2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 979-8-3503-2447-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/EMBC40787.2023.10340765

Clinical Risk Prediction Models with Meta-Learning Prototypes of Patient Heterogeneity

Lida Zhang, Rohan Khera, and Bobak J. Mortazavi, *Senior Member, IEEE*

***Abstract*— Hospitalized patients sometimes have complex health conditions, such as multiple diseases, underlying diseases, and complications. The heterogeneous patient conditions may have various representations. A generalized model ignores the differences among heterogeneous patients, and personalized models, even with transfer learning, are still limited to the small amount of training data and the repeated training process. Meta-learning provides a solution for training similar patients based on few-shot learning; however, cannot address common cross-domain patients. Inspired by prototypical networks [1], we proposed a meta-prototype for Electronic Health Records (EHR), a meta-learning-based model with flexible prototypes representing the heterogeneity in patients. We apply this technique to cardiovascular diseases in MIMIC-III and compare it against a set of benchmark models, and demonstrate its ability to address heterogeneous patient health conditions and improve the model performances from 1.2% to 11.9% on different metrics and prediction tasks.**

***Clinical relevance*— Developing an adaptive EHR risk prediction model for outcomes-driven phenotyping of heterogeneous patient health conditions.**

I. INTRODUCTION

Machine learning has increasingly focused on implementing clinical risk prediction models utilizing data from electronic health records (EHRs) to provide clinical predictions. These predictions on individual patients, in turn, support decision-making for doctors [2], [3]. However, given the large and sparse nature of EHR data [4], most models use homogenized input spaces, and static windows of observation [4], [5], despite complex health conditions requiring data representing a variety of co-morbidities, data sources (vitals and laboratory examinations), and admission lengths. Clinical subgroup assignment, for example, may better identify which patients benefit the most from treatments [6], [7]. Therefore, the variety of data representations leads to the need for general models that handle data across these complex medical scenarios.

The complexity of the health conditions of hospitalized patients has led to the development of personalized models [8], [9], and Oikonomou et al. proposed a phenomapping strategy that leverages information from all trial participants to phenotype individuals [10]. Personalized models are limited in available training data, and even with the assistance of transfer learning, it is still not optimal to train multiple models for each patient. Therefore, meta-learning [11] has been applied to EHR-based risk prediction models with limited training data to create fewer general models that apply across the varied personal settings[12], [13], but these methods pre-define each patient into one certain domain, and ignores patients' known or unknown health conditions that may result in potential cross-domain patients. Snell et al. proposed prototypical networks with a linear reinterpretation model [1] and Boniolo et al. built prediction models through patient similarity to address this limitation of meta-learning [14]; however, they do not have representative prototypes and flexible alignments for the heterogeneous patients' health conditions. Inspired by these works, we introduce meta-prototype networks to develop risk prediction models by leveraging patient heterogeneity through trainable prototypes, representations of the heterogeneous patient conditions, rather than selecting against it.

In this paper, we propose meta-prototype, a meta-learning-based adaptive EHR risk prediction model leveraging heterogeneous patient health conditions. Meta-learning is applied to train models for similar patients, and a trainable prototype network is introduced to represent flexible phenotype alignments for patients. We apply our model to a variety of patients in the MIMIC-III intensive care unit (ICU) dataset with diagnosed cardiovascular conditions, treating each cardiovascular disease as a prototype, and obtain improvements of 1.2% and 2.4% in the area under the receiver operating characteristics (AUROC), 11.9% and 3.7% in the area under Precision-Recall (AUPRC) for decompensation and in-hospital mortality respectively, and 2.2% in Cohen's Kappa score and 6.7% in the mean absolute difference (MAD) for the task of length-of-stay.

II. RELATED WORK

Machine learning has been widely applied in building risk prediction models from EHR data. Cheng et al. proposed an EHR risk prediction model by extracting meaningful features [15], and Harutyunyan et al. built a multi-task learning model for clinical prediction with time-series EHR variables [5]. These are models that work generally on patients across the EHR with homogenized input lengths and variables. Suo et al. focused on learning similarities between patients and built a personalized disease prediction model [8], and Liu et al. applied transfer learning to address the diminishing data problem in personalized models [9]. However, these personalized approaches are still limited to patients who do not have a great number of data, for example, early admitted patients. Zhang et al. proposed DynEHR, a few-shot learning approach to address the various lengths of EHR data [13], and Zhang et al. applied meta-learning to CNNs and

L. Zhang and B. Mortazavi are with the Department of Computer Science & Engineering, Texas A&M University, College Station, TX, 77843, USA (corresponding author, Lida Zhang), emails: {lidazhang, bobakm}@tamu.edu, R. Khera is with the Yale School of Medicine, New Haven, CT, 06510, USA, email{rohan.khera}@yale.edu

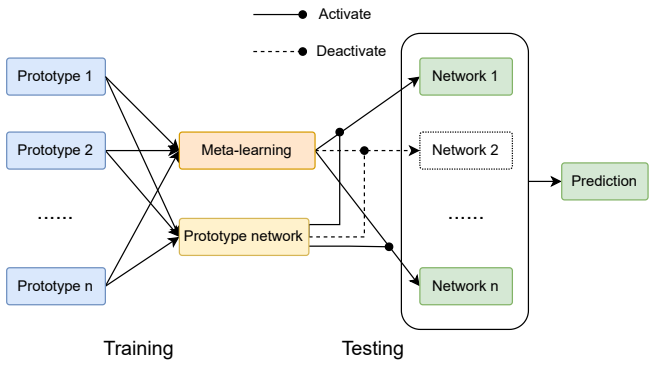

Fig. 1. Meta-prototype framework. The prototypes are trained through meta-learning, and a prototype network is trained for prototype alignment. During testing, the prototype network decides which prototype-specific network is activated for the final prediction.

Long-Short Term Memory (LSTM) for low-resource EHR models [12]. However, patients' health conditions are usually complex. Many patients are admitted to hospitals with multiple diseases or complications, and these meta-learning-based methods do not leverage this phenotypic information that come with observationally-selected variables based upon clinical courses of treatment.

III. METHODS

In this section, we introduce our work in two parts: training meta-prototype and generating risk prediction with our meta-prototype. Figure 1 illustrates a framework of our model.

A. Meta-prototype training

Given a model $\mathcal{F}$ with a feature extractor $\mathcal{F}_\theta$ and a predictor $\mathcal{F}_\eta$, θ and η are used to indicate their parameters respectively. In a time-series setting, we apply an LSTM for the feature extractor and fully-connected layers for the predictor. For a data point x and its label y, the learning cost of our model is represented as:

$$\mathcal{L}_{\theta,\eta} = \mathcal{L}(\mathcal{F}_{\theta;\eta}(x), y). \tag{1}$$

Let D be a set of prototypes. In each episode of training the meta-prototype, a subset D' of prototypes is randomly sampled ($D' \subseteq D$). For each prototype i in D', a model $\mathcal{F}_{\theta_i;\eta_i}$ is first initialized from the meta-learner $\mathcal{F}_{\theta;\eta}$, and the cost $\mathcal{L}_{\theta_i;\eta_i}$ for this prototype can be calculated according to Equation 1 on a randomly sampled support set. The model for prototype i can then be adapted to $\mathcal{F}_{\bar{\theta}_i,\bar{\eta}_i}$ from $\mathcal{L}_{\theta_i;\eta_i}$ with a few steps:

$$\bar{\theta}_i = \theta_i - \tau\nabla_{\theta_i}\mathcal{L}_{\theta_i;\eta_i}, \qquad \bar{\eta}_i = \eta_i - \tau\nabla_{\eta_i}\mathcal{L}_{\theta_i;\eta_i},$$

where τ is a learning rate.

After the adapted model is obtained, a query set from prototype i is then sampled and applied on $\mathcal{F}_{\bar{\theta}_i;\bar{\eta}_i}$ to calculate a cost $\mathcal{L}_{\bar{\theta}_i;\bar{\eta}_i}$ from Equation 1.

With the meta-learning-based training approach for the prediction models of multiple prototypes, it is still not clear what these prototypes are. Instead of using the mean of the embedded examples [1] or a certain example [16], we introduce a linear prototype network $\mathcal{F}_\phi$, a fully-connected network without a bias, as the trainable prototypes, and each column ϕ_j can represent a prototype. In each training episode, a set of data points x and their prototype label c are sampled. The representation of x is calculated from the extractor $\mathcal{F}_\theta$ (without any adaptation, in order to have a fair comparison among different prototypes), and the prototype network $\mathcal{F}_\phi$ is used to align the data to certain prototypes. The prototype network can be trained from

$$\hat{\mathcal{L}}_{\theta,\phi} = \mathcal{H}(\mathcal{F}_{\theta;\phi}(x), c),$$

where $\mathcal{H}$ denotes a cross-entropy loss function and c is a ten-class cardiovascular disease phenotype for each patient.

After collecting the prototype classification cost $\hat{\mathcal{L}}_{\theta;\phi}$ and the query set cost $\mathcal{L}_{\bar{\theta}_i;\bar{\eta}_i}$ from all the sampled prototypes D', the meta-learner $\mathcal{F}_\theta$, $\mathcal{F}_\eta$, and prototype network $\mathcal{F}_\phi$ can be optimized as:

$$\theta = \theta - \mu(\sum_i^{D'} \nabla_\theta \mathcal{L}_{\bar{\theta}_i;\bar{\eta}_i} + \nabla_\theta \hat{\mathcal{L}}_{\theta;\phi}),$$

$$\eta = \eta - \mu \sum_i^{D'} \nabla_\eta \mathcal{L}_{\bar{\theta}_i;\bar{\eta}_i}, \qquad \phi = \phi - \mu\nabla_\phi \hat{\mathcal{L}}_{\theta;\phi},$$

where μ is another learning rate.

B. Risk prediction with meta-prototype

Before making predictions, the prototype-specific network $\mathcal{F}_{\bar{\theta}_i;\bar{\eta}_i}$ for each prototype is first adapted from the trained meta-learner with their corresponding support set. Given a data point x, the prototype alignment β is calculated from $\mathcal{F}_{\theta;\phi}$, and then calculate a mask α_i for each prototypes i ($i \in D$) using Top-k [17], [18]:

$$\beta = \mathcal{F}_{\theta;\phi}(x), \quad \alpha_i = \begin{cases} 1 & \text{if } \beta_i \text{ in top } k \text{ value of all } \beta \\ 0 & \text{otherwise.} \end{cases}$$

A final prediction can be generated from the prototype masks

$$p(x) = \sum_i^D \alpha_i \cdot \mathcal{F}_{\bar{\theta}_i;\bar{\eta}_i}(x)$$

IV. EXPERIMENTS

A. Dataset and data preprocessing

Medical Information Mart for Intensive Care (MIMIC-III) is a publicly available EHR dataset [19] which collects 53,423 adult patients admitted to Beth Israel Deaconess Medical Center intensive care units (ICUs) between 2001 and 2012. We apply our proposed method meta-prototype on MIMIC-III, focusing on cardiovascular diseases. From the MIMIC-III ICD-9 diagnosis table and its HCUP CCS category [5], ten cardiovascular diseases (or conditions common to cardiovascular-related complications) are retained, as shown in Table I. We treat each disease here as a prototype when building our meta-prototype, and a patient may be aligned to one or multiple prototypes.

TABLE I
CARDIOVASCULAR CONDITION CATEGORIES

Index	Category
0	Acute and unspecified renal failure
1	Acute cerebrovascular disease
2	Acute myocardial infarction
3	Cardiac dysrhythmias
4	Chronic kidney disease
5	Congestive heart failure; nonhypertensive
6	Coronary atherosclerosis and related
7	Essential hypertension
8	Hypertension with complications
9	Shock

There are 17 charted observations and laboratory measurements selected formatting 76 features (one-hot encoding for categorical measures and numeric values for continuous measurements) [5] as the input of our model. The irregular data is split into a series of one-hour time windows without overlapping. The average values are calculated if there is more than one data point in a window, and missing data is imputed with the most recent values. In order to apply mini-batch optimization in training, zeros are padded at the end of shorter sequences.

B. Prediction tasks and evaluation

We test our model on three prediction tasks based on MIMIC-III: decompensation (rapid deterioration of patient conditions), the length of stay in the intensive care unit (ICU), and in-hospital mortality. Decompensation and in-hospital mortality are binary classification tasks. Decompensation has 13.5% of positive examples, and in-hospital mortality has 2.1%. Therefore, in addition to the evaluation metric of AUROC, we also introduce AUPRC to evaluate these two imbalanced classification tasks. The length-of-stay is framed as a multi-class classification problem [5]. Cohen's Kappa score and MAD are used to evaluate this task.

C. Model implementation and baseline models

In the experiments, we set the hidden size of the LSTM-based feature extractor $\mathcal{F}_\theta$ to be 128, and apply a one-layer fully-connected network for the predictor $\mathcal{F}_\eta$. As we discussed in the previous sections, a fully-connected network without bias is used as the prototype network $\mathcal{F}_\phi$. The dataset is split into a 70% training, a 15% validation set, and a 15% test set, with 10 repeated experiments. In each training episode, we randomly sample five prototypes and train each prototype-specific model $\mathcal{F}_{\theta_i;\eta_i}$ with five steps, and the model adaptation when making prediction has five steps as well. The prototype-specific model training has a learning rate τ of 0.005, and the training of the meta-learner has a learning rate μ of 0.0005. For the Top-k mechanism, we run hyperparameter tuning experiments and set k to be four. This study is implemented in Python 3.6, PyTorch 1.3.1, NumPy 1.18, scikit-learn 0.21 on the server of 2 Xeon 2.2GHz CPUs, 8 GTX 1080ti GPUs, and 528 GB RAM.

To understand the performance of meta-prototype, we compare our model with five baseline models: a logistic regression model with grid search for penalty and regularization strength, an attention-based transformer model [20], an LSTM model, a phased LSTM (p-LSTM) for time-series irregularity [21], and a meta-learning model [11], [13] with fixed prototypes obtained directly from cardiovascular diseases phenotype labels (MAML). The transformer model has query and value sizes of eight, two heads, two blocks, and attention size 12. The LSTM and p-LSTM models both have hidden size 128, and the MAML model is built based on the same structure of LSTM. The learning rates for deep neural network models are 0.0005.

D. Experimental results

Table II shows the results of our experiments. For MAML and our meta-prototype, we calculate the average performance from all the prototypes (diseases) and their standard deviations. From the table, our meta-prototype has great improvements on all three tasks over all baseline models. For the binary classification tasks decompensation and in-hospital mortality, our model has higher values for both AUROC and AUPRC, especially AUPRC. The significant improvement on AUPRC shows the ability of our model to address the imbalanced datasets and implies a higher sensitivity of our model in predicting at-risk patients and a potential for better performance in saving patients' lives. For length-of-stay, the higher value of the Cohen's Kappa score of our model indicates higher inter-annotator agreements between our predictions and the ground truth, and the lower MAD value additionally reinforces the lower errors of predicting the remaining length of stay in ICUs. When comparing the meta-learning-based models MAML and our meta-prototype with their base model LSTM, we can observe that MAML is sometimes even worse than the LSTM (on decompensation), showing the limitation of vanilla meta-learning in addressing the cross-domain situation, and further indicating the flexibility of meta-prototype in prototype alignment in complex situations.

Figure 2 is a heatmap of the Top-k masking in the task of in-hospital mortality. Y-axis is the ten cardiovascular prototypes, and x-axis is the predicted masking from the prototype network and Top-4 mechanism. We observe that the prototype network can predict various prototypes.

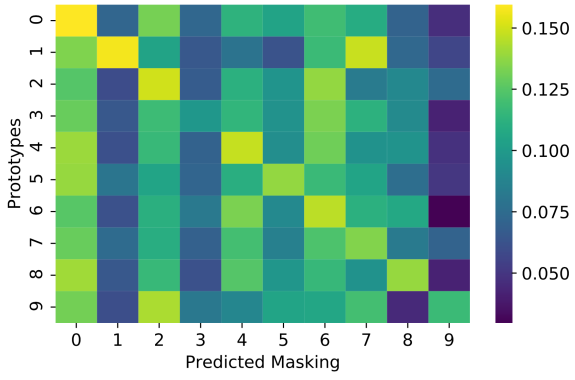

Fig. 2. A heatmap for in-hospital mortality Top-k masking. The indexes of prototypes align with Table I.

TABLE II
AVERAGE PERFORMANCE (AND STANDARD DEVIATIONS) ON MIMIC-III

Task	Decompensation		Length-of-stay		In-hospital Mortality	
Evaluation	AUROC	AUPRC	Kappa	MAD	AUROC	AUPRC
LogisticRegression	0.816 (0.016)	0.231 (0.026)	0.346 (0.008)	163.8 (10.9)	0.795 (0.011)	0.492 (0.019)
Transformer	0.837 (0.012)	0.241 (0.019)	0.371 (0.019)	160.0 (6.9)	0.829 (0.012)	0.497 (0.013)
LSTM	0.848 (0.009)	0.278 (0.012)	0.405 (0.013)	156.2 (6.4)	0.835 (0.011)	0.500 (0.010)
P-LSTM	0.836 (0.007)	0.207 (0.014)	0.382 (0.008)	152.4 (7.8)	0.834 (0.006)	0.504 (0.009)
MAML	0.837 (0.007)	0.269 (0.011)	0.404 (0.005)	152.7 (4.9)	0.836 (0.04)	0.535 (0.007)
Meta-prototype	**0.858 (0.008)**	**0.311 (0.009)**	**0.413 (0.006)**	**141.9 (5.5)**	**0.856 (0.005)**	**0.555 (0.008)**

V. LIMITATIONS AND FUTURE WORK

In this study, we evaluate our proposed model within cardiovascular diseases, and we plan to expand the experiments to other diseases, or a cross-domain setting among different types of diseases (e.g., cardiovascular and diabetes). In addition, the current prototype network is limited to a pre-defined number of prototypes and therefore needs to be re-trained if a new condition is included. In the future, we also look forward to modifying the prototype network to be flexible to growing prototypes.

VI. CONCLUSION

Patients in the hospital often have complex health conditions, such as multiple diseases, complications, or underlying diseases. A generalized model cannot represent the variation among different diseases, and personalized models are limited to the amount of training data and tedious training process. In this paper, we propose meta-prototype networks, applying meta-learning to similar patients, and then introduce a trainable prototype network to represent the prototypes. We test our meta-prototype on cardiovascular diseases in MIMIC-III, and outperform on all three prediction tasks, especially in predicting risky patients.

REFERENCES

[1] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[2] L. L. Raket, J. Jaskolowski, B. J. Kinon, J. C. Brasen, L. Jönsson, A. Wehnert, and P. Fusar-Poli, "Dynamic electronic health record detection (detect) of individuals at risk of a first episode of psychosis: a case-control development and validation study," *The Lancet Digital Health*, vol. 2, no. 5, pp. e229–e239, 2020.

[3] C. G. Walsh, K. B. Johnson, M. Ripperger, S. Sperry, J. Harris, N. Clark, E. Fielstein, L. Novak, K. Robinson, and W. W. Stead, "Prospective validation of an electronic health record–based, real-time suicide risk model," *JAMA network open*, vol. 4, no. 3, pp. e211 428–e211 428, 2021.

[4] S. N. Shukla and B. M. Marlin, "Multi-time attention networks for irregularly sampled time series," *arXiv preprint arXiv:2101.10318*, 2021.

[5] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, p. 96, 2019.

[6] D. K. McGuire, W. J. Shih, F. Cosentino, B. Charbonnel, D. Z. Cherney, S. Dagogo-Jack, R. Pratley, M. Greenberg, S. Wang, S. Huyck *et al.*, "Association of sglt2 inhibitors with cardiovascular and kidney outcomes in patients with type 2 diabetes: a meta-analysis," *JAMA cardiology*, vol. 6, no. 2, pp. 148–158, 2021.

[7] E. K. Oikonomou, M. A. Suchard, D. K. McGuire, and R. Khera, "Phenomapping-derived tool to individualize the effect of canagliflozin on cardiovascular risk in type 2 diabetes," *Diabetes care*, vol. 45, no. 4, pp. 965–974, 2022.

[8] Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, A. Zhang, and J. Gao, "Personalized disease prediction using a cnn-based similarity learning method," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017, pp. 811–816.

[9] K. Liu, X. Zhang, W. Chen, S. Alan, J. A. Kellum, M. E. Matheny, S. Q. Simpson, Y. Hu, and M. Liu, "Development and validation of a personalized model with transfer learning for acute kidney injury risk estimation using electronic health records," *JAMA Network Open*, vol. 5, no. 7, pp. e2 219 776–e2 219 776, 2022.

[10] E. K. Oikonomou, E. S. Spatz, M. A. Suchard, and R. Khera, "Individualising intensive systolic blood pressure reduction in hypertension using computational trial phenomaps and machine learning: a post-hoc analysis of randomised clinical trials," *The Lancet Digital Health*, vol. 4, no. 11, pp. e796–e805, 2022.

[11] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[12] X. S. Zhang, F. Tang, H. H. Dodge, J. Zhou, and F. Wang, "Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2487–2495.

[13] L. Zhang, X. Chen, T. Chen, Z. Wang, and B. J. Mortazavi, "Dynehr: Dynamic adaptation of models with data heterogeneity in electronic health records," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2021, pp. 1–4.

[14] F. Boniolo, G. Boniolo, and G. Valente, "Prediction via similarity: Biomedical big data and the case of cancer models," *Philosophy & Technology*, vol. 36, no. 1, p. 8, 2023.

[15] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk prediction with electronic health records: A deep learning approach," in *Proceedings of the 2016 SIAM international conference on data mining*. SIAM, 2016, pp. 432–440.

[16] J. Crabbé, Z. Qian, F. Imrie, and M. van der Schaar, "Explaining latent representations with a corpus of examples," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 154–12 166, 2021.

[17] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[18] Z. Huo, L. Zhang, R. Khera, S. Huang, X. Qian, Z. Wang, and B. J. Mortazavi, "Sparse gated mixture-of-experts to separate and interpret patient heterogeneity in ehr data," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2021, pp. 1–4.

[19] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] L. Phased, "Accelerating recurrent network training for long or event-based sequences," 2016.