Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Retrieval-aligned Tabular Foundation Models Enable Robust Clinical Risk Prediction in Electronic Health Records Under Real-world Constraints

Minh-Khoi Pham

minhkhoi.pham@adaptcentre.ie

Dublin City University

Thang-Long Nguyen Ho

Dublin City University

Thao Thi Phuong Dao

University of Science

Tai Tan Mai

Dublin City University

Minh-Triet Tran

University of Science

Marie Elizabeth Ward

St. James's Hospital

Una Geary

St. James's Hospital

Rob Brennan

University College Dublin

Nick McDonald

Trinity College Dublin

Martin Crane

Dublin City University

Marija Bezbradica

Dublin City University

Article

Keywords:

Posted Date: March 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9085469/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Retrieval-aligned Tabular Foundation Models Enable Robust Clinical Risk Prediction in Electronic Health Records Under Real-world Constraints

Minh-Khoi Pham [1 2] **Thang-Long Nguyen Ho** [1 2] **Thao Thi Phuong Dao** [3 4] **Tai Tan Mai** [1 2] **Minh-Triet Tran** [3 5]
Marie E. Ward [6] **Una Geary** [6] **Rob Brennan** [2 7] **Nick McDonald** [8] **Martin Crane** [1 2] **Marija Bezbradica** [1 2]

Abstract

Accurate clinical outcome prediction from structured electronic health records (EHRs) remains challenging due to high-dimensional heterogeneous features, severe outcome imbalance, and distribution shift—constraints that are far more pronounced than in general tabular benchmarks. Although recently developed tabular foundation models, particularly in-context and retrieval-augmented approaches, achieve strong results on generic datasets, their behavior under realistic clinical constraints remains unclear.

In this work, we present the first systematic multi-cohort EHR benchmark comparing classical machine learning, deep tabular models, and tabular in-context learning (TICL) methods, including prior-fitted network (PFN)–based approaches and their retrieval-augmented variants (RA-TICL), across small clinical datasets and large real-world ICU cohorts. Through controlled stress tests varying data scale, feature dimensionality, outcome rarity, and cross-cohort generalization, we identify regime-dependent performance patterns. PFN-based TICL models are notably sample-efficient without retraining in low-data regimes. Meanwhile classic distance-based retrieval, which has been used to adapt PFN-based models to much larger datasets, leads to degradation as feature heterogeneity and class imbalance intensify in EHR domain.

To address this limitation, we propose Attention Weighting for Aligned Retrieval Embeddings (AWARE), a task-aligned retrieval framework that improves neighborhood quality through supervised embedding retrieval learning and lightweight adapter fine-tuning. Across stress tests, AWARE yields relative AUPRC improvements that increase with data complexity, from approximately 0.3–4.0% as training size scales, to 7.0% under high feature dimensionality, and up to 12.2% under extreme outcome imbalance, demonstrating that retrieval alignment becomes increasingly critical in complex EHR settings. Collectively, our findings establish retrieval quality and retrieval–inference coupling as central bottlenecks for deploying tabular in-context learning in clinical prediction.

1. Introduction

Accurate prediction from structured electronic health records (EHRs) is central to modern clinical decision support, enabling risk stratification, resource allocation, and timely intervention. However, EHR-based prediction remains methodologically challenging due to high-dimensional heterogeneous features, severe outcome imbalance, irregular temporal structure, and cross-institutional variation (Sarwar et al., 2022).

While recent years have seen rapid advances in foundation models for computer vision and natural language processing, tabular learning, and EHR modeling in particular, has also attracted growing interest, albeit at a more gradual pace and with comparatively fewer large-scale, systematic evaluations (Van Breugel & Van Der Schaar, 2024). Most EHR foundation models have primarily relied on sequential or language-based representations that emphasize longitudinal structure and large-scale pretraining (Wornow et al., 2023). Although effective in selected settings, these paradigms often require extensive pretraining on real-world clinical data, operate over long token sequences, and impose complicated standardization of heterogeneous medical events into universal formats. Such design choices introduce (1) substantial training computational and architectural complexity, and (2) practical and governance-related constraints associated with real sensitive patient data, motivating interest in alternative paradigms that are more efficient. This motivates us to adopt

[1]School of Computing, Dublin City University, Dublin, Ireland [2]ADAPT Centre, Dublin City University, Dublin, Ireland [3]University of Science, Vietnam National University, Ho Chi Minh City, Vietnam [4]Department of Otolaryngology, Thong Nhat Hospital, Ho Chi Minh City, Vietnam [5]John von Neumann Institute, Vietnam National University, Ho Chi Minh City, Vietnam [6]St James's Hospital, Dublin, Ireland [7]School of Computer Science, University College Dublin, Dublin, Ireland [8]School of Psychology, Trinity College Dublin, Dublin, Ireland. Correspondence to: Minh-Khoi Pham <minhkhoi.pham@adaptcentre.ie>.

Not peer-reviewed

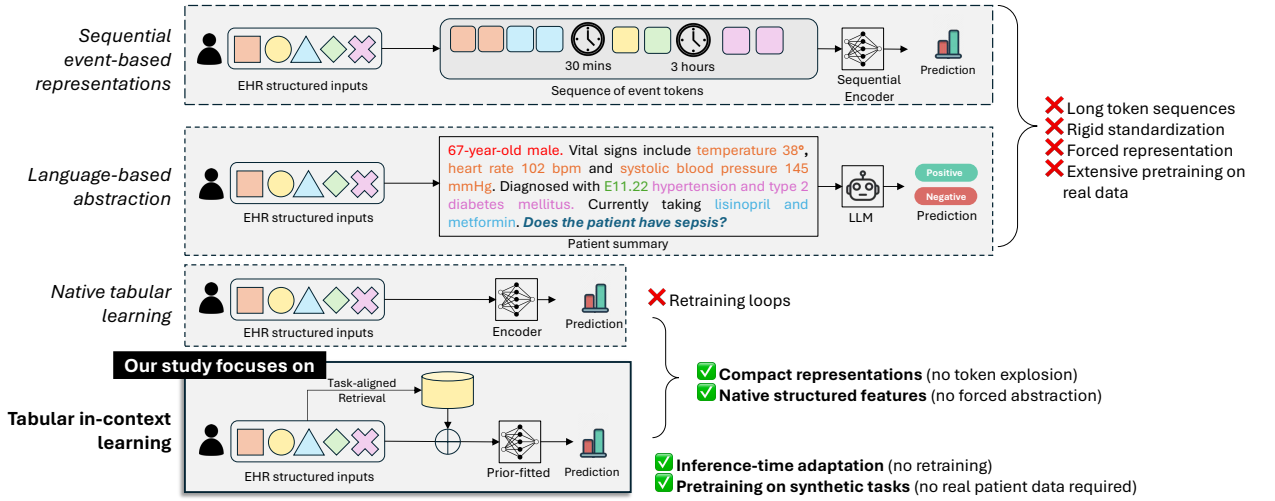


Figure 1. Conceptual comparison of major learning paradigms for EHR foundation models, illustrating differences in representation, adaptation strategy, and computational trade-offs across native tabular models, sequential event-based models, language-based abstractions, and tabular in-context learning. ✓highlights advantages while ✗highlights disavantages.

a newly alternative approach that has been introduced recently in tabular deep learning, that can address these issues of EHRs. We highlight the main differences between these EHR-based learning paradigms in Figure 1.

Recent advances in tabular in-context learning (TICL), such as TabPFN (Hollmann et al., 2025) and its retrieval-augmented variants (RA-TICL), namely LocalPFN, TabDPT (Thomas et al., 2024; Wen et al., 2025), offer a foundation-style paradigm for inference-time adaptation, relying only on synthetic data, and no task-specific retraining. For EHR-based clinical prediction, such approaches are particularly attractive because they can substantially reduce the need for repeated model development and retraining when new prediction tasks arise across cohorts or institutions (Moor et al., 2023). This enables rapid adaptation to shifting patient populations, changing outcome definitions, or newly available features, while lowering training and operational costs associated with conventional supervised learning pipelines. However, strong performance on generic tabular benchmarks might not guarantee reliable results in clinical outcome prediction from structured EHR data (Brown et al., 2025). In particular, it remains unclear how reliably TICL and retrieval-augmented TICL behave in high-dimensional, sparse, and imbalanced EHR representations where retrieval-based context construction maybe noisy or unstable (Neha et al., 2025).

In this work, we introduce an adaptive retrieval framework that learns task-aligned embeddings to improve neighborhood quality for RA-TICL. We evaluate this approach alongside classical machine learning baselines, deep tabular models, and prior-fitted network (PFN)–based TICL methods across diverse small-scale benchmarks and large real-world EHR cohorts. Our experiments isolate the effects of feature heterogeneity, class imbalance, dataset scale, and cross-cohort generalization, enabling a fine-grained analysis of when inference-time adaptation succeeds and fails, and how retrieval quality mediates this behavior. Figure 2 illustrates the diversity of the benchmark suite in terms of scale, dimensionality, rarity, and clinical domain.

This work makes the following contributions to the literature:

- **A first systematic, multi-cohort benchmark of tabular foundation models for EHR-based clinical prediction.** We evaluate classical machine learning baselines, deep tabular architectures, and TICL/RA-TICL methods across multiple real-world EHR cohorts and EHR-based supervised learning tasks, enabling controlled comparisons under shared preprocessing and evaluation protocols.

- **An adaptive retrieval framework (AWARE) for RA-TICL.** We introduce a task-aligned embedding-based retrieval strategy and a parameter-efficient adapter finetuning strategy, that both improves neighborhood quality and alignment under high-dimensional, imbalanced EHR conditions, enabling more stable retrieval-augmented in-context learning for clinical tasks.

- **A stress test of inference-time adaptation under clinical constraints.** We conduct a rigorous evaluation of these tabular foundation models under different clinical data constraints. We identify settings where TICL

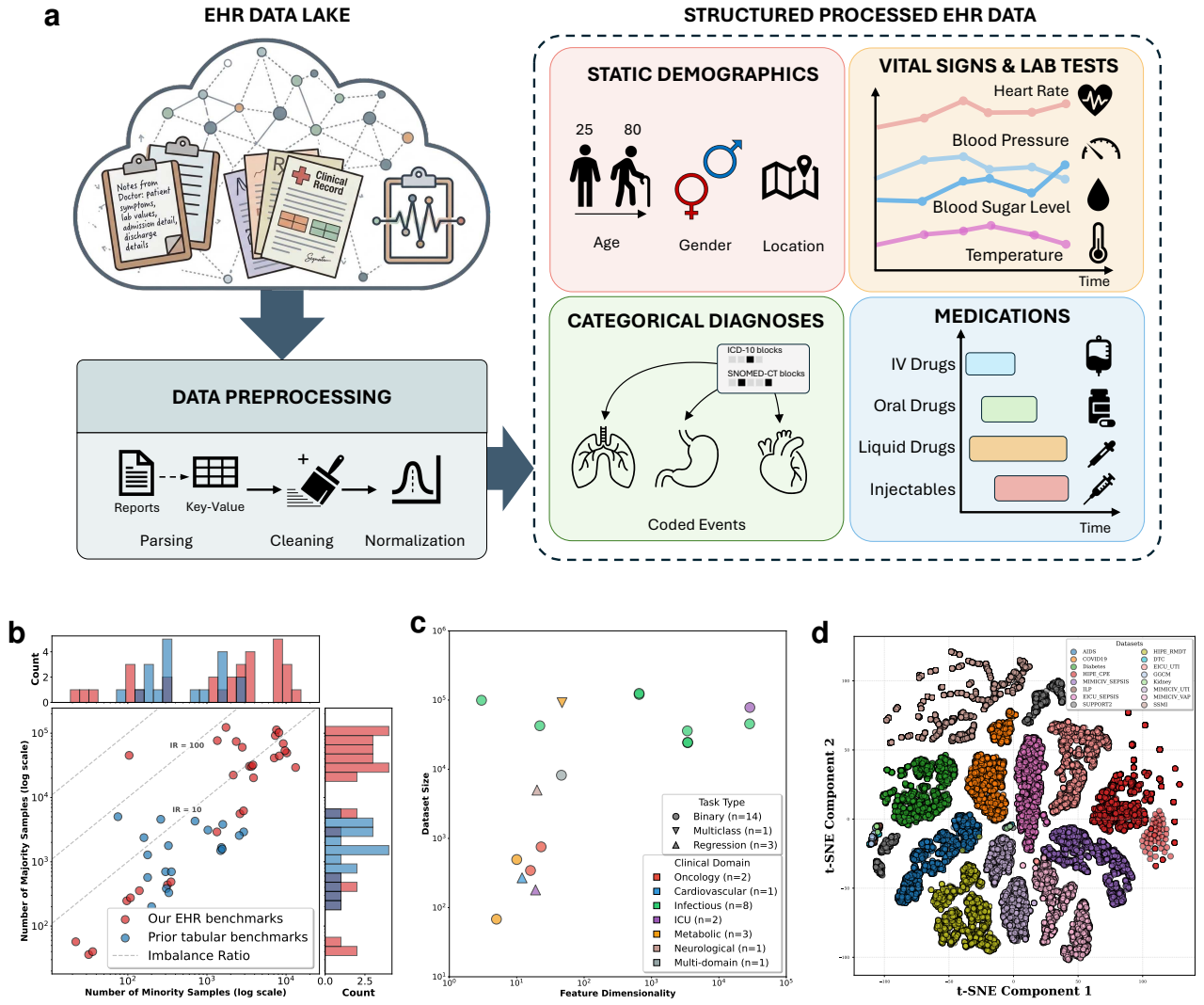


Figure 2. Landscape of EHR benchmarks evaluated in this study. The datasets span wide variation in cohort size, feature dimensionality, outcome prevalence, and clinical domain, illustrating key challenges of (a) heterogeneity, (b) rarity, (c) cross-task, cross-institutional generalizability and (d) distribution shift faced by tabular learning methods in real-world EHR settings. See detailed descriptions of reported datasets in Table A2.

is most effective (e.g., low-data and moderately imbalanced regimes) and where performance degrades as feature heterogeneity, cohort scale, and outcome rarity increase.

Taken together, our findings provide a clinically grounded evaluation of TICL and RA-TICL for EHR-based clinical prediction, discussing their strengths and limitations under realistic clinical data constraints. Based on these insights, we offer guidance for the reliable deployment of outcome prediction systems and outline directions for future method development.

2. Preliminaries and Problem Setup

Our goal in this paper is to evaluate in-context and retrieval-based tabular learning under non-sequential EHR pipelines, which we call snapshot-based[1] EHR prediction tasks. This differs from prior work (Fallahpour et al., 2025; Odgaard et al., 2024) focused on trajectory-dependent outcomes, such as length of stay, readmission, and long-horizon mortality, that require explicit sequence modeling. We focus on these tasks that operate on fixed-length, temporally-aggregated tabular representations. This setting reflects common deployment pipelines and isolates structural challenges such as feature heterogeneity, class imbalance, and context construction without introducing explicit sequence modeling. We formulate these tasks as supervised learning problems over tabular representations derived from structured EHR data.

In this setting, each patient encounter is represented by a d-dimensional feature vector $x \in \mathbb{R}^d$, comprising numerical measurements (e.g., laboratory values and vital signs), ordinal variables, and encoded categorical information such as diagnoses, procedures, medications, and demographics. Depending on the clinical task, the prediction target y may correspond to a discrete outcome label $y \in \{1, \ldots, C\}$ for classification (e.g., mortality, readmission, infection occurrence) or a continuous value $y \in \mathbb{R}$ for regression (e.g., length of stay). It should be noted that there can be multiple prediction targets for the same dataset.

Let the training dataset be defined as

$$D_{\text{train}} = \{(x^{(i)}_{\text{train}}, y^{(i)}_{\text{train}})\}_{i=1}^{N_{\text{train}}},$$

and let the test inputs be

$$X_{\text{test}} = \{x^{(i)}_{\text{test}}\}_{i=1}^{N_{\text{test}}},$$

with the objective of predicting the corresponding labels

$$Y_{\text{test}} = \{y^{(i)}_{\text{test}}\}_{i=1}^{N_{\text{test}}}.$$

[1] or also called static EHR

In this work, we use this setup to evaluate how different learning strategies behave under clinically realistic constraints such as data sparsity, heterogeneity, outcome rarity, and generalizability.

3. Background and Related Work

3.1. Challenges of EHR Data

EHRs exhibit structural properties that fundamentally shape the reliability, scalability, and clinical applicability of machine learning models (Li et al., 2020; Rasmy et al., 2021). Unlike curated tabular benchmarks, EHR data are heterogeneous, temporally structured, highly imbalanced, and subject to substantial variation across institutions and patient populations (Sarwar et al., 2022) (Figure 2).

In a real-world electronic health record, we may encounter thousands of variables, such as age, sex, comorbidities, hundreds of laboratory indices, hourly vital signs, ICD diagnosis codes, and procedural interventions(Harutyunyan et al., 2019; Miotto et al., 2016; Rajkomar et al., 2018; Shickel et al., 2017). However, if the task is to predict ICU mortality, only a relatively small subset, such as serum lactate, blood pressure, SOFA score, or mechanical ventilation status, may carry substantial prognostic value. The remaining variables are often sparse and not directly relevant to the specific prediction task. From a clinical perspective, this reflects everyday practice that every piece of information in the chart is not equally meaningful for every decision. From a machine learning perspective, this heterogeneity and noise complicate representation learning, as the model must identify clinically meaningful signals within a high-dimensional, partially irrelevant feature space (Miotto et al., 2016). For retrieval-augmented approaches in particular, which rely on identifying similar patients in feature space, the presence of many non-informative dimensions can distort similarity metrics, leading to retrieved neighbors that are mathematically close but not clinically comparable (Neha et al., 2025).

In addition to structural heterogeneity, EHR-based prediction tasks are characterized by severe outcome rarity and pervasive distribution shift. In clinical practice, many clinically significant events are relatively rare, such as hospital-acquired infections, septic shock, or mortality within a low-risk population(Moor et al., 2021; Zhu et al., 2026). This results in extreme class imbalance, where only a small fraction of patients experience the event of interest. Under such conditions, a model may achieve high aggregate accuracy simply by predicting the majority class, yet fail to detect rare but life-threatening cases(Sarwar et al., 2022). Furthermore, distribution shift is common across institutions, departments, and time periods. This is due to differences in patient demographics, treatment protocols, coding practices, or disease prevalence. A model trained in one hospital or co-

Model	Adaptation		Training		EHR Properties	
	In-context Learning	Retrieval -augmented	Pretrained	Task -adaptive	Feature Weighting	Imbalance Handling
Classical tabular and tree-based models						
Logistic Regression	✗	✗	✗	✓	✓	✗
k-Nearest Neighbors	✗	✓	✗	✗	✗	✗
XGBoost (Chen & Guestrin, 2016)	✗	✗	✗	✓	✓	✗
LightGBM (Ke et al., 2017)	✗	✗	✗	✓	✓	✗
CatBoost (Prokhorenkova et al., 2018)	✗	✗	✗	✓	✓	✗
MLP	✗	✗	✗	✓	✗	✗
Deep tabular models						
TabNet (Arik & Pfister, 2021)	✗	✗	✗	✓	✓	✗
TabTransformer (Huang et al., 2020)	✗	✗	✗	✓	✗	✗
ResNet (Gorishniy et al., 2023b)	✗	✗	✗	✓	✗	✗
TabM (Gorishniy et al., 2025)	✗	✗	✗	✓	✗	✗
Trompt (Chen et al., 2023)	✗	✗	✗	✓	✓	✗
ExcelFormer (Chen et al., 2024)	✗	✗	✗	✓	✓	✗
Tabular in-context learning models						
ModernNCA (Ye et al., 2024)	✓	✓	✗	✓	✗	✗
TabPFN (Hollmann et al., 2025)	✓	✗	✓	✗	✗	✓
kNNPFN (Thomas et al., 2024)	✓	✓	✓	✗	✗	✗
TabDPT (Ma et al., 2025)	✓	✓	✓	✗	✗	✗
AWARE + RA-TICL (Ours)	✓	✓	✓	✓	✓	✓

Table 1. Comparison of tabular learning paradigms evaluated in this study and their adaptation characteristics for EHR-based clinical prediction.

hort may therefore generalize poorly in another setting. For retrieval-based methods, class imbalance further degrades neighborhood quality, as retrieved similar cases are disproportionately drawn from the majority class, impoverishing the contextual support for rare outcomes(Nejjar et al., 2024). In clinical deployment, this poses a critical challenge because it is precisely these rare, high-risk cases that demand the most accurate and reliable prediction.

3.2. Clinical Representation Learning to EHR Foundation Models

Prediction from structured EHR features in tabular form has long been central to EHR-based clinical prediction (Jensen et al., 2012). In practice, this setting has been dominated by decision tree ensembles, most notably gradient-boosted methods such as XGBoost (Chen & Guestrin, 2016), LightGBM (Ke et al., 2017), and CatBoost (Prokhorenkova et al., 2018), which remain strong baselines for clinical outcome prediction due to their robustness to heterogeneous feature types, resilience under moderate data availability, and favorable interpretability–performance trade-offs (Grinsztajn et al., 2022; McElfresh et al., 2023). These properties have made tabular models widely adopted across diverse EHR-based supervised learning tasks and institutional settings.

Motivated by the success of deep learning in other domains, a growing body of work has explored neural architectures tailored to tabular data. Models such as TabNet (Arik & Pfister, 2021) and TabTransformer (Huang et al., 2020) introduced attention-based feature selection and contextualized embeddings, while more recent approaches, including ResNet-style architectures (Gorishniy et al., 2023b), TabM (Gorishniy et al., 2025), ExcelFormer (Chen et al., 2024), Trompt (Chen et al., 2023), and DANets(Chen et al., 2022)—aim to learn task-specific representations end-to-end in large-data regimes. A common theme across these models is the explicit modeling of feature relevance and structure: rather than treating all variables as equally informative, they incorporate mechanisms that implicitly select, weight, or group features to suppress noise from weakly relevant or redundant inputs. These models have been applied to diverse EHR-based clinical prediction tasks, including cardiovascular disease risk stratification (S et al., 2025; Sumon et al., 2025), glaucoma surgery prediction using multimodal EHR and imaging features (Koornwinder et al., 2025), Carbapenemase-Producing Enterobacteriace (CPE) infection forecasting (Pham et al., 2025a), preeclampsia risk prediction (Rabby et al., 2025), obstructive sleep apnea prediction (Chi et al., 2026), and hepatocellular carcinoma risk estimation (Park et al., 2025). Table 1 provides a con-

solidated overview of all benchmarked models in this study, situating each method by its adaptation strategy, training paradigm, and suitability for core challenges in EHR-based clinical prediction.

Despite these advances, most deep tabular models for EHR-based clinical prediction remain *task-specific*: they are trained to solve a single outcome at a time and typically require a moderate amount of labeled data for each new task (Brown et al., 2025). This paradigm limits scalability in real-world EHR settings, where institutions often face many EHR-based supervised learning tasks across heterogeneous cohorts, with constrained opportunities for repeated model development (Pang et al., 2025). With the increasing availability of large-scale EHR data and growing computational resources, recent work has therefore shifted toward *foundation-style learning* for EHRs (Khan et al., 2024; Moor et al., 2023), aiming to build models that support reuse, transfer, and adaptation across multiple tasks and clinical contexts. Existing efforts in this direction are commonly categorized into *language-based abstractions* and *sequential event-based representations* (Wornow et al., 2023).

Language-based abstractions convert structured EHR features into textual or pseudo-textual formats to enable prompt-based inference with large language models (Peng et al., 2023; Singhal et al., 2025). Although promising results have been reported on selected benchmarks (Acharya et al., 2024; Ben Shoham & Rappoport, 2024; Hegselmann et al., 2025), this approach relies heavily on prompt engineering and text abstractions that may obscure clinically relevant structure and complicate robustness and interpretability (Lee et al., 2024; Zhu et al., 2024). Sequential event-based representations operate directly on longitudinal EHR trajectories, building on earlier deep learning approaches such as DeepPatient, RETAIN, and GRU-D (Che et al., 2018; Choi et al., 2016; Miotto et al., 2016). More recent models include Transformer-based architectures (e.g., MedBERT (Rasmy et al., 2021), BEHRT (Li et al., 2020), CORE-BEHRT (Odgaard et al., 2024)) and state-space or autoregressive models capable of zero-shot prediction (Fallahpour et al., 2025; Pham et al., 2025b; Renc et al., 2024). While effective for temporally complex tasks, these approaches require large event vocabularies and rigid standardization pipelines, which can limit scalability and flexibility across institutions (Wornow et al., 2023).

More recently, prior-fitted and in-context learning approaches such as TabPFN have begun to attract attention in EHR-based clinical prediction settings (Ding et al., 2025; El-Melegy et al., 2024; Mahdi, 2025). Early clinical studies have reported competitive performance in applications such as dementia risk prediction in Parkinson's disease (Tran & Byeon, 2024) and glioblastoma prognostication (Karabacak et al., 2024), highlighting the potential of inference-time adaptation for structured EHR features. These developments motivate closer examination of *tabular in-context learning* as a complementary paradigm for EHR modeling. Recent benchmarking work has further shown that simple count-based tabular approaches, including ontology roll-ups paired with LightGBM or TabPFN, remain highly competitive with sequential transformers and large language model pipelines on structured EHR tasks, underscoring the continued strength of tabular learning and the need for systematic evaluation across datasets and modeling paradigms (Gao et al., 2025). In this work, we focus in particular on retrieval-augmented TICL, which seeks to scale inference-time conditioning to large and heterogeneous clinical cohorts by constructing compact, query-specific contexts from observed patient data.

3.3. In-Context Learning for Tabular Data

In-context learning (ICL) has recently emerged as a promising direction for tabular prediction, and is particularly relevant when models must adapt at inference time without task-specific retraining. TabPFN (Hollmann et al., 2025)[2] has demonstrated that a prior-fitted transformer, trained on synthetic tabular tasks, can adapt to new classification problems at inference time without gradient-based fine-tuning, achieving strong performance in small-sample settings. Subsequent work extended this framework with improved architectures and broader task families, including Real-TabPFN(Garg et al., 2025), TabDPT (Ma et al., 2025), and TabICL (Qu et al., 2025).

Many of these methods can be understood within the framework of *Prior-Fitted Networks*, which approximate amortized Bayesian inference over tabular datasets (Müller et al., 2021). Given a test input x and a context dataset D, PFNs learn a predictive distribution $q_\theta(y \mid x, D)$ that approximates the posterior predictive $p(y \mid x, D)$ by minimizing the expected negative log-likelihood

$$\mathcal{L}_{\text{PFN}} = \mathbb{E}_{(x,y,D)} \left[-\log q_\theta(y \mid x, D) \right]. \tag{1}$$

During training, datasets are sampled from a distribution over generative processes, encouraging the model to acquire inductive biases that support rapid adaptation to unseen tasks. Architecturally, PFN-based models represent each such encounter as a token, embedding features into dense representations, and processing them via self-attention. At deployment, predictions are obtained by conditioning on a prompt of labeled reference examples and unlabeled queries, rather than updating model parameters.

While effective on synthetic and general-purpose tabular benchmarks, the behavior of TICL methods on real-world EHR-based supervised learning tasks remains insufficiently understood. Clinical datasets are high-dimensional and of-

[2]We refer to TabPFNv2 simply as TabPFN.

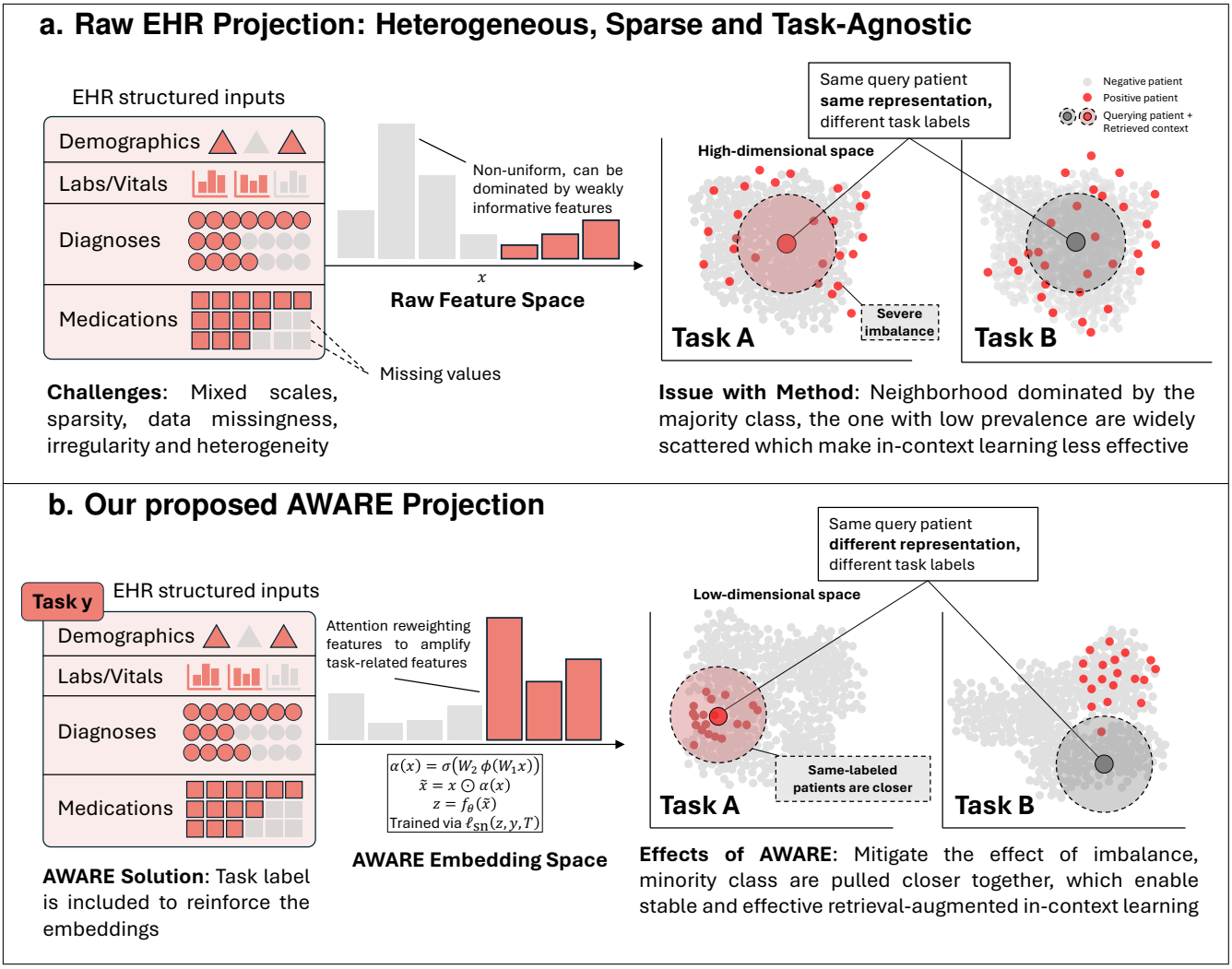


Figure 3. The challenges of raw EHR data for retrieval and the solution of task-aware alignment. (A) Raw vector is heterogeneous, sparse and task-agnostic, leading to noisy geometry and multi-task misalignment. (B) Task-aligned projection reshapes geometry for label-consistent, effective retrieval.

ten dominated by sparse, task-dependent signal (Sarwar et al., 2022), making it neither computationally feasible nor methodologically appropriate to condition on the full training cohort at inference time. Consequently, practical use of TICL for EHR-based clinical prediction requires explicit *context construction*, selecting a compact, query-specific subset of training encounters. This introduces a critical design challenge: how to retrieve context that is predictive for clinical outcome prediction in high-dimensional, imbalanced feature spaces, motivating the use of retrieval-augmented approaches (Neha et al., 2025).

3.4. Retrieval-Augmented Tabular In-Context Learning in Clinical Settings

A practical limitation of tabular in-context learning (TICL) in large EHR cohorts is that conditioning on all available training encounters is infeasible, necessitating an explicit context construction step at deployment. Retrieval augmentation addresses this limitation by selecting a small, query-specific subset of training encounters to include in the prompt (Qin et al., 2021; Zheng et al., 2023). In clinical natural language processing, retrieval-augmented prompting has been shown to improve robustness and factual accuracy of large language models without additional training (Lewis et al., 2020; Yang et al., 2025). This motivates analogous retrieval-augmented approaches for EHR representations, where retrieval induces a similarity-based inductive

bias while keeping inference-time conditioning tractable for EHR-based supervised learning tasks.

Formally, RA-TICL constructs, for each query instance $\mathbf{x}_q$, a context set

$$D_{\text{context}}(\mathbf{x}_q) = \underset{\mathbf{x}_i \in D_{\text{train}}}{\arg\,\text{top-k}}\ s(\mathbf{x}_q, \mathbf{x}_i), \tag{2}$$

where $s(\cdot,\cdot)$ denotes a similarity function over encounter representations, and $k \ll |D_{\text{train}}|$. The retrieved set D_{context} is combined with the query instance to form the in-context prompt used for inference, enabling inference-time adaptation without updating model parameters. Representative approaches include TabR (Gorishniy et al., 2023a), ModernNCA (Ye et al., 2024), and retrieval-integrated PFN variants such as LocalPFN (Thomas et al., 2024) and MixturePFN (Xu et al., 2024).

Despite their promise, RA-TICL methods have seen limited and fragmented adoption in clinical settings. Existing medical applications of retrieval-based context construction often rely on handcrafted similarity measures or task-specific pipelines rather than learned, task-aligned retrieval. Prior work has explored distance-based retrieval for clinical prediction using manually designed feature or sequence similarity metrics, including applications to cancer diagnosis, mortality, and readmission prediction (Prasetyo et al., 2024; Wang et al., 2022; Xue et al., 2025). However, we argue that such naïve retrieval in raw EHR feature space may suffer from several limitations (Figure 3a). High-dimensional and heterogeneous clinical features can distort distance-based similarity metrics, leading to retrieval dominated by weakly informative dimensions, and severe class imbalance can yield neighborhoods heavily skewed toward majority-class examples with poor label consistency. As a result, retrieved encounters may appear similar yet be uninformative for clinical outcome prediction, reducing the effectiveness and stability of retrieval-augmented in-context learning (Qin et al., 2021; Zheng et al., 2023).

These limitations indicate that effective retrieval for RA-TICL must rely on representations explicitly aligned with task-relevant structure rather than raw feature similarity. To this end, we propose AWARE (Figure 3b), a learned retrieval framework that optimizes neighborhood quality in embedding space. It yields contexts with higher label purity and greater stability under high-dimensional, heterogeneous, and imbalanced EHR conditions, thereby improving the reliability of retrieval-augmented in-context learning.

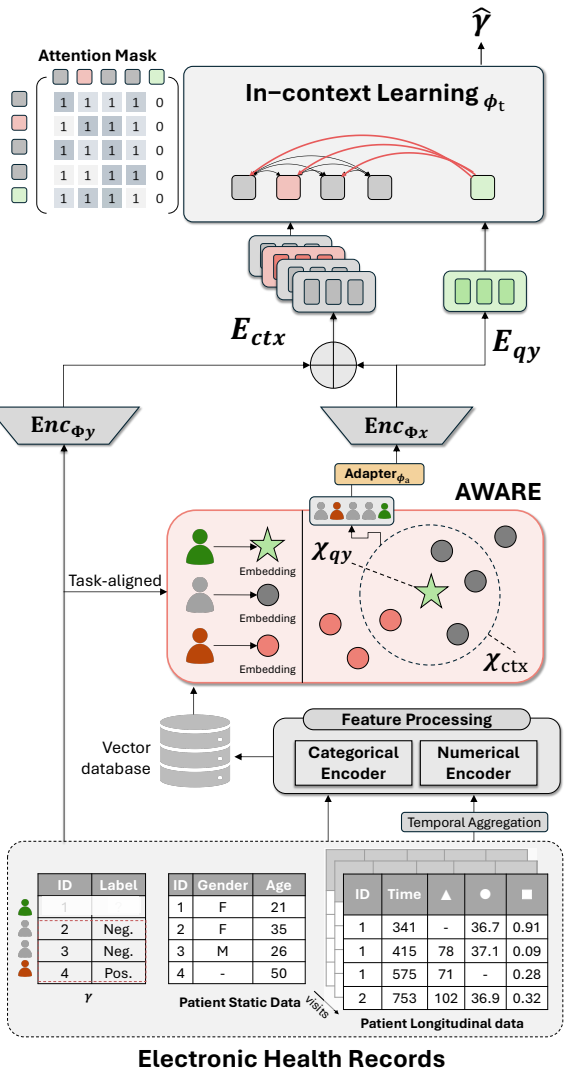


Figure 4. Overview of our designed task-aligned retrieval for retrieval-augmented in-context learning EHR model.

assess the performance and robustness of existing tabular models, particularly TICL and RA-TICL, under the structural constraints characteristics of real-world EHR data. We focus on these domain-specific challenges that are central to clinical deployment: (1) data scaling (Section 6.3.1) (2) feature heterogeneity (Section 6.3.2), (3) outcome rarity (Section 6.3.3), and (4) transferability across datasets and institutions (Section 6.3.4).

4. Methodology

This study is primarily designed as a systematic methodological evaluation of tabular learning approaches for EHR-based clinical prediction. Along with our task-aligned retrieval approach for RA-TICL methods, we also aim to

In-context learning architecture for EHRs. Figure 4 illustrates how we adapt retrieval-augmented tabular in-context learning for EHR-based clinical prediction. Following the conceptual design of PFN-based models such as

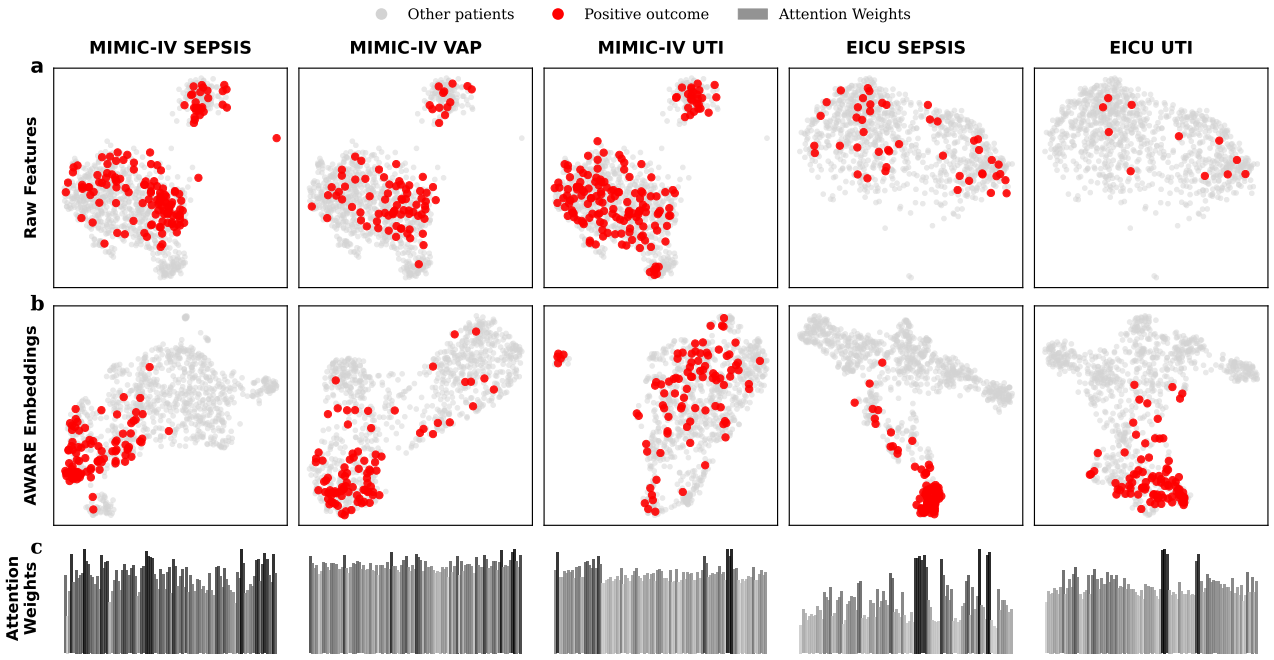


Figure 5. Visualization of patient neighborhood structure across MIMIC-IV and eICU dataset, illustrated by projecting data into 2D space. (a) In raw feature space, proximity does not reliably correspond to predictive relevance across different clinical tasks, resulting in scattered positive examples. (b) In contrast, task-aligned embedding retrieval (AWARE) reshapes the representation space such that label-consistent examples form coherent local clusters while preserving smooth structure for downstream in-context learning. (c) Black bars illustrate task-specific feature reweighting induced by the attention-based encoder, highlighting how retrieval adapts to heterogeneous clinical variables.

TabPFN and TabDPT, we retain the original in-context backbone and introduce two new modules within our AWARE framework. Structured EHR features are first encoded into vectors in a learned embedding space using an attention-based encoder trained with metric learning objectives, as described later. Then during inference time, for a querying patient instance, nearest neighbors are retrieved in this learned embedding space, forming a task-aligned context set (as formulated in Equation 2). The retrieved examples are then concatenated with the query instance to construct the in-context prompt for the underlying ICL model. The prompt are then passed through a lightweight projection adapter that maps them into a representation space compatible with the frozen backbone. The frozen backbone of TICL models are often stacked Transformer layers with attention masking designed to prevent information leakage between the query representations. The model then produces the final prediction via prior-fitted inference without any updating to the weights of the backbone.

Motivation for adaptive, task-aligned retrieval. A central limitation of existing RA-TICL methods is their reliance on fixed similarity metrics defined in the raw feature space, most commonly Euclidean or cosine distance (Thomas et al., 2024). In EHR data, patient representations are high-dimensional, sparse, and dominated by rare but clinically important events. Moreover, the same feature space often supports multiple prediction tasks (as in Figure 5a), each requiring a different notion of clinically meaningful similarity. Under these conditions, a single task-agnostic retrieval metric is unlikely to be universally appropriate and may select context examples that are feature-similar but clinically irrelevant or label-misaligned.

To address the limitations of raw feature-space retrieval, we propose AWARE (Attention Weighting for Aligned Retrieval Embeddings), a task-aligned retrieval framework that learns embedding representations optimized for neighborhood quality in RA-TICL. AWARE shapes the embedding space so that retrieved neighbors better reflect predictive relevance under heterogeneous and imbalanced EHR conditions. Retrieval is performed exclusively in the learned embedding space, while the underlying in-context model remains unchanged. Figure 5 illustrates this idea.

Beyond improving retrieval, we further introduce a lightweight adapter fine-tuning mechanism within the RA-TICL backbone. Specifically, we train parameter-efficient

adapter layers using bootstrapped contexts constructed from SNNL-based retrieval, a similar idea to Thomas et al. (Thomas et al., 2024). This second-stage alignment enables the in-context model to better internalize task-relevant neighborhood structure without modifying the pretrained backbone weights. The framework offer both geometric alignment (via SNNL) and representation alignment (via adapter), while preserving the modularity and efficiency of the original in-context framework.

Task-Aligned Embedding and Neighborhood Shaping. AWARE employs a lightweight neural encoder that combines feature-wise attention with embedding-based metric learning. Given a patient feature vector $x \in \mathbb{R}^d$, the encoder computes instance-specific attention weights

$$\alpha(x) = \sigma(W_2\, \phi(W_1 x)) \in [0,1]^d, \tag{3}$$

where $\phi(\cdot)$ is a nonlinear activation and $\sigma(\cdot)$ denotes the sigmoid function. These weights are applied via element-wise scaling,

$$\tilde{x} = x \odot \alpha(x) \tag{4}$$

where a $\odot$ b denotes represents element-wise multiplication. It performs soft, task-aligned feature selection by amplifying clinically relevant features while suppressing noisy or weakly informative ones. From the reweighted representation $\tilde{x}$, the encoder produces a low-dimensional embedding:

$$z = f_\theta(\tilde{x}), \tag{5}$$

which is used exclusively for retrieval.

To directly align neighborhood structure with clinical outcomes, the encoder are trained using the Soft Nearest Neighbor Loss (SNNL) (Frosst et al., 2019). Given a batch of b embeddings $\{z_i\}_{i=1}^b$ with labels $\{y_i\}_{i=1}^b$, the loss is defined as

$$\ell_{\text{sn}}(z, y, T) = -\frac{1}{b}\sum_{i=1}^{b} \log \frac{\sum_{j\neq i,\, y_j = y_i} \exp\left(-\frac{d(z_i,z_j)}{T}\right)}{\sum_{k\neq i} \exp\left(-\frac{d(z_i,z_k)}{T}\right)}, \tag{6}$$

where $d(\cdot,\cdot)$ denotes a distance metric and T is a temperature parameter. Minimizing Eq. 6 encourages embeddings from the same outcome class to form compact local neighborhoods while separating embeddings from different classes, yielding task-aligned similarity for retrieval in heterogeneous feature spaces. We conceptually demonstrate the design of AWARE in Figure 3b.

Adapter-based representation alignment. While SNNL-based embedding learning improves neighborhood quality during retrieval, the retrieved context must still be mapped into a representation space compatible with the frozen in-context backbone. To address this, we introduce a parameter-efficient adapter A_ψ that operates prior to the backbone model. Rather than modifying internal Transformer layers, the adapter learns a task-specific projection that maps both support and query representations into a feature space aligned with the backbone's pretrained representation.

Inspired by Xu et al. (Xu et al., 2024), we train the adapter using bootstrapped prompt simulations constructed from the downstream dataset. Let $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N_{\text{train}}}$ denote the downstream dataset. We approximate the inference-time prompt distribution by sampling $\mathcal{D}_{\text{bootstrap}} \sim p(\mathcal{D} \mid \mathcal{D}_{\text{train}})$. For large datasets ($N_{\text{train}} > 3000$), we first sample an anchor instance $(x_q, y_q) \sim p(x, y \mid \mathcal{D}_{\text{train}})$, and construct its context set via SNNL-based retrieval:

$$\mathcal{D}_{\text{context}} = \text{SNNL-Retrieval}(x_q \mid \mathcal{D}_{\text{train}}, B), \tag{7}$$

where B denotes the context size. The bootstrapped prompt is then formed as $\mathcal{D}_{\text{bootstrap}} = \mathcal{D}_{\text{context}} \cup \{(x_q, y_q)\}$. For smaller datasets, we instead sample a large subset of $\mathcal{D}_{\text{train}}$ as $\mathcal{D}_{\text{context}}$ and treat the remaining samples as pseudo-query points.

We freeze the pretrained backbone transformer $q_{\theta_0}(y \mid x, \mathcal{D})$ and optimize only the adapter parameters ψ. The adapted model is defined as

$$q_{\theta_0,\psi}(y \mid x, \mathcal{D}) = q_{\theta_0}\big(y \mid A_\psi(x), A_\psi(\mathcal{D})\big). \tag{8}$$

The adapter is trained by minimizing the negative log-likelihood over bootstrapped prompts:

$$\mathcal{L}_{\text{adapter}} = -\mathbb{E}_{\mathcal{D}_{\text{bootstrap}}} \log q_{\theta_0,\psi}\big(y_q \mid x_q, \mathcal{D}_{\text{context}}\big). \tag{9}$$

Intuitively, the frozen backbone encodes the synthetic pretraining prior $p(\mathcal{D} \mid \phi)p(\phi)$, while the adapter learns a dataset-specific correction that aligns the model with the empirical prompt distribution induced by SNNL-based retrieval, $p(\mathcal{D} \mid \mathcal{D}_{\text{train}})$. This preserves prior structure and stability while enabling task-aware representation alignment under realistic EHR context construction.

Imbalance-aware training and ensemble stabilization. Many EHR-based clinical prediction tasks exhibit severe class imbalance, where majority-class examples can dominate embedding structure and distort neighborhood quality. To mitigate this effect during AWARE training, we adopt balanced mini-batch sampling. Specifically, for a mini-batch $\mathcal{B}$, samples are drawn with probability inversely proportional to class frequency:

$$\mathbb{P}\big((x_i, y_i) \in \mathcal{B}\big) \propto \frac{1}{N_{y_i}}, \tag{10}$$

where N_{y_i} denotes the number of training instances with label y_i. This ensures approximately uniform class representation within each mini-batch, stabilizing neighborhood formation in embedding space, particularly for rare outcomes.

To further reduce variance in retrieval behavior, we employ a K-fold cross-validation ensemble of AWARE encoders (with $K = 5$ in our experiments). Let $\{\mathcal{D}_{\text{train}}^{(k)}\}_{k=1}^{K}$ denote the K disjoint training folds. For each fold k, we train an independent encoder $f_{\theta^{(k)}} : \mathbb{R}^d \to \mathbb{R}^m$ using SNNL with balanced sampling, yielding embeddings $z^{(k)} = f_{\theta^{(k)}}(x)$. At inference time, embeddings are aggregated across folds to obtain a stabilized representation:

$$\bar{z}(x) = \frac{1}{K} \sum_{k=1}^{K} f_{\theta^{(k)}}(x). \tag{11}$$

Retrieval is then performed in the ensemble-averaged embedding space using $\bar{z}(x)$ for both query and candidate support instances.

The final AWARE training objective thus consists solely of the Soft Nearest Neighbor Loss, while balanced sampling and cross-validated ensembling act as regularization mechanisms that reduce sampling variance and improve retrieval stability in high-dimensional and imbalanced EHR settings.

Data Availability The datasets analyzed in this study are publicly available or accessible through established data use agreements. MIMIC-IV is available via PhysioNet (https://physionet.org/) to credentialed users who complete the required data use training and agreements. The eICU Collaborative Research Database is available through PhysioNet under similar access requirements. The OpenML and UCI datasets used in this study are publicly accessible via their respective repositories. The privately held HIPE dataset was used under institutional data use agreements and is not publicly available due to privacy and regulatory restrictions. Access to this dataset may be considered by the data custodians upon reasonable request and subject to appropriate approvals. All data used in this study are either publicly accessible through the sources described above or governed by third-party data use agreements.

Code Availability The source code used for data preprocessing, model training, evaluation, and generation of figures in this study will be made publicly available upon publication, with documentation sufficient to reproduce the main results. To facilitate reproducibility, we provide detailed preprocessing descriptions and hyperparameter configurations in the Appendix A.4.

5. Experimental Validation

Benchmark Design. We evaluate various methods (listed in Appendix A.3) on a diverse collection of EHR datasets selected to reflect the heterogeneity, sparsity, and distribution complexity of real-world clinical data. The datasets include structured diagnosis and procedure codes, demographic variables, laboratory measurements, and time-series vital signs. To ensure a fair assessment of generalization, we exclude any datasets reported as part of the pretraining corpus of the benchmarked models. Prediction tasks span regression, binary classification, and multiclass classification, with both balanced and highly imbalanced outcome distributions. A full list of datasets can be found in the Appendix A.4.

Datasets and Tasks Choice. Our experiments draw from four complementary data sources. First, we use MIMIC-IV (Johnson et al., 2023), a large single-center critical care dataset, and derive infection-related prediction tasks such as sepsis, urinary tract infection (UTI), and ventilator-associated pneumonia (VAP). Second, we evaluate on eICU (Pollard et al., 2018), a multi-center intensive care unit (ICU) dataset spanning over 200 hospitals, which supports assessment of robustness under institutional and cohort shifts, particularly for sepsis and UTI. Third, we use a private daily-recorded cohort from a large acute hospital system, codenamed as HIPE (Pham et al., 2024; 2025a), including prediction of a rare infection such as CPE. Finally, we include a curated set of 12 OpenML/UCI clinical datasets that provide small- to medium-scale EHR-based supervised learning tasks. These datasets support a systematic study of data scale, feature heterogeneity, severe outcome rarity, and cross-cohort generalizability across diverse EHR-based clinical prediction settings.

Feature Engineering and Preprocessing. For the OpenML / UCI datasets, we retain all available features and apply standard preprocessing, including mean imputation for numerical variables, mode imputation for categorical variables, and feature normalization. Missingness is handled through this standardized imputation pipeline to ensure comparability across models and datasets. While missingness in clinical data may arise under missing-at-random (MAR) or missing-not-at-random (MNAR) mechanisms, explicitly modeling missingness patterns is beyond the scope of this methodological evaluation. Although missing data is an important topic and has been examined elsewhere in the context of EHR data (Getzen et al., 2023), we merely treat missingness as an inherent component of EHR heterogeneity and focus our analysis on how different learning paradigms behave under consistent preprocessing assumptions.

For MIMIC-IV and eICU, we follow established EHR preprocessing pipelines (Gupta et al., 2022; Wu et al., 2024) (details in Appendix A.4). Because these cohorts contain thou-

sands of highly sparse variables, dimensionality reduction is required for computational feasibility and cross-model compatibility. We therefore apply model-agnostic filtering based on variance and feature prevalence, removing near-constant variables and extremely rare features before capping the dimensionality at 500 when necessary. This procedure avoids architecture-specific inductive bias in feature selection while retaining a broad and heterogeneous representation of the EHR feature space. Time-series measurements are summarized using simple aggregation statistics (mean, minimum, maximum, and standard deviation).

Training Configuration and Hyperparameters. Most tabular models are trained and evaluated under a unified experimental protocol to ensure fair and controlled comparisons. In particular, their hyperparameters are optimized independently for each dataset using 30 validation trials, ensuring a matched tuning budget across approaches. For each selected configuration, models are retrained with three random seeds to estimate performance variability. For retrieval-based methods using AWARE, the retrieval encoders are trained independently of the downstream TICL models. All training is performed for 50 epochs using the AdamW optimizer with learning rate 10^{-3} and task-dependent batch sizes determined by memory constraints. Retrieval encoders are trained once per dataset and task and remain frozen during TICL inference, ensuring that downstream performance differences reflect retrieval quality rather than joint retraining effects. For adapter finetuning, we let it train for 5 epoch with relatively small learning rate, number of trainable parameters are only around 0.01% the number parameters of total RA-TICL backbones. For pretrained RA-TICL models, unless otherwise specified, we use a fixed context size of 1024 labeled instances to balance retrieval diversity and computational feasibility. This value is held constant across experiments to ensure comparability.

Evaluation Configuration. We use stratified patient-level train / test splits for all longitudinal EHR datasets (MIMIC-IV, eICU, and HIPE), so patients are unique across splits. Since HIPE is collected from a distinct hospital system with different coding practices and patient populations, it also serves as an external evaluation cohort relative to the public ICU datasets. Feature extraction and preprocessing use statistics computed on the training split only (e.g., normalization and imputation parameters), and these are reused unchanged for validation and test.

Evaluation Metrics. We report evaluation metrics commonly used in clinical prediction. For classification tasks, we report the area under the receiver operating characteristic curve (AUROC), which measures the probability that a randomly chosen positive instance is assigned a higher predicted score than a randomly chosen negative instance. The area under the precision–recall curve (AUPRC) summarizes the trade-off between precision and recall across decision thresholds and is particularly informative under severe class imbalance (Saito & Rehmsmeier, 2015).

For retrieval-based models, we evaluate retrieval quality using Precision@k, defined as:

$$\text{Precision@}k = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}[y_i = 1],$$

where y_i denotes the ground-truth label of the i-th retrieved instance.

For multiclass classification, we additionally report the F1-score, defined as the harmonic mean of precision and recall:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

For regression tasks, performance is measured using mean absolute error (MAE) and root mean squared error (RMSE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|, \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

6. Results

This section presents a systematic evaluation of tabular learning paradigms under clinically realistic EHR conditions. We first examine how retrieval quality affects retrieval-augmented in-context learning, focusing on the impact of AWARE-based context construction. We then analyze model behavior across key structural regimes of EHR data, including feature heterogeneity, outcome rarity, and cross-cohort generalization.

6.1. Effectiveness of AWARE for EHR Retrieval

Figure 6 presents a percentage-based ablation analysis of AWARE and its subsequent enhancements across three retrieval-based backbones (kNN, kNN-PFN, and TabDPT) and two large-scale ICU cohorts (MIMIC-IV and eICU). Performance is expressed as relative improvement over each backbone's baseline configuration, allowing direct comparison of the incremental contribution of each component to both AUROC and AUPRC.

Across architectures and datasets, the introduction of task-aligned retrieval via SNNL marks the first substantial and systematic performance increase. While attention weighting alone yields modest and occasionally inconsistent changes, SNNL consistently produces positive gains, indicating that explicit alignment of neighborhood geometry with task labels is critical for effective context construction. These effects are especially pronounced in AUPRC, where improvements are amplified under class imbalance. For example,

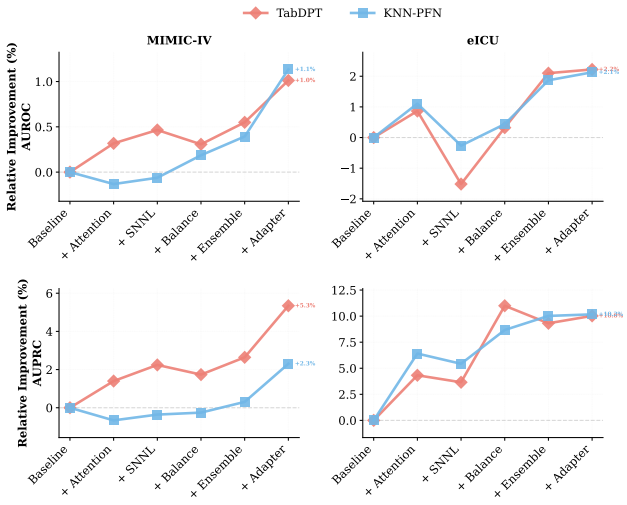


Figure 6. Relative (%) improvement over baseline as AWARE components are progressively introduced across three retrieval-based backbones on MIMIC-IV and eICU. Results are shown for AUROC (top) and AUPRC (bottom), averaged across tasks. See detailed numbers in Table A3.

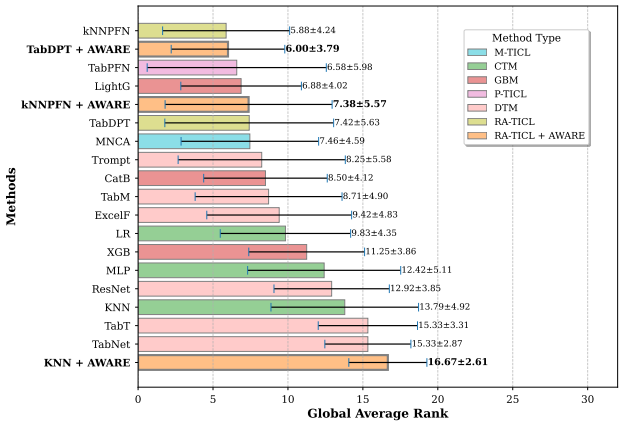


Figure 7. Aggregate performance rankings of tabular models on 12 small-medium EHR datasets, rankings are measured by average score of performance. Rankings summarize relative performance across tasks within each regime. See detailed numbers in Table A4.

on eICU, relative AUPRC gains exceed 10% for PFN-based backbones and approach 90% for the kNN backbone, reflecting substantial recovery in rare-event discrimination.

Balanced sampling further enhances performance, particularly in imbalanced tasks, reinforcing the interpretation that retrieval bias toward majority-class exemplars is a primary limiting factor in vanilla RA-TICL. Importantly, these improvements are observed across all three backbones, suggesting that gains arise from improved context quality rather than increased model capacity. Subsequent refinements, including cross-validation ensembling and adapter fine-tuning, provide additional but comparatively incremental improvements, building upon the performance foundation established by task-aligned retrieval.

Overall, the result demonstrates that AWARE's primary contribution lies in restructuring retrieval neighborhoods (Figure 5) to better support rare-event conditioning. Once retrieval alignment is corrected, downstream stabilization techniques yield consistent but secondary gains, underscoring that inference-time embedding alignment is the dominant mechanism driving robustness improvements in large-scale, heterogeneous EHR settings. The complete result table is included in the Appendix A.5.

6.2. Overall results

This section reports overall performance rankings across our collection of small-to-medium and large-scale EHR datasets. These aggregated rankings provide a high-level view of model behavior across regimes, but should be interpreted cautiously. Aggregated rankings obscure clinically relevant variation across tasks, outcome prevalences, and data characteristics, and do not capture differences in robustness, stability, or failure behavior. Complete per-dataset performance results for the models can be found in the Appendix A.5.

6.2.1. Small-medium EHR datasets

Figure 7 summarizes model performance across 12 small- to medium-scale EHR datasets. In this regime, TICL methods (e.g., TabPFN) and retrieval-augmented variants (e.g., kNNPFN, TabDPT, ModernNCA) achieve strong average AUROC rankings without requiring task-specific fine-tuning or extensive hyperparameter search (with the exception of ModernNCA). These findings align with prior work demonstrating superior sample efficiency of foundation-based tabular models on general benchmarks (Erickson et al., 2025). In smaller datasets where labeled data are limited, inference-time conditioning appears sufficient to capture predictive structure without additional optimization.

In contrast, incorporating AWARE does not consistently yield further improvements in this regime. This is expected, as AWARE is primarily designed to address challenges associated with large-scale, high-dimensional, and highly imbalanced EHR settings. When datasets are smaller and feature spaces are more compact, raw similarity-based retrieval is often adequate (TabPFN without retrieval already has comparative performance to kNNPFN). While tree-based models remain competitive and exhibit stable performance, they are not uniformly superior in our experiments, contrary to some prior claims in the EHR literature. Given the heterogeneity of feature types across these datasets (Ap-

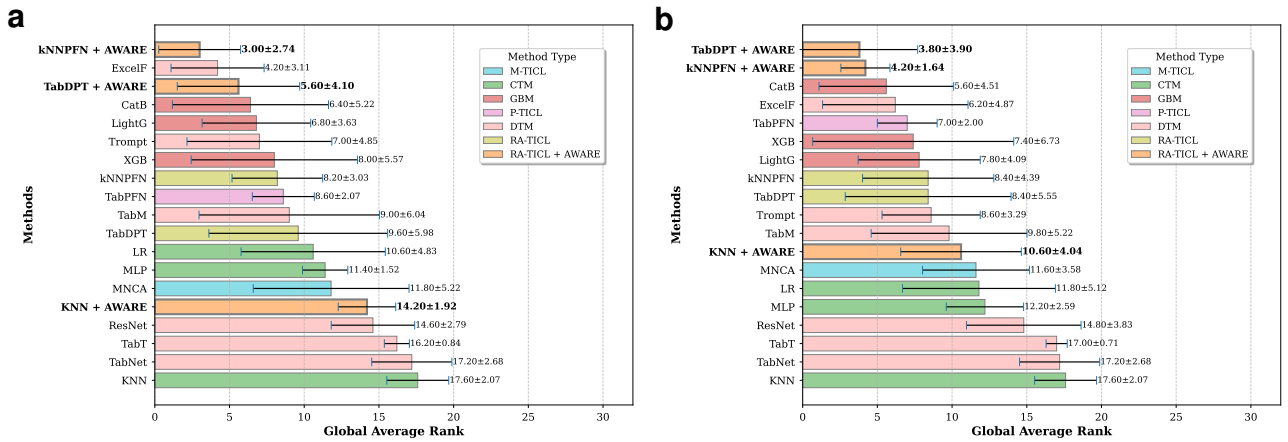


Figure 8. Aggregate performance rankings of tabular models on large-scale EHR datasets (MIMIC-IV and eICU), measured by average (a) AUROC and (b) AUPRC across tasks within each regime. See detailed numbers in Table A5.

pendix A.4), the results suggest that TICL methods provide a favorable balance of efficiency and effectiveness for small- to medium-scale clinical prediction tasks.

6.2.2. Large-scale EHR datasets

As dataset scale and feature heterogeneity increase, the performance landscape shifts substantially. Figure 8 shows that, in large-scale EHR regimes (averaging across MIMIC-IV and eICU tasks), models with explicit feature selection mechanisms (e.g., ExcelFormer, Trompt) or tree-based architectures (e.g., CatBoost, LightGBM, XGBoost) rise in the rankings relative to TICL and RA-TICL methods.

This pattern challenges the intuition that retrieval universally enhances in-context learning. Instead, it suggests that retrieval quality deteriorates in high-dimensional, sparse, and imbalanced EHR settings, where raw feature similarity becomes less aligned with predictive relevance. Consistent with this interpretation, ModernNCA exhibits a pronounced decline in ranking on large-scale cohorts, indicating sensitivity to neighborhood construction as data complexity increases.

Incorporating the proposed AWARE framework mitigates this degradation. By aligning embedding geometry with task labels and reducing majority-class dominance, AWARE enables RA-TICL models to recover performance in these challenging regimes and, in several cases, surpass competing approaches.

Several deep tabular architectures, including TabTransformer, ResNet, and TabNet, rank below simpler baselines such as logistic regression in this regime, suggesting overfitting or instability when modeling heterogeneous EHR features without strong inductive bias. In contrast, gradient-boosted tree methods remain consistently competitive, reinforcing their status as reliable baselines in complex clinical datasets. Notably, ExcelFormer and Trompt improve markedly at larger scale, likely benefiting from built-in feature filtering mechanisms that attenuate noise from weakly informative or sparse variables.

In the following subsections, we dissect these trends in greater detail. We first isolate the effects of feature dimensionality and heterogeneity, examining how weakly informative features influence neighborhood quality. We then analyze robustness under severe class imbalance and evaluate how retrieval strategies affect context construction. Together, these analyses clarify the conditions under which tabular in-context and retrieval-augmented learning succeed in EHR prediction, and when unaligned retrieval may introduce clinically meaningful degradation.

6.3. Performance Under EHR Constraints

We next investigate how core structural properties of EHR data shape model behavior. Specifically, we analyze the effects of feature heterogeneity, outcome rarity, and cross-cohort distribution shift. These controlled experiments isolate how different learning paradigms respond to increasing data complexity and real-world imbalance.

Before presenting the results, we emphasize an important methodological distinction. Several classical and deep tabular baselines achieve higher absolute performance in certain regimes; however, these models are extensively optimized through dataset-specific hyperparameter search, cross-validation, and full retraining. In contrast, TICL-based approaches operate without task-specific fine-tuning, domain adaptation, or manual feature engineering, relying instead on pretrained priors and inference-time conditioning in a few-shot manner. The comparison therefore reflects fun-

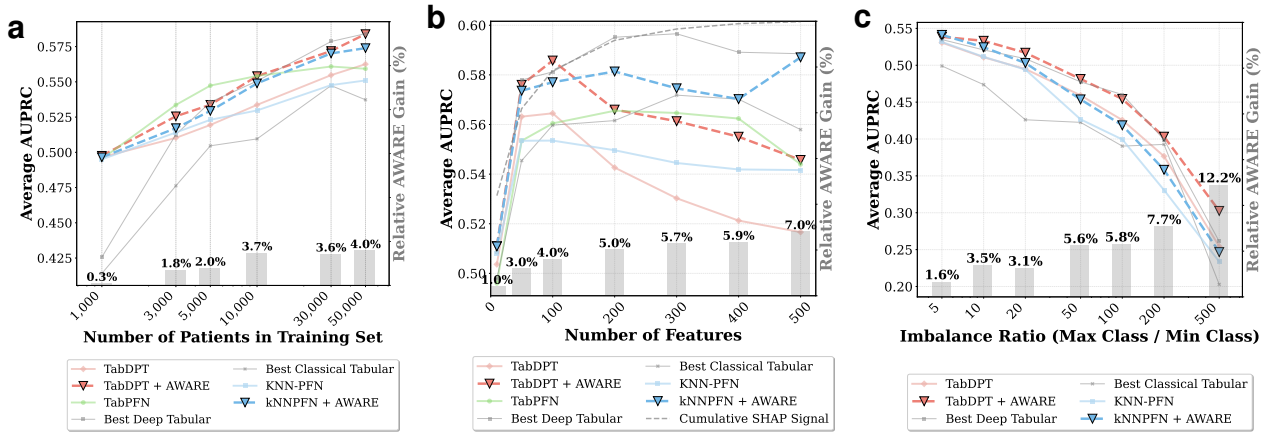


Figure 9. Performance of tabular models under three representative clinical constraints. To improve readability given the large number of models evaluated, highly competitive baselines within each paradigm are consolidated into a single curve labeled *Best Classical Model* and *Best Deep Tabular Model*, representing the strongest-performing model in each family under the respective constraint. These are compared with individual tabular in-context learning methods and their AWARE-enhanced variants (dashed lines). Performance is measured using AUPRC across all settings, while overlaid bar charts report the average relative performance gains (in %) obtained by incorporating AWARE. (a) Performance scaling with training data size. (b) Performance degradation under increasing feature heterogeneity. (c) Performance under increasing outcome imbalance.

damentally different adaptation paradigms: fully supervised optimization versus retraining-free inference-time adaptation. Importantly, while baseline TICL methods do not always achieve the top performance, integrating our proposed AWARE framework substantially improves relative AUPRC performance as the data constraints scale up.

6.3.1. Data Scale Sensitivity

We first examine how model performance evolves as the amount of labeled training data increases, a central factor in real-world clinical deployment where data availability varies substantially across institutions and tasks. While large academic medical centers may accumulate tens of thousands of patient records, many clinical prediction problems operate in low-data regimes. Understanding how different learning paradigms scale with training size therefore provides insight into their practical utility across diverse healthcare settings.

Figure 9a reports AUPRC as a function of training set size, ranging from 1,000 to 50,000 patients randomly sampled from both MIMIC-IV and eICU datasets. As expected, performance improves monotonically for all models as more training data become available. In the low-data regime around 1,000 patients, P-TICL and RA-TICL models achieve the strongest performance, consistent with their ability to leverage inductive priors and inference-time adaptation. However, performance gains for baseline TICL methods saturate earlier than for deep tabular models. Beyond approximately 10,000 patients, deep representation-learning approaches consistently outperform standard TICL variants, indicating that fixed prior-based inference becomes limiting as data availability increases.

Notably, incorporating the proposed AWARE framework into RA-TICL models (TabDPT and kNNPFN) alters this scaling behavior. As shown in Figure 9a, AWARE-enhanced variants maintain performance improvements as dataset size increases and remain competitive even at 50,000 patients. The overlaid bar charts further quantify these effects by reporting the relative performance gains (in percentage) achieved by AWARE at each sample size. These gains increase with data scale, from approximately 0.3% in the smallest regime to around 3–4% in larger datasets, indicating that AWARE not only improves absolute performance but also enhances scalability under increasing data complexity.

6.3.2. Feature Heterogeneity

We next examine sensitivity to feature heterogeneity by progressively expanding the input dimensionality from 10 to 500 features. Features are added in descending order of importance according to mean absolute SHapley Additive exPlanations) (SHAP) values, such that early increments introduce high-signal variables while later increments increasingly incorporate lower-signal and potentially noisy dimensions. As illustrated by the cumulative SHAP signal curve in Figure 9b, informative signal saturates relatively early, and subsequent feature additions primarily increase heterogeneity rather than useful information content.

Under this controlled expansion, baseline TICL models (TabPFN, TabDPT, and kNNPFN) exhibit progressively

steeper performance degradation as dimensionality increases. In contrast, the best deep and classical tabular baselines remain comparatively stable from 200 features and beyond, suggesting greater robustness to feature noise through learned representations or implicit feature selection. These results indicate that the scalability limitations of TICL methods in high-dimensional EHR settings are driven not merely by dataset size, but by the dilution of task-relevant signal within heterogeneous feature spaces.

Incorporating AWARE substantially mitigates this degradation. As shown in Figure 9b, AWARE-enhanced RA-TICL variants consistently maintain higher AUPRC across the full dimensionality range and exhibit reduced performance drop in the high-feature regime. The overlaid bar charts quantify these improvements as relative percentage gains over the corresponding baseline RA-TICL models. Notably, the gains increase monotonically with dimensionality—from approximately 1–4% at low feature counts to nearly 7% at 500 features—demonstrating that AWARE becomes increasingly beneficial as heterogeneity intensifies.

Taken together, these findings suggest that in heterogeneous EHR feature spaces, limitations of RA-TICL models arise not only from representational capacity but from context construction in noisy, high-dimensional inputs. By introducing task-aligned retrieval and implicit feature reweighting, AWARE helps adapt RA-TICL models to EHR complexity and mitigates heterogeneity-induced instability.

6.3.3. Outcome Rarity

Rare outcomes are pervasive in clinical prediction tasks, particularly for adverse events such as hospital-acquired infections, in-hospital mortality, and treatment complications (Cohen et al., 2006). To systematically evaluate robustness to outcome rarity, we adopt a controlled imbalance-scaling protocol. The training set size is fixed (10,000 samples), while the negative-to-positive ratio (imbalance ratio, IR) is progressively increased from 5 to 500 through stratified subsampling without replacement. All models are evaluated on a fixed held-out test set with the original class distribution (Zhu et al., 2026).

Figure 9c illustrates how AUPRC degrades as class imbalance intensifies. As expected, all methods exhibit monotonic performance decline as positive prevalence decreases. However, the rate and stability of this degradation differ substantially across paradigms. The best deep tabular model maintains comparatively strong absolute performance across moderate imbalance levels, while the best classical model degrades more sharply, particularly beyond IR = 50. Baseline RA-TICL models (TabDPT and kNNPFN) demonstrate competitive robustness under mild-to-moderate imbalance but experience increasingly pronounced degradation in the extreme regime (IR $\geq$ 200), reflecting the difficulty of constructing representative contextual examples when minority instances become scarce.

Incorporating AWARE significantly mitigates this effect. As shown in Figure 9c, AWARE-enhanced variants consistently remain above their baseline counterparts across all imbalance levels, with the performance gap widening as rarity increases. The overlaid bar charts quantify these improvements as relative percentage gains, which grow from approximately 1.6% at IR = 5 to over 12% at IR = 500. This monotonic increase indicates that AWARE becomes progressively more beneficial as minority-class sparsity intensifies.

These results suggest that under extreme prevalence skew, limitations of baseline RA-TICL models stem not only from representational capacity but from instability in retrieval and context construction when positive examples are scarce. By introducing task-aligned retrieval and improved minority coverage, AWARE enhances robustness in rare-event regimes while preserving the retraining-free nature of inference-time adaptation.

6.3.4. Generalizability

Across the full collection of datasets evaluated in this study, TICL and RA-TICL models maintain competitive performance relative to classical machine learning and deep tabular architectures, without task-specific fine-tuning or manual feature engineering. This holds across diverse medical domains, feature types (laboratory measurements, medications, diagnosis codes, procedures), varying levels of missingness, and heterogeneous institutional practices.

On small- and medium-scale EHR cohorts (Section 6.2.1), inference-time conditioning alone is sufficient to achieve performance comparable to or exceeding strong tree-based and neural baselines. Importantly, these results are obtained without hyperparameter search or dataset-specific adaptation, suggesting that pretrained tabular priors transfer effectively across heterogeneous clinical settings.

In large-scale ICU benchmarks (Section 6.2.2), performance remains competitive despite increased feature dimensionality and distributional variability. While certain tree-based or feature-selective deep models achieve top rankings in specific tasks, TICL-based approaches consistently remain within a narrow performance margin, demonstrating stability across tasks with differing prevalences, clinical objectives, and feature compositions. This consistency across domains indicates that inference-time conditioning provides a robust inductive bias that generalizes beyond the dataset characteristics seen during pretraining.

We further extend this evaluation to the privately held HIPE cohort, originally analyzed in prior work (Pham et al., 2024;

Method	HIPE			
	RDMT		CPE	
	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
XGB	$0.824_{\pm 0.010}$	$0.432_{\pm 0.020}$	$0.755_{\pm 0.043}$	$0.036_{\pm 0.019}$
CatB	$0.827_{\pm 0.005}$	$0.446_{\pm 0.015}$	$0.645_{\pm 0.078}$	$0.027_{\pm 0.025}$
LightG	$0.814_{\pm 0.055}$	$0.438_{\pm 0.065}$	$0.719_{\pm 0.092}$	$0.030_{\pm 0.021}$
TabNet	$0.827_{\pm 0.005}$	$0.440_{\pm 0.009}$	$0.558_{\pm 0.083}$	$0.111_{\pm 0.055}$
ResNet	$0.830_{\pm 0.001}$	$0.457_{\pm 0.006}$	$0.773_{\pm 0.030}$	$0.046_{\pm 0.030}$
TabT	$\underline{0.834}_{\pm 0.004}$	$0.461_{\pm 0.007}$	$\underline{0.781}_{\pm 0.057}$	$0.080_{\pm 0.017}$
TabPFN	$0.828_{\pm 0.002}$	$0.455_{\pm 0.004}$	$0.719_{\pm 0.051}$	$0.022_{\pm 0.010}$
kNNPFN	$0.706_{\pm 0.000}$	$0.365_{\pm 0.000}$	$0.760_{\pm 0.000}$	$0.141_{\pm 0.000}$
TabDPT	$0.709_{\pm 0.000}$	$0.362_{\pm 0.000}$	$0.746_{\pm 0.000}$	$0.017_{\pm 0.000}$
Ours (Tabular In-Context Learning Models + AWARE)				
kNNPFN	$\mathbf{0.843}_{\pm 0.005}$	$\mathbf{0.480}_{\pm 0.006}$	$\mathbf{0.782}_{\pm 0.024}$	$\mathbf{0.232}_{\pm 0.020}$
TabDPT	$\underline{0.834}_{\pm 0.005}$	$\underline{0.465}_{\pm 0.002}$	$\underline{0.781}_{\pm 0.017}$	$\underline{0.221}_{\pm 0.002}$

Table 2. Performance on HIPE prediction tasks. Results are mostly reused from Pham et al. (Pham et al., 2025a). With our new inclusion of kNNPFN and TabDPT baselines and AWARE versions.

2025a), by incorporating RA-TICL variants into the same experimental framework (Table 2). HIPE represents an extreme setting characterized by substantial feature heterogeneity, high-dimensional sparse coding, and severe outcome imbalance. In this environment, baseline retrieval-augmented TICL models (kNNPFN and TabDPT) exhibit noticeable performance degradation relative to the strongest task-specific deep models, reflecting the compounded challenges of sparsity, noisy similarity structure, and rare-event prevalence in raw EHR feature space.

However, when combined with the proposed AWARE retrieval alignment, performance improves substantially across both tasks and metrics. The AWARE variants consistently outperform their vanilla counterparts and achieve the best overall performance in most settings, with particularly pronounced gains in AUPRC for the highly imbalanced CPE task (e.g., from 0.141 to 0.232 for kNNPFN). Notably, these improvements allow AWARE-based TICL models to match or surpass the strongest deep tabular baselines such as TabTransformer and ResNet, despite requiring no task-specific retraining. These results suggest that much of the degradation observed in naïve retrieval stems from misaligned similarity structure in heterogeneous EHR spaces, and that task-aware retrieval alignment can substantially improve the effectiveness of in-context conditioning under severe imbalance.

These findings suggest that inference-time tabular conditioning remains viable even under substantial structural and statistical shift. While extreme imbalance and heterogeneity can attenuate performance, degradation is gradual rather than catastrophic. Importantly, TICL methods achieve this level of cross-cohort robustness without retraining, handcrafted feature pipelines, or domain-specific engineering.

Overall, across public ICU benchmarks and a highly heterogeneous real-world hospital cohort, TICL and RA-TICL provide strong zero-tuning baselines that generalize across institutions, medical domains, and feature regimes. Their performance profile indicates that in-context adaptation offers a practically deployable alternative to fully retrained task-specific models, particularly in settings where rapid deployment and minimal engineering overhead are critical.

7. Discussion

The recent success of foundation models and in-context learning on generic tabular benchmarks has motivated their adoption for EHR-based prediction (Bohr et al., 2025). Our results, however, demonstrate that transferring these paradigms to clinical data is neither straightforward nor uniformly beneficial. The observed limitations do not reflect shortcomings of individual models, but rather expose systematic mismatches between the assumptions underlying tabular foundation learning and the structural properties of real-world EHRs (Sarwar et al., 2022).

Recent benchmarking work has shown that simple count-based models, including ontology-based roll-ups paired with LightGBM or TabPFN, remain highly competitive against sequential transformers and mixture-of-agents LLM pipelines on structured EHR prediction tasks, often matching or outperforming more complex approaches (Gao et al., 2025). These findings underscore the continued strength of tabular learning for EHRs, while also highlighting the need for systematic, multi-dataset evaluations to understand when and why more adaptive or foundation-style methods provide meaningful advantages. While slightly similar to our work, we provide more in-depth comparison between different count-based tabular models on broader range of EHR-based clinical prediction tasks.

A central assumption implicit in many tabular in-context learning methods is that proximity in feature space corresponds to predictive relevance. In EHR data, this assumption is frequently violated: clinical feature spaces are high-dimensional and sparse, with predictive signal concentrated in task-specific subsets of variables (Scheurwegs et al., 2017). As a result, distance-based retrieval can select contextual examples that are superficially similar yet clinically irrelevant or label-misaligned for the outcome of interest (Qin et al., 2021; Zheng et al., 2023). While naïve distance-based retrieval frequently fails in heterogeneous EHR feature spaces, our AWARE framework demonstrates that explicitly learning task-aligned embeddings substantially mitigates these limitations. By reshaping neighborhood structure through local objectives, AWARE improves label alignment within retrieved contexts and stabilizes RA-TICL performance across datasets. Nonetheless, retrieval quality remains a central bottleneck: even learned embed-

dings were found to be sensitive to extreme imbalance and distribution shift. Future work may therefore explore tighter coupling between retrieval and prediction, such as jointly optimizing embedding and in-context inference or propagating gradients from the downstream PFN model into the retrieval space.

A related assumption is that increasing data scale naturally improves reliability. While TICL models benefit from strong inductive priors in low-data regimes, larger EHR cohorts do not consistently compensate for feature heterogeneity, sparsity, or outcome imbalance. Inference-time adaptation trades bias for variance, yielding favorable sample efficiency but increasing sensitivity to context composition and noise (Thomas et al., 2024; Xu et al., 2024). In clinical decision support settings, such instability may manifest as inconsistent predictions or degraded performance on rare but high-impact outcomes.

Sequential EHR models address many of these challenges by explicitly modeling temporal structure and long-range dependencies (Li et al., 2020; Wornow et al., 2025). However, these advantages come at substantial cost, including long tokenized sequences, rigid standardization pipelines, and extensive pretraining requirements. Native tabular learning, including TICL, offers a modular and training-efficient alternative that supports heterogeneous feature integration and inference-time adaptation. The goal of this work is therefore not to replace sequential modeling, but to characterize when tabular in-context learning provides a viable and practically attractive solution, and when its limitations necessitate more expressive architectures.

More broadly, our findings highlight the risks of extrapolating performance gains from synthetic or non-clinical benchmarks to medical prediction tasks. Benchmark-driven evaluation alone is insufficient for clinical AI, as models that perform well on curated tabular datasets may fail under the distribution shift, imbalance, and temporal complexity characteristic of real-world healthcare data (Bohr et al., 2025). Without explicit stress testing under these conditions, claims of generality risk obscuring clinically meaningful failure modes.

From a practical perspective, tabular foundation models are most appropriate in constrained settings with limited labeled data, moderate feature dimensionality, and relatively stable patient populations. In such regimes, incorporating task-aligned retrieval through AWARE can further enhance robustness by mitigating context mismatch without requiring retraining of the underlying model. In large, heterogeneous cohorts or strongly temporal tasks, models with explicit feature selection or dedicated sequence architectures may offer safer and more predictable behavior, despite reduced flexibility.

Potential use in clinical workflows The application of AI analysis to understand and predict patient outcomes is an evolving sphere of research. Too much 'noise' from the data and not enough disease- and patient-specific data to support accurate clinical decision making can undermine its utility. Complementing understanding of data points on linear patient pathways with more ability to deep dive into multiple data streams (e.g. vital signs, lab results, procedures undertaken) and changing patterns in these data could help clinicians in designing, implementing and quality assuring care plans. In this context, PFN-based models pretrained entirely on synthetic tabular data offer an additional advantage for healthcare deployment, as they avoid the need for large-scale pretraining on sensitive patient records while still providing strong inductive priors for downstream clinical prediction tasks. The EU AI Act calls for a 'Fundamental Rights Impact Assessment' (FRIA) before deploying high-risk AI systems in healthcare involving patient data which includes ensuring that AI-enabled tools that support clinical decision making are completely transparent to the clinician and do not introduce any bias for the patient. It is incumbent on researchers to avoid situations where AI-supervised learning tools with large data sets are skewed towards more common conditions, increasing the risk that rare conditions are not recognised.

8. Limitations

This study has several limitations that need consideration. First, our evaluation focuses on structured EHR data represented in tabular form, often constructed from temporally aggregated or engineered features. Although clinical data are inherently longitudinal (Choi et al., 2016; Lipton et al., 2015), many practical machine learning pipelines collapse rich time-series trajectories into summary statistics (e.g., mean, minimum, maximum, or standard deviation over predefined windows) to ensure compatibility with conventional tabular architectures. While this representation enables controlled comparison across learning paradigms and improves computational tractability, it introduces a fundamental trade-off between temporal fidelity and modeling simplicity. Fine-grained patient dynamics and sequential dependencies may be obscured, potentially affecting in-context adaptation and retrieval behavior, as models operate on compressed summaries rather than full clinical trajectories. Accordingly, our conclusions are most directly applicable to tabular and semi-tabular EHR representations rather than end-to-end longitudinal or multimodal modeling.

Second, although we evaluate multiple large-scale ICU datasets alongside a diverse set of smaller clinical cohorts, all datasets are retrospective and drawn from a limited number of healthcare systems. As with most EHR studies, unobserved confounding, coding practices, and institutional bi-

ases may influence model behavior. While our cross-dataset analyses partially address generalizability, prospective validation and broader multi-institutional studies are necessary to assess deployment-level robustness. In addition, our current evaluation focuses primarily on discrimination metrics; a more complete assessment for clinical deployment should incorporate calibration analyses of predicted risk estimates.

Third, although we treat missingness as an inherent component of EHR heterogeneity and apply consistent preprocessing across methods, we do not explicitly model missing-at-random or missing-not-at-random mechanisms. Because missingness in clinical data is often structured and clinically informative, this simplification may limit our ability to assess robustness under varying testing, ordering, and documentation practices. Methods that explicitly model or leverage informative missingness patterns may exhibit different behavior than observed in our comparisons.

Fourth, our analysis of retrieval strategies considers representative similarity mechanisms rather than an exhaustive exploration of all possible retrieval or metric-learning designs. Although task-aligned and learned similarity metrics improve retrieval quality relative to naïve baselines, they do not eliminate retrieval-induced instability. Extreme imbalance and distribution shift remain open challenges for retrieval-based inference in clinical settings.

Finally, this study should be interpreted as methodological benchmarking rather than evidence of clinical utility or deployment readiness. We do not evaluate downstream decision-making impact, clinician interaction, real-time workflow integration, or prospective performance in live clinical environments. Accordingly, our findings provide guidance for model development and comparative evaluation under controlled conditions, but do not constitute validation of clinical effectiveness.

9. Conclusion

This study provides a systematic methodological assessment of tabular in-context and retrieval-augmented learning for EHR-based clinical prediction. We characterize when and why foundation-style tabular learning succeeds or fails under the structural constraints of real-world clinical data, and we propose a new system (AWARE) as a principled retrieval adaptation to address context mismatch in RA-TICL. Across diverse datasets and tasks, we show that inference-time adaptation offers strong sample efficiency in low-resource settings, but remains highly sensitive to feature heterogeneity, retrieval quality, outcome rarity, and institutional distribution.

To address this, we introduced AWARE, an adaptive framework that consists of (1) an improved retrieval approach that can attend to features and learn task-aligned similarity in an embedding space to construct more predictive contexts for RA-TICL and (2) parameter-efficient adapter finetuning that learns to align the representations of EHR to downstream ICL models. By introducing AWARE, we demonstrate that retrieval quality can be substantially improved without modifying the underlying in-context model. This highlights retrieval space design, not model capacity, as a critical determinant of performance in EHR-based RA-TICL.

Taken together, this work positions tabular foundation models as complementary tools rather than general-purpose solutions for EHR modeling. Our findings further suggest that embedding-level adaptation of similarity, as implemented in AWARE, represents a practical pathway toward improving inference-time learning under clinical constraints. Progress toward reliable and clinically useful AI systems will require domain-aware representations, task-specific notions of similarity, and evaluation protocols that prioritize robustness, stability, and safety alongside predictive performance.

Acknowledgements

This research was conducted with the financial support of Taighde Éireann-Research Ireland under Grant Agreement No. 13/RC/2106_P2 at ADAPT, the Research Ireland Centre for AI-Driven Digital Content Technology at DCU funded through the Research Ireland Research Centres Programme. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any author-accepted manuscript version arising from this submission.

CRediT authorship contribution statement

Minh-Khoi Pham: Conceptualization, Methodology, Software, Visualization, Validation, Formal analysis, Investigation, Writing. **Thang-Long Nguyen Ho**: Methodology, Software, Validation, Formal analysis, Investigation, Writing. **Tai Tan Mai**: Visualization, Validation, Formal analysis, Writing. **Thao Thi Phuong Dao**: Visualization, Validation, Formal analysis, Writing. **Minh-Triet Tran**: Visualization, Validation, Formal analysis, Writing. **Marie E. Ward**: Formal analysis, Writing. **Una Geary**: Formal analysis, Writing. **Rob Brennan**: Formal analysis, Writing. **Nick McDonald**: Formal analysis, Writing. **Marija Bezbradica**: Project administration, Supervision, Writing. **Martin Crane**: Project administration, Supervision, Writing.

Competing Interests

The authors declare no competing interests.

References

Acharya, A., Shrestha, S., Chen, A., et al. Clinical risk prediction using language models: benefits and considerations. *Journal of the American Medical Informatics Association*, 31(9):1856–1864, 02 2024. ISSN 1527-974x. doi:10.1093/jamia/ocae030.

Arik, S. Ö. and Pfister, T. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 6679–6687, 2021. doi:10.1609/aaai.v35i8.16826.

Ben Shoham, O. and Rappoport, N. Cpllm: Clinical prediction with large language models. *PLOS Digital Health*, 3(12):1–15, December 2024. doi:10.1371/journal.pdig.0000680.

Bohr, A., Altstidl, T., Eskofier, B., and Salin, E. Perspective: Leveraging domain knowledge for tabular machine learning in the medical domain. In Chang, S., Hulsebos, M., Liu, Q., Chen, W., and Sun, H. (eds.), *Proceedings of the 4th Table Representation Learning Workshop*, pp. 143–155, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-268-8. doi:10.18653/v1/2025.trl-1.11.

Borzooei, S. and Tarokhian, A. Differentiated thyroid cancer recurrence. UCI Machine Learning Repository, 2023.

Brown, K. E., Yan, C., Li, Z., et al. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. *Journal of the American Medical Informatics Association*, 32(5):811–822, 2025. doi:10.1093/jamia/ocaf038.

Che, Z., Purushotham, S., Cho, K., et al. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018. doi:10.1038/s41598-018-24271-9.

Chen, A. and Chen, D. O. Simulation of a machine learning enabled learning health system for risk prediction using synthetic patient data. *Scientific Reports*, 12(1):17917, October 2022. ISSN 2045-2322. doi:10.1038/s41598-022-23011-4.

Chen, J., Liao, K., Wan, Y., et al. Danets: Deep abstract networks for tabular data classification and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3930–3938, 2022. doi:10.1609/aaai.v36i4.20309.

Chen, J., Yan, J., Chen, Q., et al. ExcelFormer: A neural network surpassing GBDTs on tabular data, July 2024.

Chen, K.-Y., Chiang, P.-H., Chou, H.-R., et al. Trompt: towards a better deep neural network for tabular data. In *Proceedings of the 40th International Conference on Machine Learning*, Icml'23. JMLR.org, 2023. doi:10.5555/3618408.3618581.

Chen, T. and Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, August 2016. doi:10.1145/2939672.2939785.

Chi, H., Dabbs-Brown, A., and Jurek-Loughrey, A. Obstructive sleep apnea prediction: A comprehensive review and comparative study. *Machine Learning*, 115(29), 2026. doi:10.1007/s10994-025-06937-4.

Chicco, D. and Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, 20(1):16, 2020a. doi:10.1186/s12911-020-1023-5.

Chicco, D. and Jurman, G. Sepsis survival minimal clinical records. UCI Machine Learning Repository, 2020b.

Choi, E., Bahadori, M. T., Sun, J., et al. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, volume 29, 2016. doi:10.5555/3157382.3157490.

Clore, J., Cios, K., DeShazo, J., et al. Diabetes 130-us hospitals for years 1999-2008. UCI Machine Learning Repository, 2014.

Cohen, G., Hilario, M., Sax, H., et al. Learning from imbalanced data in surveillance of nosocomial infection. *Artificial intelligence in medicine*, 37(1):7–18, 2006. doi:10.1016/j.artmed.2005.03.002.

Ding, Y., Ren, J., Lu, J., et al. Longitudinal progression prediction of alzheimer's disease with tabular foundation model. *arXiv preprint arXiv:2508.17649*, 2025. doi:10.48550/arXiv.2508.17649.

El-Melegy, M., Mamdouh, A., Ali, S., et al. Prostate cancer diagnosis via visual representation of tabular data and deep transfer learning. *Bioengineering*, 11(7), 2024. ISSN 2306-5354. doi:10.3390/bioengineering11070635.

Erickson, N., Purucker, L., Tschalzev, A., et al. Tabarena: A living benchmark for machine learning on tabular data. *arXiv preprint arXiv:2506.16791*, 2025. doi:10.48550/arXiv.2506.16791.

Fallahpour, A., Alinoori, M., Ye, W., et al. Ehrmamba: Towards generalizable and scalable foundation models for electronic health records. In Hegselmann, S., Zhou, H., Healey, E., Chang, T., Ellington, C., Mhasawade,

V., Tonekaboni, S., Argaw, P., and Zhang, H. (eds.), *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pp. 291–307. Pmlr, 15–16 Dec 2025. doi:10.48550/arXiv.2405.14567.

Frosst, N., Papernot, N., and Hinton, G. Analyzing and improving representations with the soft nearest neighbor loss. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2012–2020. Pmlr, 09–15 Jun 2019. doi:10.48550/arXiv.1902.01889.

Gao, J., Rosenthal, M., Wolpin, B., et al. Count-based approaches remain strong: A benchmark against transformer and LLM pipelines on structured EHR. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*, 2025. doi:10.48550/arXiv.2511.00782.

Garg, A., Ali, M., Hollmann, N., et al. Real-tabpfn: Improving tabular foundation models via continued pre-training with real-world data. *arXiv preprint arXiv:2507.03971*, 2025. doi:10.48550/arXiv.2507.03971.

Getzen, E., Ungar, L., Mowery, D., Jiang, X., and Long, Q. Mining for equitable health: Assessing the impact of missing data in electronic health records. *Journal of biomedical informatics*, 139:104269, 2023. doi:10.1016/j.jbi.2022.104269.

Gorishniy, Y., Rubachev, I., Kartashev, N., et al. Tabr: Unlocking the power of retrieval-augmented tabular deep learning. *CoRR*, abs/2307.14338, 2023a. doi:10.48550/arxiv.2307.14338.

Gorishniy, Y., Rubachev, I., Khrulkov, V., et al. Revisiting Deep Learning Models for Tabular Data, July 2023b.

Gorishniy, Y., Kotelnikov, A., and Babenko, A. Tabm: Advancing tabular deep learning with parameter-efficient ensembling. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. doi:10.48550/arXiv.2410.24210.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. doi:10.5555/3600270.3600307.

Gupta, M., Gallamoza, B., Cutrona, N., et al. An extensive data processing pipeline for MIMIC-IV. In Parziale, A., Agrawal, M., Joshi, S., Chen, I. Y., Tang, S., Oala, L., and Subbaswamy, A. (eds.), *Machine Learning for Health, ML4H 2022, 28 November 2022, New Orleans, Lousiana, USA & Virtual*, volume 193 of *Proceedings of Machine Learning Research*, pp. 311–325. Pmlr, 2022. doi:10.48550/arXiv.2204.13841.

Harutyunyan, H., Khachatrian, H., Kale, D. C., et al. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019. doi:10.1038/s41597-019-0103-9.

Hegselmann, S., von Arnim, G., Rheude, T., et al. Large language models are powerful EHR encoders. *CoRR*, abs/2502.17403, 2025. doi:10.48550/arxiv.2502.17403.

Hollmann, N., Müller, S., Eggensperger, K., et al. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second, September 2023.

Hollmann, N., Müller, S., Purucker, L., et al. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, January 2025. ISSN 1476-4687. doi:10.1038/s41586-024-08328-6.

Huang, X., Khetan, A., Cvitkovic, M., et al. TabTransformer: Tabular Data Modeling Using Contextual Embeddings. *arXiv preprint arXiv:2012.06678*, December 2020. doi:10.48550/arXiv.2012.06678.

Jensen, P. B., Jensen, L. J., and Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012. doi:https://doi.org/10.1038/nrg3208.

Johnson, A. E. W., Bulgarelli, L., Shen, L., et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, January 2023. ISSN 2052-4463. doi:10.1038/s41597-022-01899-x.

Karabacak, M., Jagtiani, P., Di, L., et al. Advancing precision prognostication in neuro-oncology: Machine learning models for data-driven personalized survival predictions in idh-wildtype glioblastoma. *Neuro-oncology advances*, 6(1):vdae096, 2024. doi:10.1093/noajnl/vdae096.

Ke, G., Meng, Q., Finley, T., et al. Lightgbm: A highly efficient gradient boosting decision tree. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Khan, W., Leem, S., See, K. B., et al. A comprehensive survey of foundation models in medicine. *CoRR*, abs/2406.10729, 2024. doi:10.48550/arxiv.2406.10729.

Knaus, W. and Lynn, J. Study to understand prognoses and preferences for outcomes and risks of treatment (support) and hospitalized elderly longitudinal project (help), 1989-1997, 2020.

Koornwinder, A., Zhang, Y., Ravindranath, R., et al. Multimodal artificial intelligence models predicting glaucoma progression using electronic health records and retinal nerve fiber layer scans. *Translational Vision Science & Technology*, 14(3):27–27, 2025. doi:10.1167/tvst.14.3.27.

Lee, S. A., Jain, S., Chen, A., et al. Emergency Department Decision Support using Clinical Pseudo-notes, April 2024.

Lewis, P., Perez, E., Piktus, A., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Li, Y., Rao, S., Solares, J. R. A., et al. BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, 10: 7155, April 2020. ISSN 2045-2322. doi:10.1038/s41598-020-62922-y.

Lipton, Z. C., Kale, D. C., Elkan, C., et al. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015. doi:10.48550/arXiv.1511.03677.

Ma, J., Thomas, V., Hosseinzadeh, R., et al. TabDPT: Scaling Tabular Foundation Models on Real Data, July 2025.

Mahdi, A. A. Diagnosing patient stroke status using modern ai after dataset balancing: A comprehensive comparative study. *Journal of Scientific Reports*, 9(1):219–228, 2025. doi:10.58970/jsr.1105.

McElfresh, D., Khandagale, S., Valverde, J., et al. When do neural nets outperform boosted trees on tabular data? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Nips '23, Red Hook, NY, USA, 2023. Curran Associates Inc. doi:10.5555/3666122.3669459.

McGilchrist, C. and Aisbett, C. Regression with frailty in survival analysis. *Biometrics*, pp. 461–466, 1991. doi:10.2307/2532138.

Miotto, R., Li, L., Kidd, B. A., et al. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6: 26094, 2016. doi:10.1038/srep26094.

Moor, M., Rieck, B., Horn, M., et al. Early prediction of sepsis in the icu using machine learning: a systematic review. *Frontiers in medicine*, 8:607952, 2021. doi:10.3389/fmed.2021.607952.

Moor, M., Banerjee, O., Abad, Z. S. H., et al. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, April 2023. ISSN 1476-4687. doi:10.1038/s41586-023-05881-4.

Müller, S., Hollmann, N., Arango, S. P., et al. Transformers can do bayesian inference. *arXiv preprint arXiv:2112.10510*, 2021. doi:10.48550/arXiv.2112.10510.

Neha, F., Bhati, D., and Shukla, D. K. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. *AI*, 6(9):226, 2025. doi:10.3390/ai6090226.

Nejjar, I., Ahmed, F., and Fink, O. IM-context: In-context learning for imbalanced regression tasks. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. doi:10.48550/arXiv.2405.18202.

Odgaard, M. F., Klein, K. V., Sillesen, M., et al. CORE-BEHRT: A carefully optimized and rigorously evaluated BEHRT. In Deshpande, K., Fiterau, M., Joshi, S., Lipton, Z., Ranganath, R., and Urteaga, I. n. (eds.), *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of *Proceedings of Machine Learning Research*. Pmlr, 16–17 Aug 2024. doi:10.48550/arXiv.2404.15201.

Pang, C., Park, J., Jiang, X., et al. Cehr-xgpt: A scalable multi-task foundation model for electronic health records. *arXiv preprint arXiv:2509.03643*, 2025. doi:10.48550/arXiv.2509.03643.

Park, S., Jeong, H., Kim, S. S., et al. Transformer-based risk estimation of hcc in patients with chronic hepatitis b. In *International Workshop on PRedictive Intelligence In MEdicine*, pp. 221–233. Springer, 2025. doi:10.1007/978-3-032-07904-6_20.

Peng, C., Yang, X., Chen, A., et al. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210, 2023. doi:10.1038/s41746-023-00958-w.

Pham, M.-K., Mai, T. T., Crane, M., et al. Forecasting Patient Early Readmission from Irish Hospital Discharge Records Using Conventional Machine Learning Models. *Diagnostics*, 14(21):2405, January 2024. ISSN 2075-4418. doi:10.3390/diagnostics14212405.

Pham, M.-K., Mai, T. T., Crane, M., et al. Explainable ai for infection prevention and control: modeling cpe acquisition and patient outcomes in an irish hospital with transformers. *BMC Medical Informatics and Decision Making*, 25(1):1–16, 2025a. doi:10.1186/s12911-025-03214-1.

Pham, M.-K., Mai, T. T., Dao, T. T. P., et al. Zero-shot Detection of Nosocomial Infections using Clinical Foundation Models on Structured Electronic Patient Records. *techRxiv*, July 2025b. doi:10.36227/techrxiv.175021755.55836808/v1.

Pollard, T. J., Johnson, A. E., Raffa, J. D., et al. The eicu collaborative research database, a freely available multicenter database for critical care research. *Scientific data*, 5(1):1–13, 2018. doi:10.1038/sdata.2018.178.

Prasetyo, S. Y., Nabiilah, G. Z., Sihotang, E. F. A., et al. Exploring cosine and euclidean metrics in k-nearest neighbors for breast cancer diagnosis. In *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*, pp. 330–335. Ieee, 2024. doi:10.1109/ICICoS62600.2024.10636883.

Prokhorenkova, L., Gusev, G., Vorobev, A., et al. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018. doi:10.5555/3327757.3327770.

Qin, J., Zhang, W., Su, R., et al. Retrieval & interaction machine for tabular data prediction. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Kdd '21, pp. 1379–1389. Acm, August 2021. doi:10.1145/3447548.3467216.

Qu, J., Holzmüller, D., Varoquaux, G., et al. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data, February 2025.

Rabby, F., Rahman, M. M., Das, R., et al. A parameter-efficient deep learning model for preeclampsia prediction using diverse datasets in low-resource settings. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, pp. 1–6, 2025. doi:10.1109/qpain66474.2025.11172026.

Rajkomar, A., Oren, E., Chen, K., et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018. doi:10.1038/s41746-018-0029-1.

Ramana, B. and Venkateswarlu, N. Ilpd (indian liver patient dataset). UCI Machine Learning Repository, 2022.

Rasmy, L., Xiang, Y., Xie, Z., et al. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):1–13, May 2021. ISSN 2398-6352. doi:10.1038/s41746-021-00455-y.

Renc, P., Jia, Y., Samir, A. E., et al. Zero shot health trajectory prediction using transformer. *npj Digital Medicine*, 7(1):256, September 2024. ISSN 2398-6352. doi:10.1038/s41746-024-01235-0.

S, N. V., T, M., and Devi, S. P. Integrating machine learning and deep learning approaches for accurate cardiovascular disease prediction from electronic health records. In *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, pp. 1090–1096, 2025. doi:10.1109/icmsci62561.2025.10894568.

Saito, T. and Rehmsmeier, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015. doi:doi.org/10.1371/journal.pone.0118432.

Sarwar, T., Seifollahi, S., Chan, J., et al. The secondary use of electronic health records for data mining: Data characteristics and challenges. *ACM Comput. Surv.*, 55(2), January 2022. ISSN 0360-0300. doi:10.1145/3490234.

Scheurwegs, E., Cule, B., Luyckx, K., et al. Selecting relevant features from the electronic health record for clinical code prediction. *Journal of biomedical informatics*, 74: 92–103, 2017. doi:10.1016/j.jbi.2017.09.004.

Shickel, B., Tighe, P. J., Bihorac, A., et al. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017. doi:10.1109/jbhi.2017.2767063.

Singhal, K., Tu, T., Gottweis, J., et al. Toward expert-level medical question answering with large language models, 2025.

Sumon, M. S. I., Islam, M. S. B., Rahman, M. S., et al. Cardiotabnet: a novel hybrid transformer model for heart disease prediction using tabular medical data. *Health Information Science and Systems*, 13(1):44, 2025. doi:10.1007/s13755-025-00361-7.

Tasci, E., Camphausen, K., Krauze, A. V., et al. Glioma grading clinical and mutation features. UCI Machine Learning Repository, 2022.

Thomas, V., Ma, J., Hosseinzadeh, R., et al. Retrieval & fine-tuning for in-context tabular models. *Advances in Neural Information Processing Systems*, 37:108439–108467, 2024. doi:10.52202/079017-3442.

Tran, V. Q. and Byeon, H. Predicting dementia in parkinson's disease on a small tabular dataset using hybrid lightgbm–tabpfn and shap. *Digital Health*, 10:20552076241272585, 2024. doi:10.1177/20552076241272585.

Tsanas, A., Little, M., McSharry, P., et al. Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests. *Nature Precedings*, pp. 1–1, 2009. doi:10.1038/npre.2009.3920.1.

Van Breugel, B. and Van Der Schaar, M. Position: why tabular foundation models should be a research priority, 2024.

Walonoski, J., Kramer, M., Nichols, J., et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, March 2018. ISSN 1527-974x. doi:10.1093/jamia/ocx079.

Wang, N., Wang, M., Zhou, Y., et al. Sequential data–based patient similarity framework for patient outcome prediction: Algorithm development. *J Med Internet Res*, 24(1):e30720, January 2022. ISSN 1438-8871. doi:10.2196/30720.

Wen, X., Zheng, S., Xu, Z., et al. Scalable in-context learning on tabular data via retrieval-augmented large language models. *arXiv preprint arXiv:2502.03147*, 2025. doi:10.48550/arXiv.2502.03147.

Wornow, M., Xu, Y., Thapa, R., et al. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1): 1–10, July 2023. ISSN 2398-6352. doi:10.1038/s41746-023-00879-8.

Wornow, M., Bedi, S., Hernandez, M. A. F., et al. Context Clues: Evaluating Long Context Models for Clinical Prediction Tasks on EHRs, March 2025.

Wu, S., Xue, L., Legido-Quigley, H., et al. Understanding factors influencing the length of hospital stay among non-severe covid-19 patients: A retrospective cohort study in a fangcang shelter hospital. *Plos one*, 15(10):e0240959, 2020. doi:10.1371/journal.pone.0240959.

Wu, Z., Dadu, A., Tustison, N., et al. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*, 2024.

Xu, D., Cirit, O., Asadi, R., et al. Mixture of In-Context Prompters for Tabular PFNs, May 2024.

Xue, Z., Liang, Z., Rajaraman, S., et al. Detecting oral cancer using tabular deep learning. In *2025 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pp. 1–6, 2025. doi:10.1109/coins65080.2025.11125786.

Yang, R., Ning, Y., Keppo, E., et al. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems*, 2(1):2, 2025. doi:10.1038/s44401-024-00004-1.

Ye, H.-J., Yin, H.-H., Zhan, D.-C., et al. Revisiting nearest neighbor for tabular data: A deep tabular baseline two decades later. *arXiv preprint arXiv:2407.03257*, 2024. doi:10.48550/arXiv.2407.03257.

Zheng, L., Li, N., Chen, X., et al. Dense representation learning and retrieval for tabular data prediction. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Kdd '23, pp. 3559–3569. Acm, August 2023. doi:10.1145/3580305.3599305.

Zhu, M., Liu, Y., Luo, Z., et al. Bridging data gaps of rare conditions in icu: a multi-disease adaptation approach for clinical prediction. *npj Digital Medicine*, 2026. doi:10.1038/s41746-025-02176-y.

Zhu, Y., Wang, Z., Gao, J., et al. Prompting Large Language Models for Zero-Shot Clinical Prediction with Structured Longitudinal Electronic Health Record Data, February 2024.

A. Appendix

A.1. Performance on Synthetic EHR Data.

We evaluated representative models on fully synthetic EHR datasets generated using Synthea (Walonoski et al., 2018), an open-source patient population simulator that produces realistic yet fully synthetic longitudinal health records based on rule-driven disease modules and publicly available epidemiological statistics. The evaluated tasks include stroke and lung cancer prediction (Chen & Chen, 2022). Across all evaluated methods—logistic regression, XGBoost, TabPFN and TabDPT—predictive performance approached ceiling levels, as seen in Table A1. This behavior is expected given the rule-based generative process underlying Synthea, in which disease progression and outcomes are governed by explicit state-transition logic and probabilistic modules. As a result, target labels are often directly inferable from a small subset of observable features, limiting the need for complex representation learning or retrieval-based inference.

Model	**Lung Cancer ↑**	**Stroke ↑**
Classical and Ensemble Models		
Logistic Regression	$0.993_{\pm 0.00}$	$0.992_{\pm 0.00}$
XGBoost	$0.999_{\pm 0.00}$	$0.999_{\pm 0.00}$
Tabular In-Context Learning Models		
TabPFN	$1.0_{\pm 0.00}$	$0.999_{\pm 0.00}$
TabDPT	$0.992_{\pm 0.00}$	$0.978_{\pm 0.00}$

Table A1. AUROC performance on fully synthetic EHR data generated using Synthea for two binary prediction tasks: lung cancer and stroke. Performance across model classes approaches ceiling levels due to the rule-based generative process underlying Synthea. Results are intended as an upper-bound sanity check rather than evidence of real-world generalization.

Within this saturated regime, PFN-based models achieve performance comparable to strong classical baselines without task-specific retraining or feature engineering. While this does not indicate superiority over established methods, it demonstrates that PFNs are compatible with simulator-driven clinical data and do not exhibit inductive bias mismatches that hinder learning of disease logic. Importantly, this compatibility suggests that the limitations of PFN-based and RA-TICL methods observed on real-world EHR datasets are unlikely to stem from representational capacity alone, but instead arise from challenges inherent to clinical data as described in the main text.

A.2. Efficiency Trade-offs

Figure A1 summarizes the trade-offs between computational efficiency and predictive performance across model classes on large-scale EHR datasets. We report relative ranking scores (higher is better) against measured training and inference time. Hyperparameter tuning time is excluded, and for models requiring supervised training, training time reflects convergence with early stopping. For PFN-based and other ICL models, which do not require task-specific training, training time is effectively negligible.

Training Efficiency. In terms of training cost, ICL-based models exhibit a distinct advantage. TabPFN and TabDPT incur no dataset-specific training time, yet achieve mid-to-high performance rankings, placing them favorably among all evaluated models. This contrasts with classical k-nearest neighbors, which is also training-free but exhibits substantially lower predictive rankings. Gradient-boosted models (e.g., XGBoost, LightGBM) occupy an intermediate regime, requiring modest training time while achieving competitive performance. In contrast, deep tabular and retrieval-based models such as ModernNCA and TabR incur significantly higher training costs while achieving lower or comparable rankings. The highest-ranked deep tabular models (e.g., ExcelFormer, Trompt) require the great amount of training time and computational resources, reflecting the cost of end-to-end representation learning on large EHR cohorts. Our AWARE adapters are the most notable where they achieve both highest rankings while being training time-efficient.

Inference Efficiency. Inference-time costs reveal a different pattern. Classical linear and tree-based models remain the fastest at inference, with per-patient prediction times substantially lower than all other model classes. Among neural models, shallow architectures such as MLPs and ResNet-style tabular models achieve moderate inference efficiency. PFN-based

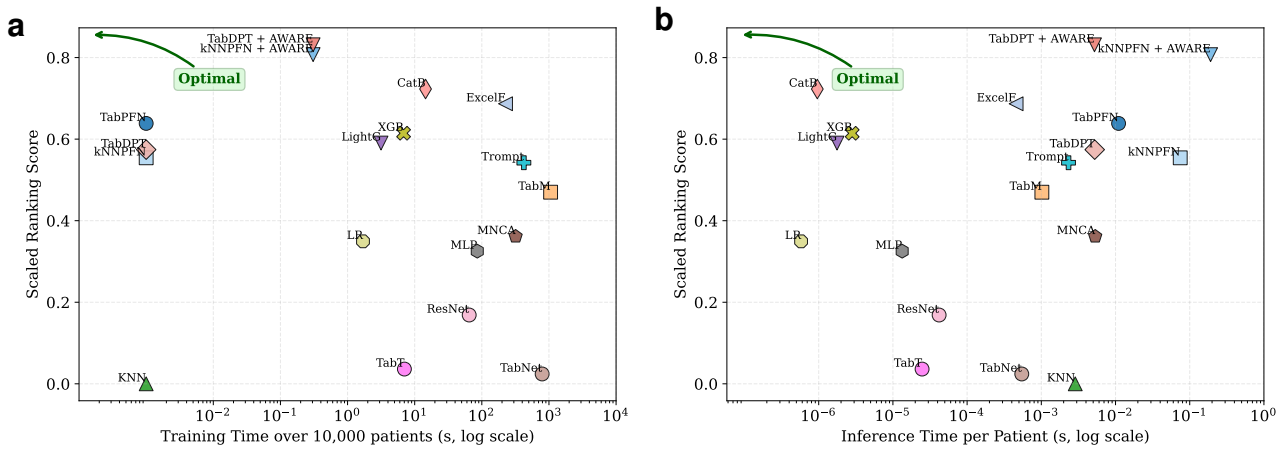


Figure A1. Training and inference efficiency trade-offs across model classes on large-scale EHR datasets. The left panel reports relative predictive ranking versus training time (measured until convergence with early stopping), while the right panel reports ranking versus per-patient inference time. Hyperparameter tuning and pretraining costs are excluded. Models based on in-context learning incur no task-specific training cost.

ICL models, including TabPFN and TabDPT, exhibit inference times comparable to other transformer-based tabular models and retrieval-augmented approaches. Retrieval-based variants, particularly kNNPFN, incur the highest inference cost due to explicit neighbor search and context construction, especially with AWARE incoporated. Importantly, despite slower inference relative to classical models, PFN-based approaches remain competitive with deep tabular models that require both training and inference overhead.

Implications for Clinical Use. Taken together, these results highlight a key efficiency trade-off for in-context learning in EHR prediction. PFN-based models offer a favorable balance between predictive performance and training efficiency by eliminating the need for task-specific retraining, which can be advantageous in settings with limited computational resources or rapidly changing clinical tasks. However, this benefit is partially offset by increased inference-time cost, particularly for retrieval-augmented variants. These findings suggest that ICL-based approaches are best suited for scenarios where retraining is infeasible or costly, while inference latency remains acceptable, rather than as drop-in replacements for lightweight classical models in latency-critical clinical workflows.

A.3. Model Descriptions

This section provides brief descriptions of the models included in our benchmark, focusing on their architectural characteristics and relevance to EHR prediction tasks.

TabTransformer (Huang et al., 2020) models categorical features using contextual embeddings derived from Transformer layers, enabling richer interactions between categorical variables. It has been shown to outperform standard multilayer perceptrons (MLPs) on tabular data and to exhibit robustness to noisy or missing categorical inputs.

TabNet (Arik & Pfister, 2021) is a deep learning architecture for tabular data that performs instance-wise feature selection through a sequential attention mechanism. By selecting a sparse subset of features at each decision step, TabNet provides a degree of interpretability while maintaining competitive predictive performance.

ExcelFormer (Chen et al., 2024)is a Transformer-based architecture designed for tabular data, incorporating a semi-permeable attention mechanism and gated linear units. It also introduces interpolation-based data augmentation strategies (Hid-Mix and Feat-Mix) to improve robustness under irregular target functions.

Trompt (Chen et al., 2023) is a prompt-inspired neural architecture for tabular learning that derives sample-specific feature importance through learned prompts. By dynamically adjusting feature relevance at inference time, Trompt aims to bridge

the performance gap between deep tabular models and tree-based methods.

TabM (Gorishniy et al., 2025) is a parameter-efficient deep tabular model that ensembles multiple MLP submodels with shared weights. This design improves training efficiency and generalization while mitigating issues such as dead neurons commonly observed in deep networks.

ModernNCA (Ye et al., 2024) revisits classical Neighborhood Components Analysis (NCA) and integrates modern optimization and representation learning techniques to support differentiable nearest-neighbor retrieval for tabular prediction tasks.

TabPFN (Hollmann et al., 2023) is a prior-fitted network pretrained on a large corpus of synthetically generated tabular datasets derived from causal models. It performs classification via in-context learning, enabling inference-time adaptation without task-specific fine-tuning.

kNNPFN (Thomas et al., 2024) (also referred to as TabPFN-kNN) augments the TabPFN prior-fitted architecture with an explicit k-nearest-neighbor retrieval step at inference time. For each query instance, a subset of similar training examples is selected using a distance metric in feature space, and only this retrieved context is provided to the PFN model. This strategy reduces the computational burden associated with conditioning on large training sets while introducing a similarity-based inductive bias. TabPFN-kNN thus represents a simple yet effective retrieval-augmented extension of PFN-based in-context learning, enabling scalable inference-time adaptation on larger tabular datasets.

TabDPT (Ma et al., 2025) is a retrieval-augmented tabular in-context learning model that extends the TabPFN paradigm to larger and more heterogeneous datasets. Instead of conditioning on the full training set, TabDPT constructs a query-specific context via retrieval and processes the retrieved examples jointly with the query through a Transformer encoder. By combining pretrained PFN initialization with explicit context construction, TabDPT enables scalable inference-time adaptation while maintaining the feature-order invariance and amortized Bayesian inference properties of PFN-based models.

A.4. Dataset Preprocessing and Feature Extraction

We conduct experiments using two publicly available ICU EHR datasets: MIMIC-IV (v3.1) and eICU-CRD (v2.0). For both datasets, analyses are restricted to adult patients and constructed at the ICU-stay level.

In MIMIC-IV, patients aged $\geq$ 18 years are included if ICU stays satisfy duration requirements for hospital-acquired infection prediction. To ensure infections are acquired during hospitalization rather than present at admission, only ICU stays of at least 48 hours are considered. A 24-hour observation window is used to extract patient features, followed by a 24-hour prediction gap, after which the model predicts the onset of hospital-acquired infections. Outcomes include three clinically relevant infections commonly monitored in critical care settings: sepsis, urinary tract infection (UTI), and ventilator-associated pneumonia (VAP). These outcomes are identified using diagnosis codes and clinical event records mapped to standardized infection definitions where available.

In eICU-CRD, adult patients aged 18–89 years with ICU stays between 12 hours and 10 days are included. Clinical events recorded during the first 12 hours of ICU admission are used as the observation window. Prediction targets similarly focus on hospital-acquired infections, including sepsis and urinary tract infection, defined using diagnosis codes, treatment indicators, and infection-related clinical variables when available in the dataset.

Structured features are extracted from multiple EHR sources, including diagnoses, treatments or procedures, medications, laboratory measurements, vital signs, and physiological variables. In MIMIC-IV, diagnosis and procedure codes are encoded using frequency counts, medications are aggregated by mean administration rate and dosage, and charted measurements are treated as continuous variables summarized within fixed temporal buckets (e.g., 2-hour intervals). In eICU-CRD, diagnosis and treatment information is parsed from hierarchical strings and standardized across hospitals. Laboratory measurements and vital signs are normalized and summarized using minimum, maximum, mean, and standard deviation over the observation window, with additional physiological variables derived from APACHE-related features.

Variable-length time series are converted into fixed-length representations through temporal aggregation. Missing continuous measurements in MIMIC-IV are handled via forward- and backward-filling followed by median imputation. In eICU-CRD, missing categorical variables are imputed using the mode, while continuous variables are normalized and missing values are

set to zero after normalization. The resulting datasets consist of fixed-length feature matrices with one row per ICU stay, enabling consistent input representations across all evaluated models.

Dataset Name	#Rows	#Features	Task	Imbalance Ratio	Cohort	Feature Types
Synthetic EHR datasets						
Synthea-Stroke (Chen & Chen, 2022)	16,055	793	Binary	–	Multi-domain	DEM, VIT, LAB, CODE
Synthea-LungCancer (Chen & Chen, 2022)	4,384	773	Binary	–	Oncology	DEM, VIT, LAB, CODE
Small-to-medium real-world EHR datasets						
Heart Failure (Platelets Counts) (Chicco & Jurman, 2020a)	299	12	Regression	–	Cardiovascular	DEM, LAB
OXF – Parkinsons Telemonitoring (Tsanas et al., 2009)	5,875	22	Regression	–	Neurological	DEM, PHYS, TMP
TunedIT – ICU	200	19	Binary	–	ICU	DEM, VIT, LAB
SUPPORT2 (Knaus & Lynn, 2020)	9,105	46	Binary	2.13	Multi-domain	DEM, VIT, LAB, ADMIN
Sepsis Survival Minimal (Chicco & Jurman, 2020b)	110,341	4	Binary	12.57	Infectious	DEM
Differentiated Thyroid Cancer (Borzooei & Tarokhian, 2023)	383	16	Binary	2.55	Oncology	DEM, LAB
Glioma Grading (Clinical + Mutation) (Tasci et al., 2022)	839	23	Binary	1.38	Oncology	DEM, GEN
AIDS Classification 50,000	50,000	23	Binary	2.22	Infectious	DEM, LAB
Indian Liver Patient (Ramana & Venkateswarlu, 2022)	583	11	Binary	2.54	Metabolic	DEM, LAB
COVID-19 Hospital Treatment Plan (Wu et al., 2020)	318,438	18	Multiclass	31.69	Infectious	DEM, ADMIN
Diabetes130US (Clore et al., 2014)	101,766	49	Binary	4.83	Metabolic	DEM, LAB
Kidney (McGilchrist & Aisbett, 1991)	76	6	Binary	1.12	Metabolic	DEM, LAB
Large-scale longitudinal real-world EHR datasets						
MIMIC-IV (SEPSIS) (Johnson et al., 2023)	35,885	3,400	Binary	8.408	Infectious	DEM, VIT, LAB, TMP
MIMIC-IV (UTI) (Johnson et al., 2023)	24,145	3,440	Binary	5.226	Infectious	DEM, VIT, LAB, TMP
MIMIC-IV (VAP) (Johnson et al., 2023)	24,495	3,440	Binary	10.38	Infectious	DEM, VIT, LAB, TMP
eICU (SEPSIS) (Pollard et al., 2018)	120,442	655	Binary	14.91	Infectious	DEM, VIT, LAB, TMP
eICU (UTI) (Pollard et al., 2018)	124,828	655	Binary	70.65	Infectious	DEM, VIT, LAB, TMP
HIPE (CPE) (Pham et al., 2025a)	45,333	28,566	Binary	434.89	Infectious	DEM, CODE, ADMIN, TMP
HIPE (RDMT) (Pham et al., 2024)	78,142	28,779	Binary	7.48	ICU	DEM, CODE, ADMIN, TMP

Table A2. Summary of EHR datasets used in this study, spanning diverse clinical cohorts, task types, and feature modalities. Feature type abbreviations: DEM = demographics; VIT = vital signs; LAB = laboratory results; CODE = medical codes / diagnoses; GEN = genetic / molecular features; PHYS = physiological signals; ADMIN = administrative variables; TMP = temporal features.

A.5. Supplementary Benchmark Results

Model	MIMIC-IV						eICU			
	SEPSIS		VAP		UTI		SEPSIS		UTI	
	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
kNN Backbone										
KNN (baseline)	$0.866_{\pm 0.000}$	$0.446_{\pm 0.000}$	$0.774_{\pm 0.000}$	$0.235_{\pm 0.000}$	$0.637_{\pm 0.000}$	$0.249_{\pm 0.000}$	$0.869_{\pm 0.000}$	$0.348_{\pm 0.000}$	$0.815_{\pm 0.000}$	$0.064_{\pm 0.000}$
+ Attention-Weighted Retrieval	$0.860_{\pm 0.013}$	$0.429_{\pm 0.038}$	$0.775_{\pm 0.002}$	$0.268_{\pm 0.005}$	$0.634_{\pm 0.002}$	$0.241_{\pm 0.001}$	$0.889_{\pm 0.003}$	$0.402_{\pm 0.020}$	$0.824_{\pm 0.008}$	$0.074_{\pm 0.005}$
+ SNNL	$\mathbf{0.893}_{\pm 0.017}$	$\mathbf{0.500}_{\pm 0.042}$	$0.780_{\pm 0.012}$	$0.256_{\pm 0.012}$	$0.647_{\pm 0.004}$	$0.246_{\pm 0.007}$	$0.889_{\pm 0.013}$	$0.407_{\pm 0.039}$	$0.761_{\pm 0.023}$	$0.054_{\pm 0.008}$
+ Balanced Sampling	$0.880_{\pm 0.020}$	$0.461_{\pm 0.024}$	$0.780_{\pm 0.012}$	$0.265_{\pm 0.008}$	$0.646_{\pm 0.015}$	$0.243_{\pm 0.011}$	$\mathbf{0.968}_{\pm 0.001}$	$\mathbf{0.708}_{\pm 0.008}$	$0.875_{\pm 0.021}$	$0.112_{\pm 0.013}$
+ 5-fold CV Ensemble	$0.886_{\pm 0.007}$	$0.481_{\pm 0.014}$	$\mathbf{0.785}_{\pm 0.010}$	$\mathbf{0.274}_{\pm 0.009}$	$\mathbf{0.659}_{\pm 0.001}$	$\mathbf{0.255}_{\pm 0.005}$	$0.966_{\pm 0.002}$	$0.671_{\pm 0.012}$	$\mathbf{0.908}_{\pm 0.009}$	$\mathbf{0.122}_{\pm 0.004}$
kNNPFN Backbone										
kNNPFN (baseline)	$0.908_{\pm 0.000}$	$0.545_{\pm 0.000}$	$0.786_{\pm 0.000}$	$0.274_{\pm 0.000}$	$0.662_{\pm 0.000}$	$0.254_{\pm 0.000}$	$0.970_{\pm 0.000}$	$0.700_{\pm 0.000}$	$0.895_{\pm 0.000}$	$0.100_{\pm 0.000}$
+ Attention Weighting	$0.907_{\pm 0.003}$	$0.541_{\pm 0.009}$	$0.788_{\pm 0.004}$	$0.274_{\pm 0.004}$	$0.658_{\pm 0.001}$	$0.251_{\pm 0.002}$	$0.978_{\pm 0.001}$	$0.733_{\pm 0.007}$	$0.908_{\pm 0.002}$	$0.118_{\pm 0.001}$
+ SNNL	$0.909_{\pm 0.002}$	$0.547_{\pm 0.006}$	$0.784_{\pm 0.003}$	$0.271_{\pm 0.006}$	$0.661_{\pm 0.004}$	$0.252_{\pm 0.003}$	$0.974_{\pm 0.002}$	$0.733_{\pm 0.007}$	$0.886_{\pm 0.028}$	$0.110_{\pm 0.019}$
+ Balanced Sampling	$0.910_{\pm 0.004}$	$0.542_{\pm 0.011}$	$0.786_{\pm 0.007}$	$0.268_{\pm 0.000}$	$0.665_{\pm 0.001}$	$\mathbf{0.261}_{\pm 0.007}$	$0.977_{\pm 0.001}$	$\mathbf{0.744}_{\pm 0.006}$	$0.897_{\pm 0.032}$	$0.125_{\pm 0.022}$
+ 5-fold CV Ensemble	$0.911_{\pm 0.001}$	$0.545_{\pm 0.005}$	$0.789_{\pm 0.004}$	$0.275_{\pm 0.006}$	$\mathbf{0.666}_{\pm 0.001}$	$0.257_{\pm 0.002}$	$\mathbf{0.979}_{\pm 0.001}$	$0.738_{\pm 0.003}$	$0.921_{\pm 0.008}$	$\mathbf{0.142}_{\pm 0.003}$
+ Adapter Finetuning	$\mathbf{0.918}_{\pm 0.001}$	$\mathbf{0.562}_{\pm 0.004}$	$\mathbf{0.799}_{\pm 0.000}$	$\mathbf{0.279}_{\pm 0.007}$	$\mathbf{0.666}_{\pm 0.001}$	$0.257_{\pm 0.002}$	$\mathbf{0.979}_{\pm 0.001}$	$0.739_{\pm 0.003}$	$\mathbf{0.926}_{\pm 0.002}$	$\mathbf{0.142}_{\pm 0.003}$
TabDPT Backbone										
TabDPT (baseline)	$0.905_{\pm 0.000}$	$0.528_{\pm 0.000}$	$0.784_{\pm 0.000}$	$0.274_{\pm 0.000}$	$0.658_{\pm 0.000}$	$0.257_{\pm 0.000}$	$0.961_{\pm 0.000}$	$0.659_{\pm 0.000}$	$0.882_{\pm 0.000}$	$0.103_{\pm 0.000}$
+ Attention Weighting	$0.908_{\pm 0.002}$	$0.540_{\pm 0.006}$	$0.788_{\pm 0.002}$	$0.279_{\pm 0.001}$	$0.659_{\pm 0.001}$	$0.254_{\pm 0.001}$	$0.968_{\pm 0.000}$	$0.684_{\pm 0.004}$	$0.891_{\pm 0.001}$	$0.112_{\pm 0.003}$
+ SNNL	$0.907_{\pm 0.001}$	$0.539_{\pm 0.007}$	$0.788_{\pm 0.003}$	$\mathbf{0.286}_{\pm 0.007}$	$0.663_{\pm 0.005}$	$0.258_{\pm 0.005}$	$0.969_{\pm 0.002}$	$0.693_{\pm 0.012}$	$0.846_{\pm 0.040}$	$0.097_{\pm 0.017}$
+ Balanced Sampling	$0.906_{\pm 0.001}$	$0.538_{\pm 0.006}$	$0.784_{\pm 0.010}$	$0.278_{\pm 0.010}$	$\mathbf{0.664}_{\pm 0.005}$	$\mathbf{0.261}_{\pm 0.007}$	$0.968_{\pm 0.004}$	$\mathbf{0.717}_{\pm 0.007}$	$0.881_{\pm 0.011}$	$0.129_{\pm 0.004}$
+ 5-fold CV Ensemble	$0.910_{\pm 0.001}$	$0.547_{\pm 0.004}$	$0.790_{\pm 0.003}$	$0.285_{\pm 0.002}$	$0.661_{\pm 0.000}$	$0.254_{\pm 0.003}$	$\mathbf{0.972}_{\pm 0.002}$	$0.699_{\pm 0.006}$	$0.910_{\pm 0.010}$	$0.135_{\pm 0.007}$
+ Adapter Finetuning	$\mathbf{0.916}_{\pm 0.003}$	$\mathbf{0.565}_{\pm 0.006}$	$\mathbf{0.791}_{\pm 0.003}$	$\mathbf{0.286}_{\pm 0.003}$	$\mathbf{0.664}_{\pm 0.001}$	$0.264_{\pm 0.003}$	$\mathbf{0.972}_{\pm 0.002}$	$0.700_{\pm 0.008}$	$\mathbf{0.912}_{\pm 0.013}$	$\mathbf{0.139}_{\pm 0.002}$

Table A3. Ablation study of AWARE and subsequent enhancements on large-scale ICU EHR tasks (MIMIC-IV and eICU). Each row incrementally adds one component to the backbone, isolating its contribution to predictive performance. Results are reported as AUROC and AUPRC (mean ± std over 3 seeds) across six clinical prediction tasks.

Model	Regression			Classification								
	HF ↓	OXF-PT ↓	TIT ↓	SUPPORT2 ↑	SSMI ↑	DTC ↑	GGCM ↑	AIDS ↑	ILP ↑	COVID19 ↑	Diabetes ↑	Kidney ↑
Classical and Ensemble Models												
LR	$1.360_{\pm 0.000}$	$0.413_{\pm 0.000}$	$1.070_{\pm 0.000}$	$0.952_{\pm 0.000}$	$0.692_{\pm 0.000}$	$0.984_{\pm 0.000}$	$0.880_{\pm 0.002}$	$0.710_{\pm 0.000}$	$0.755_{\pm 0.002}$	$0.849_{\pm 0.000}$	$0.850_{\pm 0.000}$	$0.812_{\pm 0.000}$
KNN	$1.357_{\pm 0.010}$	$0.238_{\pm 0.007}$	$1.119_{\pm 0.000}$	$0.884_{\pm 0.021}$	$0.567_{\pm 0.027}$	$0.938_{\pm 0.020}$	$0.821_{\pm 0.022}$	$0.648_{\pm 0.008}$	$0.696_{\pm 0.005}$	$0.843_{\pm 0.011}$	$0.837_{\pm 0.020}$	$0.833_{\pm 0.072}$
CatB	$1.367_{\pm 0.019}$	$0.400_{\pm 0.159}$	$1.139_{\pm 0.027}$	$0.982_{\pm 0.001}$	$0.691_{\pm 0.001}$	$0.987_{\pm 0.003}$	$0.848_{\pm 0.007}$	$0.713_{\pm 0.007}$	$0.729_{\pm 0.024}$	$0.854_{\pm 0.001}$	$0.865_{\pm 0.003}$	$0.792_{\pm 0.157}$
LightG	$1.345_{\pm 0.020}$	$0.514_{\pm 0.192}$	$1.025_{\pm 0.018}$	$0.980_{\pm 0.002}$	$0.691_{\pm 0.001}$	$\mathbf{0.990}_{\pm 0.003}$	$0.873_{\pm 0.013}$	$0.714_{\pm 0.003}$	$0.730_{\pm 0.029}$	$0.854_{\pm 0.001}$	$0.860_{\pm 0.002}$	$0.938_{\pm 0.062}$
XGB	$1.387_{\pm 0.040}$	$0.675_{\pm 0.000}$	$1.237_{\pm 0.057}$	$0.980_{\pm 0.002}$	$0.691_{\pm 0.000}$	$0.944_{\pm 0.070}$	$0.866_{\pm 0.015}$	$0.706_{\pm 0.001}$	$0.714_{\pm 0.035}$	$0.850_{\pm 0.001}$	$0.863_{\pm 0.002}$	$0.906_{\pm 0.031}$
MLP	$1.336_{\pm 0.024}$	$0.323_{\pm 0.007}$	$1.103_{\pm 0.062}$	$0.969_{\pm 0.002}$	$0.675_{\pm 0.013}$	$0.909_{\pm 0.025}$	$0.873_{\pm 0.009}$	$0.705_{\pm 0.004}$	$0.637_{\pm 0.142}$	$0.854_{\pm 0.005}$	$0.856_{\pm 0.005}$	$0.562_{\pm 0.062}$
Deep Tabular Models												
ResNet	$1.572_{\pm 0.000}$	$0.307_{\pm 0.000}$	$1.424_{\pm 0.000}$	$0.972_{\pm 0.000}$	$0.692_{\pm 0.000}$	$0.914_{\pm 0.039}$	$0.845_{\pm 0.014}$	$0.707_{\pm 0.006}$	$0.678_{\pm 0.039}$	$0.852_{\pm 0.007}$	$0.857_{\pm 0.003}$	$0.854_{\pm 0.036}$
TabM	$1.755_{\pm 0.000}$	$0.237_{\pm 0.000}$	$1.094_{\pm 0.047}$	$0.975_{\pm 0.002}$	$0.692_{\pm 0.001}$	$0.929_{\pm 0.009}$	$\mathbf{0.892}_{\pm 0.027}$	$0.716_{\pm 0.000}$	$0.698_{\pm 0.032}$	$0.852_{\pm 0.009}$	$0.860_{\pm 0.010}$	$0.833_{\pm 0.036}$
ExcelF	$1.357_{\pm 0.039}$	$0.373_{\pm 0.035}$	$1.166_{\pm 0.162}$	$0.979_{\pm 0.004}$	$0.692_{\pm 0.000}$	$0.869_{\pm 0.145}$	$0.874_{\pm 0.017}$	$0.716_{\pm 0.002}$	$0.684_{\pm 0.065}$	$0.856_{\pm 0.006}$	$0.862_{\pm 0.004}$	$0.729_{\pm 0.157}$
TabNet	$1.434_{\pm 0.070}$	$0.379_{\pm 0.047}$	$3.721_{\pm 3.349}$	$0.972_{\pm 0.002}$	$0.684_{\pm 0.001}$	$0.690_{\pm 0.101}$	$0.718_{\pm 0.116}$	$0.709_{\pm 0.003}$	$0.643_{\pm 0.038}$	$0.849_{\pm 0.008}$	$0.859_{\pm 0.003}$	$0.667_{\pm 0.295}$
TabT	$1.353_{\pm 0.004}$	$0.704_{\pm 0.002}$	$1.251_{\pm 0.003}$	$0.962_{\pm 0.004}$	$0.662_{\pm 0.001}$	$0.920_{\pm 0.011}$	$0.844_{\pm 0.027}$	$0.698_{\pm 0.015}$	$0.687_{\pm 0.015}$	$0.849_{\pm 0.007}$	$0.817_{\pm 0.004}$	$0.771_{\pm 0.036}$
Trompt	$1.370_{\pm 0.008}$	$0.711_{\pm 0.004}$	$1.258_{\pm 0.010}$	$0.982_{\pm 0.001}$	$0.692_{\pm 0.001}$	$0.948_{\pm 0.012}$	$0.883_{\pm 0.012}$	$\mathbf{0.718}_{\pm 0.001}$	$0.734_{\pm 0.040}$	$0.857_{\pm 0.003}$	$0.859_{\pm 0.006}$	$0.917_{\pm 0.144}$
MNCA	$1.396_{\pm 0.028}$	$0.291_{\pm 0.000}$	$1.063_{\pm 0.077}$	$0.974_{\pm 0.002}$	$0.692_{\pm 0.000}$	$0.937_{\pm 0.018}$	$0.881_{\pm 0.004}$	$0.710_{\pm 0.003}$	$0.716_{\pm 0.006}$	$\mathbf{0.859}_{\pm 0.001}$	$0.863_{\pm 0.002}$	$0.958_{\pm 0.036}$
Tabular In-Context Learning Models												
TabPFN	$\mathbf{1.335}_{\pm 0.003}$	$0.113_{\pm 0.011}$	$1.223_{\pm 0.003}$	$\mathbf{0.984}_{\pm 0.000}$	$\mathbf{0.693}_{\pm 0.000}$	$0.984_{\pm 0.000}$	$0.883_{\pm 0.000}$	$0.713_{\pm 0.000}$	$0.750_{\pm 0.000}$	$0.811_{\pm 0.010}$	$0.842_{\pm 0.006}$	$\mathbf{1.000}_{\pm 0.000}$
kNNPFN	$1.358_{\pm 0.003}$	$\mathbf{0.040}_{\pm 0.001}$	$1.146_{\pm 0.027}$	$\mathbf{0.984}_{\pm 0.000}$	$0.686_{\pm 0.000}$	$\mathbf{0.990}_{\pm 0.000}$	$0.881_{\pm 0.001}$	$0.713_{\pm 0.000}$	$0.755_{\pm 0.004}$	$0.853_{\pm 0.002}$	$\mathbf{0.866}_{\pm 0.001}$	$0.958_{\pm 0.036}$
TabDPT	$1.353_{\pm 0.000}$	$0.626_{\pm 0.000}$	$1.230_{\pm 0.000}$	$\mathbf{0.984}_{\pm 0.000}$	$0.690_{\pm 0.000}$	$0.981_{\pm 0.000}$	$0.883_{\pm 0.003}$	$0.705_{\pm 0.000}$	$0.772_{\pm 0.003}$	$\mathbf{0.859}_{\pm 0.000}$	$0.862_{\pm 0.001}$	$\mathbf{1.000}_{\pm 0.000}$
Ours (Tabular In-Context Learning Models + AWARE)												
KNN	$1.394_{\pm 0.000}$	$0.417_{\pm 0.023}$	$1.253_{\pm 0.000}$	$0.931_{\pm 0.018}$	$0.689_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.709_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.847_{\pm 0.005}$	$0.722_{\pm 0.326}$	$0.500_{\pm 0.000}$
kNNPFN	$1.364_{\pm 0.001}$	$0.070_{\pm 0.007}$	$\mathbf{1.001}_{\pm 0.003}$	$\mathbf{0.984}_{\pm 0.004}$	$0.684_{\pm 0.000}$	$0.984_{\pm 0.000}$	$0.881_{\pm 0.000}$	$0.713_{\pm 0.000}$	$\mathbf{0.783}_{\pm 0.000}$	$0.852_{\pm 0.001}$	$0.846_{\pm 0.002}$	$0.812_{\pm 0.000}$
TabDPT	$1.346_{\pm 0.000}$	$0.107_{\pm 0.003}$	$1.043_{\pm 0.000}$	$0.981_{\pm 0.001}$	$0.691_{\pm 0.000}$	$0.981_{\pm 0.000}$	$0.886_{\pm 0.000}$	$0.711_{\pm 0.001}$	$0.749_{\pm 0.000}$	$0.856_{\pm 0.002}$	$0.849_{\pm 0.006}$	$\mathbf{1.000}_{\pm 0.000}$

Table A4. Benchmark coverage of models across small-to-medium scale EHR datasets, sourced from OpenML and UCI repositories. F1 score is reported for Diabetes and COVID19 while AUROC is reported for others.

Model	MIMIC-IV						eICU			
	SEPSIS		VAP		UTI		SEPSIS		UTI	
	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
Classical and Ensemble Models										
LR	$0.897_{\pm 0.001}$	$0.507_{\pm 0.001}$	$0.773_{\pm 0.000}$	$0.236_{\pm 0.001}$	$0.662_{\pm 0.005}$	$0.258_{\pm 0.013}$	$0.929_{\pm 0.001}$	$0.540_{\pm 0.002}$	$0.913_{\pm 0.000}$	$0.115_{\pm 0.001}$
KNN	$0.841_{\pm 0.000}$	$0.395_{\pm 0.000}$	$0.756_{\pm 0.020}$	$0.228_{\pm 0.016}$	$0.588_{\pm 0.000}$	$0.200_{\pm 0.000}$	$0.823_{\pm 0.000}$	$0.303_{\pm 0.000}$	$0.732_{\pm 0.000}$	$0.046_{\pm 0.000}$
CatB	$0.899_{\pm 0.003}$	$0.516_{\pm 0.016}$	$0.756_{\pm 0.012}$	$0.240_{\pm 0.013}$	$\underline{0.664}_{\pm 0.010}$	$0.262_{\pm 0.010}$	$\mathbf{0.982}_{\pm 0.001}$	$\underline{0.760}_{\pm 0.004}$	$\underline{0.942}_{\pm 0.002}$	$0.152_{\pm 0.005}$
LightG	$0.910_{\pm 0.012}$	$0.551_{\pm 0.041}$	$0.777_{\pm 0.018}$	$0.242_{\pm 0.046}$	$0.649_{\pm 0.003}$	$0.246_{\pm 0.003}$	$\mathbf{0.982}_{\pm 0.002}$	$0.757_{\pm 0.022}$	$0.908_{\pm 0.050}$	$0.114_{\pm 0.034}$
XGB	$0.910_{\pm 0.009}$	$0.545_{\pm 0.039}$	$0.744_{\pm 0.025}$	$0.206_{\pm 0.024}$	$0.635_{\pm 0.034}$	$0.234_{\pm 0.026}$	$\mathbf{0.982}_{\pm 0.001}$	$\mathbf{0.762}_{\pm 0.002}$	$0.920_{\pm 0.000}$	$\underline{0.162}_{\pm 0.045}$
MLP	$0.901_{\pm 0.007}$	$0.521_{\pm 0.011}$	$0.769_{\pm 0.008}$	$0.235_{\pm 0.020}$	$0.640_{\pm 0.014}$	$0.247_{\pm 0.018}$	$0.965_{\pm 0.003}$	$0.665_{\pm 0.024}$	$0.898_{\pm 0.034}$	$0.101_{\pm 0.029}$
Deep Tabular Models										
ResNet	$0.883_{\pm 0.007}$	$0.484_{\pm 0.008}$	$0.733_{\pm 0.039}$	$0.204_{\pm 0.022}$	$0.603_{\pm 0.008}$	$0.219_{\pm 0.010}$	$0.971_{\pm 0.001}$	$0.703_{\pm 0.009}$	$0.898_{\pm 0.001}$	$0.094_{\pm 0.013}$
TabM	$0.870_{\pm 0.001}$	$0.463_{\pm 0.005}$	$0.768_{\pm 0.013}$	$0.256_{\pm 0.012}$	$0.650_{\pm 0.005}$	$0.247_{\pm 0.010}$	$\mathbf{0.982}_{\pm 0.000}$	$0.755_{\pm 0.006}$	$0.938_{\pm 0.001}$	$0.127_{\pm 0.003}$
ExcelF	$0.915_{\pm 0.002}$	$\mathbf{0.565}_{\pm 0.002}$	$\underline{0.795}_{\pm 0.006}$	$0.255_{\pm 0.009}$	$0.653_{\pm 0.018}$	$0.255_{\pm 0.018}$	$0.976_{\pm 0.001}$	$0.673_{\pm 0.029}$	$\mathbf{0.945}_{\pm 0.007}$	$\mathbf{0.177}_{\pm 0.004}$
TabNet	$0.896_{\pm 0.006}$	$0.513_{\pm 0.016}$	$0.650_{\pm 0.084}$	$0.147_{\pm 0.034}$	$0.544_{\pm 0.034}$	$0.179_{\pm 0.017}$	$0.959_{\pm 0.006}$	$0.631_{\pm 0.037}$	$0.599_{\pm 0.110}$	$0.019_{\pm 0.006}$
TabT	$0.888_{\pm 0.001}$	$0.488_{\pm 0.001}$	$0.732_{\pm 0.007}$	$0.200_{\pm 0.010}$	$0.609_{\pm 0.019}$	$0.219_{\pm 0.016}$	$0.955_{\pm 0.005}$	$0.614_{\pm 0.008}$	$0.891_{\pm 0.001}$	$0.083_{\pm 0.002}$
Trompt	$0.906_{\pm 0.002}$	$0.521_{\pm 0.003}$	$0.784_{\pm 0.015}$	$0.262_{\pm 0.016}$	$0.625_{\pm 0.015}$	$0.232_{\pm 0.010}$	$\underline{0.979}_{\pm 0.000}$	$0.739_{\pm 0.006}$	$0.940_{\pm 0.010}$	$0.139_{\pm 0.045}$
MNCA	$0.892_{\pm 0.004}$	$0.514_{\pm 0.008}$	$0.724_{\pm 0.006}$	$0.208_{\pm 0.015}$	$0.626_{\pm 0.012}$	$0.227_{\pm 0.009}$	$0.975_{\pm 0.001}$	$0.723_{\pm 0.013}$	$0.931_{\pm 0.002}$	$0.132_{\pm 0.008}$
Tabular In-Context Learning Models										
TabPFN	$0.905_{\pm 0.001}$	$0.541_{\pm 0.006}$	$0.780_{\pm 0.005}$	$0.266_{\pm 0.013}$	$0.657_{\pm 0.003}$	$0.255_{\pm 0.002}$	$0.968_{\pm 0.001}$	$0.695_{\pm 0.011}$	$0.918_{\pm 0.002}$	$0.139_{\pm 0.002}$
kNNPFN	$0.907_{\pm 0.001}$	$0.538_{\pm 0.006}$	$0.782_{\pm 0.004}$	$0.266_{\pm 0.010}$	$0.661_{\pm 0.005}$	$0.258_{\pm 0.003}$	$0.969_{\pm 0.002}$	$0.691_{\pm 0.013}$	$0.901_{\pm 0.009}$	$0.104_{\pm 0.006}$
TabDPT	$0.906_{\pm 0.002}$	$0.529_{\pm 0.004}$	$0.782_{\pm 0.003}$	$0.270_{\pm 0.003}$	$0.663_{\pm 0.008}$	$0.262_{\pm 0.006}$	$0.960_{\pm 0.002}$	$0.657_{\pm 0.003}$	$0.884_{\pm 0.003}$	$0.106_{\pm 0.004}$
Ours (Tabular In-Context Learning Models + AWARE)										
KNN	$0.878_{\pm 0.025}$	$0.502_{\pm 0.021}$	$0.758_{\pm 0.035}$	$0.239_{\pm 0.011}$	$0.632_{\pm 0.002}$	$0.242_{\pm 0.006}$	$0.963_{\pm 0.005}$	$0.695_{\pm 0.036}$	$0.892_{\pm 0.020}$	$0.144_{\pm 0.006}$
kNNPFN	$\mathbf{0.918}_{\pm 0.001}$	$\underline{0.562}_{\pm 0.004}$	$\mathbf{0.799}_{\pm 0.000}$	$0.279_{\pm 0.007}$	$\mathbf{0.666}_{\pm 0.001}$	$0.257_{\pm 0.002}$	$\underline{0.979}_{\pm 0.001}$	$0.739_{\pm 0.003}$	$0.926_{\pm 0.002}$	$0.142_{\pm 0.003}$
TabDPT	$\underline{0.916}_{\pm 0.003}$	$\mathbf{0.565}_{\pm 0.006}$	$0.791_{\pm 0.003}$	$\underline{0.286}_{\pm 0.003}$	$\underline{0.664}_{\pm 0.001}$	$\underline{0.264}_{\pm 0.003}$	$0.972_{\pm 0.002}$	$0.700_{\pm 0.008}$	$0.912_{\pm 0.013}$	$0.139_{\pm 0.002}$

Table A5. Performance on hospital-acquired infection prediction tasks on large-scale EHR datasets, including MIMIC-IV and eICU. For each task and each model, we extensively ran hyperparameters tuning for 30 rounds.