Dual Knowledge-Guided Data Augmentation for Robust Clinical Prediction Models

Sangwoo Moon[1], Peong Gang Park[2,3], Naye Choi[4], Ji Hyun Kim[5,6], Seon Hee Lim[7], Joo Hoon Lee[8], Hee Sun Baek[9], Min Hyun Cho[9], Keum Hwa Lee[2], Jae Il Shin[2,3], Kyoung Hee Han[10], Jeong Yeon Kim[11], Ji Yeon Song[7], Eun Mi Yang[12], Seong Heon Kim[4,6,13], Yo Han Ahn[4,6,13], Hee Gyung Kang[4,6,13], and Eujin Park[14,*]

[1]Department of Computer Science and Engineering, College of Engineering, Seoul National University, Seoul, Republic of Korea
[2]Division of Pediatric Nephrology, Severance Children's Hospital, Seoul, South Korea
[3]Department of Pediatrics, Yonsei University College of Medicine, Seoul, South Korea
[4]Department of Pediatrics, Seoul National University Children's Hospital, Seoul, South Korea
[5]Department of Pediatrics, Seoul National University Bundang Hospital, Seongnam, South Korea
[6]Department of Pediatrics, Seoul National University College of Medicine, Seoul, South Korea
[7]Department of Pediatrics, Pusan National University Children's Hospital, Pusan National University, School of Medicine, Busan, South Korea
[8]Department of Pediatrics, Children's Hospital, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea
[9]Department of Pediatrics, Kyungpook National University, School of Medicine, Daegu, South Korea
[10]Department of Pediatrics, Jeju National University College of Medicine, Jeju, South Korea
[11]Department of Pediatrics, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea
[12]Department of Pediatrics, Chonnam National University Hospital and School of Medicine, Gwangju, South Korea
[13]Kidney Research Institute, Seoul National University Medical Research Center, Seoul, South Korea
[14]Department of Pediatrics, Korea University Guro Hospital, Seoul, South Korea
[*]eujinpark@korea.ac.kr

ABSTRACT

The successful adoption of medical Artificial Intelligence (AI) models is critically hampered by the domain shift problem, particularly in the challenging Single-Source Domain Generalization (SSDG) scenario prevalent in data-scarce pediatric medicine. Existing augmentation methods, such as standard Mixup and Input Masking, fail to leverage the rich structural information and expert knowledge embedded within clinical data, leading to models that overfit to source-domain specifics (*i.e.*, shortcuts). This study proposes a Dual Knowledge-Guided Data Augmentation Framework to systematically enhance model robustness by embedding clinical expertise into the learning process. Our framework introduces two novel components: 1) Similarity-guided Mixup, which generates clinically plausible virtual data by selectively interpolating between patients with high similarity in key clinical features; and 2) Group-based Masking, which simulates realistic data imperfections by concurrently masking clinically related feature groups. We validated this framework using the multicenter KNOW-pedCKD cohort for pediatric chronic kidney disease, training exclusively on a single source domain and testing on three unseen target domains. The proposed method significantly improved the clinically critical metric of Recall. This research validates that embedding domain knowledge into data augmentation offers a promising and practical pathway toward developing generalizable and trustworthy medical AI models that can reliably operate across heterogeneous clinical environments. The codes are available at https://github.com/msw6468/KNOW_pedCKD_SDG.

Introduction

Advancements in Artificial Intelligence (AI) and machine learning have shown immense potential across numerous fields. In the medical domain, in particular, clinical decision support systems (CDSS) based on imaging signals and Electronic Health Records (EHR) data are emerging as core technologies[1–3]. These systems play a crucial role in enhancing patient safety and treatment efficacy by assisting physicians in diagnoses and minimizing potential errors during the therapeutic process.

At the heart of this technological progress are predictive models that learn meaningful patterns from given sets of training data. Such models are actively developed to predict disease risk, simulate progression, and forecast treatment responses using

biological signals, laboratory results, imaging, and genetic data. Among these applications, AI holds particular promise in the management of chronic diseases such as chronic kidney disease (CKD). CKD has a high global prevalence and represents a major public health concern. Its progression to end-stage kidney disease is associated with substantial comorbidities and imposes a significant clinical and socioeconomic burden, as patients ultimately require costly kidney replacement therapies (KRTs) such as dialysis or kidney transplantation[4,5]. Therefore, early prediction of disease progression and timely intervention to slow its course are of paramount importance, and AI-based predictive models can serve as powerful tools in this context[6–8].

Despite this bright outlook, numerous challenges stand in the way of the successful implementation of medical AI models in real clinical settings. One of the most critical obstacles is the domain shift problem[9]. Domain shift refers to the phenomenon where an AI model, trained on data from a specific institution (the source domain), exhibits substantial degradation in predictive performance when applied to another institution (the target domain). This is a critical issue that undermines the robustness and trustworthiness of medical AI systems and constitutes a major technical barrier to their widespread adoption. To address this limitation, recent research has focused on Domain Generalization[10] and robust learning approaches that enhance model stability across diverse environments.

Although children account for only a small proportion of all CKD patients, their prolonged disease course and lifelong complications necessitate greater reliance on AI-based approaches to support clinical decision-making. However, pediatric data for CKD patients are inherently more difficult to collect than adult data and remain limited in both quantity and quality[11]. These constraints underscore the particular importance of single-source domain generalization and robust learning approaches in achieving reliable and generalizable model performance.

This study aimed to address the challenge of domain generalization in clinical prediction, where a model trained on data from a single source must maintain robust performance when applied to unseen target domains. The following sections provide a literature review covering Single-Source Domain Generalization, Mixup, and Input Masking.

Single Source Domain Generalization

Single-Source Domain Generalization (SSDG) has been particularly active in the field of Computer Vision[12,13]. The core idea in vision is to decompose image data into two components: Content and Style. Here, content refers to the 'domain-invariant' features, such as the essential shape and structure of an object, which remain consistent across domains. On the other hand, style refers to the domain-variant features, such as the color palette, texture, and lighting of an image, which change depending on the domain.

Based on this approach, researchers have employed strategies that artificially diversify the style of an image to train the model to focus solely on the content. Successful data augmentation techniques have been proposed that manipulate the frequency characteristics of an image using the Fourier Transform[14–17] or apply deep learning-based style transfer[18] techniques to render a single image as if it were created in various artistic styles.

However, these computer vision-based SSDG approaches face a clear limitation: they cannot be directly applied to the tabular dataset. While the separation of content and style has been successful in computer vision, this often relies on some prior knowledge about what constitutes stylistic variation. In contrast, for the tabular data in a single-source setting, we inherently lack the prior knowledge about inter-hospital differences needed to define what constitutes domain-invariant features versus domain-variant features. Consequently, the existing SSDG studies are difficult to apply to tabular datasets, which necessitated the introduction of robust learning or regularization techniques for domain generalization.

Mixup

Mixup[19] methods regularize deep networks by creating novel training instances from multiple samples. The foundational method, Mixup, generates virtual samples through linear interpolation of two data points and their labels. Manifold Mixup[20] applies interpolation within the hidden layers.

Another popular approach, CutMix[21], combines samples by cutting a patch from one image and pasting it onto another, with labels mixed proportionally to the patch area. Subsequent methods have refined this patch-based strategy. SaliencyMix[22] leverages saliency detection to ensure the cropped region is informative, while ResizeMix[23] simplifies the process by resizing a source image into a small patch.

However, these methods have a fundamental problem: it combines patient data randomly without any consideration for the clinical context[19,20] or leverage characteristics of vision domain[21–23]. For example, mixing the data of a stable, early-stage patient with that of a late-stage patient on the verge of requiring renal replacement therapy can create a clinically implausible virtual patient. Such unrealistic data can act as 'noise' in the model training process and degrade performance, which strongly suggests the need for the proper approaches.

Input Masking

Input Masking[24], also known as Input Dropout, is another widely used technique to enhance model robustness. It operates by randomly setting a fraction of the input features to zero (or another baseline value) during training. This prevents the

model from becoming overly reliant on any single feature, forcing it to learn a more distributed and robust representation of the data. However, like standard Mixup, conventional input masking has limitations in a clinical setting. It typically applies masking to all features with a uniform probability, assuming that feature dropout is random. In reality, missing data in clinical environments often follows specific, non-random patterns. For instance, a set of related lab values from a single blood test may be missing concurrently. The failure of standard masking to account for these real-world patterns motivates our proposal for a knowledge-guided approach, which simulates more realistic missing data scenarios.

Methods

This section details the proposed benchmark and methodology. Figure 1 presents an overview of both the benchmark and the method.

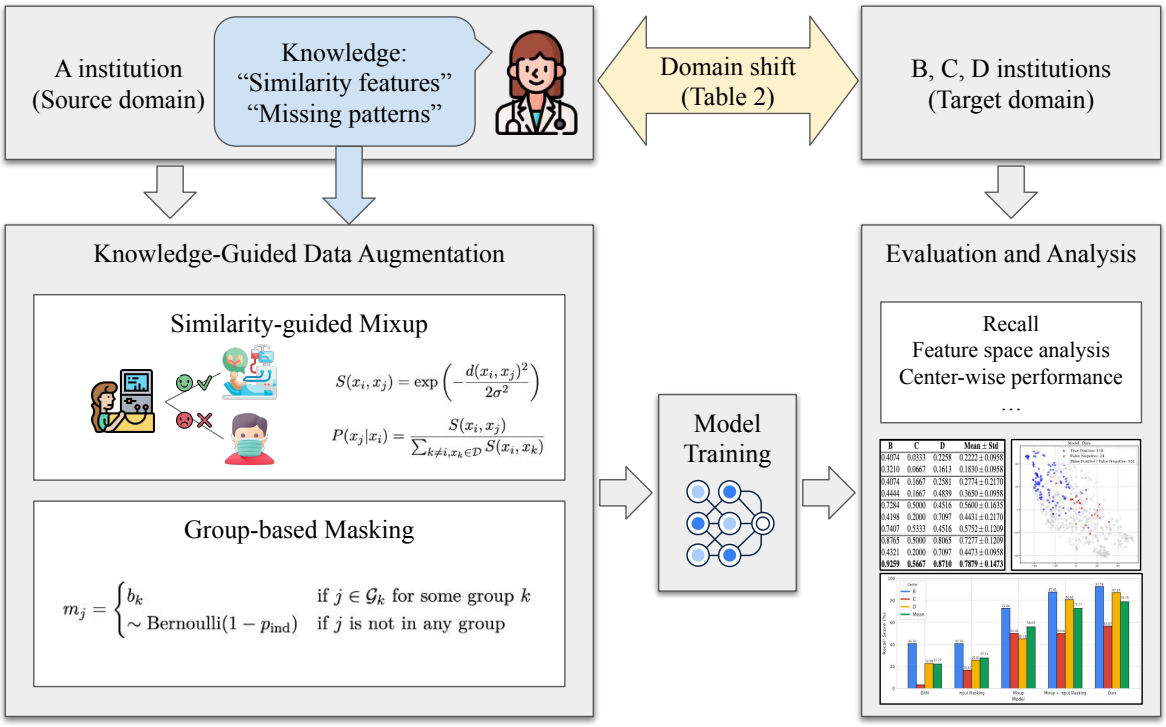

Figure 1. Overview of our framework. We presented a benchmark for evaluating model generalization performance under domain shift conditions (*i.e.*, single-source domain generalization). In this setting, we evaluated and analyzed the performance of models trained using clinical knowledge-guided data augmentation.

Problem setting

The primary goal our study falls under the framework of Domain Generalization (DG), aiming to train a robust model f_θ on a known source domain $\mathscr{D}_s$ that can maintain high performance when applied to an unseen target domain $\mathscr{D}_t$. Specifically, we address the challenging Single-Source Domain Generalization (SSDG) setting, where the model is exclusively trained on data from a single domain $\mathscr{D}_s$, with no access to data from the target domains or any other external source.

We formally denote the source and target datasets as $\mathscr{D}_s$ and $\mathscr{D}_t$, respectively. The score challenge arises because these domain are sampled from distinct probability distributions ($P_s \neq P_t$). Each data point $(\mathbf{x}_i, y_i) \in \mathscr{D}$ consists of a D-dimensional feature vector $\mathbf{x}_i \in \mathbb{R}^D$ and a binary label $y_i \in \{0, 1\}$. In this setting, we train a model $f_\theta : \mathbb{R}^D \to \{0, 1\}$ on the source domain $\mathscr{D}_s$ and evaluate its performance on the unseen target domain $\mathscr{D}_t$.

Proposed Method 1: Similarity-guided Mixup

Prior Mixup research largely focused on two approaches: random selection for pixel-wise interpolation or leveraging image-specific prior knowledge for augmentation. The former approach is inadequate for clinical data as it performs interpolation without regard for a patient's medical status, potentially generating clinically implausible virtual samples. Furthermore, the latter, which utilizes visual priors (*e.g.*, saliency or patch area), is fundamentally ill-suited for application to tabular datasets.

Our proposed method overcomes this limitation by defining sample similarity based on a clinically meaningful feature set, making it effective for medical tabular data. We introduce Similarity-guided Mixup, a technique that first identifies a subset of clinically significant features using prior domain knowledge. Mixup augmentation is then performed exclusively between patient samples that exhibit high similarity within this predefined clinical feature space.

We define a subset of features, $\mathscr{F}_{\text{clinical}} \subset \{1,2,\ldots,D\}$, which are selected based on domain expertise as being most critical to the clinical task. Let $\phi : \mathbb{R}^D \to \mathbb{R}^{|\mathscr{F}_{\text{clinical}}|}$ be a projection function that extracts only these clinically meaningful features from a given sample vector $\mathbf{x}$.

Distance Calculation The distance $d(\mathbf{x}_i, \mathbf{x}_j)$ between two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ is measured exclusively within this projected feature space using the Euclidean norm (L_2 norm):

$$d(\mathbf{x}_i, \mathbf{x}_j) = ||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||_2 \tag{1}$$

Similarity Score This distance is subsequently used to calculate a similarity score, which forms the basis for our probabilistic sampling. We formalize this using a Gaussian kernel to convert distance into a similarity score $S(\mathbf{x}_i, \mathbf{x}_j)$:

$$S(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{x}_j)^2}{2\sigma^2}\right) \tag{2}$$

where σ is a hyperparameter that controls the sensitivity of the similarity score.

Sampling Probability Finally, the probability $P(\mathbf{x}_j|\mathbf{x}_i)$ of selecting anchor sample $\mathbf{x}_j$ from the dataset $\mathscr{D}$ to form a Mixup pair with $\mathbf{x}_i$ is given by normalizing these similarity scores across all possible pairs:

$$P(\mathbf{x}_j|\mathbf{x}_i) = \frac{S(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{k:\mathbf{x}_k \in \mathscr{D}} S(\mathbf{x}_i, \mathbf{x}_k)} \tag{3}$$

This ensures that samples that are clinically more similar have a higher probability of being mixed, thereby generating more plausible and effective augmented data.

Interpolation Once the pair $(\mathbf{x}_i, \mathbf{x}_j)$ is selected according to $P(\mathbf{x}_j|\mathbf{x}_i)$, the augmented sample $(\mathbf{x}', y')$ is generated using the conventional Mixup formula:

$$\mathbf{x}' = \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j \tag{4}$$

$$y' = \lambda y_i + (1-\lambda) y_j \tag{5}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$.

Proposed Method 2: Group-based Masking

Conventional Input Masking assumes missingness occurs randomly and independently across features. Our group-based masking incorporates domain knowledge by modeling the correlated missing patterns often observed in clinical data, thereby generating more realistic data imperfections to enhance model robustness.

Identifying Correlated Missing Patterns Based on clinical consultation (*e.g.*, specific lab panels or concurrent measurement difficulties like in pediatric BP), features known to be concurrently missing are grouped into sets $\mathscr{G} = \{\mathscr{G}_1, \mathscr{G}_2, \ldots\}$.

Masking Procedure The binary mask vector $m \in \{0,1\}^D$ applied to the input feature vector $x \in \mathbb{R}^D$ is generated based on feature groups $\{\mathscr{G}_1, \ldots, \mathscr{G}_K\}$ defined by clinical knowledge. For each group G_k, a single binary variable $b_k \sim \text{Bernoulli}(1 - p_{\text{group}})$ is sampled.

$$m_j = \begin{cases} b_k & \text{if } j \in \mathscr{G}_k \text{ for some group } k \\ \sim \text{Bernoulli}(1-p_{\text{ind}}) & \text{if } j \text{ is not in any group} \end{cases} \tag{6}$$

Here, p_{group} is the probability of group masking, and p_{ind} is the probability of individual feature masking. The final masked vector is $\mathbf{x}' = \mathbf{x} \otimes m$. The final augmented input $\tilde{\mathbf{x}}$ is the result of applying both correlated (Group) and independent (Random) masking mechanisms, simulating real-world data heterogeneity and missingness structure.

Dataset and evaluation

For the validation of our proposed methodology, this study utilized data from the KNOW-pedCKD study (ClinicalTrials.gov: NCT02165878; registered June 11, 2014), a multicenter prospective observational cohort comprising Korean pediatric patients with CKD. Data were collected from teaching hospitals associated with seven major pediatric nephrology centers in South Korea. Figure 2 illustrates the study protocol, including the specific collection period, patient inclusion, and exclusion criteria utilized for the KNOW-pedCKD dataset.

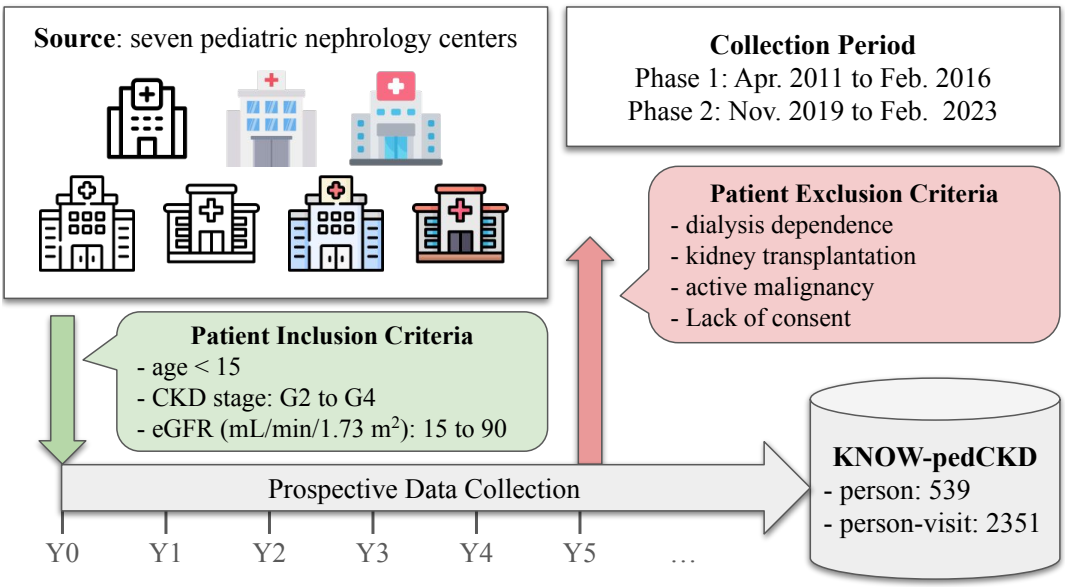

Figure 2. **Overview of KNOW-pedCKD dataset collection**

Table 1. **Demographics and clinical characteristics**

Characteristic, unit	A (Source)	B (Target1)	C (Target2)	D (Target3)
Sex, Male	333 (30.69%)	144 (32.88%)	102 (29.91%)	111 (35.35%)
Age, years	12.85 ± 5.91	12.10 ± 6.35	13.69 ± 5.45	14.43 ± 5.11
eGFR, mL/min/1.73m^2	58.05 ± 29.24	57.39 ± 32.09	61.90 ± 26.32	89.85 ± 29.52
Systolic BP, mmHg	111.36 ± 12.49	113.98 ± 15.36	114.63 ± 11.39	117.76 ± 15.76
Diastolic BP, mmHg	66.65 ± 10.93	68.86 ± 12.32	70.01 ± 9.18	66.42 ± 11.57
Hemoglobin, g/dL	13.14 ± 1.88	12.51 ± 2.06	13.58 ± 1.86	13.20 ± 1.79
Reticulocyte, %	1.48 ± 1.25	1.34 ± 0.87	1.63 ± 0.63	1.25 ± 0.49
Potassium, mmol/L	4.40 ± 0.52	4.28 ± 0.61	4.40 ± 0.43	4.36 ± 0.46
Chloride, mmol/L	105.52 ± 3.52	105.45 ± 3.53	103.21 ± 3.15	104.48 ± 2.89
Calcium, mg/dL	9.60 ± 0.54	9.29 ± 0.63	9.43 ± 0.52	9.43 ± 0.64
Phosphate, mg/dL	4.47 ± 0.81	4.86 ± 4.19	4.29 ± 0.70	4.20 ± 0.87
Albumin, g/dL	4.26 ± 0.42	3.99 ± 0.56	4.36 ± 0.39	4.36 ± 0.49
UPCR, mg/mg	1.19 ± 2.91	2.01 ± 7.68	0.71 ± 0.95	0.72 ± 1.79
Calcium phosphate binder	279 (27.41%)	80 (18.31%)	59 (17.61%)	10 (3.25%)
Iron supplements				
- Oral	218 (20.15%)	62 (14.16%)	29 (8.53%)	18 (5.73%)
- Intravenous	6 (0.55%)	4 (0.91%)	0 (0.00%)	0 (0.00%)
- Intravenous and oral	4 (0.37%)	21 (4.79%)	0 (0.00%)	0 (0.00%)
- Not use	854 (78.93%)	351 (80.14%)	311 (91.47%)	296 (94.27%)
ESA				
- Epoetin alfa	9 (0.83%)	13 (2.97%)	0 (0.00%)	2 (0.64%)
- Epoetin beta	86 (7.96%)	21 (4.79%)	0 (0.00%)	1 (0.32%)
- Darbepoetin alfa	13 (1.20%)	3 (0.68%)	0 (0.00%)	3 (0.96%)
- CERA	2 (0.19%)	4 (0.91%)	0 (0.00%)	0 (0.00%)
- Not use	971 (89.82%)	397 (90.64%)	340 (100.00%)	308 (98.09%)
ACE inhibitor	304 (29.09%)	144 (33.18%)	135 (40.42%)	208 (66.45%)
ARB	384 (36.75%)	60 (13.82%)	30 (8.98%)	91 (29.07%)

Table 2. Statistical significance test results for features between the source and target domains. P-values are reported for continuous variables, and chi-square test P-values are reported for categorical variables. A significant difference in distribution was observed for most features.

Characteristic, unit	A (Source)	B (Target1)	C (Target2)	D (Target3)
Count (% of patients)	1085 (15.48%)	438 (18.49%)	341 (8.80%)	314 (9.87%)
Sex	-	0.41	0.79	0.12
Age, years	-	0.02	0.04	0
eGFR, mL/min/1.73m^2	-	0.35	0.01	0
Systolic BP, mmHg	-	0.01	0	0
Diastolic BP, mmHg	-	0	0	0.31
Hemoglobin, g/dL	-	0	0	0.28
Reticulocyte, %	-	0.04	0	0.04
Potassium, mmol/L	-	0	0.53	0.05
Chloride, mmol/L	-	0.66	0	0
Calcium, mg/dL	-	0	0	0
Phosphate, mg/dL	-	0.01	0	0
Albumin, g/dL	-	0	0	0
UPCR, mg/mg	-	0	0.07	0
Calcium phosphate binder	-	0	0	0
Iron supplements	-	0.33	0	0
ESA	-	0.70	0	0
ACE inhibitor	-	0.12	0	0
ARB	-	0	0	0.01

The core research problem is defined as a classification task focused on predicting outcomes for patients in an unseen institution (target domain). Specifically, the task is to predict whether a patient's condition would worsen (defined as Kidney Replacement Therapy (KRT), death, or a $\geq 20\%$ decline in eGFR) within the subsequent year, based on their current visit data.

To measure Single-Source Domain Generalization (SSDG) performance, we sequentially named the institutions with the largest patient populations as A, B, C, and D. We designated the largest institution, A, as the source domain, and utilized the remaining three institutions (B, C, and D) as the target domains to build our benchmark. Patient records, collected sequentially during annual visits, were segmented into person-visit units (representing a one-year interval). In preprocessing, standardization was applied to continuous features to minimize the impact of scale differences on model training. Categorical features were processed using one-hot encoding to prevent the model from assuming an arbitrary ordinal relationship or magnitude between distinct categories (*e.g.*, assigning a value of 1, 2, 3 to categories would imply $3 > 1$, which is incorrect for nominal data). Missing data during model training were handled via Multiple Imputation by Chained Equations (MICE), utilizing an imputer previously trained on the raw training dataset. Table 1 shows the raw feature distribution for each domain, while Table 2 presents the statistical significance difference between the source domain (A) and the target domains (B, C, and D). Specifically, p-value analysis confirmed strict differences in feature distributions across the datasets, statistically validating the existence of a domain shift between institutions.

Model evaluation employed both AUROC and Recall score. The model at the epoch with the highest AUROC on the validation set was selected as the optimal checkpoint, as AUROC provides a comprehensive, threshold-independent measure of overall model discrimination. However, Recall was chosen for reporting the final performance across different methods because, in a clinical setting with critical conditions like kidney failure, maximizing the detection of true positive cases is often prioritized to ensure minimal missed intervention opportunities.

Baselines

We compared the performance of our proposed method against several baselines: For non-augmentation-based baseline, Empirical Risk Minimization (ERM), Invariant Risk Minimization (IRM[25]). As IRM is inherently designed for multi-source domain generalization, we adapted it to our SSDG setting by creating pseudo-domains by clustering patients within the source domain (A), thereby simulating a multi-source environment for training. For augmentation based methods as non-knowledge guided baselines, we select Mixup[19], Manifold Mixup[20], and Input Masking.

Implement details

Data preprocessing techniques, including standardization, one-hot encoding, and imputation, were all implemented using the scikit-learn package. The model architecture employed was a Multi-Layer Perceptron (MLP). Considering the input size, the network was configured with two hidden layers consisting of 32 nodes and 16 nodes, respectively, with the ReLU activation function applied to the output of each node. Training was carried out for 2,000 epochs using the ADAM[26] optimizer with a fixed learning rate of 0.0001. Consistent with our evaluation strategy, the final performance was reported using the best model, which was selected based on the most balanced AUROC achieved across all training epochs.

In the Similarity-Guided Mixup, clinical similarity between patients was defined using features identified as critical in the widely used Pediatric Estimated Time to KRT Calculator. This web-based clinical prediction tool was developed and adapted from the Chronic Kidney Disease in CHIldren (CKiD) study, a large, prospective, multicenter cohort established in the United States and Canada[27]. The calculator predicts time to KRT based on key clinical parameters, including urine protein to creatinine ratio (UPCR), blood pressure, estimated glomerular filtration rate (eGFR), hemoglobin level, serum albumin, chloride, and bicarbonate concentrations[28].

For Group-based Masking, we established predefined feature groups based on clinical correlations and practice: ('Diastolic BP', 'Systolic BP') were grouped because they are measured concurrently, and ('ARB', 'ACE inhibitor') were grouped because they belong to the same RAAS-inhibitor drug class.

Results

Table 3. Recall Performance Comparison in the Single-Source Domain Generalization Setting. We present the center-wise performance across target domains (B, C and D) and the overall mean ($\pm$ standard deviation) across all centers, demonstrating that our proposed method achieves the highest Recall performance compared to all baselines.

MLP	Method	B	C	D	Mean $\pm$ Std
W/o Augmentation	ERM	0.4074	0.0333	0.2258	0.2222 ± 0.0958
	IRM	0.3210	0.0667	0.1613	0.1830 ± 0.0958
Masking only	Input Masking	0.4074	0.1667	0.2581	0.2774 ± 0.2170
	Group-based Masking	0.4444	0.1667	0.4839	0.3650 ± 0.0958
Mixup only	Mixup	0.7284	0.5000	0.4516	0.5600 ± 0.1635
	Manifold Mixup	0.4198	0.2000	0.7097	0.4431 ± 0.2170
	Similarity-guided Mixup	0.7407	0.5333	0.4516	0.5752 ± 0.1209
Mixup + Masking	Mixup + Input Masking	0.8765	0.5000	0.8065	0.7277 ± 0.1209
	Manifold Mixup + Masking	0.4321	0.2000	0.7097	0.4473 ± 0.0958
	Ours	**0.9259**	**0.5667**	**0.8710**	$\mathbf{0.7879 \pm 0.1473}$

Overall Performance Comparison. Table 3 presents the performance results in terms of Recall. We observe that the proposed method significantly improves performance. Our method achieved mean Recall of 0.7879, outperforming the best baseline (Mixup + Masking) by a substantial margin of 6.20 percentage points. This indicates that our approach is particularly effective at identifying patients at risk of condition deterioration across unseen domains.

Ablation of Proposed Components Table 3 also presents an ablation study on each component of our proposed method. Relative to the ERM, Group-based Masking achieved a greater performance improvement than the conventional Input Masking technique, and Similarity-guided Mixup also demonstrated a larger gain compared to the conventional Mixup method. Furthermore, our combined method achieved the highest performance improvement in the Mixup + Masking combination, confirming that each component contributes to performance enhancement and exhibits a significant synergistic effect.

Feature Space Analysis. To qualitatively analyze how our proposed method improves the model's feature representation, we visualized the latent space of the test set using t-SNE[29] in Figure 3. The t-SNE visualizations demonstrate that our proposed framework learns a more structured and discriminative feature space compared to the baselines. Compared to baseline methods like ERM, our method's True Positives (blue dots) and False Negatives (red 'x' marks) are observed to be relatively better clustered. Most critically, our approach achieves a dramatic reduction in False Negatives , with only 24 misclassified positive cases—a considerable improvement over the 93 observed in ERM and 54 in Mixup. This reduction confirms that the Dual Knowledge-Guided Augmentation framework successfully forces the model to learn a robust decision boundary that accurately captures the inherent structure of the data, thereby maximizing the identification of high-risk patients across unseen target domains.

Center-wise analysis. When examining the center-specific recall scores in Figure 4, we observed that performance generally rose across all hospitals, in line with the overall performance improvement. Notably, C achieved a substantial gain when Mixup

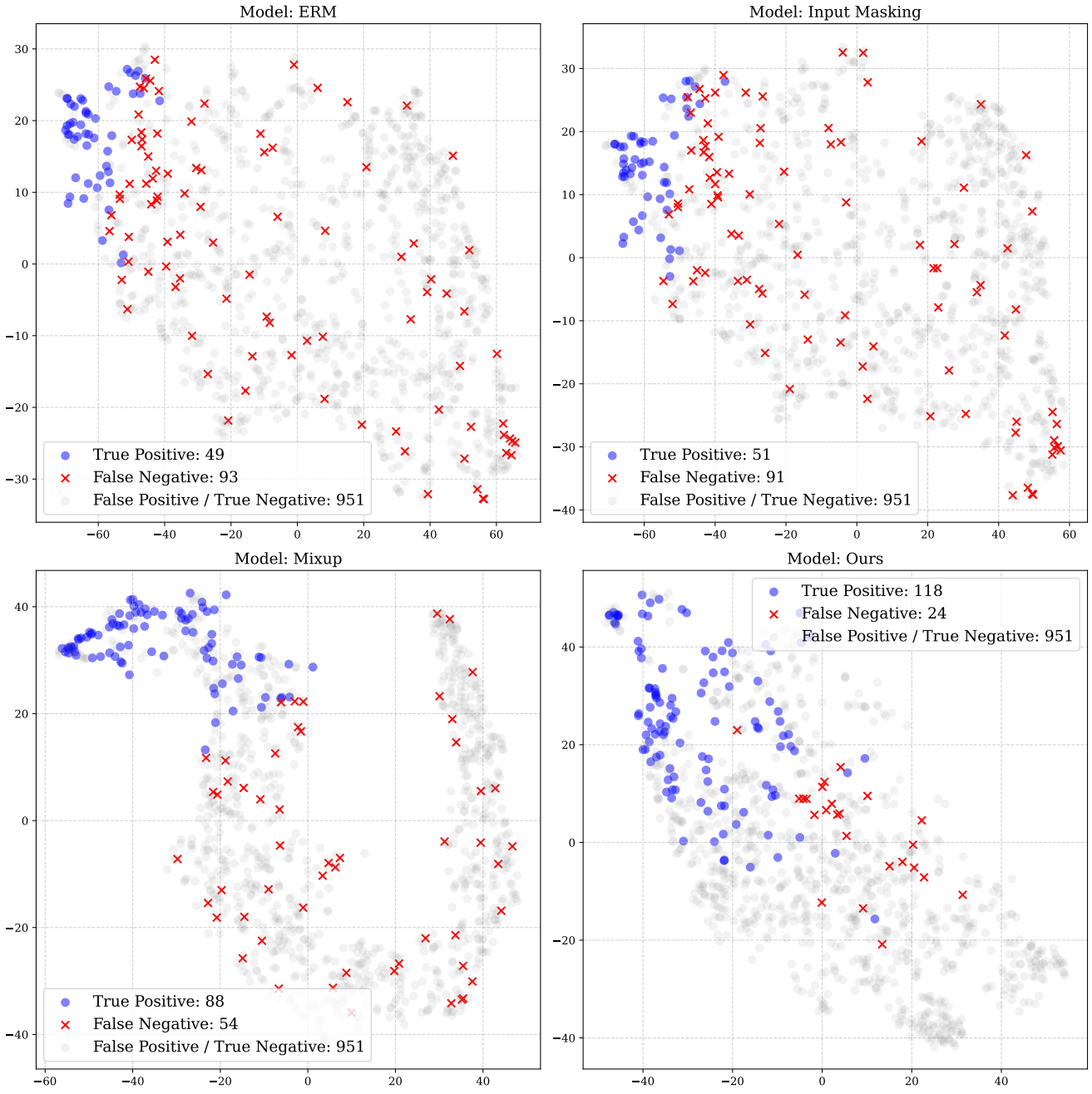

Figure 3. Feature Space Analysis via t-SNE Visualization. The plot visualizes the effect of each method by showing True Positives (TP) and False Negatives (FN) against the backdrop of the negative class (faded for clarity). Our method yields a highly structured feature space with robust separation between positive and negative classes, resulting in a decision boundary that is appropriately positioned to effectively maximize True Positive identification.

was incorporated, suggesting that Mixup's data augmentation, by increasing the number of samples in the minority class, was a contributing factor.

Discussion

This study addressed the challenge of domain generalization in clinical prediction, where a model trained on data from a single source must maintain robust performance on unseen target domains. To this end, we proposed two novel knowledge-guided data augmentation techniques, Similarity-guided Mixup and Group-based Masking, that explicitly incorporate clinical context

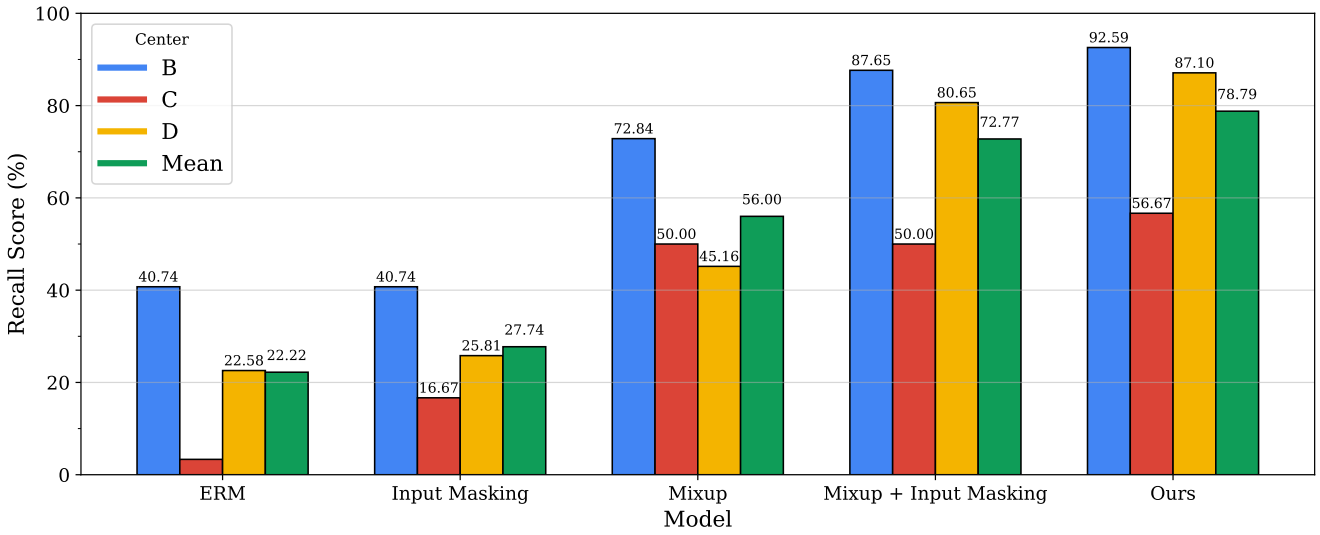

Figure 4. Center-wise Recall Analysis Demonstrating Robustness to Domain Shift. The bar plot presents the recall score for each method across the individual target domains (B, C, and D), alongside the overall mean recall score. This center-wise breakdown highlights the robustness and generalization capability of our proposed method, which achieves the highest mean recall (78.79%) and consistently high performance across all unseen target institutions

into the learning process. Experimental results demonstrated that these methods substantially outperformed existing baselines, including ERM, IRM, Mixup, and Input Masking, particularly in terms of recall, thereby validating its efficacy of our approach.

The key factor underlying the success of our framework was the integration of clinical knowledge into the data augmentation process. Conventional Mixup techniques often generate unrealistic noise and degrade performance by indiscriminately interpolating data from clinically disparate patients. In contrast, our Similarity-guided Mixup selectively interpolated data between clinically similar patients, producing plausible and meaningful augmented samples. This approach functioned as an effective regularization mechanism, enhancing data diversity while preserving the intrinsic structure of the original data distribution.

The Group-based Masking method further strengthened model robustness by mimicking realistic missing data patterns commonly observed in clinical practice (*e.g.*, the concurrent absence of related clinical values). Unlike conventional masking methods that apply dropout uniformly across features, our technique encouraged the model to learn stable representations under authentic data imperfections, improving its ability to handle missingness in real-world target domains.

The comparatively limited performance of baseline models, particularly IRM, can be explained by their inability to capture the complex inter-institutional variations that characterize real-world clinical data. Pseudo-domains generated from within a single dataset failed to reproduce the heterogeneity in patient demographics, measurement equipment, and data recording protocols observed across institutions. This finding underscores that a central challenge in domain generalization is not merely to enforce invariance but to simulate the true diversity between domains.

This study makes several notable contributions. First, it empirically highlights the importance of embedding domain knowledge into the medical AI models, particularly for tabular data where standard augmentation techniques often underperform. Second, it presents a concrete and practical methodology for building generalizable models from single-source data, a setting that reflects the data accessibility constraints commonly encountered in health care AI research. Together, these contributions demonstrate that medical knowledge-driven regularization can meaningfully enhance both model robustness and clinical interpretability.

Several limitations should be acknowledged. First, our validation was confined to a specific pediatric chronic kidney disease dataset (KNOW-pedCKD). Future studies should assess whether the proposed methods are generalized to other disease contexts or adult populations. Second, the identification of 'clinically significant features' and their grouping relied on expert knowledge, which may introduce subjectivity. Future work could incorporate data-driven approaches for automated feature discovery. Third, our analysis utilized only a single source domain because of limited sample sizes from other institutions, leaving the impact of varying source domains unexamined. Finally, we emphasized that the primary contribution of this study lies in the proposal of a novel knowledge-guided regularization framework, rather than a definitive solution to the single-source domain generalization (SSDG) problem itself. We utilized the challenging SSDG setting as a rigorous testbed to demonstrate our method's effectiveness in enhancing model robustness. Future research could investigate the integration of our method with

other dedicated domain generalization techniques to further address the core SSDG challenge.

Future research will focus on validating the generalizability of the proposed framework across diverse clinical datasets, automating the identification of clinically significant features, and extending the methodology to Multi-Source Domain Generalization settings. By integrating domain knowledge with advanced augmentation techniques, this study demonstrates a promising pathway toward developing medical AI models that are both more robust and more trustworthy across heterogeneous clinical environments.

References

1. Rajkomar, A., Oren, E., Chen, K. *et al.* Scalable and accurate deep learning with electronic health records. *npj Digit. Medicine* (2018).
2. Lakhani, P. & Sundaram, B. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* (2017).
3. Sutton, R. T. *et al.* An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digit. Medicine* (2020).
4. Ng, D. K. *et al.* Incidence of initial renal replacement therapy over the course of kidney disease in children. *Am. J. Epidemiol.* (2019).
5. Xie, Y. *et al.* Analysis of the global burden of disease study highlights the global, regional, and national trends of chronic kidney disease epidemiology from 1990 to 2016. *Kidney Int.* (2018).
6. Islam, M. A., Majumder, M. Z. H. & Hussein, M. A. Chronic kidney disease prediction based on machine learning algorithms. *J. Pathol. Informatics* (2023).
7. Ghosh, S. K. & Khandoker, A. H. Investigation on explainable machine learning models to predict chronic kidney diseases. *Sci. Reports* (2024).
8. Moon, S. *et al.* Development of a prediction tool for kidney function decline in children with chronic kidney disease. *Kidney Res. Clin. Pract.* (2025).
9. Kanakasabapathy, M. K., Thirumalaraju, P., Kandula, H. *et al.* Adaptive adversarial neural networks for the analysis of lossy and domain-shifted datasets of medical images. *Nat. Biomed. Eng.* (2021).
10. Guo, L. L., Pfohl, S. R., Fries, J. *et al.* Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Sci. Reports* (2022).
11. Harada, R., Hamasaki, Y., Okuda, Y., Hamada, R. & Ishikura, K. Epidemiology of pediatric chronic kidney disease/kidney failure: learning from registries and cohort studies. *Pediatr. Nephrol.* (2022).
12. Xu, Q. *et al.* Simde: A simple domain expansion approach for single-source domain generalization. In *CVPR Workshops* (2023).
13. Cugu, I., Mancini, M., Chen, Y. & Akata, Z. Attention consistency on visual corruptions for single-source domain generalization. In *CVPR Workshops* (2022).
14. Xu, Q., Zhang, R., Zhang, Y., Wang, Y. & Tian, Q. A fourier-based framework for domain generalization. In *CVPR* (2021).
15. Zhao, H. *et al.* Morestyle: relax low-frequency constraint of fourier-based image reconstruction in generalizable medical image segmentation. In *MICCAI* (2024).
16. Liu, C., Cao, Y., Su, X. & Zhu, H. Universal frequency domain perturbation for single-source domain generalization. In *Proceedings of the 32nd ACM International Conference on Multimedia* (2024).
17. Huang, J., Guan, D., Xiao, A. & Lu, S. Fsdr: Frequency space domain randomization for domain generalization. In *CVPR* (2021).
18. Zhou, K., Yang, Y., Qiao, Y. & Xiang, T. Domain generalization with mixstyle. *ICLR* (2021).
19. Zhang, H., Cisse, M., Dauphin, Y. N., Lopez-Paz, D. *et al.* mixup: Beyond empirical risk minimization. In *ICLR* (2018).
20. Verma, V. *et al.* Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning* (2019).
21. Yun, S. *et al.* Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV* (2019).
22. Uddin, A. F. M., Monira, M., Shin, W., Chung, T. & Bae, S. H. Saliencymix: A saliency guided data augmentation strategy for better regularization. *ICLR* (2021).

23. Qin, J. *et al.* Resizemix: Mixing data with preserved object information and true labels. *arXiv preprint arXiv:2012.11101* (2020).

24. Balasubramanian, S. & Feizi, S. Towards improved input masking for convolutional neural networks. In *ICCV* (2023).

25. Arjovsky, M., Bottou, L., Gulrajani, I. & Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

26. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *Int. Conf. on Learn. Represent. (ICLR)* (2015).

27. Wong, C. J., Moxey-Mims, M., Jerry-Fluker, J., Warady, B. A. & Furth, S. L. Ckid (ckd in children) prospective cohort study: a review of current findings. *Am. J. Kidney Dis.* (2012).

28. Ng, D. K. *et al.* Development of an adaptive clinical web-based prediction tool for kidney replacement therapy in children with chronic kidney disease. *Kidney Int.* (2023).

29. Maaten, L. V. D. & Hinton, G. Visualizing data using t-sne. *J. machine learning research* (2008).

Acknowledgements

We acknowledge the help of the following individuals at the Medical Research Collaborating Center (Biomedical Research Institute of Seoul National University Hospital): Heejung Ahn, Sungkyung Kim (data management), Jayoun Kim, and Nanhee Park (biostatistics).

Author contributions statement

S.M. and E.P. conceptualized the study. S.M. conducted the methodology and wrote the original draft. E.P. reviewed and edited the manuscript and supervised the work. All authors contributed to data curation, investigation, and resources. All authors read and approved the final manuscript.

Additional information

Ethical consent

This study was reviewed and approved by the Institutional Review Board of Seoul National University Hospital (No. H-1906-068-1041).

Informed consent

Informed consent was obtained from participants or their guardians, as appropriate, based on their ages.

Competing interests

No potential conflict of interest relevant to this article was reported. This work was supported by the research program of the National Institute of Health (NIH) under research project No. 2025E110100. Additional support was provided by a grant from Korea University Medicine (No. K2427191).