ExpertODE: Continuous Diagnosis Prediction with Expert Enhanced Neural Ordinary Differential Equations

1st Hengyu Zhang
College of Computer and Data Science
Fuzhou University
Fuzhou, China
hengyuzhang07@gmail.com

2nd Hang Lv
College of Computer and Data Science
Fuzhou University
Fuzhou, China
lvhangkenn@gmail.com

3rd Yanchao Tan*
College of Computer and Data Science
Fuzhou University
Fuzhou, China
yctan@fzu.edu.cn

4th Guofang Ma
School of Computer Science
Zhejiang Gongshang University
Zhejiang, China
maguofang199l@zju.edu.cn

5th Fan Wang
College of Computer Science
Zhejiang University
Zhejiang, China
fanwang97@zju.edu.cn

6th Carl Yang
Department of Computer Science
Emory University
Atlanta, United States
j.carlyang@emory.edu

***Abstract*—Continuous diagnosis prediction based on multimodal electronic health records (EHRs) of patients is a promising yet challenging task for AI in healthcare. Existing studies ignore abundant domain knowledge of diseases (e.g., specific medical terms and their interrelations) in textual EHRs, which fails to accurately predict disease progression and assist in sequential diagnosis prediction. To this end, we first propose an Expert enhanced neural Ordinary Differential Equations (ExpertODE) framework for continuous diagnosis prediction. In particular, we first propose a novel Mixture of Language Experts (MoLE) module to enhance disease embeddings with domain knowledge. Furthermore, we propose a Contrastive Neural Ordinary Differential Equation (CNODE) module to continuously model temporal correlations of disease progression, and implement a unified contrastive learning framework to jointly optimize the domain-based MoLE module and the temporal-based CNODE module. Extensive experiments on two real-world textual EHR datasets show significant performance gains brought by our ExpertODE, yielding average improvements of 3.91% for diagnosis prediction over state-of-the-art competitors.**

***Index Terms*—Diagnosis Prediction, Language Experts, Contrastive Neural ODE**

I. Introduction

Multimodal electronic health records (EHRs) are valuable data sources for researchers to construct continuous diagnosis prediction and assist clinical decision-making [13], [20]. These data encompass various formats, including texts (clinical notes, diagnostic records), signals (sensor records), images (ultrasound scans), environmental data, and behavioral data.

Among them, textual data has garnered attention for the predictive models in medical applications. This is largely due to the rich **domain knowledge** encapsulated in texts (e.g., specific terms from semantic domain and their interrelations from clinical domain). Many recent studies leverage deep learning models to excavate the semantic interplay in textual EHRs, such as Recurrent Neural Networks [1], Convolutional Neural Networks [19], Attention-based mechanisms [9], and Transformers [21]. However, several challenges remain in delving into abundant textual EHR data:

Challenge I. *How to effectively utilize domain knowledge for accurate disease representations?* As shown in Fig. 1, the semantic relations between Allergic Asthma and Bronchial Asthma can be understood by general language experts due to the phrase overlapping of "Asthma". However, different asthma can correspond to different syndromes under the knowledge of clinical experts. For example, Allergic Asthma exhibits a notable degree of stability over time, while Bronchial Asthma phenotypes often demonstrate a trend toward syndromic exacerbation. Although general Language Models (LMs) equipped with a broad spectrum of knowledge can model the semantic similarities among disease names (e.g., BERT [5] and GPT-2 [14]), they fall short in capturing specialized knowledge extracted from medical corpus. Since clinical experts are not readily available in every situation and manual collection is often very costly, the existing methods fail to automatically obtain domain knowledge of diseases for accurate disease representations.

Challenge II. *How to model continuous disease progression together with the domain knowledge?* Although the clinical data are generally discrete with patients' irregular visits in EHRs (shown in Fig. 1), the majority of disease progressions and alterations in a patient's physical state occur continuously [4], [17]. Therefore, it is essential to utilize sporadic, partially observed patient visits for continuous modeling of diagnosis prediction. However, how to jointly optimize continuous diagnosis prediction based on the disease embeddings learned with domain knowledge remains unknown.

*Corresponding author

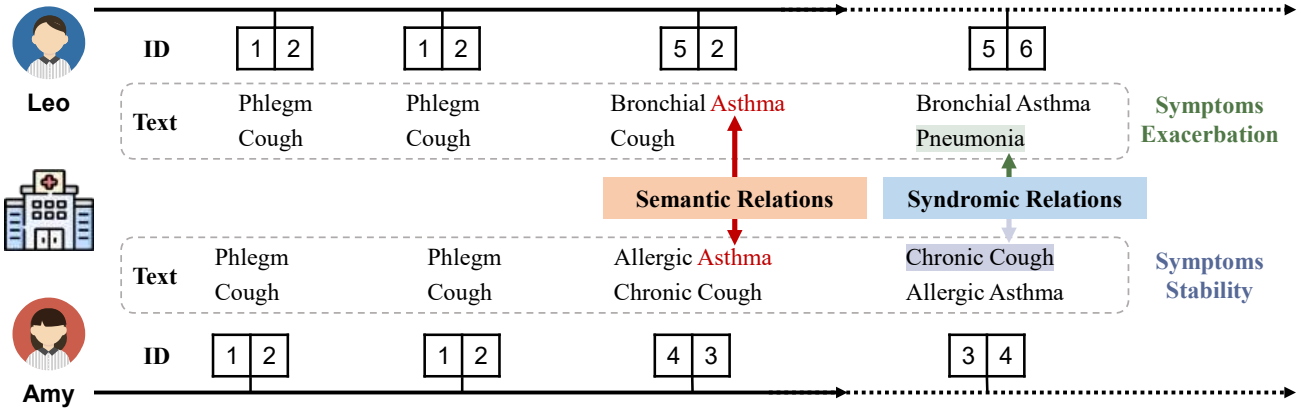


Fig. 1: An toy example depicting semantic and syndromic relations in disease progression using textual EHR data.

To address the above challenges, we propose a novel continuous diagnosis prediction model with Expert enhanced neural Ordinary Differential Equations (ExpertODE). In particular, we first propose a Mixture of Language Experts (MoLE) module that harnesses LMs from both general and clinical domains. The general domain LMs (e.g., GPT-2) offer a wide comprehension of common semantic patterns with large parameter scales, while clinical LMs (e.g., Clinical-BERT) provide specialized knowledge extracted from medical contexts with smaller parameter scales. Then, we propose a Contrastive Neural Ordinary Differential Equation (CNODE) module. We propose to leverage neural Ordinary Differential Equations (ODEs) to continuously model disease progression under patients' irregular visits, where a unified contrastive learning framework is designed to bridge continuous disease progression together with the domain knowledge of diseases.

The main contributions of this work are summarized as: (1) *Novel integration of domain experts.* We propose an effective MoLE module that integrates both clinical and general language experts to enrich disease embeddings. (2) *Effective model design.* ExpertODE is the first unified framework to jointly capture domain-based semantic relations and continuous disease progression for diagnosis prediction. (3) *Extensive experiments on real-world datasets.* Extensive experimental results against state-of-the-art approaches demonstrate the superiority of our proposed model.

II. Related Work

Language Models for General and Clinical Domains. With the massive corpora and powerful computation resources for pre-training, the general LMs can be categorized into three types: encoder-only LMs (e.g., BERT [5]), decoder-only LMs (e.g., GPT-2 [14]), and encoder-decoder LMs (e.g., T5 [15]). Although the general LMs have been widely applied to various NLP tasks, they may fail to solve specialized domain problems, especially in the clinical domain. To learn rich domain knowledge, existing studies have tried to pre-train the LMs on a large medical corpus from scratch [21]. For example, Clinical-BERT [22] is pre-trained on MIMIC-III [6] including multimodal EHRs. However, these clinical LMs may ignore the comprehension of common semantic patterns and lack generalization capabilities.

Dynamics Modeling for Diagnosis Prediction. EHR data can be observed as the sequence of patients' visits for diagnosis prediction [1]. Recently, deep learning has been widely adopted to model the dynamics in EHR data. For example, Dipole [10] applied bidirectional long-short-term memory networks and attention mechanisms to predict patient visit information. Timeline [1] utilized time-aware attention mechanisms in RNNs for health event predictions. Chet [7] designed a context-aware dynamic graph learning method to learn disease combinations and disease development schemes. However, they ignore the timestamps of the visits and fail to capture the continuous dynamics of disease progression. To better leverage irregular timestamps, recent studies [2] proposed to capture temporal dependencies via learning time embeddings. For instance, HiTANet [9] designed time interval vectors to model irregular time gaps between successive visits. Procare [18] leveraged neural ODEs to capture the continuous-time disease progression. However, they do not consider the domain knowledge of diseases.

III. Methodology

A. Method Overview

Our goal is to provide continuous diagnosis prediction based on multimodal EHRs via modeling accurate disease representations and continuous-time dynamics of the patient's health state (shown in Fig. 2). To leverage domain knowledge, we obtain the disease representation $\boldsymbol{X}_d$ for the d-th disease from $\boldsymbol{X}_d^G$ and $\boldsymbol{X}_d^M$, which are the embeddings from the general and clinical language experts, respectively. To capture the dynamics across irregular visit sequences, our process starts with initializing an ID embedding $\boldsymbol{I}_d$ for d-th disease. We then aggregate these into sequential visit embeddings $\{\boldsymbol{v}_{i,k}\}_{k=0}^{M_i-1}$, each associated with timestamps $\{t_{i,k}\}_{k=0}^{M_i-1}$, to form the comprehensive patient embedding $\boldsymbol{P}_i$ for patient p_i. M_i denotes the number of visits of patient p_i. Finally, we predict the last diagnosed diseases in the visit v_{i,M_i} at the timestamp t_{i,M_i} (i.e., diagnosis prediction $\hat{\boldsymbol{V}}_{i,M_i}$ and $\tilde{\boldsymbol{V}}_{i,M_i}$).

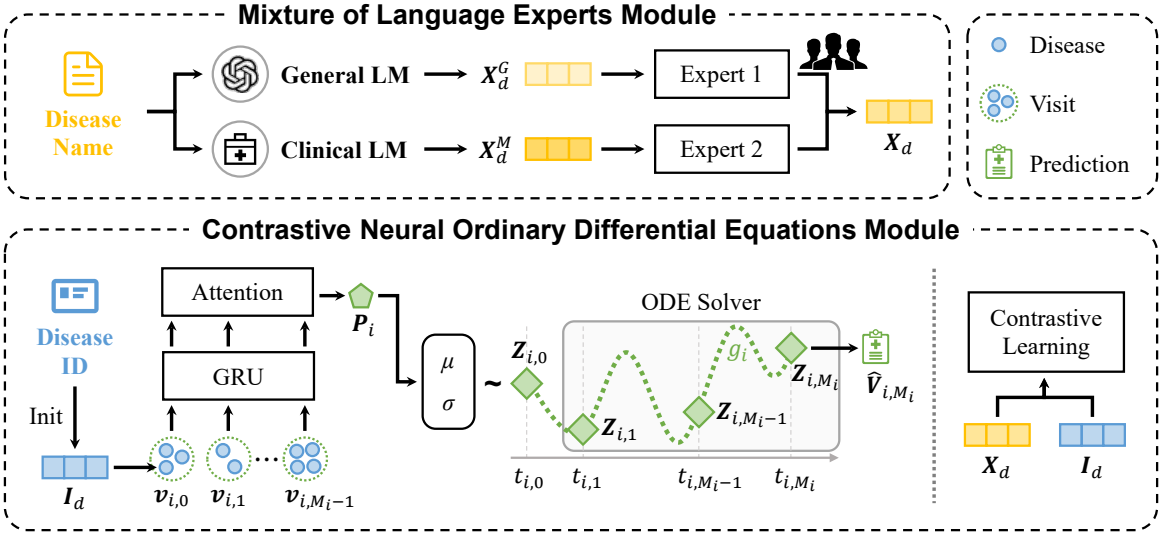


Fig. 2: The overall framework of Expert enhanced neural Ordinary Differential Equations (ExpertODE).

B. Mixture of Language Experts

As highlighted in the introduction section, accurate disease representations are crucial for effective diagnosis prediction. This task is particularly challenging due to the complex semantic relations and syndromic relations among diseases shown in Fig. 1. There is a pressing need for an automated mechanism that can seamlessly integrate rich domain knowledge from textual EHRs, such as specific terms from the semantic domain and their clinical interrelations, without the necessity for manual curation or supervision.

Inspired by the success of representation learning based on LMs, we propose to leverage language experts to incorporate external domain knowledge, where a novel Mixture of Language Experts (MoLE) module can harness distinct LMs from both general and clinical domains. In this way, the general language expert with a much larger parameter scale can capture the shared semantics provided by the term "Asthma" (shown in Fig. 1), whereas the clinical expert with a smaller parameter scale can distinguish the conditions based on disease progression patterns and associated syndromes.

Specifically, we adopt a gating mechanism to aggregate knowledge from various LMs (i.e., Clinical-BERT and GPT-2). Note that, these language models can be flexibly replaced according to the specific situation. Then, we can learn the importance of different LMs for each particular disease. The ensemble representation for a disease $\boldsymbol{X}_d$ is calculated as:

$$\boldsymbol{X}_d = \text{MoLE}(\boldsymbol{X}_d^G, \boldsymbol{X}_d^M) = \omega \cdot \boldsymbol{X}_d^G + (1-\omega) \cdot \boldsymbol{X}_d^M, \quad (1)$$

where $\boldsymbol{X}_d^G$ is the embedding from the general LM, $\boldsymbol{X}_d^M$ is the embedding from the clinical LM, and ω are the weights assigned to the respective embeddings by the MoLE module, ensuring a balanced integration.

C. Contrastive Neural Ordinary Differential Equations

When constructing disease progression from EHRs, it is needed to address the challenges posed by the discrete and irregular nature of the observed data. Traditional models that consider only discrete visit entries fail to capture the continuity inherent in disease progression. As shown in Fig. 1, without recognizing the evolving nature of diseases, we can not distinguish patient Leo's trajectory (i.e., from Bronchial Asthma to Pneumonia) from patient Amy's progression (i.e., remaining Allergic Asthma).

To capture the temporal dynamics of patient visit sequences for diagnosis prediction, we introduce a novel approach integrating a Gated Recurrent Unit (GRU) based encoder with an attention mechanism to generate patient embeddings, followed by a neural ODE model to evolve these embeddings over time.

Specifically, we first initialize an ID embedding $\boldsymbol{I}_d$ for d-th disease. Then, we obtain a visit representation via $\boldsymbol{v}_{i,k} = \sum_{d \in D_{i,k}} \boldsymbol{I}_d$, where $D_{i,k}$ is the set of diseases diagnosed in the k-th visit of patient i. Furthermore, we obtain patient embedding $\boldsymbol{P}_i$ via processing the input sequence of patient visits $\{\boldsymbol{v}_{i,k}\}_{k=0}^{M_i-1}$ with a GRU-based attention mechanism:

$$\boldsymbol{P}_i = \text{Attention}(\text{GRU}(\{\boldsymbol{v}_{i,k}\}_{k=0}^{M_i-1}). \quad (2)$$

The GRU [8] is adept at capturing temporal dependencies, while the attention mechanism [9] weighs the importance of each visit, allowing the model to focus on the most relevant information. This process yields a consolidated patient state representation $\boldsymbol{P}_i$.

Subsequently, we adopt a latent ODE model to describe a continuous process by a neural ODE in the latent space. The mathematical formulation of neural ODE is written as:

$$\frac{d\boldsymbol{Z}(t)}{dt} = f_{ode}(\boldsymbol{Z}(t), t; \theta), \quad (3)$$

where the evolved latent state $\boldsymbol{Z}_t$ at any time t is used for predicting the likelihood of future disease onset.

To obtain $\boldsymbol{Z}_t$, it requires solving an ODE initial value problem via $\boldsymbol{Z}_0$. With widely adopted reparameterization trick [17], we derive the initial state $\boldsymbol{Z}_{i,0}$ by sampling from

a Gaussian distribution with mean μ and variance σ, both of which are yielded by a neural network predicated on the patient embedding $\boldsymbol{P}_i$. In this way, we have the distribution q_ϕ approximates the posterior of $\boldsymbol{Z}_{i,0}$ given $\boldsymbol{P}_i$ as follows:

$$\boldsymbol{Z}_{i,0} \approx q_\phi(\boldsymbol{Z}_{i,0}|\boldsymbol{P}_i) = \mathcal{N}(\mu(\boldsymbol{P}_i), \sigma(\boldsymbol{P}_i)^2). \tag{4}$$

This initial state $\boldsymbol{Z}_{i,0}$ is then evolved through time using an ODE solver, reflecting the continuous progression of the patient's health state over time. The ODE function g_i models the continuous-time dynamics of the patient's health state:

$$\boldsymbol{Z}_{i,M_i} = \text{ODESolve}_\eta(g_i, \boldsymbol{Z}_{i,0}, t_0, t_1, \ldots, t_{M_i}), \tag{5}$$

where $\boldsymbol{Z}_{i,M_i}$ represents the patient's state at time t_{i,M_i} and η is a hyperparameter to control the step size of neural ODEs.

Based on the state $\boldsymbol{Z}_{i,M_i}$ that captures the evolution of the patient's condition throughout their clinical history, we can make continuous predictions for patients as follows:

$$\hat{\boldsymbol{V}}_{i,M_i} = p_\theta(\boldsymbol{Z}_{i,M_i}) = \sigma(\boldsymbol{W}_i \boldsymbol{Z}_{i,M_i} + \boldsymbol{b}_i). \tag{6}$$

The training objective is formulated to maximize the Evidence Lower Bound (ELBO) by jointly training the encoder and the generative model. The ELBO is given by:

$$\begin{aligned} \mathcal{L}_{\text{ELBO}} = &- \mathbb{E}_{q_\phi(\boldsymbol{Z}_{i,0}|\boldsymbol{P}_i)}[\log p_\theta(\boldsymbol{Z}_{i,M_i})] \\ &+ \text{KL}[q_\phi(\boldsymbol{Z}_{i,0}|\boldsymbol{P}_i)||p(\boldsymbol{Z}_{i,0})], \end{aligned} \tag{7}$$

where p_θ is the likelihood of the patient embedding given the latent state, q_ϕ is the approximated posterior distribution, and KL is the Kullback-Leibler divergence between the approximated posterior and the prior distribution of the initial latent state. This objective encourages the model to learn embeddings that are informative of the patient's future health outcomes, providing a dynamic and continuous approach to disease diagnosis.

Since the diagnosis prediction is a multi-label classification task, we use a dense layer with a softmax function to calculate the predicted probability. Specifically, the prediction $\tilde{\boldsymbol{V}}_{i,M_i}$ is based on the inferred patient status $\boldsymbol{Z}_{i,M_i}$, and the objective function for diagnosis prediction $\mathcal{L}_{pred}$ is given by:

$$\tilde{\boldsymbol{V}}_{i,M_i} = \text{softmax}(\text{MLP}(\boldsymbol{Z}_{i,M_i})), \tag{8}$$

$$\begin{aligned} \mathcal{L}_{pred} = &-\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{V}_{i,M_i} \log\left(\tilde{\boldsymbol{V}}_{i,M_i}\right) \\ &+ (1 - \boldsymbol{V}_{i,M_i}) \log\left(1 - \tilde{\boldsymbol{V}}_{i,M_i}\right), \end{aligned} \tag{9}$$

where $\boldsymbol{V}_{i,M_i}$ is the ground-truth of patient i's M_i-th diagnosis and N is the number of patients.

Furthermore, to effectively integrate the semantic view provided by the LMs and the temporal view from the neural ODEs, we employ contrastive learning to align two modules:

$$\mathcal{L}_{con} = \sum_d -\log \frac{\exp(\cos(\boldsymbol{I}_d, \boldsymbol{X}_d))}{\sum_{d'} \exp(\cos(\boldsymbol{I}_d, \boldsymbol{X}_{d'}))}, \tag{10}$$

where $\boldsymbol{I}_d$ is the disease embedding learned by CNODE, $\boldsymbol{X}_d$ is the disease embedding learned by MoLE, and cos denotes the cosine similarity. This contrastive loss function encourages the alignment of embeddings from both the CNODE and MoLE pathways, thereby synergizing the semantic and temporal insights for a comprehensive disease representation.

The final objective function of the proposed ExpertODE is given as follows:

$$\min_{\boldsymbol{I}_d, \boldsymbol{X}_d} \mathcal{L}_{ELBO} + \mathcal{L}_{pred} + \lambda \mathcal{L}_{con}, \tag{11}$$

where λ is a hyperparameter to control the contrastive loss.

TABLE I: Statistics of the datasets used in our experiments.

Dataset	MIMIC-III	eICU
# of patients	2,371	23,828
# of visits	7,279	59,908
Avg. visits per patient	3.07	2.51
# of unique ICD-9 codes	4,880	2,591
Avg. # of diagnosis codes per visit	13.39	4.22
Max # of diagnosis codes per visit	39.0	95.0

IV. Experiments

In this section, we evaluate our proposed ExpertODE framework focusing on four research questions: **RQ1**: How does ExpertODE perform in comparison to state-of-the-art diagnosis prediction methods? **RQ2**: What are the effects of different model components? **RQ3**: How do the hyperparameters affect the prediction performance and how to choose optimal values? **RQ4**: How does ExpertODE improve the modeling of diseases with general and clinical LMs?

A. Experimental Settings

Datasets and Evaluation Protocols. We use two real-world EHR datasets to verify the effectiveness of compared methods, i.e., MIMIC-III [6] and eICU [12]. Both datasets are fully anonymized and carefully sanitized before our access. We chose patients who made at least two visits for both datasets. The statistics are summarized in Table I. ICD-9 is the official disease code [7], [17]. For evaluation metrics, we use Recall@k and NDCG@k that are consistent with [7], [8].

Methods for Comparison. We adopt 11 representative state-of-the-art methods as baselines for the performance comparison with our proposed ExpertODE: (1) interaction modeling methods: **GRAM** [3], **KAME** [11], **MHM** [13], **TAdaNet** [16], and **CGL** [8]; (2) dynamic modeling methods: **RETAIN** [2], **Dipole** [10], **Timeline** [1], **HiTANet** [9], **Chet** [7], and **ProCare** [18].

Implementation Details. We split the dataset randomly according to patients into training/validation/test sets (i.e., 1660/237/474 on MIMIC-III and 16679/2383/4766 on eICU), which is consistent with [7], [8]. We optimize the compared baselines with standard Adam and tune all hyperparameters on training sets through grid search. In particular, η in $\{0.01, 0.05, 0.10, 0.50\}$ and λ in $\{0.5, 1.0, 1.5\}$. We set the embedding dimension D as 128 and the batch size as 128 for all compared methods on MIMIC-III and eICU. We carefully tune the hyperparameters of baselines as suggested in the original papers to achieve their best performance.

TABLE II: Experimental results on two benchmark EHR datasets with Recall and NDCG. The best performances are highlighted in **boldface** and the second runners are underlined.

Method	Recall@5	NDCG@5	Recall@10	NDCG@10	Recall@5	NDCG@5	Recall@10	NDCG@10
	MIMIC-III				eICU			
RETAIN	0.1510	0.4188	0.2134	0.3537	0.3213	0.3428	0.3901	0.3605
Dipole	0.1442	0.3999	0.2038	0.3378	0.3071	0.3274	0.3727	0.3452
GRAM	0.1429	0.4059	0.2112	0.3510	0.3049	0.3318	0.3862	0.3576
Timeline	0.1487	0.4123	0.2100	0.3482	0.3175	0.3376	0.3840	0.3548
KAME	0.1353	0.3992	0.2055	0.3070	0.2887	0.3268	0.3759	0.3126
MHM	0.1383	0.4080	0.2128	0.3481	0.2954	0.3340	0.3893	0.3547
TAdaNet	0.1433	0.4114	0.2172	0.3568	0.3056	0.3371	0.3972	0.3642
HiTANet	0.1502	0.4166	0.2122	0.3518	0.3204	0.3413	0.3881	0.3584
CGL	0.1538	0.4265	0.2173	0.3602	0.3379	0.3624	0.4298	0.3872
Chet	0.1636	0.4403	0.2312	0.3719	0.3495	0.3604	0.4228	0.3790
ProCare	0.1885	0.5004	0.2687	0.4271	0.4121	0.4097	0.5004	0.4352
ExpertODE $_{M\text{-}}$	0.1908	0.5101	0.2704	0.4310	0.4157	0.4291	0.5047	0.4502
ExpertODE $_{C\text{-}}$	0.1888	0.5112	0.2643	0.4269	0.4038	0.4095	0.4981	0.4378
ExpertODE	**0.1963**	**0.5243**	**0.2759**	**0.4439**	**0.4245**	**0.4341**	**0.5105**	**0.4548**

TABLE III: Hyperparameter Studies on MIMIC-III and eICU.

Param.	Recall@5	NDCG@5	Recall@5	NDCG@5
	MIMIC-III		eICU	
η = 0.01	0.1921	0.5146	0.4213	0.4301
η = 0.05	**0.1963**	**0.5243**	0.4228	0.4319
η = 0.10	0.1949	0.5130	**0.4245**	**0.4341**
η = 0.50	0.1904	0.5096	0.4136	0.4266
λ = 0.5	0.1896	0.5170	0.4160	0.4148
λ = 1.0	**0.1963**	**0.5243**	**0.4245**	**0.4341**
λ = 1.5	0.1947	0.5141	0.4224	0.4296

TABLE IV: Different MoLE Weights ω of exampled diseases learned by our proposed ExpertODE on MIMIC-III.

ICD-9	Disease name	ω
252.02	Hyperparathyroidism, unspecified	0.51
252.00	Secondary Hyperparathyroidism	0.55
272.00	Pure Hypercholesterolemia	0.30
401.00	Malignant Essential Hypertension	0.37
790.22	Impaired Glucose Tolerance Test	0.41

B. Overall Performance Comparison (RQ1)

We compare the continuous diagnosis prediction results of the proposed ExpertODE framework to those of the baseline models. Table II shows the Recall@k and NDCG@k on MIMIC-III and eICU datasets with k={5, 10}. We have the following observations.

ExpertODE consistently outperforms all baselines across all metric on both datasets. This answers RQ1, showing that our proposed ExpertODE that captures domain-aware semantic relations and continuous disease progression is capable of continuous diagnosis prediction. Compared with the second-best performance, the performance gains of ExpertODE ranges from 2.02% with Recall@10 on eICU to 5.96% achieved with NDCG@5 on eICU.

Specifically, ExpertODE outperforms CGL and Chet, which predict health events through transition functions on disease graphs. It demonstrates the effectiveness of modeling the continuously evolving nature of disease progression. Although ProCare can capture disease severity, interaction, and continuous progression, it does not integrate the rich, semantic understanding that ExpertODE's MoLE module provides through its language experts. It further validates the significance of leveraging abundant domain knowledge from EHRs.

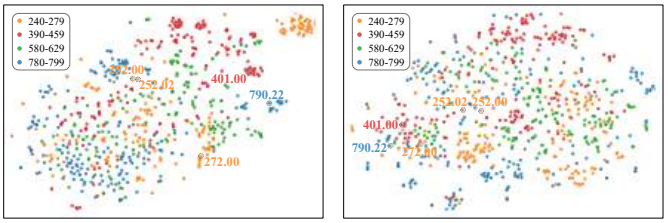
C. Ablation Studies (RQ2)

To better understand our proposed techniques, we conduct ablation studies as follows: ExpertODE $_{M\text{-}}$ removes the Mixture of Language Experts (MoLE) module from ExpertODE, and ExpertODE $_{C\text{-}}$ removes the Contrastive Neural Ordinary Differential Equation (CNODE) module from ExpertODE.

As shown in Table II, compared with ExpertODE $_{M\text{-}}$, ExpertODE leads to performance gains ranging from 1.02% (achieved in NDCG@10 on eICU) to 2.99% (achieved in NDCG@10 on MIMIC-III), where ExpertODE $_{M\text{-}}$ fails to fully utilize the domain knowledge for accurate disease embeddings. Furthermore, the performance gains of ExpertODE over ExpertODE $_{C\text{-}}$ ranges from 2.49% (with Recall@10 on eICU) to 6.01% (with NDCG@5 on eICU), where ExpertODE can capture latent continuous-time dynamics in patient data. The results also affirm the effectiveness of our CNODE module for handling patients' irregular visits.

D. Effect of Hyperparameters (RQ3)

As shown in Table III, η controls the step size for different sampling frequencies of the neural ODEs. The optimal η value on MIMIC-III is about 0.05, and the optimal η value on eICU is about 0.1. Since the average number of visits is larger on MIMIC-III, it's reasonable to use a smaller sampling frequency for modeling the health status trajectory of patients. Moreover, λ controls the weight of contrastive loss. ExpertODE achieves the best performance with λ = 1.0. Too small λ will cause inadequate modeling disease embeddings and fail to fully leverage domain knowledge extracted from general and clinical LMs, while too large λ will likely ignore the continuous-time dynamics and cause a decrease in the performance of disease progression.



(a) GPT-2 (b) Clincal-BERT

Fig. 3: Visualizations of disease embeddings learned by different language experts. Best viewed in color.

E. Case Studies (RQ4)

To provide more insights into the advantages of ExpertODE in modeling disease embeddings, we provide five exampled diseases on MIMIC-III. The detailed information of diseases learned by ExpertODE is presented in Table IV and the corresponding visualizations of embeddings are shown in Fig. 3. As shown in Fig. 3(a), for Hyperparathyroidism, unspecified (252.02) and Secondary Hyperparathyroidism (252.00), the general language expert (e.g., GPT-2) can capture the shared semantics provided by the term "Hyperparathyroidism" and accurately model the similarity between them (e.g., they both belong to the ICD-9 of 240-279). Therefore, the MoLE weights ω of them are higher than 0.5, which means the importance of general language experts more than clinical ones. Furthermore, as shown in Fig. 3(b), although Pure Hypercholesterolemia (272.00), Malignant Essential Hypertension (401.00), and Impaired Glucose Tolerance Test (790.22) do not have the shared semantics, the clinical language expert (e.g., Clinical-BERT) can capture the medical correlation. Possessing rich domain knowledge, it identifies these conditions as syndromic diseases. Consequently, this understanding brings their representations closer in the embedding space and corresponds to low MoLE weights ω.

V. Conclusion

In this paper, we propose to make diagnosis predictions based on the patients' irregular visits and domain knowledge of diseases. Specifically, we propose a novel expert enhanced continuous model (ExpertODE) with two pivotal techniques, which capture complex dependencies between continuous diagnosis prediction optimization and domain-based diagnostic textual representation. Extensive quantitative experiments demonstrate the clear advantages of our ExpertODE over the state-of-the-art baselines towards the precise diagnosis, which is further consolidated with our real case study results.

VI. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants (No.6230071268), the Natural Science Foundation of Zhejiang Province (LQ23F020007), Zhejiang Provincial Department of Agriculture and Rural Affairs Project (2024SNJF044), and Zhejiang Gongshang University "Digital+" Disciplinary Construction Management Project (SZJ2022B001). Carl Yang was not supported by any fund from China.

References

[1] T. Bai, S. Zhang, B. L. Egleston, and S. Vucetic. Interpretable representation learning for healthcare via capturing disease progression through time. In *SIGKDD*, pages 43–51, 2018.

[2] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NeurIPS*, 2016.

[3] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun. Gram: graph-based attention model for healthcare representation learning. In *SIGKDD*, pages 787–795, 2017.

[4] T. Dang, J. Han, T. Xia, E. Bondareva, C. Siegele-Brown, J. Chauhan, A. Grammenos, D. Spathis, P. Cicuta, and C. Mascolo. Conditional neural ode processes for individual disease progression forecasting: a case study on covid-19. In *SIGKDD*, pages 3914–3925, 2023.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 2016.

[7] C. Lu, T. Han, and Y. Ning. Context-aware health event prediction via transition functions on dynamic disease graphs. In *AAAI*, volume 36, pages 4567–4574, 2022.

[8] C. Lu, C. K. Reddy, P. Chakraborty, S. Kleinberg, and Y. Ning. Collaborative graph learning with auxiliary text for temporal event prediction in healthcare. In *IJCAI*, 2021.

[9] J. Luo, M. Ye, C. Xiao, and F. Ma. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *SIGKDD*, pages 647–656, 2020.

[10] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *SIGKDD*, pages 1903–1911, 2017.

[11] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *CIKM*, 2018.

[12] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 2018.

[13] Z. Qiao, Z. Zhang, X. Wu, S. Ge, and W. Fan. Mhm: Multi-modal clinical data based hierarchical multi-label diagnosis prediction. In *SIGIR*, pages 1841–1844, 2020.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[15] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[16] Q. Suo, J. Chou, W. Zhong, and A. Zhang. Tadanet: Task-adaptive network for graph-enriched meta-learning. In *SIGKDD*, pages 1789–1799, 2020.

[17] Y. Tan, C. J. Yang, X. Wei, C. Chen, W. Liu, L. Li, J. Zhou, and X. Zheng. Metacare++: Meta-learning with hierarchical subtyping for cold-start diagnosis prediction in healthcare data. In *SIGIR*, pages 449–459, 2022.

[18] Y. Tan, Z. Zhou, L. Yu, W. Liu, C. Chen, G. Ma, X. Hu, V. S. Hertzberg, and C. Yang. Enhancing personalized healthcare via capturing disease severity, interaction, and progression. *ICDM*, 2023.

[19] Z. Wang, R. Wen, X. Chen, S. Cao, S. Huang, B. Qian, and Y. Zheng. Online disease diagnosis with inductive heterogeneous graph convolutional networks. In *WWW*, pages 3349–3358, 2021.

[20] R. Xu, M. K. Ali, J. C. Ho, and C. Yang. Hypergraph transformers for ehr-based clinical predictions. *AMIA*, 2023:582, 2023.

[21] Y. Xu, K. Yang, C. Zhang, P. Zou, Z. Wang, H. Ding, J. Zhao, Y. Wang, and B. Xie. Vecocare: Visit sequences-clinical notes joint learning for diagnosis prediction in healthcare data. In *IJCAI*, pages 4921–4929, 2023.

[22] B. Yan and M. Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *AAAI*, volume 36, pages 2982–2990, 2022.