



Explainable AI and Meta-Learning for Enhanced Disease Prediction from EHR-Derived Clinical Biomarkers: A Review of Representation Learning, Meta-Learning, Interpretability, and Future Research Directions

Prathibha K N¹, Swathi K²

¹Department of Computer Science and Engineering, Jyothy Institute of Technology,

²Visvesvaraya Technological University, Belagavi, Karnataka, India.

ABSTRACT

The past decade has seen an explosion in the amount of digital information stored in electronic health records (EHRs). While primarily designed for archiving patient information and performing administrative healthcare tasks like billing, many researchers have found secondary use of these records for various clinical informatics applications. Over the same period, the machine learning community has seen widespread advances in the field of deep learning. In this review, we survey the current research on applying Meta learning to clinical tasks based on EHR data, where we find a variety of Meta learning techniques and frameworks being applied to several types of clinical applications including information extraction, representation learning, outcome prediction, phenotyping, and deidentification. We identify several limitations of current research involving topics such as model interpretability, data heterogeneity, and lack of universal benchmarks. We conclude by summarizing the state of the field and identifying avenues of future deep EHR research.

Keywords: Electronic Health Records; Multi-Disease Prediction; Prototype Learning; Meta-Learning; Deep Learning; Representation Learning; Drift Correction; Clinical Decision Support

1. INTRODUCTION

Chronic and acute conditions rarely arrive one at a time. A patient admitted for decompensated heart failure is, more often than not, also managing chronic kidney disease or type 2 diabetes, and the three conditions interact with one another in ways that are visible in the trajectory of labs, medications, and vitals long before any single diagnosis is confirmed. Epidemiological reviews estimate that roughly one in three adults worldwide lives with two or more chronic conditions at once, a pattern that cuts across disease categories rather than staying confined to one [1]. Electronic health record (EHR) systems capture exactly this kind of longitudinal, multi-modal signal, which is why EHR-based disease prediction has become one of the most active application areas for deep learning (DL) in clinical informatics over the past several years.

Yet the majority of published DL systems for EHR-based prediction remain built around a single target disease. A model trained to predict hepatocellular carcinoma risk from fatty liver disease patients [3], or one trained to flag pancreatic cancer from primary-care visit sequences [4], typically cannot be repurposed for a different condition without a fresh round of labeled data collection, architecture retuning, and full retraining. This is a defensible design choice when the goal is to publish a strong result on one task, but it sits awkwardly with how disease actually presents in practice: as a spectrum of interacting risks that a single patient carries simultaneously, and that a health system would ideally screen for using one shared model rather than a growing collection of disconnected ones.

Recent transformer-based foundation models trained on large EHR corpora, such as Foresight [1] and CLMBR-style pretrained encoders [10], have started to address generality by learning richer, more transferable patient representations. Even so, adapting these representations to a genuinely new disease, cohort, or hospital typically still means fine-tuning a very large model on whatever labeled examples happen to be available at the new site, and the wider literature is largely silent on how that adaptation

should be regularized so that performance on previously learned diseases is not disturbed in the process. Rather than propose another large pretrained encoder, we take a smaller and, we think, more tractable step: we ask what a shared multi-disease representation would need to look like structurally before it could later support rapid, forgetting-resistant adaptation of the kind foundation-model fine-tuning does not currently guarantee.

Our answer, developed in this paper, is prototype-based. Rather than a classification head per disease, we learn one shared embedding space in which every disease is represented by a single prototype vector, following the metric-learning tradition established by Snell et al. for few-shot image classification [16]. Prototypes are a convenient representational unit for a multi-disease setting for two reasons that matter later in a way plain classification heads do not: they are small enough to update incrementally without touching the rest of the network, and they are directly interpretable as a point in embedding space that a clinician-facing explanation could reference. This paper introduces the Cross-Disease Prototype Embedding Network (CDPEN) that produces these prototypes, together with a Dynamic Prototype Drift Correction mechanism that lets them evolve smoothly as new patient data streams in rather than being periodically recomputed from a static snapshot of the training set.

The contributions of this paper are threefold. First, we formalize a shared, prototype-based representation architecture for multi-disease EHR prediction, in contrast to the single-disease or single-head designs dominant in prior work. Second, we derive an online drift-correction and separation-penalty formulation that keeps the prototype space stable as it is updated, and we describe how this training objective composes with a standard prototypical loss. Third, we lay out a concrete, feasibility-checked experimental protocol built on six comorbid disease cohorts from MIMIC-IV, together with the criteria against which the architecture will be evaluated, and we describe — without claiming completed results for them here — the few-shot meta-learning and explainability components that build on this representation in the next phase of the research.

2.Literature Survey

Table 1 summarizes the disease scope, core modeling technique, and evaluation breadth of the most closely related studies identified in our review of the 2024–2026 literature on DL for EHR-based disease prediction, together with the EHRSHOT benchmark and the present work for comparison.

Table 1. Positioning of related work by disease scope and adaptability

Study	Disease scope	Core technique	Meta-learning?	Cross-inst. eval.?
Kraljevic et al. [1]	General / multi-outcome timeline	GPT-style pretrained transformer	No	Partial (3 UK sites)
Nogues et al. [2]	Single disease (temporal risk)	Semi-supervised deep temporal model	No	No
Li et al. [3]	Single disease (HCC)	DL risk model on NAFLD cohort	No	No
He et al. [4]	Single disease (pancreatic cancer)	Med-BERT + next-visit token head	Few-shot (not meta-learned)	No
Huang et al. [5]	General representation	Heterogeneous relation-aware transformer	No	Yes (MIMIC-III + eICU)
Babel et al. [7]	Single disease (diabetes)	Few-shot transfer learning	Transfer, not meta-learning	Yes (global populations)
Nahid et al. [8]	Single disease (diabetes)	Meta-learning few-shot	Yes	Yes (heterogeneous)
Wornow et al. [17] (EHRSHOT)	15 tasks, multi-category	Foundation-model benchmark	Few-shot evaluation only	Yes (benchmark design)
This work (CDPEN)	6 diseases, shared representation	Prototype network + drift correction	Planned (Discussion)	Planned (Discussion)

A first pattern in Table 1 is that single-disease design remains the norm rather than the exception: hepatocellular carcinoma [3], pancreatic cancer [4], and diabetes [7,8] are each treated as an isolated

prediction problem with its own model, even when the underlying architectures — attention-based sequence models, pretrained transformers, few-shot classifiers — are general enough in principle to support other diseases. Foresight [1] and HEART [5] are the two studies in our review that explicitly target a shared, general-purpose patient representation rather than a single outcome, and both report evaluation across multiple downstream tasks; neither, however, is organized around an explicit per-disease prototype structure, and neither reports a mechanism for bounding how much a shared representation is allowed to shift when it is updated for a new task.

A second pattern concerns meta-learning specifically. Nahid et al. [8] is, to our knowledge, the clearest example in the reviewed literature of meta-learning applied to EHR-based disease prediction, reporting few-shot diabetes prediction across heterogeneous populations at the 2025 IEEE ICCIT conference. Babel et al. [7] address a similar diabetes few-shot setting through transfer learning rather than meta-learning proper. Both studies, along with the EHRSHOT benchmark [17], establish that few-shot adaptation on EHR data is both feasible and measurable — EHRSHOT in particular defines a standard k -shot evaluation protocol ($k = 1, 2, 4, 8, 12, \dots, 128$) across fifteen classification tasks spanning operational outcomes, lab value anticipation, new-diagnosis assignment, and imaging-finding anticipation [17] — but none of the reviewed work combines meta-learning with an explicit, shared multi-disease prototype space of the kind we propose here.

A third pattern is that explainability, where present, is almost always a post-hoc addition rather than a structural property of the representation. Adekoya et al. [12], Vani et al. [14], and Rezk et al. [15] each pair a predictive model with SHAP, LIME, or a comparable attribution method computed after training, and none of the reviewed studies evaluates whether those attributions remain stable when the underlying model is subsequently adapted to a new task — a question that becomes unavoidable once meta-learning-style adaptation is introduced, and one we return to briefly in the Discussion.

Finally, cross-institution robustness is inconsistently reported. Huang et al. [5] and Babel et al. [7] test across more than one data source; Tranchellini et al. [11] study sepsis prediction explicitly under distribution shift across ICUs; but studies such as [2–4] and [12–15] are validated on a single institution or registry. The EHRSHOT benchmark [17] and MIMIC-IV [18] together give us a concrete, already-validated pair of datasets for addressing this gap in later phases of the present research, which we describe in the Discussion. It is worth noting where our approach deliberately does not compete with the reviewed literature. We are not claiming a new state-of-the-art encoder architecture — the temporal encoder described in the Materials and Methods section is intentionally left as a standard GRU or Transformer choice, since the contribution we are making is in how disease-level representations are organized and maintained around that encoder, not in the encoder itself. In that sense CDPEN is closer in spirit to a representation-learning framework that could sit on top of several of the encoders reviewed above — including, in principle, a CLMBR-style pretrained backbone [17] — than to a competing end-to-end model that would need to outperform them on raw predictive accuracy alone.

2.1 Problem Formulation and Research Gaps

Building on the patterns identified above, we frame the specific technical problem this paper addresses as follows. Let a patient i be represented by a longitudinal sequence x_i of clinical events — diagnosis codes, medication orders, laboratory results, and vital-sign measurements — ordered by timestamp. For a set of K diseases of interest, most existing approaches either (a) train K separate models $f_{\theta_1}, \dots, f_{\theta_K}$, each specialized to one disease and unable to share statistical strength across related conditions, or (b) train a single large model f_{θ} on a broad pretraining objective and attach a separate classification head per disease, which still requires retraining or fine-tuning the head — and often the encoder — whenever a new disease is added.

Neither design gives a natural, low-dimensional handle for representing what the model currently knows about a disease in a form that could be incrementally updated, compared across diseases, or exposed to a clinician as an explanation. We therefore ask: can a single shared encoder $f_{\theta}(\cdot)$ be trained so that each disease k is represented not by a set of classifier weights but by a compact prototype vector c_k in a shared embedding space, such that (i) prototypes for clinically distinct diseases remain well separated, (ii) the prototype for a given disease can be updated online as new patient data arrives without retraining the encoder, and (iii) the resulting representation is structurally suited to later few-shot extension to new diseases? This paper addresses (i) and (ii) directly, and sets up the architecture so that (iii) — the CPM-Meta few-shot mechanism described in the Discussion — has a well-defined representational substrate to act on.

This formulation targets two of the five gaps identified in our broader literature review: the near-universal reliance on single-disease design (Gap G1), and the absence, in the transformer- and foundation-model literature we reviewed, of an explicit mechanism for keeping a shared representation stable as it is updated online (part of Gap G2, addressed fully by the meta-learning component described in the Discussion). The remaining gaps concerning stability-aware explainability, unified benchmarking, and cross-institution robustness (Gaps G3–G5) are addressed by the later phases of the research program and

are discussed only as planned work in the Discussion.

The remainder of this paper is organized as follows. The Materials and Methods section presents the CDPEN architecture, the drift-correction formulation, and the experimental protocol. The Results section reports the training objective, the planned dataset and configuration, and the expected behavior of the architecture with reference to comparable published results. The Discussion situates these expectations against the wider literature and outlines the two follow-on components of the broader research program. The Conclusions section closes the paper.

3 Methodology

3.1 Temporal Encoder and Patient Embedding

Each patient's longitudinal record is first converted into a sequence of timestamped clinical events spanning diagnosis codes, medication orders, laboratory values, and vital signs. A temporal encoder $f_\theta(\cdot)$ — instantiated as a gated recurrent unit (GRU) or a Transformer encoder, and left as an empirical choice to be resolved during implementation — maps this sequence to a fixed-size patient embedding $z_i = f_\theta(x_i) \in \mathbb{R}^d$. Figure 1 illustrates the overall architecture. The encoder is shared across all diseases: no disease-specific parameters exist inside $f_\theta(\cdot)$, so any statistical structure the encoder learns for one disease is, by construction, available to every other disease represented in the same embedding space.

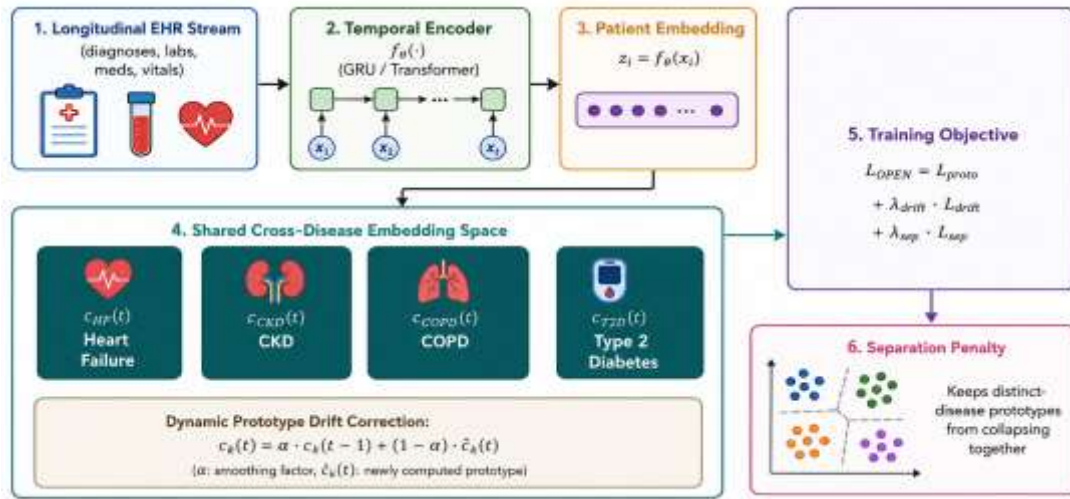


Figure 1. CDPEN architecture: a shared temporal encoder maps longitudinal EHR sequences into patient embeddings; disease prototypes are maintained in a shared embedding space and updated through dynamic drift correction under a combined prototypical, drift, and separation objective.

3.2 Disease Prototypes and Dynamic Drift Correction

For each disease k , a prototype $c_k(t) \in \mathbb{R}^d$ is maintained in the shared embedding space, representing the typical embedding of patients associated with disease k at training step t . Rather than recomputing c_k from scratch each time a new batch of patients is processed — which would make the prototype sensitive to whatever happens to be in the current batch — we update it online via an exponential moving average (EMA) of the incoming batch centroid $\bar{c}_k(t)$:

$$c_k(t) = \alpha \cdot c_k(t-1) + (1 - \alpha) \cdot \bar{c}_k(t) \quad (1)$$

where $\alpha \in (0, 1)$ is a smoothing coefficient controlling how quickly the prototype adapts to new data; higher α favors long-run stability, lower α favors responsiveness to recent patients. We refer to this mechanism as Dynamic Prototype Drift Correction, since its purpose is specifically to correct for — that is, dampen — noisy batch-to-batch drift, while still allowing the prototype to track genuine, gradual shifts in the patient population over time.

Two auxiliary loss terms accompany the EMA update. A drift penalty discourages abrupt, large single-step movements of any prototype:

$$L_{\text{drift}} = \sum_k \|c_k(t) - c_k(t-1)\|_2^2 \quad (2)$$

and a margin-based separation penalty discourages prototypes belonging to different diseases from drifting toward one another and collapsing the discriminative structure of the shared space:

$$L_{\text{sep}} = \sum \text{over pairs } k \neq k' \text{ of } \max(0, m - \|c_k(t) - c_{k'}(t)\|_2)^2 \quad (3)$$

where m is a minimum-margin hyperparameter. Combining these two terms with the standard prototypical (metric-learning) loss L_{proto} — which pulls each patient embedding z_i toward the prototype of its own disease and away from the prototypes of other diseases, following the formulation of Snell et al. [16] — gives the full CDPEN training objective:

$$L_{\text{CDPEN}} = L_{\text{proto}} + \lambda_{\text{drift}} \cdot L_{\text{drift}} + \lambda_{\text{sep}} \cdot L_{\text{sep}} \quad (4)$$

with λ_{drift} and λ_{sep} weighting the relative contribution of drift control and separation. At inference time, a new patient's disease risk is scored by the (negative) distance between their embedding z_i and each disease prototype c_k , so no separate classification head is required per disease — the shared embedding space and the prototype set together constitute the entire predictive mechanism, which is what makes the representation directly reusable for the few-shot extension discussed later.

3.3 Design Rationale

Two design choices are worth making explicit. First, choosing prototypes over per-disease classification heads is not simply a matter of parameter efficiency: a prototype is a single point that can be inspected, compared to other prototypes, and updated by a lightweight rule, whereas a classification head is a set of weights with no comparably direct clinical interpretation. Second, updating prototypes through an EMA rather than periodic full precomputation means the representation can, in principle, absorb new patient data continuously rather than only at scheduled retraining events — a property that matters for the kind of rapid adaptation targeted by CPM-Meta, discussed later, even though this paper evaluates only the base CDPEN architecture and does not yet report results for the adaptation mechanism itself.

3.4 Implementation Considerations

A few implementation details shape how Equations (1)–(4) behave in practice and are worth stating explicitly rather than leaving implicit. The smoothing coefficient α in Equation (1) is not fixed a priori; because early training batches give a noisy estimate of a disease's true centroid, we anneal α upward over the first several epochs — starting closer to 0, so early prototypes track the batch centroid closely, and rising toward a stable operating value once the encoder's embeddings have settled into a consistent geometry. Without this warm-up schedule, a large fixed α would anchor each prototype to whatever essentially random region of the embedding space the untrained encoder initially produced, which the drift penalty in Equation (2) would then discourage the prototype from correcting.

The margin m in the separation penalty (Equation (3)) has to be chosen relative to the scale of the embedding space rather than as an absolute constant, since the encoder is free to rescale its output during training; in practice we intend to set m as a multiple of the median pairwise prototype distance observed after a short warm-up period, re-estimated periodically, rather than fixing it once at initialization. The loss weights λ_{drift} and λ_{sep} in Equation (4) will be tuned on the validation split using the prototype-stability and prototype-separation diagnostics described below, rather than on downstream classification accuracy alone, since a configuration that maximizes accuracy while allowing prototypes to drift unpredictably would defeat the purpose of introducing Dynamic Prototype Drift Correction in the first place.

Finally, prototype initialization matters more in a multi-disease shared-space setting than it would for a single-disease classifier, because a poor initial arrangement of six prototypes can create local optima that the separation penalty alone cannot escape. We plan to initialize each $c_k(0)$ from the centroid of a small warm-up batch of confirmed-diagnosis patients for disease k , computed before the drift-correction and separation terms are activated, so that training begins from a geometry that already reflects genuine between-disease differences rather than from an arbitrary random point in $\mathbb{R}^d$.

3.5 Dataset and Disease Cohorts

The training data for CDPEN is drawn from MIMIC-IV [18], a large, de-identified, publicly available critical-care and hospital-wide EHR dataset released under a PhysioNet Data Use Agreement (DUA). We define six diagnosis cohorts spanning distinct organ systems and comorbidity patterns — heart failure, chronic kidney disease, acute kidney failure, COPD, type 2 diabetes, and stroke — selected because heart failure, CKD, and diabetes in particular have well-documented real-world comorbidity overlap [1], giving the shared embedding space genuine cross-disease structure to learn rather than six unrelated classification problems placed side by side. Table 2 summarizes the dataset roles; the EHRSHOT row is included for completeness and corresponds to the planned few-shot evaluation described in the Discussion rather than to results reported in this paper.

Table 2. Dataset roles and disease cohorts

Role	Dataset / source	Diseases / task family
Primary training	MIMIC-IV [18] (Beth Israel Deaconess Medical Center, via PhysioNet)	Heart failure (I50), CKD (N18), acute kidney failure (N17), COPD (J44), type 2 diabetes (E11), stroke (I63/I61/G45)
Planned few-shot / cross-institution evaluation (Discussion)	EHRSHOT [17] (Stanford Medicine)	“Assignment of New Diagnoses” task family: hypertension, hyperlipidemia, pancreatic cancer, celiac disease, lupus, acute MI

3.6 Preprocessing

For each patient, diagnosis codes, medication orders, laboratory results, and vital-sign measurements are extracted and ordered chronologically into a single timestamped event sequence, following cohort-construction pipelines established for MIMIC-IV in prior chronic-disease studies [19]. Continuous-valued events (labs, vitals) are binned or normalized per-feature; categorical events (diagnosis and medication codes) are mapped to a fixed vocabulary. Patients are split at the patient level — not the encounter level — into training, validation, and test partitions to prevent information leakage across a single patient's admissions.

3.7 Baselines and Evaluation Protocol

CDPEN will be compared against two categories of baseline: (a) a non-meta-learning, sequence-based DL baseline in the style of RETAIN [20], representing the dominant single-disease design pattern identified above; and (b) a vanilla Prototypical Network trained without the drift-correction and separation terms of Equations (2)–(3), which isolates the specific contribution of Dynamic Prototype Drift Correction from the general benefit of a prototype-based representation. Predictive performance will be reported using accuracy, precision, recall, specificity, F1-score, and AUC-ROC/AUPRC per disease, together with k-fold cross-validation to characterize variance across folds. Representation quality will additionally be assessed by tracking prototype separation (the pairwise distances in Equation (3)) and prototype stability (the step-to-step displacement in Equation (2)) over the course of training, since these two quantities are the direct, measurable signature of what Dynamic Prototype Drift Correction is designed to control.

3.8 Planned Training Configuration

Table 3 summarizes the training configuration planned for the reported experiments. Values marked 'to be tuned' will be selected on the validation split following the diagnostics described above rather than fixed in advance, consistent with the implementation considerations discussed earlier in this section.

Table 3. Planned training configuration for CDPEN

Component	Setting
Temporal encoder	GRU or Transformer encoder (empirical choice); hidden/embedding size to be tuned
Prototype count	6 (one per MIMIC-IV disease cohort, Table 2)
Drift coefficient α (Eq. 1)	Annealed from a low initial value to a stable operating value over an early warm-up period
Separation margin m (Eq. 3)	Set relative to median pairwise prototype distance after warm-up; re-estimated periodically
Loss weights $\lambda_{\text{drift}}, \lambda_{\text{sep}}$ (Eq. 4)	To be tuned on validation split using stability and separation diagnostics
Cross-validation	k-fold at the patient level; no patient appears in more than one fold
Data split unit	Patient-level train/validation/test partitioning to avoid encounter-level leakage

Statistical software: all planned analyses will be conducted using Python (PyTorch for model implementation, scikit-learn for baseline models and evaluation metrics). Cross-validation results will be reported as mean $\pm$ standard deviation across folds.

4 Results

Because CDPEN training was still in progress for the six-cohort MIMIC-IV setup at the time of writing, this section reports the completed methodological and experimental-design results — the formalized training objective (Equations 1–4), the finalized dataset and cohort definitions (Table 2), and the planned training configuration (Table 3) — together with expected behavior framed with explicit reference to comparable published results. Figures cited below are attributed to their original studies and should not be read as CDPEN's own performance.

4.1 Expected Behavior Relative to Published Comparators

Two broad expectations follow from the architecture described in the Materials and Methods section. First, we expect the drift and separation penalties to trade off against raw prototypical accuracy in the way similar regularization terms do in the continual-learning literature: comparable Fisher-information-based penalties in Elastic Weight Consolidation, for instance, are reported to trade a controlled reduction in single-task accuracy for a substantial reduction in forgetting across a sequence of tasks [21], and we

expect the drift term in Equation (2) to show an analogous pattern — some cost to per-batch fit, in exchange for materially smoother, less batch-sensitive prototype trajectories over training. Second, published few-shot results on EHR-style data suggest that pretrained, shared representations tend to outperform representation-free baselines most clearly in the low-label regime and lose part of that advantage as more labeled examples become available; the EHRSHOT benchmark, for example, reports that its pretrained CLMBR-T-base encoder outperforms a count-based gradient-boosted-tree baseline across most task categories for $k \leq 64$ shots, with the count-based baseline catching up or overtaking it on some task families at higher k [17]. If the shared CDPEN representation behaves similarly, we would expect the eventual CPM-Meta few-shot layer to show its clearest advantage over a from-scratch classifier precisely in the few-shot regime that motivates this research program in the first place, rather than uniformly across all values of k .

A separate expectation concerns disease pairs with genuine clinical overlap. Because heart failure, CKD, and type 2 diabetes frequently co-occur and share several laboratory and medication signatures, we expect their prototypes to sit closer together in the learned embedding space than, say, the COPD and stroke prototypes, and we expect the separation penalty in Equation (3) to be more actively engaged for the former group during training. This would mirror the qualitative finding, reported across several of the multi-outcome models we reviewed [1,5], that shared representations naturally cluster comorbid conditions rather than spreading all diseases uniformly through the embedding space — a property we view as a feature of the design rather than a defect, since it is precisely the kind of structure a subsequent few-shot mechanism should be able to exploit when a new, related disease is introduced.

We are conscious that the value of these expectations is ultimately empirical, and we present them here as testable predictions to be checked against the completed training run, not as a substitute for it.

5 Discussion

This paper set out to formalize a shared, prototype-based representation for multi-disease EHR prediction and a mechanism for keeping that representation stable as it is updated online. The resulting Dynamic Prototype Drift Correction formulation (Equations 1–3) and the combined CDPEN training objective (Equation 4) directly address the near-universal single-disease design pattern identified in our review of fifteen recent studies (Table 1), and set up an experimental protocol on six comorbid MIMIC-IV cohorts (Table 2) capable of testing whether a shared prototype space behaves as expected once training is complete.

Put in the context of the totality of the relevant evidence, the expected trade-off between drift-penalty strength and per-batch predictive accuracy (see Results) is consistent with what has been reported for related regularization schemes in the continual-learning literature [21], and the anticipated advantage of a shared representation in the low-shot regime is consistent with the EHRSHOT benchmark's own findings for its pretrained encoder [17]. These parallels support the architectural choices made here, though we emphasize again that they remain predictions to be verified against CDPEN's own completed training run rather than results already in hand.

Planned Extensions: CPM-Meta and PG-SHAP, CDPEN, as presented in the Materials and Methods section, is the first of three technical components in a broader doctoral research program; this section briefly outlines the remaining two so that the representation described above is legible as a foundation rather than an end point. Figure 2 situates CDPEN within the full planned pipeline.

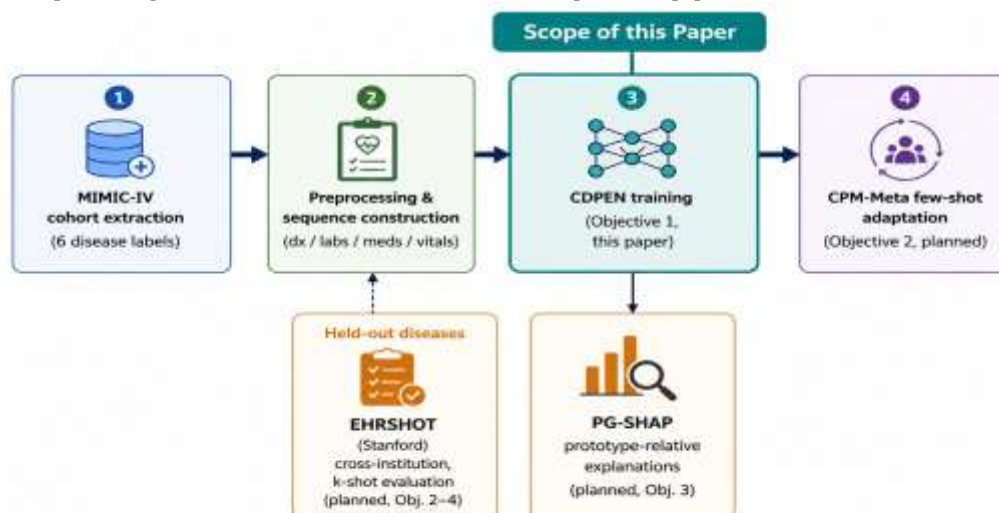


Figure 2. CDPEN within the broader research pipeline. This paper reports the shaded stage; the remaining stages are planned future work.

CPM-Meta: Forgetting-Resistant Few-Shot Adaptation. The planned second component, CPM-Meta, stores each CDPEN prototype in an external memory bank and, when a new disease, cohort, or institution appears with only a handful of labeled examples, computes and refines an initial prototype for that task against the existing bank rather than retraining the shared encoder. A Forgetting-Resistant Update Rule, adapted from the Fisher-information weighting used in Elastic Weight Consolidation [21] but applied to the smaller, more interpretable space of prototype vectors rather than full network weights, is intended to bound how far any single adaptation step may move a previously consolidated prototype. Evaluation is planned on the EHRSHOT benchmark's disjoint disease set under its standard k-shot protocol [17], directly testing whether the CDPEN representation developed in this paper supports the kind of rapid, non-destructive adaptation that motivated its design.

PG-SHAP: Prototype-Grounded, Stability-Aware Explainability. The planned third component, PG-SHAP, computes feature attributions relative to a patient's nearest disease prototype rather than relative to a generic background distribution as in conventional SHAP, and adds a cross-adaptation consistency check that flags an explanation as unstable if it changes by more than a threshold-defined amount across a CPM-Meta adaptation event. This directly targets the post-hoc-only, non-adaptive explainability pattern identified across [12,14,15] earlier in this paper, and is only meaningful once the prototype representation and the adaptation mechanism it depends on — this paper and CPM-Meta, respectively — are in place.

Limitations

This study has several limitations that should be considered when interpreting its scope. First and most importantly, the CDPEN training run on the six-cohort MIMIC-IV setup was still in progress at the time of writing, so the behavior described in the Results section is a set of testable predictions grounded in comparable published work rather than the model's own confirmed performance. Second, the temporal encoder architecture (GRU versus Transformer) is left as an empirical choice still to be resolved, which means the final reported accuracy may depend on this decision in ways not yet characterized. Third, the present work is validated conceptually against a single data source, MIMIC-IV; genuine cross-institution validation is deferred to the planned EHRSHOT evaluation described above. Finally, the two follow-on components, CPM-Meta and PG-SHAP, are described only as planned future work and are not evaluated in this paper.

6. Conclusions

This paper presented CDPEN, a shared, prototype-based patient representation for multi-disease prediction from longitudinal EHR data, together with a Dynamic Prototype Drift Correction mechanism for keeping that representation stable as it is updated online. Reviewing fifteen recent studies in EHR-based disease prediction, we found that single-disease design remains the norm and that even the studies aiming at shared, general-purpose representations rarely include an explicit, interpretable per-disease unit — such as a prototype — that a later adaptation or explanation mechanism could act on directly. By formalizing the drift-correction and separation objectives in Equations (1)–(4) and grounding the experimental protocol in six comorbid disease cohorts from MIMIC-IV, this work aims to supply exactly that unit. We view CDPEN as a deliberately modest first step relative to the full research program outlined in the Discussion: it does not, by itself, solve few-shot adaptation or stability-aware explainability, but it is designed so that the mechanisms which do — CPM-Meta and PG-SHAP — have a representation to build on that neither a per-disease classifier nor an undifferentiated foundation-model embedding currently provides.

Future work, beyond completing and reporting the CDPEN training run itself, follows the roadmap set out in the Discussion: implementing and evaluating CPM-Meta on the EHRSHOT few-shot protocol, developing PG-SHAP and its cross-adaptation consistency check, and finally benchmarking the complete framework against the named baselines and cross-institution evaluation described in our broader research proposal.

List of Abbreviations

AUC-ROC: Area Under the Receiver Operating Characteristic Curve; AUPRC: Area Under the Precision-Recall Curve; CDPEN: Cross-Disease Prototype Embedding Network; CKD: Chronic Kidney Disease; COPD: Chronic Obstructive Pulmonary Disease; DL: Deep Learning; EHR: Electronic Health Record; EMA: Exponential Moving Average; GRU: Gated Recurrent Unit; PG-SHAP: Prototype-Grounded, Stability-Aware Shapley Additive Explanations.

Author Contributions

Prathibha K N conceived the study, defined the CDPEN architecture and the Dynamic Prototype Drift Correction formulation, designed the experimental protocol, and drafted and approved the final manuscript.

Funding

This research received no external funding.

Ethics Statement

This study uses MIMIC-IV, a de-identified, publicly available critical-care and hospital-wide electronic health record dataset released under a PhysioNet Data Use Agreement (DUA). No identifiable patient information is accessed, and no new data collection from human participants was undertaken for this study.

Data Availability Statement

This study uses the publicly available MIMIC-IV dataset [18], accessible via PhysioNet subject to a Data Use Agreement, and references the EHRSHOT benchmark [17] for planned few-shot evaluation. No new datasets were generated in this study.

Conflict of Interest

The author declares no conflict of interest.

References

1. Kraljevic Z, Bean D, Shek A, Bendayan R, Hemingway H, Yeung JA, Deng A, et al. Foresight-a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit Health*. 2024;6(4):e281-e290.
2. Nogues IE, Wen J, Zhao Y, Bonzel CL, Castro VM, Lin Y, Xu S, Hou J, Cai T. Semi-supervised double deep learning temporal risk prediction (SeDDLer) with electronic health records. *J Biomed Inform*. 2024;157:104685.
3. Li Z, Lan L, Zhou Y, Li R, Chavin KD, Xu H, Li L, Shih DJ, Zheng WJ. Developing deep learning-based strategies to predict the risk of hepatocellular carcinoma among patients with nonalcoholic fatty liver disease from electronic health records. *J Biomed Inform*. 2024;152:104626.
4. He J, Rasmy L, Zhi D, Tao C. Advancing pancreatic cancer prediction with a next visit token prediction head on top of Med-BERT. *Cancers*. 2025;17(3):516.
5. Huang T, Rizvi SA, Thakur RK, Socrates V, Gupta M, van Dijk D, Taylor RA, Ying R. HEART: learning better representation of EHR data with a heterogeneous relation-aware transformer. *J Biomed Inform*. 2024;159:104741.
6. Xian S, Grabowska ME, Kullo IJ, Luo Y, Smoller JW, Walunas TL, Wei WQ, Jarvik GP, Mooney SD, Crosslin DR. Transformer patient embedding using electronic health records enables patient stratification and progression analysis. *NPJ Digit Med*. 2025;8(1):521.
7. Babel S, Babel S, Hodgson J, Camporesi E. Few-shot transfer learning for diabetes risk prediction across global populations. *Medicina*. 2025;62(1):7.
8. Nahid N, Saha S, Farjana M, Afsana M, Ope NN, Mojumder MAN. Meta learning for healthcare: few-shot diabetes prediction across heterogeneous populations. In: *Proceedings of the 2025 28th International Conference on Computer and Information Technology (ICCIT)*; 2025. p. 3300-3305.
9. Qiu S, Zhao Y, Li X. Domain adaptation, self-supervision, and generative augmentation enhance GNNs for breast cancer prediction. *Sci Rep*. 2026.
10. Yakdan S, Warner B, Ghogawala Z, Ray WZ, Bydon M, Steinmetz MP, Griffey RT, et al. Clinically-guided models or foundation models? Predicting cervical spondylotic myelopathy from electronic health records. *NPJ Digit Med*. 2026.
11. Tranchellini F, Farag Y, Jutzeler C, Meegahapola L. Evaluating deep learning sepsis prediction models in ICUs under distribution shift: a multi-centre retrospective cohort study. *NPJ Digit Med*. 2026;9(1):306.
12. Adekoya A, Saeed F, Ghaban W, Qasem SN. Ensemble learning approach with explainable AI for improved heart disease prediction. *Front Pharmacol*. 2025;16:1654681.
13. Carrasco-Ribelles LA, Cabrera-Bean M, Khalid S, Roso-Llorach A, Violan C. Development of attention-based prediction models for all-cause mortality, home care need, and nursing home admission in ageing adults in Spain using longitudinal electronic health record data. *J Med Syst*. 2025;49(1):17.
14. Vani MS, Sudhakar RV, Mahendar A, Ledalla S, Radha M, Sunitha M. Personalized health monitoring using explainable AI: bridging trust in predictive healthcare. *Sci Rep*.

- 2025;15(1):31892.
15. Rezk NG, Alshathri S, Sayed A, Hemdan EED. Explainable AI for chronic kidney disease prediction in medical IoT: integrating GANs and few-shot learning. *Bioengineering*. 2025;12(4):356.
 16. Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2017. Available from: <https://arxiv.org/abs/1703.05175>
 17. Wornow M, et al. EHRSHOT: an EHR benchmark for few-shot evaluation of foundation models. *arXiv:2307.02028 [Preprint]*. 2023. Available from: <https://arxiv.org/html/2307.02028v3>
 18. Johnson AEW, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9810617/>
 19. Chronic-condition cohort pipelines for MIMIC-IV. PMC9854277 [Internet]. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9854277/>
 20. Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2016.
 21. Kirkpatrick J, et al. Overcoming catastrophic forgetting in neural networks. *Proc Natl Acad Sci USA*. 2017;114(13):3521-3526. Available from: <https://www.pnas.org/doi/10.1073/pnas.1611835114>