Deep Stable Representation Learning on Electronic Health Records

Yingtao Luo†, Zhaocheng Liu‡, Qiang Liu§,*
†Carnegie Mellon University, Pittsburgh, USA
yingtaoluo@cmu.edu
‡Kuaishou Technology, Beijing, China
lio.h.zen@gmail.com
§Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
qiang.liu@nlpr.ia.ac.cn

Abstract—Deep learning models have achieved promising disease prediction performance of the Electronic Health Records (EHR) of patients. However, most models developed under the I.I.D. hypothesis fail to consider the agnostic distribution shifts, diminishing the generalization ability of deep learning models to Out-Of-Distribution (OOD) data. In this setting, spurious statistical correlations between procedures and diagnoses that may change in different environments will be exploited, which can cause sub-optimal performances of deep learning models and spurious correlation between historical EHR and future diagnosis. To address this problem, we propose to use a causal representation learning method called Causal Healthcare Embedding (CHE). CHE aims at eliminating the spurious statistical relationship by removing the dependencies between diagnoses and procedures. We introduce the Hilbert-Schmidt Independence Criterion (HSIC) to measure the degree of independence between the embedded diagnosis and procedure features. Based on causal view analyses, we perform the sample weighting technique to get rid of such spurious relationship for the stable learning of EHR across different environments. Moreover, our proposed CHE method can be used as a flexible plug-and-play module to enhance existing deep learning models on EHR. Extensive experiments on two public datasets and five state-of-the-art baselines unequivocally show that CHE can improve the prediction accuracy of deep learning models on out-of-distribution data by a large margin. In addition, the interpretability study shows that CHE could successfully leverage causal structures to reflect a more reasonable contribution of historical records for predictions.

Index Terms—Healthcare informatics, causal inference, electronic health records, out-of-distribution

I. INTRODUCTION

Healthcare predictive model for healthcare disease diagnosis based on Electronic Health Records (EHR) is a key engine for improving the quality of clinical care. The comprehensive patient information (such as demographics, diagnoses, and procedures) in EHR provides valuable assistance for personal health status tracking and monitoring [1]–[5]. To predict the future diagnoses based on a patient’s historical EHR, many deep learning models [6]–[8] are proposed with promising accuracy to discover the statistical correlations in the training distribution for predictions. Despite their great successes, the challenge of the out-of-distribution (OOD) problem has not yet been fully addressed. The I.I.D. hypothesis that most models are built upon does not hold true for practical situations due to the inevitable distribution shifts such as data selection bias and confounding factors [9]–[11]. We present the spurious correlation between diagnoses and procedures as an example. The diagnoses and procedures are often correlated as clinicians select treatments according to the patients’ current and historical diagnosis records based on medical experience and knowledge. However, the patients’ demographics and insurance information may vary a lot in the training and test datasets, causing the subtle correlation between diagnoses and procedures to vary in different environments. We argue that the EHR prediction models may be misled by the subtle dependency between diagnoses and procedures, resulting in spurious correlations between historical EHR and future prediction that are unstable when confronting the OOD data in practice. As a result, the learned statistical correlations cannot guarantee to be as effective on inference as on the training dataset.

In the following, we present the causal view analyses to discuss how the correlation between diagnoses and procedures can cause the spurious relationship for model prediction. As shown in Fig. 1, the causal diagram of future diagnosis prediction in EHR consists of two sequences of features, i.e. “past diagnoses” (X) and “past procedures” (Y). By the ignorability assumption in causality [12], any other potential

Fig. 1. The causal diagram of diagnosis prediction in EHR.
confounders are considered uncorrelated to both diagnoses and procedures. Here, we discuss the case where the $X$ and $Y$ are correlated, thus the causal effect of each feature cannot be estimated accurately. In EHR, each diagnosis has an impact on the current and future procedures. At the same time, both diagnoses and procedures influence the “future diagnosis” ($Z$). Because doctors give treatments based on the same medical knowledge, diagnoses and procedures are strongly correlated. Due to the strong correlation of $X \rightarrow Y$, it is hard for deep learning models to learn a stable relationship of $X \rightarrow Z$ and/or $Y \rightarrow Z$. As an instance, patients diagnosed with diabetes can take insulin, and diabetes may cause puffiness. With the strong correlation between diabetes and insulin, a machine learning model has a great chance to learn that insulin causes puffiness. Moreover, the correlation between the two variables may be different in various data distributions, which causes difficulty for model generalization to Out-Of-Distribution (OOD) data. For example, procedures may vary among different insurance types and only some of them can cover insulin drugs, which may result in diet control treatment or non-insulin drug treatments such as Exenatide and Liraglutide for some patients with the same diagnoses. Therefore, models trained by one type of insurances may not always generalize to new insurances.

Recently, a strand of variable decorrelation technique [13], [14] is proposed for linear models. Its basic notion is to remove the dependencies between variables through a sample weighting method and make the correlation structure between each variable and the prediction free of the confounding factors of other variables. In Fig. 1, the arrow from each diagnosis $X$ to each procedure $Y$ will be removed, which leaves the causal diagram with independent $X$ and $Y$ to estimate their contributions. While the concept of variable decorrelation is tempting for healthcare systems, how to extend it to a high-dimensional deep learning model with sequential data can be difficult. First, the nonlinear correlation cannot be measured and eliminated by linear methods. Second, the sample weighting should be redefined to accommodate the sequential data that any past diagnoses can have an impact on a future procedure along the time. It is vital to efficiently remove the dependencies of all combinations of diagnoses and procedures without excessive computational complexity.

In this paper, we propose a causal representation learning method for sequential diagnosis prediction in EHR, called Causal Healthcare Embedding (CHE). To address the two challenges, first, we use Hilbert Schmidt Independence Criterion (HSIC) [15] that measures the norm of cross-covariance from $X$ to $Y$, i.e. the degree of dependence between $X$ and $Y$, which can align with nonlinear neural models. Second, treatments can be represented by latent factors as an alternative for estimating causality. While it is computationally expensive to calculate the binary sample weighting for all sequential combinations of treatments [16], we minimize the HSIC on the two low-dimensional latent representations of diagnoses and procedures to remove the dependency between $X$ and $Y$. Without spurious relationships caused by unstable correlations, deep learning models can exploit the causation between each feature and the prediction, which also improves the model generalization. We perform CHE on multiple baseline models for predicting the future diagnosis and show by extensive experiments that it can improve these models by large margins. To be noted, our proposed CHE method is a plug-and-play module for the diagnosis prediction task. It can be easily incorporated with various diagnosis prediction models.

II. RELATED WORK

A. EHR Data Mining

The mining of EHR is essential for improving the healthcare management of patients. Many tasks that aim at improving healthcare quality can be identified as EHR data mining, such as risk prediction [1], [3], [4], [17], [18], disease progression [19], [20] and diagnosis prediction [7], [21]. Recurrent Neural Networks are naturally suitable, and Long Short-Term Memory (LSTM) [2], [20] has been widely applied. Attention-based and transformer-based models [5]–[8], [22], [23] have also become popular in EHR mining.

B. Counterfactual Prediction

Counterfactual learning is an important direction of research in causal inference [12]. Counterfactual learning can estimate the probability of counterfactual events and eventually identify the unbiased causal relationships between events. The existing counterfactual learning approaches usually reweight samples based on propensity scores [24], [25], which indicate the probabilities of observation. Under the binary treatment setting, balancing the sample weights in the loss function can remove confounding bias to make causal prediction [16], [26].

C. Variable Decorrelation

Stable learning methods perform variable decorrelation for learning causal features in models from biased data, which is a type of causal learning where there is no implicit treatments and the distribution of unobserved samples is unknown [13]. Existing stable learning methods are mostly linear models decorrelating features of samples to make the feature distribution closer to independently identically distribution [14], [27].

III. PRELIMINARY

In the EHR data, we have a set of patients $V = \{v_1, v_2, ..., v_{|V|}\}$, and patient $v_i$ has $t_i$ visits. Diagnoses and procedures are both represented in International Classification of Diseases, Ninth Revision (ICD-9)\(^1\) medical codes, where we have $M$ unique diagnosis medical codes and $N$ unique procedure medical codes. For each patient $v_i$ with $j$ visits, there exists a historical diagnosis sequence $D^i_j = [d^i_{1j}, d^i_{2j}, ..., d^i_{ji}]$ and a historical procedure sequence $P^i_j = [p^i_{1j}, p^i_{2j}, ..., p^i_{ji}]$. Each diagnosis and procedure are $M$-dimensional multi-hot vector and $N$-dimensional multi-hot vector respectively, which means that $d^i_{lj} \in \{0, 1\}^M$ and $p^i_{lj} \in \{0, 1\}^N$, where $1 \leq j \leq t_i$. In this work, we would like to predict future diagnoses, i.e., predicting what diseases a patient will have in the future, based

\(^1\)https://www.cdc.gov/nchs/icd/icd9.htm
on historical EHR. Specifically, in this work, given \( D^j_i \) and \( P^j_i \), we need to predict future diagnosis \( d^j_{i+1} \). We encode diagnoses \( D^i,j \) and procedures \( P^i,j \) into an embedding space by any encoders as \( \mathbf{E}_D^{i,j} = \text{Encoder}(D^i,j) \), \( \mathbf{E}_P^{i,j} = \text{Encoder}(P^i,j) \). As illustrated in Fig. 2, we plan to remove \( \mathbf{E}_D^{i,j} \to \mathbf{E}_P^{i,j} \) so that the contribution of \( D^i,j \) and \( P^i,j \) to predicting \( d^j_{i+1} \) can be free of the interference of unstable \( D^i,j \to P^i,j \).

**IV. METHODOLOGY**

In this section, we introduce the sample weighting method with independence testing statistics to conduct causal disease diagnosis prediction, which is a plug-and-play module that aligns well with deep learning models.

**A. Hilbert Schmidt Independence Criterion**

The removal of dependencies between \( \mathbf{E}_D^{i,j} \) and \( \mathbf{E}_P^{i,j} \) is at the core of sample weighting. To measure the dependency for optimization, we introduce HSIC [15], an independence testing statistics as the Hilbert-Schmidt norm of the cross-covariance operator between the distributions in Reproducing Kernel Hilbert Space (RKHS). As has been proved [28],

\[
\Sigma_{DP} = 0 \iff D \perp P,
\]

which means that if \( \mathbf{E}_D^{i,j} \) cannot be transformed into \( \mathbf{E}_P^{i,j} \) via a nonlinear operator, the two variables are independent.

The squared Hilbert-Schmidt norm of the cross-covariance operator \( \Sigma_{DP} \) can be approximated by the unbiased calculation in the embedding space as

\[
\text{HSIC}(\mathbf{E}_D, \mathbf{E}_P) = \frac{1}{|V|(t^i - 1)} \sum_{i=1}^{t^i - 1} \sum_{j=1}^{V} \text{HSIC}_{local}(\mathbf{E}_D^{i,j}, \mathbf{E}_P^{i,j}).
\]

(2)

Specifically, if \( r \) denotes the hidden dimensionality, we can calculate the HSIC of \( \mathbf{E}_D^{i,j} \in \mathbb{R}^r \) and \( \mathbf{E}_P^{i,j} \in \mathbb{R}^r \) by

\[
\text{HSIC}_{local}(\mathbf{E}_D^{i,j}, \mathbf{E}_P^{i,j}) = \frac{1}{(r^i - 1)^2} \text{Tr}(K_{D}JK_{P}J),
\]

(3)

where \( \text{Tr} \) is the trace of a matrix, \( J = I - 1/r \) with \( I \) as an \( r \)-order identity matrix. \( K_D \) and \( K_P \) are any kernel matrices. We can consider RBF kernel to calculate

\[
K_d(x_1, x_2) = \exp(-\frac{||x_1 - x_2||^2}{\sigma^2}),
\]

(4)

where \( x_1, x_2 \in \mathbb{E}_D^{i,j} \) represent the values in different dimensions of the latent representation. Therefore, \( x_q \in \mathbb{R}^r \) for \( q \in [1, ..., r] \). Similarly, there is \( K_p(x_1, x_2) \) where \( x_1, x_2 \in \mathbb{E}_P^{i,j} \) represent different dimensions of the latent representation. The kernel tricks of \( K_d, K_p \in \mathbb{R}^{r \times r} \) can approximately calculate HSIC rapidly. In this way, for each time’s visit by each patient, the cross-covariance in the embedding space from diagnoses to procedures can be measured by HSIC.

**B. Loss Functions**

Inspired by feature decorrelation techniques [27], [29], we propose to minimize HSIC by the sample weighting technique to mitigate the dependency between diagnoses and procedures in the embedded space. We use \( \mathbf{W} \) to denote sample weights, where the weight for patient \( i \) at the \( j \)-th visit is denoted as \( \omega^j_i \). We denote the weighted samples as \( \mathbf{WE}_D \) and \( \mathbf{WE}_P \), where the weighted samples of patient \( i \) at the \( j \)-th visit are denoted as \( \omega^j_i \mathbf{E}_D^{i,j} \) and \( \omega^j_i \mathbf{E}_P^{i,j} \).

To minimize the correlation between diagnoses and procedures, we propose to optimize \( \mathbf{W} \) with HSIC as follows

\[
\mathbf{W}^* = \arg \min_{\mathbf{W}} \text{HSIC} (\mathbf{WE}_D, \mathbf{WE}_P).
\]

(5)

Meanwhile, we define the cross-entropy loss for the classification task of the diagnosis prediction as

\[
\text{Enc}, \text{Prd} = \arg \min_{\text{Enc, Prd}} \sum_{i=1}^{V} \sum_{j=1}^{t^i-1} \omega^j_i \mathbf{L}^j_i,
\]

(6)

where

\[
\mathbf{L}^j_i = L \left( \text{Prd}(\text{Enc}(D^j_i), \text{Enc}(P^j_i)), d^j_{i+1} \right),
\]

(7)

where \( L \) denotes the cross-entropy loss function. \( \text{Enc} \) represents the encoder that maps diagnoses and procedures into the embedding space. \( \text{Prd} \) represents the final prediction layer that maps the latent representation into the one-hot probability vector. The architectures of \( \text{Enc} \) and \( \text{Prd} \) depend on the base model our method is used upon. \( L \) is also based on the specific loss function used in the base model.

**C. Model Optimization**

We iteratively optimize the weighted loss and the HSIC by

\[
\text{Enc}_{n+1}, \text{Prd}_{n+1} = \arg \min_{\text{Enc, Prd}} \sum_{i=1}^{V} \sum_{j=1}^{t^i-1} \omega^j_i \mathbf{L}^j_i,
\]

(8)

and

\[
\mathbf{W}(n + 1) = \arg \min_{\mathbf{W}} \epsilon \cdot \text{HSIC} (\text{Diag, Proc}),
\]

(9)

where

\[
\text{Diag} = \mathbf{W}(n)\text{Enc}_{n+1}(D^j_i), \text{Proc} = \mathbf{W}(n)\text{Enc}_{n+1}(P^j_i).
\]

(10)

\( \text{Enc}_n, \text{Prd}_n \) and \( \mathbf{W}(n) \) indicates encoder, final prediction layer and sample weights at the \( n \)-th iteration, and \( \mathbf{W}(0) \) is initially set as ones. \( \epsilon \) is a coefficient that balances the learning rates for updating the neural network and sample weights.
Eq. (16) and Eq. (17) are optimized iteratively, meaning that we first optimize the neural network and then optimize the HSIC for each iteration. Every two subsequences $E_{d_j}^{i,j}$ and $E_{p_j}^{i,j}$ of length $j$ are fed into a neural network to calculate the cross-entropy loss, and sample weights are multiplied to the loss to update the model parameters. Then, we use the updated model to calculate the HSIC of $E_{d_j}^{i,j}$ and $E_{p_j}^{i,j}$ obtained by the encoder part of the model. The sample weighting reassigns the importance of each sample when calculating the loss function to remove the dependency between features.

**D. Complexity Analysis**

The time complexity of calculating HSIC only grows with the hidden dimensionality $r$. By naive algorithms, the multiplication of $K_d$ and $K_p$ is $O(r^3)$, and the calculation of trace is also $O(r^3)$. For deep learning models, $r$ is a hyperparameter and is thus trivial. The calculation of overall HSIC is $O(|V|t)$ if we denote $t = \max(t')$ for $\forall i$, which is linearly proportional to the number of visits in the data. This is acceptably efficient. On the other hand, the number of treatments for each timestamp is $M$, the number of unique ICD-9 codes for diagnosis. Considering the combination of ICD-9 codes in a sequence, the total number of treatments can be as many as $M^t$, which makes the traditional counterfactual weighting to estimate propensity scores very expensive.

**V. EXPERIMENTS**

In this section, we conduct extensive experiments to verify the effectiveness of our proposed CHE method.

**A. Datasets**

We evaluate our proposed sequential stable learning method on two real-world datasets [30]: MIMIC-III\(^2\) and MIMIC-IV\(^3\). Patients with less than three admission records are excluded.

\[^2\]https://physionet.org/content/mimiciii/1.4/
\[^3\]https://physionet.org/content/mimiciv/0.4/

**• MIMIC-III Dataset** The average number of visits for the 1970 selected patients is 3.69, the average number of codes in a visit is 13.23, the total number of unique ICD-9 codes in diagnoses is 3320, and the total number of unique ICD-9 codes in procedures is 988.

**• MIMIC-IV Dataset** The average number of visits for the 10023 selected patients is 4.64, the average number of codes in a visit is 14.12, the total number of unique ICD-9 codes in diagnoses is 6274, and the total number of unique ICD-9 codes in procedures is 1970.

**B. Baseline Models**

We apply our method on the following baselines for the evaluation of diagnosis prediction accuracy: LSTM, RETAIN [6], Dipole [2], Concare [8], StageNet [5]. We denote above models as BaseModels, and we incorporate them with the CHE method as CHE+BaseModels. All models are used with adaptation to our task where only historical diagnoses and procedures are available. We implement our method with the Mindspore framework. To reflect the performance of counterfactual learning methods, we also conduct experiment on Permutation Weighting (PW) [16] for comparison. PW conducts permutation on observed features for calculating propensity scores. While the historical EHR in the dataset is regarded as positive samples, we randomly generate 10x larger number of negative samples that do not exist in the dataset and estimate their propensity scores. We also incorporate PW with the above BaseModels, and name them as PW+BaseModels.

**C. Settings**

Considering the insurance type, such as Medicare, Medicaid and Private, may affect procedures for similar diagnoses, we evaluate the performance when training and test data are divided by the type of insurances to simulate the scenario of OOD generalization. Here, we divide all the Medicare data into the training and validation set in a 0.7:0.3 ratio and use the Private/Other (MIMIC-III/MIMIC-IV) data as the test set.
Common hyperparameters used by all models in the experiments include learning rate, batch size, hidden dimension, dropout rate. These hyperparameters are tuned to obtain the optimal evaluation metrics on the validation set for each model. We apply early-stopping so the training will stop if the validation metrics do not increase in twenty epochs and the test performance will be recorded.

D. Evaluation Metrics

We adopt the topk accuracy and normalized discounted cumulative gain (NDCG) to evaluate the diagnosis prediction performance. We use the same accuracy@k metric used in prior works [2], [22] as the correct medical ICD-9 codes ranked in topk divided by min(k, |y_i|), where |y_i| is the number of ICD-9 codes in the (t+1)-th visit. NDCG@k further considers the normalization of gains and the ranking of correct medical codes. In our experiments, we use k ∈ \{10, 20\}. We also provide an averaged metric of all the four metrics on two datasets to reflect the overall comparison to baselines.

E. Performance Comparison

In this subsection, we conduct performance comparison among BaseModels, PW+BaseModels and CHE+BaseModels. In Table I, we show performance comparison with out-of-distribution data. Because we use Medicare insurance data for training and Private/Other insurance data for testing, the training and test sets are not I.I.D. Therefore, this poses additional challenges to the generalization of deep learning models. The results show that CHE has relatively greater improvements on all metrics against BaseModels than the results with random data division, with relative increase of the averaged metric of NDCG@k and ACC@k by 10.06%. And the significant test shows that, CHE+BaseModel significantly outperforms BaseModel and PW+BaseModel. This demonstrates our claim that the proposed CHE encourages models to rely on causal features and estimate their contributions to the prediction independently, as causal features are always useful to make disease diagnosis predictions regardless of the data distribution shifts. Moreover, the results also show that the counterfactual PW approach does not always increase the prediction accuracy, which might be due to the inaccurate estimation of propensity scores as discussed in Section 4.4.

F. Visualization

To understand whether the proposed CHE can make the model to learn causal features, we visualize the contribution of each feature to the feature of the prediction of future diagnosis. We know that the cause of diabetic retinopathy is diabetes. For patients with background diabetic retinopathy (ICD-9 code 36021), an ideal model should rely on related diseases such as diabetes to make prediction. In Table II, we show the contributions of diagnosis D_j in the MIMIC-III dataset and the feature interpretations, contributions of the features to the prediction, in CHE+Dipole and Dipole. Specifically, we apply gradient backpropagation for calculating feature interpretations [31–33]. The contributions of diagnosis d_j, and procedure p_j, for predicting d_{j+1} (j' ≤ j) can be calculated as \( \frac{\partial d_{j+1}}{\partial E_{d,j'}} = \frac{\partial d_{j+1}}{\partial E_{p,j'}} (E_{d,j'})^T \) and \( \frac{\partial d_{j+1}}{\partial E_{p,j'}} = \frac{\partial d_{j+1}}{\partial d_{j'}} (E_{p,j'})^T \) respectively, where \( d_{j+1} = \text{Prd} \{ \text{Enc}(D_j), \text{Enc}(P_j) \} \) is the prediction as in the loss function in Eq. (7).
In this example, the diagnosis sequence is \{4280, 25856\}, \{99592, 4280, 25060, 3572, V5681, V1251, 99662, 40391, 03811, 25050, 36201, 5856\}, \{03811, 5856, 99681, 42832\}. The future diagnosis to be predicted is backgound diabetic retinopathy (36021). The first visit contains two diseases that appear frequently among people, i.e. congestive heart failure (4280) and end stage renal disease (5856). The second visit contains some highly related features, such as diabetic retinopathy (25050), diabetes with neurological manifestations (25060), and polynuropathy in diabetes (3572). In the third visit, the complications of transplanted kidney (99681) might be related. Compared with Dipole, CHE+Dipole pays more attention to causal features, i.e., the second visit with many highly related diagnoses. Moreover, the contributions of diagnosis and procedure are less correlated.

VI. CONCLUSION

In this work, we focus on the unstable learning problem of EHR data mining. This is caused by the strong correlation between diagnoses and procedures in EHR, and it is hard for deep learning-based models to learn their causal relationships in future diagnosis. Accordingly, we propose a CHE method to learn causal representations for diagnosis prediction models via removing dependencies between diagnoses and procedures by weighting technique. We demonstrate by extensive experiments on the sequential diagnosis and procedure features as examples that CHE can significantly improve the performances of diagnosis prediction models. The visualizations demonstrate that CHE presents more causal predictions than baselines.

VII. ACKNOWLEDGEMENT

This work is partially sponsored by CAAI-Huawei Mind-Spore Open Fund.

REFERENCES

[1] Y. Cheng, F. Wang, P. Zhang, and J. Hu, “Risk prediction with electronic health records: A deep learning approach,” in SDM, 2016, pp. 432–440.
[2] 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 KDD, 2017, pp. 1903–1911.
[3] T. Ma, C. Xiao, and F. Wang, “Health-atm: A deep architecture for multifaceted patient health record representation and risk prediction,” in SDM, 2018, pp. 261–269.
[4] X. S. Zhang, F. Tang, H. H. Dudge, J. Zhou, and F. Wang, “Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records,” in KDD, 2019, pp. 2487–2495.
[5] J. Gao, C. Xiao, Y. Wang, W. Tang, L. M. Glass, and J. Sun, “Stagenet: Stage-aware neural networks for health risk prediction,” in The Web Conference, 2020, pp. 530–540.
[6] 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, pp. 3512–3520.
[7] J. Gao, X. Wang, Y. Wang, Z. Yang, J. Gao, J. Wang, W. Tang, and X. Xie, “Cump: Co-attention memory networks for diagnosis prediction in healthcare,” in ICDM, 2019, pp. 1036–1041.
[8] L. Ma, C. Zhang, Y. Wang, W. Ruan, J. Wang, W. Tang, X. Ma, X. Gao, and J. Gao, “Concare: Personalized clinical feature embedding via capturing the healthcare context,” in AAAI, 2020, pp. 833–840.
[9] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola, “Correcting sample selection bias by unlabeled data,” NeurIPS, pp. 601–608, 2006.
[10] D. Hendrycks and T. Dietterich, “Benchmarking neural network robustness to common corruptions and perturbations,” in ICLR, 2018.
[11] Y. Bengio, T. Delius, N. Rahaman, N. R. Ke, S. Lachapelle, O. Bili- niak, A. Goyal, and C. Pal, “A meta-transfer objective for learning to disentangle causal mechanisms,” in ICLR, 2019.
[12] J. Pearl, “Causal inference in statistics: An overview,” in Statistics Surveys, vol. 3, pp. 96–146, 2009.
[13] K. Kuang, P. Cui, S. Athey, R. Xiong, and B. Li, “Stable prediction across unknown environments,” in KDD, 2018, pp. 1617–1626.
[14] K. Kuang, R. Xiong, P. Cui, S. Athey, and B. Li, “Stable prediction with model misspecification and agnostic distribution shift,” in AAAI, 2020, pp. 4485–4492.
[15] J. Gao, J. Xue, K. Fukuniza, C. H. Too, L. Song, B. Schölkopf, A. J. Smola et al., “A kernel statistical test of independence,” in NeurIPS, 2007, pp. 585–592.
[16] D. Arbour, D. Dimmery, and A. Sondhi, “Permutation weighting,” in ICML, 2021, pp. 331–341.
[17] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu, “Boosting deep learning risk prediction with generative adversarial networks for electronic health records,” in ICDM, 2017, pp. 787–792.
[18] Y. Luo, C. Xu, Y. Liu, W. Liu, S. Zheng, and J. Bian, “Learning differential operators for interpretable time series modeling,” in Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 1192–1201.
[19] E. Choi, N. Du, R. Chen, L. Song, and J. Sun, “Constructing disease network and temporal progression model via context-sensitive hawkes process,” in ICDM, 2015, pp. 721–726.
[20] M. Ye, J. Luo, C. Xiao, and F. Ma, “Lasan: Modeling long-term dependencies and short-term correlations with hierarchical attention for risk prediction,” in CIKM, 2020, pp. 1753–1762.
[21] E. Choi, Z. Xu, Y. Li, M. Dusenberry, G. Flores, E. Xue, and A. Dai, “Learning the graphical structure of electronic health records with graph convolutional transformer,” in AAAI, 2020, pp. 606–613.
[22] X. Zhang, B. Qian, S. Cao, Y. Li, H. Chen, Y. Zheng, and I. Davidson, “Inprem: An interpretable and trustworthy predictive model for healthcare,” in KDD, 2020, pp. 450–460.
[23] X. Peng, G. Long, T. Shen, S. Wang, and J. Jiang, “Sequential diagnosis prediction with transformer and ontological representation,” in 2021 IEEE International Conference on Data Mining (ICDM). IEEE, 2021, pp. 489–498.
[24] N. Hassanpour and R. Greiner, “Learning disentangled representations for counterfactual regression,” in ICLR, 2019.
[25] M. J. Lopez and R. Gutman, “Estimation of causal effects with multiple treatments: a review and new ideas,” Statistical Science, pp. 432–454, 2017.
[26] N. Hassanpour and R. Greiner, “Counterfactual regression with importance sampling weights,” in IJCAI, 2019, pp. 5880–5887.
[27] Z. Shen, P. Cui, T. Zhang, and K. Kunag, “Stable learning via sample reweighting,” in AAAI, 2020.
[28] Z. Shen, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, “Deep stable learning for out-of-distribution generalization,” in CVPR, 2021, pp. 5312–5322.
[29] Z. Shen, P. Cui, J. Liu, T. Zhang, B. Li, and Z. Chen, “Stable learning via differentiated variable decorrelation,” in KDD, 2020, pp. 2185–2193.
[30] 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, vol. 3, no. 1, pp. 1–9, 2016.
[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in ICCV, 2017, pp. 618–626.
[32] D. Smilkov, N. Thorat, B. Kim, F. Viegas, and M. Wattenberg, “Smoothgrad: removing noise by adding noise,” in ICML, 2017.
[33] Q. Liu, Z. Liu, H. Zhang, Y. Chen, and J. Zhu, “Mining cross features for financial credit risk assessment,” in CIKM, 2021.