Tree of Concepts: Interpretable Continual Learners in Non-Stationary Clinical Domains

Dongkyu Cho DONGKYU.CHO@NYU.EDU
Computer Science Department
New York University
New York, NY, USA

Xiyue Li XIYUE.LI@NYULANGONE.ORG
Department of Population Health
NYU Grossman School of Medicine
New York, NY, USA

Samrachana Adhikari SAMRACHANA.ADHIKARI@NYULANGONE.ORG
Department of Population Health
NYU Grossman School of Medicine
New York, NY, USA

Rumi Chunara RUMI.CHUNARA@NYU.EDU
School of Global Public Health
Computer Science Department
New York University
New York, NY, USA

Abstract

Continual learning updates models under distribution shift while limiting performance loss on previously seen data, but clinical use also motivates inspectable interfaces that can be compared across model versions. We propose Tree of Concepts, which uses a shallow decision tree to define a fixed vocabulary of root-to-leaf rules and sequentially updates a neural concept predictor and label head with replay. The association between each concept ID and its rule is fixed; predicted concept assignments and the concept-to-label mapping remain adaptive. Across the evaluated health-related tabular protocols, Tree of Concepts attains the highest or tied mean past- and current-slice scores among the compared methods. These comparisons are descriptive point estimates rather than evidence of statistical superiority. Leaf-assignment agreement with the frozen tree ranges from 0.81 to 0.91, indicating that a fixed vocabulary does not imply invariant instance-level explanations. The results motivate further evaluation of fixed rule vocabularies as longitudinally comparable interfaces for continually updated tabular models.

1. Introduction

Machine learning systems deployed in practice operate under distributional change. Data distributions evolve across time, deployment sites, and user populations in ways that cannot be fully anticipated during training. Continual learning has emerged as a central paradigm to address this, aiming to adapt models to new data (plasticity) while preserving previously acquired knowledge (stability) (Kirkpatrick et al., 2017). Many continual-learning

approaches use deep models updated through fine-tuning or online training. At the same time, many deployment settings impose a second requirement: interpretability (Rymarczyk et al., 2023). In high-stakes domains such as healthcare, finance, and public policy, models may also need inspectable decision interfaces that support human review (Rudin, 2019). In clinical use, this requirement is especially important because predictions can influence high-consequence decisions.

These requirements can be in tension. Hard partitions in shallow decision trees can be sensitive to feature shift and threshold crossings and may require explicit updating as new data arrive. Deeper models, conversely, provide trainable representations that can support adaptation, but their latent variables are often difficult to interpret. Jointly achieving continual adaptation and a longitudinally comparable explanation interface remains challenging, especially in clinical settings where shifts are common and safety requirements are strict (Goetz et al., 2024).

In this work, we investigate whether a fixed rule vocabulary can coexist with adaptive prediction. We introduce **Tree of Concepts**, which defines one rule concept for each root-to-leaf path of an initial shallow tree and separates this fixed vocabulary from adaptive input-to-concept and concept-to-label mappings. We use the term *rule concept* operationally for a supervised tree-leaf region; we do not assume that each region is an independently meaningful clinical construct. The construction guarantees that concept ID ℓ denotes the same rule after every update; it does not guarantee that a given example is always assigned to the same concept. We evaluate on UCI Heart Disease (Janosi et al., 1989), CDC Diabetes Health Indicators (Teboul, 2020), and MIMIC-III (Johnson et al., 2016) under continual-learning protocols. Among the evaluated methods, Tree of Concepts yields the highest or tied reported mean stability and plasticity scores; these comparisons are descriptive point estimates.

Generalizable Insights about Machine Learning in the Context of Healthcare The results motivate a general design hypothesis for clinical machine learning. Clinical datasets can exhibit distribution shift and cohort heterogeneity, which remain relevant under sequential model updates. A fixed dictionary of directly inspectable tree rules may make model versions easier to compare while internal predictive mappings adapt. The present experiments evaluate this hypothesis through aggregate predictive and rule-fidelity metrics; they do not establish clinical usefulness or safety.

2. Related Work

Continual learning. Continual learning studies how to update models on sequential tasks or drifting streams while limiting catastrophic forgetting (Kirkpatrick et al., 2017). Representative families include regularization methods that protect parameters important for past tasks (Kirkpatrick et al., 2017; Aljundi et al., 2018; Kang et al., 2022), rehearsal methods that retain and replay exemplars (Rebuffi et al., 2017), and expansion-based approaches that add capacity to trade off stability and plasticity (Yan et al., 2021; Wang et al., 2022). Rehearsal is a standard strong baseline, but performance depends on buffer size and replay design (Rebuffi et al., 2017; Zhou et al., 2024), and replay-based training can dominate compute relative to raw storage in some regimes (Prabhu et al., 2023; Chavan et al., 2023; Harun et al., 2023; Cho et al., 2026). Much of the literature targets deep ar-

chitectures in vision and language, whereas tabular continual learning is comparatively less developed even though structured covariates are ubiquitous in operational systems. Recent studies examine continual learning for tabular data and show that mixed feature types, sparse interactions, and strong tree-based baselines can yield a different stability–plasticity profile than in perceptual domains (Lourenço et al., 2025; García-Santaclara et al., 2024). The gap matters for clinical risk modeling under temporal, demographic, and site shift. Clinical continual-learning work includes hospital risk models (Amrollahi et al., 2022; Lee and Lee, 2020) and medical imaging (Qazi et al., 2026). Our work focuses specifically on fixed rule interfaces for non-stationary clinical tabular data.

Concept bottleneck models. Concept bottleneck models (CBMs) were introduced as a framework that predicts human-interpretable concepts before predicting the final label, enabling concept-level explanations and test-time interventions; task performance depends on the setting and supervision (Koh et al., 2020). Since then, a central issue in the CBM literature has been the concept labeling bottleneck: standard CBMs typically assume a predefined concept vocabulary together with dense concept annotations, which are expensive to obtain and difficult to scale. Several subsequent works address this limitation from different angles. Post-hoc CBMs show that a pretrained black-box model can be converted into a concept-based predictor after training and can reduce the task-performance gap associated with concept supervision in the evaluated settings (Yuksekgonul et al., 2023). Label-free CBMs remove the need for labeled concept data and align model representations with candidate concept descriptions (Oikarinen et al., 2023), while Discover-then-Name relaxes the need to specify the concept set in advance by discovering and naming concepts learned by the model (Rao et al., 2024). More recent work has continued to extend CBMs toward practical settings, including clinical knowledge-guided variants in medicine and continual-learning variants that aim to preserve semantic consistency of concepts across tasks (Pang et al., 2024; Rymarczyk et al., 2023; Yu et al., 2025). The cited variants span clinical knowledge-guided, visual class-incremental, and language-guided settings. In contrast, our setting is non-stationary clinical tabular learning, where we derive the bottleneck from shallow decision-tree rules over structured features. The resulting ID-to-rule dictionary remains fixed across continual updates; whether individual rule regions are clinically meaningful requires separate validation.

3. Problem Formulation

In this section, we define the core problem addressed in this paper: interpretable continual learning under distribution shift. Specifically, we study supervised learning over a non-stationary stream of data slices $D_1, \ldots, D_T$, where each finite slice $D_t = \{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^{n_t}$ is sampled from a slice-specific distribution P_t, and the distributions may differ across slices. At step t, the learner updates a model h_t using D_t (and optional replay data) without retraining from scratch.

Continual learning favors models that can adapt to new slices while retaining past knowledge. Interpretability, in contrast, often favors shallow models with fixed and explicit threshold logic. Flexible adaptive models may use latent representations that are harder to inspect, whereas fixed shallow structures may be sensitive to shift. Our goal is to satisfy both requirements at once.

Objective. For each step t, we want high *plasticity* on current data and, for $t \geq 2$, high *stability* on previous data:

$$\text{Plasticity}_t = \mathcal{P}_t := \mathcal{M}(h_t, D_t), \tag{1}$$

$$\text{Stability}_t = \mathcal{S}_t := \frac{1}{t-1} \sum_{j=1}^{t-1} \mathcal{M}(h_t, D_j), \tag{2}$$

where $\mathcal{M}$ is a task metric (e.g., AUROC, accuracy). We seek models with simultaneously high $\mathcal{P}_t$ and $\mathcal{S}_t$ over the full sequence.

Beyond performance, we require a longitudinally comparable rule vocabulary. Let $r_{\ell,t}$ denote the root-to-leaf predicate associated with concept ID ℓ at update t. Tree of Concepts enforces

$$r_{\ell,t} = r_{\ell,1} \qquad \text{for every concept ID } \ell \text{ and update } t. \tag{3}$$

This invariant fixes the ID-to-rule mapping. It does not require the predicted concept distribution $\hat{c}_t(x)$, top assignment $\hat{z}_t(x)$, concept-to-label head, or final prediction to remain unchanged for a given input.

4. Proposed Method: Tree of Concepts

4.1. Notation

We consider supervised learning under a nonstationary data stream segmented into slices $D_1, \ldots, D_T$, where slice t provides $D_t = \{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^{n_t}$ sampled from a slice-specific distribution P_t, with features $x \in \mathbb{R}^d$ and labels $y \in \{1, \ldots, K\}$. Slices may index time windows, deployment sites, or evolving patient cohorts, and the distributions P_t may differ across slices. We write $\mathcal{R}_t(\theta) = \mathbb{E}_{(x,y)\sim P_t}[\ell(\hat{y}_\theta(x), y)]$ for the risk on slice t. Our goal is to learn a sequence of models that adapt to new slices while limiting performance regressions on previously seen slices and expose a fixed rule interface whose correspondence to the model's predicted concepts can be measured.

4.2. Main idea

We seek continual learning under distribution shift while maintaining an inspectable rule interface. For this, we define a fixed set of tree-leaf rule concepts using a shallow decision tree, and train a concept bottleneck model that predicts these concepts from raw features and then predicts labels from the resulting concept representation. Continual updates act on the trainable concept extractor and concept-to-label predictor, while the tree-defined concepts remain fixed. Thus the rule definition associated with each concept ID remains unchanged, although instance-level assignments and predictions may change.

4.3. Decision-Tree-Derived Rule Vocabulary

We train a shallow CART decision tree $\mathcal{T}$ on the initial training slice (Breiman et al., 1984) and treat its leaves as a discrete rule vocabulary. Each leaf corresponds to a root-to-leaf path rule, a conjunction of feature threshold tests, which yields an explicit description of the region of feature space represented by that leaf.

The decision tree serves as an inspectable rule scaffold that provides automatically generated rule-region targets without requiring manual concept annotation. Importantly, we freeze $\mathcal{T}$ after initialization so that each concept retains a fixed rule definition over time, which supports direct comparison across model updates.

Concept targets. Let $\{1, \ldots, L\}$ denote the leaf set of $\mathcal{T}$. For any input x, routing through the tree yields a leaf index

$$z(x) = \text{leaf}_{\mathcal{T}}(x) \in \{1, \ldots, L\}. \tag{4}$$

We use $z(x)$ as a pseudo-label for concept learning in every slice. Since $\mathcal{T}$ is fixed, concept supervision is available for both current and replayed data without additional labeling.

4.4. Concept Bottleneck Models for continual adaptation

We instantiate a two-stage predictor consisting of a concept extractor and a label head. The concept extractor learns a trainable mapping from raw features to the tree-defined concept space, and the head learns the downstream mapping from concepts to labels.

Decision trees provide explicit path rules but impose hard partitions that can be sensitive to covariate shift (Moshkovitz et al., 2021). The concept bottleneck model (Koh et al., 2020) complements the tree by learning a soft, trainable concept predictor that can be updated across sequential feature distributions while remaining anchored to the fixed rule vocabulary. The label head parameterizes the concept-to-outcome mapping; label supervision backpropagates through the soft concept vector and therefore updates both the head and concept extractor, without changing the ID-to-rule definitions.

Model. LeafNet first produces concept logits $a_\phi(x) \in \mathbb{R}^L$ and then a normalized distribution over leaf concepts,

$$\hat{c}(x) = \text{softmax}(a_\phi(x)), \qquad \hat{c}(x) \in \Delta^{L-1}. \tag{5}$$

The label head operates only on this normalized bottleneck. With label logits $s_\psi(x) = W_\psi \hat{c}(x) + b_\psi$, it produces

$$\hat{y}(x) = \text{softmax}(s_\psi(x)). \tag{6}$$

We use $\hat{z}(x) = \arg\max_\ell \hat{c}_\ell(x)$ to report the top concept for rule-fidelity evaluation. Softmax scores are treated as model scores rather than calibrated uncertainty estimates.

Training objective. We train LeafNet to match the frozen tree concept assignment and train the head to predict the task label. For a sample (x, y) we optimize

$$\begin{aligned} \mathcal{L}_{\text{concept}}(x) &= \text{CE}(a_\phi(x),\, z(x))\,, \\ \mathcal{L}_{\text{label}}(x, y) &= \text{CE}(s_\psi(x),\, y)\,, \end{aligned} \tag{7}$$

$$\mathcal{L}(x, y) = \mathcal{L}_{\text{label}}(x, y) + \lambda_{\text{c}}\, \mathcal{L}_{\text{concept}}(x). \tag{8}$$

where λ_{c} controls the strength of anchoring to the tree-defined rule vocabulary. Setting $\lambda_{\text{c}} = 0$ removes concept supervision while retaining end-task supervision. The concept loss encourages agreement with the fixed tree-derived rule vocabulary, while the label loss drives task accuracy.

4.5. Continual learning updates

We freeze the tree $\mathcal{T}$ after the initial phase and update the trainable parameters (ϕ, ψ) sequentially over slices. We maintain a bounded replay buffer M containing samples from past slices. At slice t, each minibatch is constructed by mixing current data and replayed data, for example by sampling $B_t \subset D_t$ and $B_M \subset M$ and concatenating them into the training minibatch B. We minimize $\sum_{(x,y)\in B} \mathcal{L}(x, y)$, computing $z(x)$ on the fly by routing through the frozen tree. After training on D_t, we update M by inserting a subset of D_t (optionally balanced by label and/or concept index) and evicting older items when capacity is exceeded. Replay supplies earlier-slice examples during optimization; in the reported experiments, replay variants generally have higher mean past-slice scores.

Summary Decision tree leaves provide a fixed set of rule-defined concepts, yielding a fixed ID-to-rule vocabulary and automatic concept supervision. The concept bottleneck model supplies continual learning capacity by learning a trainable concept extractor and an updatable concept-to-label head, both constrained by the fixed concept interface. This combination preserves an inspectable rule interface while enabling continual updates under distribution shift; the reported agreement metrics quantify, but do not guarantee, instance-level fidelity to that interface.

5. Experiments

5.1. Experimental setting

We evaluate Tree of Concepts under a *continual slice* protocol: the model is trained sequentially on ordered data subsets that define the evaluation slices, then evaluated on held-out test partitions for the current and previously seen slices. We report *plasticity* as the task metric on the current slice and *stability* as the mean of the same metric over all previously seen subsets. Unless noted, all methods use the same split, features, and (when enabled) the same replay buffer size and mixing ratio between current and replayed mini-batches.

Open-access tabular datasets. We use two publicly available tabular datasets. UCI Heart Disease (Janosi et al., 1989) is binary classification (presence of heart disease) with the usual 14 attributes (13 inputs + label). CDC Diabetes Health Indicators (Teboul, 2020) is 3-class classification (healthy / pre-diabetes / diabetes) with 21 attributes (20 inputs + label). For Heart Disease, we choose fixed age boundaries to obtain three age-ordered slices with approximately balanced sample sizes: T_1: age ≤ 52 ($n = 119$; 83 negative / 36 positive), T_2: 53–59 ($n = 93$; 43 negative / 50 positive), and T_3: age ≥ 60 ($n = 91$; 38 negative / 53 positive). For CDC Diabetes we partition by age into four slices 18–34, 35–49, 50–64, 65+.

MIMIC-III cohort and task. We further evaluate on MIMIC-III (v1.4) (Johnson et al., 2016), a de-identified ICU database with controlled access (credentialing and data-use approval required). We formulate one prediction row per ICU stay with a binary in-hospital mortality label derived from admission/discharge information (hospital-expire indicator and death timestamp for the indexed admission). The input is a *fixed tabular schema* shared across all slices: static demographics (e.g., age, sex), admission and service context, comorbidity summaries encoded from ICD-9 (e.g., Elixhauser components or grouped counts), and early-trajectory summaries of labs and vitals from the first 24 hours after ICU admission

(mean/min/max where applicable, plus *missingness indicators* so shifts in measurement frequency do not change feature dimension). Time-varying inputs are computed from the first-24-hour laboratory and vital-sign summaries described above.

MIMIC-III continual protocols. We study two slice constructions for sequential evaluation in critical care data. (A) Shifted-timestamp partition: because released dates are shifted independently by patient, protocol (A) is a synthetic sequential partition of disjoint groups ordered by those timestamps, not a calendar-time or practice-drift protocol. (B) Demographic shift: slices are disjoint patient cohorts ordered from younger to older by age bucket, holding the same label and feature definition. In both protocols, we report AUROC for mortality; the stated preprocessing pipeline and aggregation windows are frozen across slices. Dataset-level cohort size, class prevalence, and train/validation/test counts are summarized in Appendix Table 6.

Baselines and metrics. Baselines are a CART Decision Tree (Breiman et al., 1984), XGBoost (Chen and Guestrin, 2016), and FT-Transformer (Gorishniy et al., 2021), trained with the same slice schedule. For *replay* variants, we store a bounded buffer of past-slice examples and mix them into each update using the same nominal replay capacity across methods. Tree of Concepts uses replay for continual updates to the concept extractor and label head while keeping tree-defined concepts fixed (see Section 4). We report: (i) a pooled-data reference trained jointly on all slices; (ii) Avg. Past-Task (stability); and (iii) Avg. Current-Task (plasticity). Open-access datasets use AUROC (Heart) and Macro-F1 (CDC); MIMIC uses AUROC. Open-access experiments are repeated with five random seeds; MIMIC-III experiments are repeated with three random seeds. We report mean $\pm$ standard error of the mean (SE) across runs.

Interpreting pooled vs. continual scores. The pooled-data reference uses a different training regime (joint access to all slices) than the continual metrics; absolute values are therefore not directly comparable across these columns. We focus on relative ordering under a fixed protocol and on the stability-plasticity tradeoff (past vs. current task); the displayed replay variants have higher mean past-task scores in these experiments.

5.2. Results

We first report results on the publicly available tabular datasets, followed by evaluation on the restricted-access MIMIC-III ICU cohort under two sequential protocols.

Open-access dataset results. In Tables 1 and 2, replay variants have higher mean past-task scores than their corresponding no-replay baselines, and the Decision Tree has the lowest mean past-task score among the evaluated baselines. Among the evaluated methods, Tree of Concepts has the highest or tied mean past- and current-task scores on both datasets. These point-estimate comparisons do not establish statistical superiority.

MIMIC-III results. We evaluate MIMIC-III mortality under (A) the shifted-timestamp partition and (B) demographic shift, and report both protocols to describe performance under two sequential slicing schemes.

Table 1: UCI Heart Disease (binary classification, AUROC). We report the mean and standard error over five runs.

Method	Replay	Pooled ref.	Avg. Past-Task	Avg. Current-Task
No Replay				
Decision Tree	✗	0.83 ± 0.02	0.56 ± 0.04	0.72 ± 0.03
XGBoost	✗	0.89 ± 0.01	0.75 ± 0.03	0.84 ± 0.02
FT-Transformer	✗	0.88 ± 0.02	0.73 ± 0.03	0.83 ± 0.02
Replay				
Decision Tree	✓	–	0.63 ± 0.04	0.73 ± 0.03
XGBoost	✓	–	0.78 ± 0.02	0.85 ± 0.02
FT-Transformer	✓	–	0.76 ± 0.02	0.84 ± 0.03
Ours				
Tree of Concepts	✓	0.89 ± 0.01	0.80 ± 0.02	0.85 ± 0.02

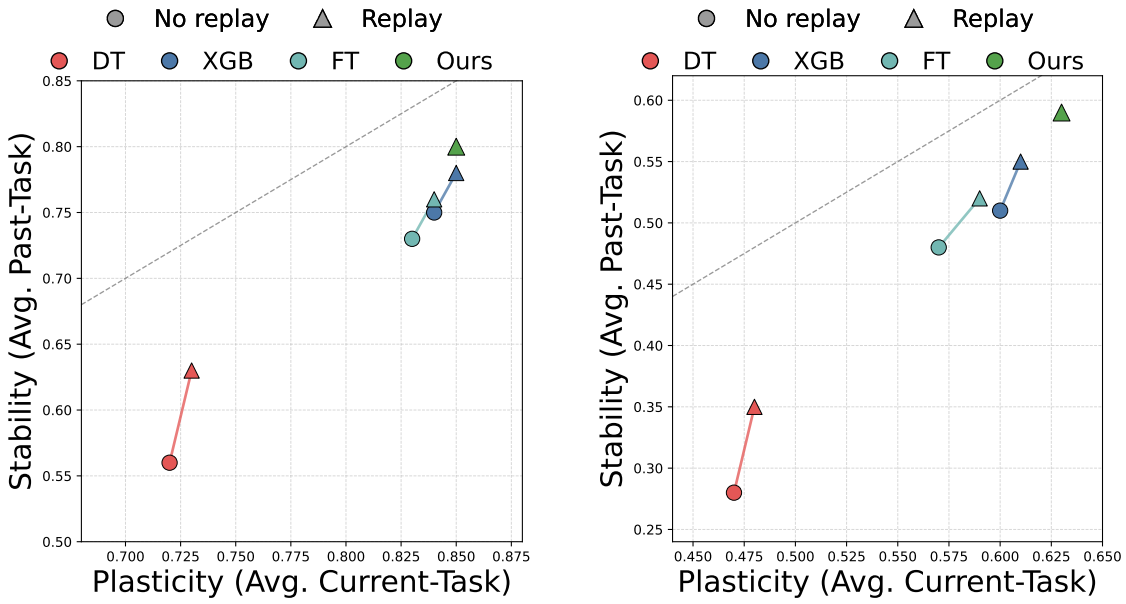
Table 2: CDC Diabetes Health Indicators (3-class classification, Macro-F1). We report the mean and standard error over five runs.

Method	Replay	Pooled ref.	Avg. Past-Task	Avg. Current-Task
No Replay				
Decision Tree	✗	0.58 ± 0.02	0.28 ± 0.05	0.47 ± 0.04
XGBoost	✗	0.65 ± 0.02	0.51 ± 0.03	0.60 ± 0.02
FT-Transformer	✗	0.64 ± 0.01	0.48 ± 0.04	0.57 ± 0.03
Replay				
Decision Tree	✓	–	0.35 ± 0.05	0.48 ± 0.03
XGBoost	✓	–	0.55 ± 0.03	0.61 ± 0.02
FT-Transformer	✓	–	0.52 ± 0.03	0.59 ± 0.03
Ours				
Tree of Concepts	✓	0.66 ± 0.01	0.59 ± 0.02	0.63 ± 0.02

Aggregate fidelity to the fixed rule interface. The tree is frozen after initialization, so each concept ID always corresponds to the same rule. At each slice, we compare (1) the concept ID assigned by the frozen tree, $z(x)$, and (2) the top concept ID predicted by the concept model, $\hat{z}(x) = \arg\max_\ell \hat{c}_\ell(x)$. We report leaf-assignment agreement (higher is better), rule-fidelity gap (absolute performance difference when using predicted concepts versus tree-assigned concepts; lower is better), and high-score leaf disagreement (disagreement among samples whose maximum concept score exceeds 0.8; lower is better). These are cross-sectional fidelity measures, not longitudinal measures of invariant patient-level explanations.

Table 5 shows higher agreement on UCI/CDC and lower agreement on the MIMIC protocols. The reported fidelity gaps range from 0.012 to 0.033; leaf agreement of 0.81 on MIMIC protocol (B) corresponds to 0.19 disagreement with literal tree routing. We report aggregate indicators here; per-rule clinical case studies were not evaluated.

Observation. Across the evaluated baseline blocks, replay variants have higher reported mean Avg. Past-Task scores than the corresponding no-replay variants; mean Avg. Current-Task changes by 0.01–0.02. These mean differences are observed with bounded replay memories whose retained fractions vary by dataset. In every baseline block, the Decision Tree



(*a*) UCI Heart Disease. (*b*) CDC Diabetes Health Indicators.

Figure 1: Stability–plasticity trade-off on open-access tabular datasets (mean across five runs; tables give mean ± SE). The x-axis is plasticity (Avg. Current-Task), and the y-axis is stability (Avg. Past-Task). Segments link no-replay and replay for each baseline.

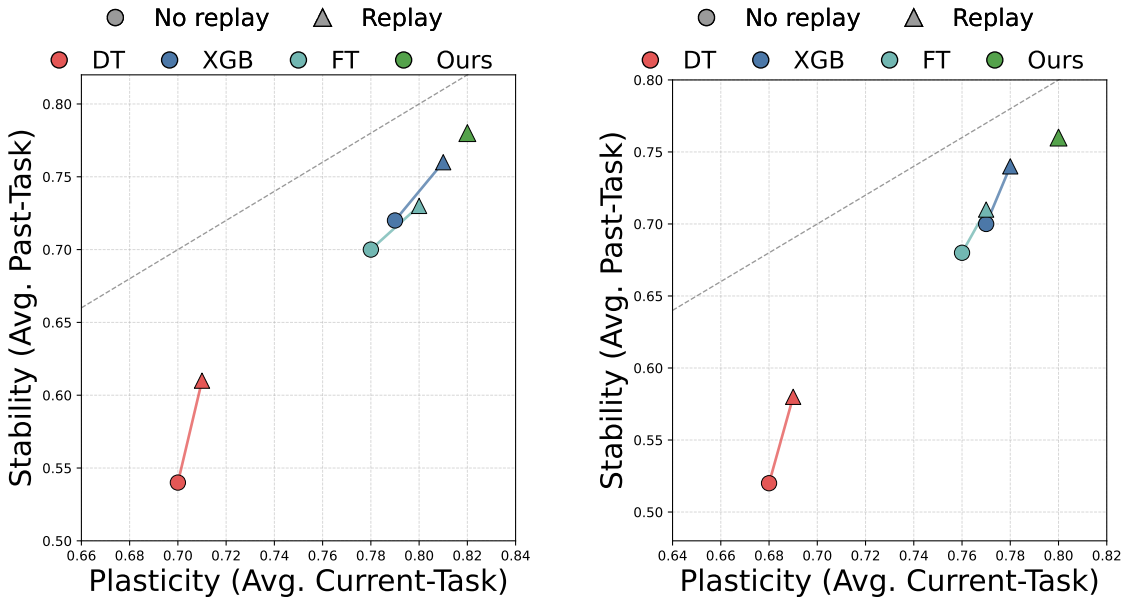
Table 3: MIMIC-III in-hospital mortality (AUROC), continual protocol (A) shifted-timestamp partition based on released deidentified admission timestamps. We report the mean and standard error over three runs.

Method	Replay	Pooled ref.	Avg. Past-Task	Avg. Current-Task
No Replay				
Decision Tree	✗	0.78 ± 0.02	0.54 ± 0.04	0.70 ± 0.03
XGBoost	✗	0.85 ± 0.01	0.72 ± 0.03	0.79 ± 0.02
FT-Transformer	✗	0.84 ± 0.02	0.70 ± 0.03	0.78 ± 0.03
Replay				
Decision Tree	✓	–	0.61 ± 0.04	0.71 ± 0.03
XGBoost	✓	–	0.76 ± 0.02	0.81 ± 0.02
FT-Transformer	✓	–	0.73 ± 0.03	0.80 ± 0.02
Ours				
Tree of Concepts	✓	0.86 ± 0.01	0.78 ± 0.02	0.82 ± 0.02

has the lowest mean Avg. Past-Task, and XGBoost has higher displayed mean past- and current-task scores than FT-Transformer; Tree of Concepts has the highest or tied mean past- and current-task scores among the evaluated methods. Because the two MIMIC-III protocols use different slice definitions and evaluation sets, we report their results separately rather than interpret their absolute scores as a direct difficulty comparison. Table 5 quantifies correspondence to the fixed rule dictionary but does not establish clinical trustworthiness or invariant instance-level explanations.

Table 4: MIMIC-III in-hospital mortality (AUROC), continual protocol (B) demographic shift by age bucket. We report the mean and standard error over three runs.

Method	Replay	Pooled ref.	Avg. Past-Task	Avg. Current-Task
No Replay				
Decision Tree	✗	0.77 ± 0.01	0.52 ± 0.04	0.68 ± 0.03
XGBoost	✗	0.84 ± 0.02	0.70 ± 0.03	0.77 ± 0.02
FT-Transformer	✗	0.83 ± 0.01	0.68 ± 0.03	0.76 ± 0.02
Replay				
Decision Tree	✓	–	0.58 ± 0.04	0.69 ± 0.03
XGBoost	✓	–	0.74 ± 0.03	0.78 ± 0.02
FT-Transformer	✓	–	0.71 ± 0.03	0.77 ± 0.03
Ours				
Tree of Concepts	✓	0.85 ± 0.01	0.76 ± 0.02	0.80 ± 0.02



(*a*) Protocol (A): shifted-timestamp partition.

(*b*) Protocol (B): demographic shift.

Figure 2: Stability–plasticity trade-offs on MIMIC-III mortality prediction under protocols (A) and (B); points are means across three runs (see tables for ± SE).

6. Discussion

Technical implications. The main technical hypothesis of this work is that a fixed rule dictionary can coexist with adaptive continual prediction. Tree of Concepts separates fixed concept-ID-to-rule definitions from trainable input-to-concept and concept-to-label mappings. Among the evaluated methods, the full Tree of Concepts configuration has the highest or tied reported mean past- and current-task scores across the tabular protocols. These are descriptive point estimates, not evidence of statistical superiority. The fixed

Table 5: Aggregate fidelity of predicted tree-leaf rule concepts to the frozen rule interface (mean ± SE; five runs for UCI/CDC and three runs for MIMIC-III).

Setting	Leaf agreement ↑	Rule-fidelity gap ↓	High-score disagreement ↓
UCI Heart Disease	0.91 ± 0.01	0.012 ± 0.004	0.039 ± 0.009
CDC Diabetes	0.86 ± 0.02	0.020 ± 0.006	0.056 ± 0.011
MIMIC-III (A) Shifted timestamp	0.83 ± 0.02	0.027 ± 0.008	0.070 ± 0.013
MIMIC-III (B) Demographic	0.81 ± 0.02	0.033 ± 0.009	0.078 ± 0.014

dictionary makes ID-to-rule definitions directly comparable across updates, but neither concept assignments nor the label head are invariant across updates.

Clinical implications. In healthcare, model updates are not only a statistical problem but also a workflow and governance problem. A fixed rule dictionary could make model versions easier to compare. However, the reported aggregate fidelity metrics do not establish clinical usability, trust, or safety. Leaf-assignment agreement ranges from 0.81 to 0.91, corresponding to disagreement rates of 0.09–0.19 with literal tree routing. Case-level audits, longitudinal evaluation on fixed patients, calibration analysis, and clinician studies are required before drawing deployment conclusions.

Limitations. This study has several limitations. First, the empirical evidence is limited to health-related tabular benchmarks and does not capture prospective clinical deployment. Second, we use the term *concept* operationally for supervised tree-leaf regions; their clinical coherence has not been validated. Fixing the ID-to-rule mapping does not fix instance-level assignments, and aggregate leaf agreement does not measure longitudinal explanation stability. Third, the quality and coverage of the rule interface depend on the initial tree. Fourth, our continual setup assumes a fixed feature schema and preprocessing pipeline, which may fail under changes in coding, instrumentation, or data capture. Fifth, the baseline suite is limited to three model families and simple replay; it omits regularization-based continual learning, stronger replay methods, adaptive online trees, and continual concept models. Sixth, reported differences are descriptive means from three or five runs without paired significance tests. Seventh, age-defined slices have non-overlapping support on the age feature, and retaining age as an input can reveal slice identity. Eighth, average past/current metrics omit worst-slice behavior, transfer, calibration, patient-level explanation trajectories, and concept-frequency-adjusted fidelity. Ninth, the evidence is aggregate rather than based on case-level rule audits with clinician adjudication. Finally, replay may be constrained by privacy, governance, or storage restrictions, and extension to images, waveforms, or text requires different concept definitions.

Future directions. Several next steps follow naturally from these limitations. The most immediate are independently reproducible cohort protocols, further MIMIC tasks or cohorts, external institutional validation, and subgroup analyses. Another important direction is to evaluate rule quality with clinicians and to report case-level trajectories rather than relying only on aggregate predictive and fidelity metrics. More broadly, future work could study how to version and update the rule vocabulary when medical knowledge, coding practices, or patient populations change while preserving explicit mappings to prior concepts.

7. Concluding Remarks

This work studies a modular continual-learning design in which the definitions of root-to-leaf rules remain fixed while concept prediction and label mapping adapt sequentially. Among the evaluated methods and protocols, Tree of Concepts attains the highest or tied reported mean past- and current-slice scores. These descriptive comparisons do not establish statistical superiority, clinical meaningfulness of the tree-leaf rules, or invariant patient-level explanations. They motivate further study of fixed rule vocabularies as longitudinally comparable interfaces for continually updated tabular models.

Acknowledgments

We thank NYU High Performance Computing (HPC) for providing computational resources and technical support for the simulations and experiments. We are grateful for the support provided through the NSF grants 1845487 and 2129076. This work was also in part supported by the Google Research Award program's TPU compute grant.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018.

Fatemeh Amrollahi, Supreeth P Shashikumar, Andre L Holder, and Shamim Nemati. Leveraging clinical data across healthcare institutions for continual learning of predictive risk models. *Scientific reports*, 12(1):8380, 2022.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

Vivek Chavan, Paul Koch, Marian Schlüter, and Clemens Briese. Towards realistic evaluation of industrial continual learning scenarios with an emphasis on energy consumption and computational footprint. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11506–11518, 2023.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Dongkyu Cho, Taesup Moon, Rumi Chunara, Kyunghyun Cho, and Sungmin Cha. Forget forgetting: Continual learning in a world of abundant memory, 2026. URL https://arxiv.org/abs/2502.07274.

Pablo García-Santaclara, Bruno Fernández-Castro, and Rebeca P. Díaz-Redondo. Overcoming catastrophic forgetting in tabular data classification: A pseudorehearsal-based approach, 2024. URL https://arxiv.org/abs/2407.09039.

Lea Goetz, Nabeel Seedat, Robert Vandersluis, and Mihaela van der Schaar. Generalization—a key challenge for responsible ai in patient-facing clinical applications. *NPJ Digital Medicine*, 7(1):126, 2024.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pages 18932–18943, 2021.

Md Yousuf Harun, Jhair Gallardo, Tyler L. Hayes, and Christopher Kanan. How efficient are today's continual learning algorithms? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2431–2436, June 2023.

Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C52P4X.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Haeyong Kang, Rusty John Lloyd Mina, Sultan Rizky Hikmawan Madjid, Jaehong Yoon, Mark Hasegawa-Johnson, Sung Ju Hwang, and Chang D Yoo. Forget-free continual learning with winning subnetworks. In *International Conference on Machine Learning*, pages 10734–10750. PMLR, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models, 2020. URL https://arxiv.org/abs/2007.04612.

Cecilia S Lee and Aaron Y Lee. Clinical applications of continual learning machine learning. *The Lancet Digital Health*, 2(6):e279–e281, 2020.

Afonso Lourenço, João Gama, Eric P. Xing, and Goreti Marreiros. Bridging streaming continual learning via in-context large tabular models, 2025. URL https://arxiv.org/abs/2512.11668.

Michal Moshkovitz, Yao-Yuan Yang, and Kamalika Chaudhuri. Connecting interpretability and robustness in decision trees through separation. In *International Conference on Machine Learning*, pages 7839–7849. PMLR, 2021.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models, 2023. URL https://arxiv.org/abs/2304.06129.

Winnie Pang, Xueyi Ke, Satoshi Tsutsui, and Bihan Wen. Integrating clinical knowledge into concept bottleneck models, 2024. URL https://arxiv.org/abs/2407.06600.

Ameya Prabhu, Hasan Abed Al Kader Hammoud, Puneet K Dokania, Philip HS Torr, Ser-Nam Lim, Bernard Ghanem, and Adel Bibi. Computationally budgeted continual

learning: What does matter? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3698–3707, 2023.

Mohammad Areeb Qazi, Anees Ur Rehman Hashmi, Santosh Sanjeev, Ibrahim Almakky, Numan Saeed, Camila Gonzalez, and Mohammad Yaqub. Continual learning in medical imaging: a survey and practical analysis. *ACM Computing Surveys*, 58(8):1–25, 2026.

Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery, 2024. URL https://arxiv.org/abs/2407.14499.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning, 2017. URL https://arxiv.org/abs/1611.07725.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Dawid Rymarczyk, Joost van de Weijer, Bartosz Zieliński, and Bartłomiej Twardowski. Icicle: Interpretable class incremental continual learning, 2023. URL https://arxiv.org/abs/2303.07811.

Alex Teboul. Diabetes health indicators dataset. Kaggle, 2020. URL https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset. Accessed: 2026-04-13.

Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *European conference on computer vision*, pages 398–414. Springer, 2022.

Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3014–3023, 2021.

Lu Yu, Haoyu Han, Zhe Tao, Hantao Yao, and Changsheng Xu. Language guided concept bottleneck models for interpretable continual learning, 2025. URL https://arxiv.org/abs/2503.23283.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models, 2023. URL https://arxiv.org/abs/2205.15480.

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9851–9873, 2024.

Table 6: Dataset-level cohort summary.

Dataset	N_{source}	N_{used}	Train / Val / Test	Outcome prevalence
UCI Heart Disease	303	≈ 300	180 / 60 / 60	$\approx 46\%$ positive
CDC Diabetes	253,680	250,000	150,000 / 50,000 / 50,000	$\approx 74\%/16\%/10\%$
MIMIC-III (A,B)	53,423	50,000	30,000 / 10,000 / 10,000	$\approx 11.5\%$ in-hospital mortality

Appendix A.

Detailed Experimental Settings

Data preprocessing. All tabular inputs are preprocessed with a fixed pipeline shared across methods: (i) continuous variables are z-score normalized using first-slice training-partition statistics, (ii) categorical variables are one-hot encoded, and (iii) missing values are imputed with training medians/modes and accompanied by missingness indicators. For continual experiments, preprocessing parameters are initialized on the first slice's training partition and then kept fixed to avoid leakage across future slices. This preserves a forward-only deployment order without using future-slice preprocessing statistics and keeps preprocessing identical across methods.

Slice construction. For UCI Heart Disease and CDC Diabetes, slices follow the demographic partitions in Section 5.1. For UCI Heart Disease, we use approximately 300 observations and select 180 / 60 / 60 observations for the train / validation / test partitions. For MIMIC-III, each sample corresponds to one ICU stay, and features are extracted from the first 24 hours. We evaluate two slice protocols: (A) a shifted-timestamp partition and (B) demographic cohorts. Protocol (A) divides ICU stays into four groups ordered by released deidentified admission timestamp; patient-specific shifts preclude cross-patient calendar chronology.

Model and optimization details. The frozen concept scaffold is a shallow CART-style decision tree trained only on the initial slice's training partition (max depth $= 4$). LeafNet (concept extractor) is a 2-layer MLP with hidden widths $(128, 64)$, ReLU activations, and dropout 0.1. The label head is a linear layer applied to the normalized concept-probability vector; raw concept logits are used only by the concept cross-entropy loss. We optimize with Adam (learning rate 10^{-3}, weight decay 10^{-5}, batch size 256), with early stopping on validation performance (patience $= 8$, max epochs $= 80$). Unless stated otherwise, $\lambda_c = 1.0$ in Eq. (8) of the main text.

Replay setup. Replay minibatches mix current and memory samples at a 1:1 ratio. Memory is class-balanced when possible. Default replay capacities are 2,048 samples for open-access datasets and 8,192 for MIMIC-III. Open-access experiments are repeated over five random seeds, while MIMIC-III experiments are repeated over three random seeds; we report mean $\pm$ standard error.

Table 7: Component ablations for Tree of Concepts (mean ± SE, 5 runs).

Setting	UCI Past	UCI Current	CDC Past	CDC Current	Concept Agreement
Full model (ours)	0.80 ± 0.02	0.85 ± 0.02	0.59 ± 0.02	0.63 ± 0.02	0.89 ± 0.02
w/o replay	0.71 ± 0.03	0.84 ± 0.03	0.46 ± 0.03	0.61 ± 0.03	0.86 ± 0.02
w/o concept loss ($\lambda_c = 0$)	0.67 ± 0.04	0.82 ± 0.03	0.43 ± 0.04	0.58 ± 0.03	0.60 ± 0.04
Refresh tree each slice	0.73 ± 0.03	0.82 ± 0.02	0.49 ± 0.03	0.60 ± 0.03	0.69 ± 0.04

Table 8: Replay memory ablation on MIMIC-III protocol (A), the shifted-timestamp partition (mean ± SE, 3 runs).

Replay capacity	Avg. Past-Task	Avg. Current-Task	Concept Agreement
0 (no replay)	0.57 ± 0.04	0.78 ± 0.02	0.77 ± 0.03
512	0.68 ± 0.03	0.80 ± 0.02	0.81 ± 0.02
2,048	0.75 ± 0.03	0.81 ± 0.02	0.83 ± 0.02
8,192	0.78 ± 0.02	0.82 ± 0.02	0.84 ± 0.02

Reproducibility Details

For transparency, we summarize the available dataset, cohort, slice, preprocessing, model, replay, evaluation, split-size, and seed-count information in the manuscript and appendix. Access to MIMIC-III follows its credentialing and data-use requirements.

Ablation Studies

Ablation A: component comparisons. Table 7 reports component ablations for replay, fixed tree definitions, and concept supervision. Each ablation has a lower reported mean past-slice score than the full model; the no-concept-loss variant has the lowest reported mean concept-agreement score.

Ablation B: replay memory size. In Table 8, the reported mean past-slice and concept-agreement scores increase monotonically with replay capacity, while successive mean increments become smaller.