Hyperbolic Diffusion Model For Hierarchical EHR Data Augmentation And Risk Prediction

By Abhijeet Sahdev (as4673, Group 15), Dept. of Computer Science, NJIT

Instructor: Prof. Mengjia Xu, Dept. of Data Science, NJIT

Report, Code

Contents

Background

- Electronic health records (EHR) are hierarchical and irregular time series [1].
  - Each patient has a chronological sequence of visits (admissions, ED visits, follow-ups).
  - Each visit is represented as a set of ICD diagnosis and procedure codes.
  - Visits vary in length, frequency, and sparsity, creating irregular trajectories.
  - These trajectories reflect evolving clinical states critical for risk modeling

- Standard Euclidean models flatten this structure and distort clinical relationships.

- ICD diagnosis and procedure codes live in deep tree-like ontologies [2].
  - ICD codes lie in a tree-structured ontology (chapters → categories → subcodes).
  - Semantically related codes are close in the ICD graph, but not in Euclidean space.
  - This motivates hyperbolic embeddings, where tree-like data is naturally preserved [3-6].
  - Code co-occurrence across patients builds a clinical similarity graph [7]

Motivation

- MedDiffusion models EHR in Euclidean space, thus ignores manifold curvature [8].
- Synthetic trajectories often lose hierarchical consistency of ICD codes.
- Geometry-mismatched synthetic data can hurt downstream risk prediction.
- Hence, there's a need for geometry-aware augmentation pipeline that respects ontology and time [3-6].

Problems to Address / Objectives

- Structural misalignment in synthetic EHR
- Geometry-agnostic generative models
- Poor generalization

Thus, my goals are:

- Build curvature-aware generative + predictive model [3-6].
- Align hyperbolic code geometry with graph diffusion distances [3, 7].
- Improve heart failure risk discrimination (a study) vs. MedDiffusion baseline [8].
- Provide a reproducible, end-to-end pipeline from EHR to synthetic data and risk scores built on MIMIC-III v1.4 [1].

Methodology

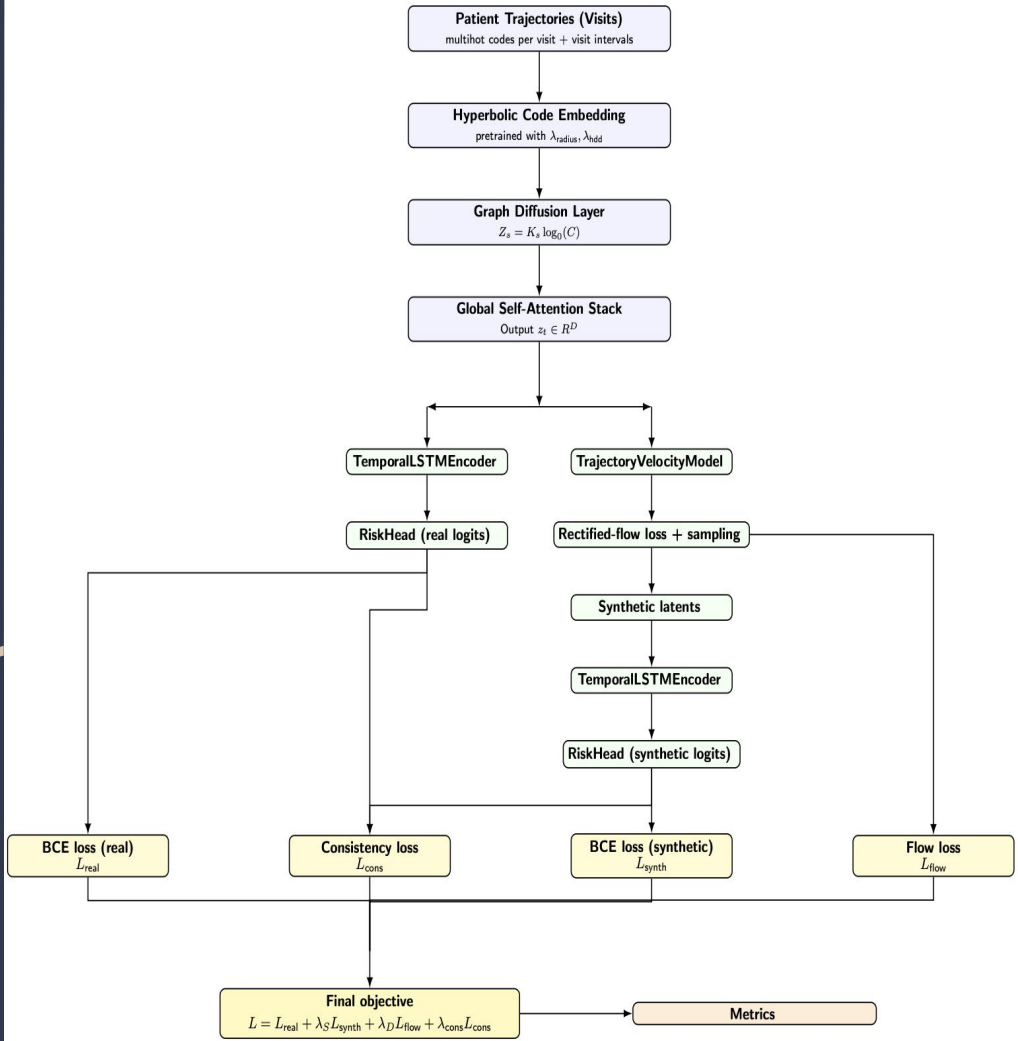

Fig 1 HyperMedDiff-Risk Model Architecture

Methodology

- Cohort Construction

- MIMIC-III v1.4 critical-care database [1].
- Predict incident heart failure (HF) at discharge from longitudinal pre-index history.
- Follows RETAIN and MedDiffusion cohort designs [9,8].
- Positive Cohort (HF Cases)
    a. Adults (age ≥ 18) with ICD-9 428.* HF diagnosis [8,9] .
    b. Index admission = first HF-coded admission.
    c. Standard exclusions:
        i. <18 years, ELECTIVE admissions,
        ii. <2 lifetime visits,
        iii. in-hospital death during index stay.
- Controls : No HF diagnosis at any point (no ICD-9 428.x codes) [8].

Methodology

- Cohort Construction

- For each HF patient select up to two controls matching:
  - Exact gender & race (White, Black, Hispanic, Asian)
  - Exact age
  - Comparable visit count: control's lifetime visits $\in [N,N+4][N,N+4]$.
    i. Prevents models from exploiting sequence length differences
- Observation Window
  a. Use a one-year history before the index date for model inputs.
  b. Collect all inpatient visits in this window; extract diagnosis & procedure codes.
  c. Remove E-codes to reduce noise.

Methodology

- Cohort Construction

- Trajectory Representation
  a. Each patient: sequence of visits $\{V_p\}=(V_{p,1},\ldots,V_{p,Kp})$.
  b. Each visit: multi-hot vector over a global vocabulary of ICD-9 + PCS codes.
  c. Vocabulary built from entire MIMIC history for full clinical coverage.
- Matching MedDiffusion's visit statistics requires a **one-year window**.
  - A strict 6-month window cannot reach their reported 2.61 visits/patient. MedDiffusion's statistics reflect a broader lookback despite referring to a 6-month window in the paper.

Metric	Ours	MedDiffusion
Positive cases (HF)	2,835	2,820
Negative controls	4,566	4,702
Avg. visits / patient	2.62	2.61
Avg. codes / visit	13.39	13.06
Unique ICD-9 tokens	4,844	4,874

Table 1 Cohort Metrics

Methodology

- Hyperbolic Embeddings

- ICD codes are embedded in hyperbolic space B^d using Poincaré ball embeddings [3].
- Pretrained with radius regularization and HDD loss to capture clinical hierarchy and co-occurrence structure [3].
  - We scan the cohort to build a co-occurrence adjacency matrix, normalize it into a transition matrix, and run multi-step random walks.
  - For each code, diffusion outputs across steps are concatenated into a diffusion signature (f_i).
  - HDD aligns hyperbolic distances with these diffusion profiles.
- Final hyperbolic embeddings are frozen for downstream training.

$$\mathcal{L}_{\text{pre}} = \lambda_{\text{radius}} \underbrace{\frac{1}{N} \sum_i (\|c_i\|_{\mathbb{B}} - r^\star)^2}_{\mathcal{L}_{\text{radius}}} + \lambda_{\text{hdd}} \underbrace{\mathbb{E}_{i,j} \left(\|f_i - f_j\|_2 - d_{\mathbb{B}}(c_i, c_j) \right)^2}_{\mathcal{L}_{\text{HDD}}}$$

Methodology

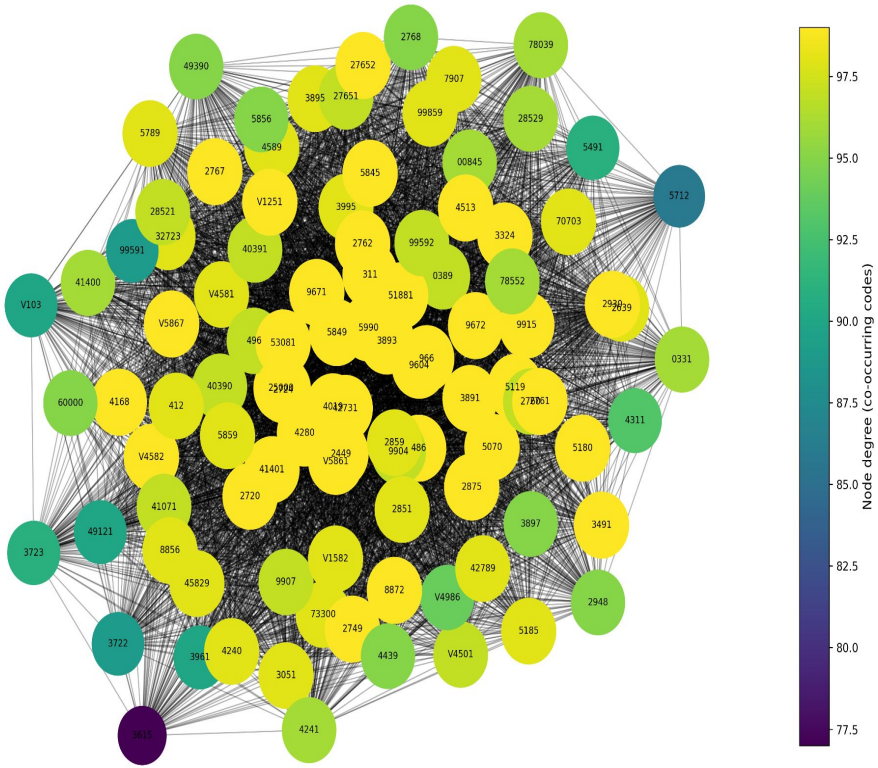

Fig 2: MIMIC III Heart Failure Hypergraph Top ICD 9 Code Co-Occurrences

Methodology

- Hyperbolic Graph Encoder

Stage 1: Hyperbolic → Tangent Diffusion Encoding

- Each visit is represented as a multi-hot ICD code vector.
- Codes are retrieved from the hyperbolic embedding table.
- Embeddings are log-mapped to the tangent space at the origin to enable linear diffusion [7].
- Graph diffusion kernels propagate information across clinically related codes.
- Outputs from multiple diffusion scales are concatenated → a curvature-aware visit feature capturing local + global ICD neighborhoods.

Methodology

- Hyperbolic Graph Encoder

Stage 2: Global Self-Attention over Codes

- Diffused code representations are refined with stacked self-attention layers with feed forward neural network and layer norm.
- Attention contextualizes each code with respect to the entire vocabulary, not just graph neighbors.
- Enhances modeling of:
  - Comorbidities,
  - Diagnosis–procedure interactions,
  - Chronic disease patterns and long-range dependencies.
- Produces globally contextualized, diffusion-aware code embeddings.

Methodology

- Hyperbolic Graph Encoder

Stage 3: Time-Aware Visit Pooling using temporal encodings consistent with prior EHR sequence models [9]

- Code-level representations are pooled in tangent space to form a single embedding per visit.
- A temporal encoding modulates each visit by its inter-visit gap Δt, capturing clinically meaningful irregular follow-ups.
- Produces a sequence of visit-level embeddings $Z_{data} \in R^{B \times L \times d}$ integrating:
  - Hyperbolic ICD geometry,
  - Multi-hop graph diffusion structure,
  - Longitudinal temporal dynamics.
  - Here, B is the batch size, L is the sequence length and d is the embedding dimension.

Methodology

- ## Rectified Flow Trajectory Model

- Replaces stochastic DDPM with a deterministic rectified-flow transport field [11].
- Given real visit embeddings z_1 and noise samples z_0 form linear interpolants z_t.
- Train a velocity field $v_\theta(z_t, t, h_{<t})$:o predict the direction needed to move $z_0 \rightarrow z_1$.
- Optimize

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,z_0,z_1} \left\| v_\theta(z_t, t, h_{<t}) - (z_1 - z_0) \right\|_2^2$$

  evaluated only on real visit positions.
- During generation, we integrate the velocity field forward from pure noise, producing stable synthetic trajectories conditioned on the evolving patient state.
- Benefits:
  - Deterministic velocity transport (no stochastic drift)
  - Avoids DDPM noise-induced instability
  - Generates geometry-consistent trajectories in hyperbolic tangent space

Methodology

- Temporal risk encoder and prediction head

- Both real and flow-generated visit sequences are processed by a single-layer TemporalLSTMEncoder with masking [9].
- The LSTM produces per-visit hidden states $h_{p,t'}$ each summarizing all prior visits.
- A pooled representation $h_p = \text{LSTMPool}(h_{p,1:L})$ serves as the final patient summary vector.
- The RiskHead (linear layer + sigmoid) outputs logits $\hat{y}_p$ and predicted HF probabilities $\sigma(\hat{y}_p)$ [8].
- Mirrors the MedDiffusion backbone, but operates on richer, curvature-aware visit embeddings, improving discrimination and temporal modeling.

Methodology

- Final Loss

- Real Risk BCE $\mathcal{L}_{\text{real}} = \frac{1}{B}\sum_{p=1}^{B}\text{BCE}(\hat{y}_p, y_p)$
- Synthetics Risk BCE $\mathcal{L}_{\text{synth}} = \frac{1}{B}\sum_{p=1}^{B}\text{BCE}(\hat{y}_p^{\text{synth}}, y_p)$
- Feature Consistency (our new term)

$$\mathcal{L}_{\text{cons}} = \text{MSE}\left(h_{\text{real}}, h_{\text{synth}}\right)$$

- Finally,

$$\mathcal{L}_{\text{HyperMedDiff}} = \mathcal{L}_{\text{real}} + \lambda_S \mathcal{L}_{\text{synth}} + \lambda_D \mathcal{L}_{\text{flow}} + \lambda_{\text{consistency}} \mathcal{L}_{\text{cons}}.$$

Experiments

- Given the imbalance in the dataset as shown in Table 1 Cohort Metrics, we use the following metrics:
  - Area under Precision Recall Curve (AUPRC)
  - Cohen's Kappa
  - Embedding Correlation (for geometry)
- The baseline configuration is diffusion steps = [1, 2, 4, 8], embed dim = 128, dropout = 0.2, λ_{HDD} = 0.02, λ_{radius} = 0.003, λ_S = 1.0, λ_D = 1.0, $\lambda_{consistency}$ = 0.1, learning rate = 10^{-4} and training epochs = 100.
-

Experiment	Modification Relative to Baseline
01_Baseline	None
02_NoDiffusion	diffusion_steps = [1]
03_LocalDiff	diffusion_steps = [1, 2]
04_GlobalDiff	diffusion_steps = [1, 2, 4, 8, 16]
05_NoHDD	$\lambda_{HDD} = 0.0$
06_StrongHDD	$\lambda_{HDD} = 0.1$
07_HighDropout	dropout = 0.5
08_SmallDim	embed_dim = 64
09_DiscrimOnly	$\lambda_S = 0.0$
10_GenFocus	$\lambda_S = 2.0$

Table 2 : Ablation Study Configurations

Discussion

- Geometry shapes structure, not necessarily accuracy but not great for interpretability [3-6].
- Graph diffusion depth controls the manifold scale.
- Hyperbolic alignment enables controllable geometry.
- Generative supervision modulates stability [11].

Run	Experiment	AUPRC	Kappa	Corr
0	MedDiffusion (paper)	0.7064	0.4526	N/A
1	Base	0.7991	0.5046	0.8385
2	02_NoDiffusion	0.7919	0.5046	0.8897
3	03_LocalDiff	0.8054	0.5046	0.8533
4	04_GlobalDiff	0.8058	0.5046	0.8380
5	05_NoHDD	0.8291	0.5046	−0.0021
6	06_StrongHDD	0.8022	0.5046	0.9071
7	07_HighDropout	0.8063	0.5046	0.8127
8	08_SmallDim	0.7928	0.5046	0.7374
9	09_DiscrimOnly	0.8125	0.5046	0.8261
10	10_GenFocus	0.8115	0.5046	0.8419

Table 3 : Ablation results for HyperMedDiff-Risk on MIMIC-III HF prediction.

Comparison Summary

- **Performance Gains:** HyperMedDiff-Risk surpasses the MedDiffusion AUPRC (0.7064) in every configuration [8]. The baseline reaches 0.7991, and several variants exceed 0.81, showing that hyperbolic graph-diffusion encoding and rectified-flow transport yield a more informative latent structure than the original Euclidean DDPM design.
- **Calibration & Agreement:** All supervised models maintain $\kappa \approx 0.50$, outperforming MedDiffusion (0.4526). Diffusion and geometric variants shift Corr without affecting κ, indicating that encoder geometry reshapes the embedding structure while leaving overall predictive agreement stable [3-7].

Future Work

- Hyperparameter tuning.
- Implementation of this model on other risk prediction tasks specified in MedDiffusion [8].
- Implementation of a purely generative model with a decoder that performs well on recall.
- Expand hyperbolic/graph-based modeling to other hierarchical health ontologies [3–6].

Weaknesses

- NoHDD variant attains strong AUPRC but loses geometric interpretability [3].
- LSTM bottleneck relative to more expressive encoders [10].
- Limited to MIMIC-III HF cohort [1].

Acknowledgements

- I thank Yiran Huang for her assistance in setting up and debugging the NJIT Wulver HPC environment, which enabled the large-scale diffusion and ablation experiments conducted in this work.
- I am grateful to Prof. Mengjia Xu for her guidance
throughout the development of the project.
- I also thank Sarang Patil for formulating the initial idea that evolved into the HyperMedDiff-Risk architecture.
- Finally, I acknowledge the use of generative AI tools, including ChatGPT and Codex, for proofreading, experimental verification and cross-checking mathematical consistency during adaptation of the MedDiffusion pipeline.

Citations

[1] Johnson, A. E. W., et al. “MIMIC-III, a freely accessible critical care database.” Scientific Data, 3(1), 2016.

[2] World Health Organization. “ICD-9-CM: International Classification of Diseases, Ninth Revision.” WHO, 1977.

[3] Nickel, M., & Kiela, D. “Poincaré Embeddings for Learning Hierarchical Representations.” NeurIPS, 2017.

[4] Ganea, O., et al. “Hyperbolic Neural Networks.” NeurIPS, 2018.

[5] Chami, I., et al. “Hyperbolic Graph Convolutional Neural Networks.” NeurIPS, 2019.

[6] Sala, F., et al. “Representation Tradeoffs for Hyperbolic Embeddings.” ICML, 2018.

[7] Coifman, R. R., & Lafon, S. “Diffusion Maps.” Applied and Computational Harmonic Analysis, 21(1), 2006.

[8] Y. Zhong, S. Cui, J. Wang, X. Wang, Z. Yin, Y. Wang, H. Xiao, M. Huai, T. Wang, and F. Ma, “Meddiffusion: Boosting health risk
prediction via diffusion-based data augmentation,” 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11469648/

[9] Choi, E., et al. “RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention.” NeurIPS, 2016.

[10] Vaswani, A., et al. “Attention is All You Need.” NeurIPS, 2017.

[11] Liu, X., et al. “Flow Straight and Fast: Learning to Generate via Rectified Flow.” ICML, 2022.