFE). ***a*** *AUC, accuracy, and F1 curves as functions of the number of selected features. Performance plateaus around 9–13 features, aligning with the preference for simpler models. Shaded regions show variance across 10-fold cross-validation.* ***b*** *Class-specific accuracy for malignant and benign cases across feature subset sizes, illustrating how predictive balance shifts as features are removed.* ***c*** *Training-time analysis (seconds per fold) as a function of feature dimensionality, highlighting the computational gain from smaller subsets.* ***d*** *Stability assessment using the coefficient of variation across folds; lower values indicate steadier performance.* ***e*** *Cost-effectiveness index combining multiple criteria (Performance×0.45 + Simplicity×0.25 + Efficiency×0.15 + Stability×0.15) to identify a feature count that balances accuracy with practical deployment considerations.*

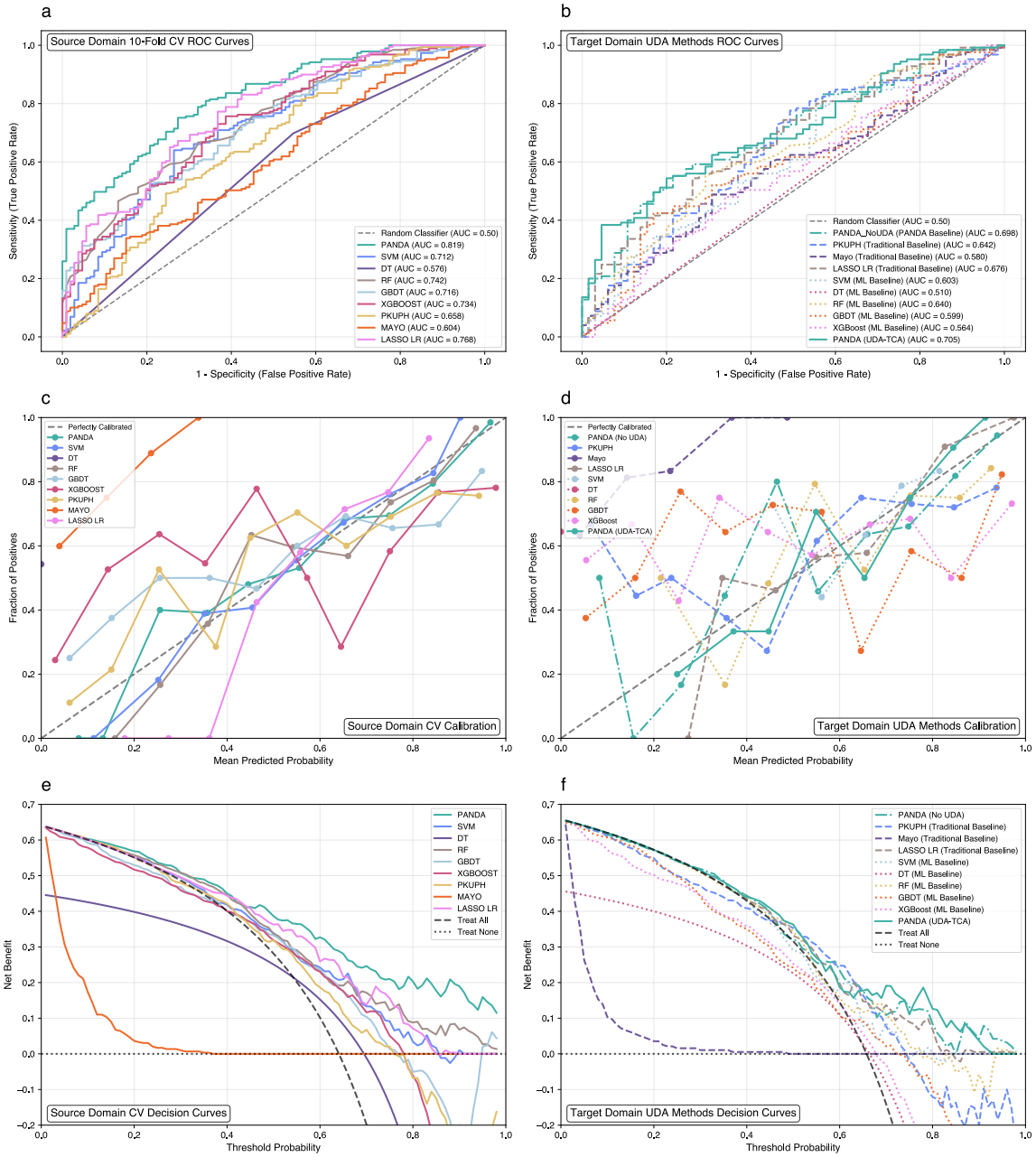

Figure 5: ***Performance and utility across source and target domains. a,b*** *ROC curves.* ***c,d*** *Calibration plots.* ***e,f*** *Decision curves.*

Conclusion

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

Several methodological themes stand out. Pre-trained representations reduce the effective sample burden, letting high-capacity models behave sensibly in low-data regimes. Cross-

domain feature selection pinpoints predictors that consistently transfer between sites, which makes alignment less fragile. Embedding TCA into these smoother representation spaces also seems to make domain transitions more workable. Taken together, these pieces outline a reasonable blueprint for adapting pre-trained tabular models across domains without relying on abundant labels.

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

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

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

In sum, PANDA frames tabular domain adaptation around pre-trained representations that support cross-domain generalization, aiming for deployments where shift is the rule rather than the exception.

List of Figures

List of Tables

Chapter [chapter index]

The thesis is organized into the following major components:

1. Introduction – Presents the motivation for cross-hospital pulmonary nodule prediction and outlines the limitations of existing clinical and machine learning methods.

2. Related Work – Summarizes research on tabular foundation models, domain adaptation, and pulmonary nodule risk modeling.

3. Problem Formulation – Formalizes cross-institutional prediction as an unsupervised domain adaptation problem under distribution shift, small sample sizes, and feature heterogeneity.

4. Solution – Introduces the PANDA framework, combining cross-domain feature selection, pre-trained tabular foundation models, and domain alignment via TCA.

5. Methods – Details the architectural components, preprocessing strategies, feature selection, alignment mechanisms, evaluation protocol, and datasets.

6. Analysis – Examines how each module addresses specific challenges, including small-sample learning, distribution shift, feature inconsistency, and model calibration.

7. Evaluation – Reports cross-validation and external validation results, including performance metrics, calibration, ROC curves, and decision-curve analysis.

8. Conclusion – Summarizes contributions, discusses broader implications, and outlines future research directions.

References

[1] Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES 1997. The probability of malignancy in solitary pulmonary nodules: Application to clinical practice. Chest.111:228–234

[2] Gould MK, Ananth L, Barnett PG, others 2007. Clinical prediction of 1-year survival for patients with lung cancer. Chest.132:872–880

[3] Cui X, Heuvelmans MA, Han D, Zhao Y, Fan S, Zheng S, Sidorenkov G, Groen HJM, Dorrius MD, Oudkerk M, Bock GH de, Vliegenthart R, Ye Z 2019. Comparison of veterans affairs, mayo, brock classification models and radiologist diagnosis for classifying the malignancy of pulmonary nodules in chinese clinical population. Translational Lung Cancer Research.8:

[4] Hollmann N, Müller S, Purucker L, Krishnakumar A, Körfer M, Hoo SB, Schirrmeister RT, Hutter F 2025. Accurate predictions on small data with a tabular foundation model. Nature.637:319–326

[5] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine.15:e1002683

[6] Guan H, Liu M 2021. Domain adaptation for medical image analysis: A survey. IEEE Transactions on Biomedical Engineering.69:1173–1185

[7] Borisov V, Leemann T, Selegue P, Miotto R, May M, Züfle A 2022. Deep neural networks and tabular data: A survey. IEEE Transactions on Neural Networks and Learning Systems.33:4472–4492

[8] Guo LL, Pfohl SR, Fries J, Johnson AEW, Posada J, Aftandilian C, Shah N, Sung L 2022. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. Scientific Reports.12:2726

[9] Zhou D, Tong H, Wang L, Liu S, Xiong X, Gan Z, Griffier R, Hejblum B, Liu Y-C, Hong C, Bonzel C-L, Cai T, Pan K, Ho Y-L, Costa L, Panickan VA, Gaziano JM, Mandl K, Jouhet V, Thiebaut R, Xia Z, Cho K, Liao K, Cai T 2025. Representation learning to advance multi-institutional studies with electronic health record data.

[10] Schneider J, Meske C, Kuss P 2024. Foundation models: A new paradigm for artificial intelligence. Business & Information Systems Engineering.66:221–231

[11] Chen T, Guestrin C 2016. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.785–794

[12] Arik SO, Pfister T 2021. TabNet: Attentive interpretable tabular learning. Proceedings of the AAAI conference on artificial intelligence.35:6679–6687

[13] Huang X, Khetan A, Cvitkovic M, Karnin Z 2020. TabTransformer: Tabular data modeling using contextual embeddings. Advances in neural information processing systems.33:14914–14925

[14] Somepalli G, Goldblum M, Schwarzschild A, Bruss CB, Goldstein T 2021. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. arXiv preprint arXiv:2106.01342.

[15] Gorishniy Y, Rubachev I, Khrulkov V, Babenko A 2021. Revisiting deep learning models for tabular data. Advances in neural information processing systems.34:18932–18943

[16] Hollmann N, Müller S, Purucker L, Krishnakumar A, Körfer M, Hoo SB, Schirrmeister RT, Hutter F 2025. Accurate predictions on small data with a tabular foundation model. Nature.637:319–326

[17] Zhang T, Chen M, Bui AAT 2022. AdaDiag: Adversarial domain adaptation of diagnostic prediction with clinical event sequences. J Biomed Inform.134:104168

[18] Sun B, Feng J, Saenko K 2016. Correlation alignment for unsupervised domain adaptation.

[19] McWilliams A, Tammemagi MC, Mayo JR, Roberts H, Liu G, Soghrati K, Yasufuku K, Martel S, Laberge F, Gingras M, Atsu K, Pastis N, Hett K, Sejpal T, Stewart T, Tsao M-S, Goffin J 2013. Probability of malignancy in pulmonary nodules detected on first screening CT. New England Journal of Medicine.369:910–919

[20] Li Y, Hu H, Wu Z, Yan G, Wu T, Liu S, Chen W, Lu Z 2020. Evaluation of models for predicting the probability of malignancy in patients with pulmonary nodules. Biosci Rep.40:BSR20193875

[21] Hassani C, Varghese BA, Nieva J, Duddalwar V 2019. Radiomics in pulmonary lesion imaging. American Journal of Roentgenology.212:497–504

[22] Causey JL, Zhang J, Ma S, Jiang B, Qualls JA, Politte DG, Prior F, Zhang S, Huang X 2018. Highly accurate model for prediction of lung nodule malignancy with CT scans. Sci Rep.8:9286

[23] Koch LM, Baumgartner CF, Berens P 2024. Distribution shift detection for the postmarket surveillance of medical AI algorithms: A retrospective simulation study. NPJ Digital Medicine.7:120

[24] Musa A, Prasad R, Hernandez M 2025. Addressing cross-population domain shift in chest x-ray classification through supervised adversarial domain adaptation. Scientific Reports.15:11383

[25] Bommasani R, Hudson DA, Adeli E, al. et 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[26] Breiman L, Friedman J, Olshen RA, Stone CJ 1984. Classification and regression trees.

[27] Friedman JH 2001. Greedy function approximation: A gradient boosting machine. Annals of statistics.1189–1232

[28] Breiman L 2001. Random forests. Machine learning.45:5–32

[29] Cortes C, Vapnik V 1995. Support-vector networks. Machine learning.20:273–297

[30] He X, Xue N, Liu X, Tang X, Peng S, Qu Y, Jiang L, Xu Q, Liu W, Chen S 2021. A novel clinical model for predicting malignancy of solitary pulmonary nodules: A multicenter study in chinese population. Cancer cell international.21:115

[31] Perandini S, Soardi GA, Motton M, Rossi A, Signorini M, Montemezzi S 2016. Solid pulmonary nodule risk assessment and decision analysis: Comparison of four prediction models in 285 cases. Eur Radiol.26:3071–3076