Dual Knowledge-Guided Data Augmentation for Robust Clinical Prediction Models

Sangwoo Moon[1] **and Eujin Park**[2,*]

[1]Department of Computer Science and Engineering, College of Engineering, Seoul National University, Seoul, 08826, Republic of Korea
[2]Department of Pediatrics, Korea University Guro Hospital, Seoul, 08300, Republic of Korea
[*]eujinpark@korea.ac.kr

ABSTRACT

ABSTRACT

Please note: Abbreviations should be introduced at the first mention in the main text – no abbreviations lists. Suggested structure of main text (not enforced) is provided below.

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 a core technology.[1–3]. These systems play a crucial role in enhancing patient safety and treatment efficacy by helping physicians' diagnoses and reducing potential errors during the therapeutic process.

At the heart of this technological progress are predictive models that learn meaningful patterns from a given set of training data. Models are actively being developed to predict disease risk, simulate progression, and forecast treatment responses using biological signals, test results, imaging, and genetic data. Among these, the role of AI is especially critical in the management of chronic diseases such as Chronic Kidney Disease (CKD). CKD has a high global prevalence and imposes a significant burden on patients and society, as its progression to end-stage renal disease requires high-cost treatments such as dialysis or kidney transplantation. Therefore, early prediction of disease progression and timely intervention to slow its progression are of paramount importance, and AI-based predictive models can serve as powerful tools to this end[4–6].

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[7]. To address this, research is actively focusing on Domain Generalization[8] and related fields such as robust learning. Domain shift refers to the phenomenon where an AI model, trained on data from a specific hospital (the source domain), experiences a significant degradation in predictive performance when applied to data from a different hospital (the target domain). This is a critical issue that undermines the robustness and trustworthiness of AI models, representing the largest technical hurdle to the widespread adoption of medical AI technology.

The reasons why the domain shift problem is particularly severe in medical data are multifaceted.

Heterogeneity of Patient Populations. Each hospital has a distinct distribution of demographic characteristics (age, sex, ethnicity), socioeconomic background, and underlying comorbidities in its primary patient population. For example, tertiary referral hospitals have a higher proportion of patients with severe and rare diseases, whereas primary care clinics show a data distribution centered on patients with mild and chronic conditions.

Variations in Clinical Practices. For the same disease, subtle differences exist between hospitals—and even between individual physicians—in diagnostic criteria, preferred types of tests, prescription patterns, and data recording practices. These variations cause shifts in the data's statistical distribution and can lead to the model overfitting to the specific clinical protocols of a hospital.

Differences in Measurement Environments. Discrepancies in the technical infrastructure used by hospitals, such as blood analysis equipment, imaging devices (CT, MRI), and EHR solutions, also introduce subtle variations in the data. The model is at risk of learning these device- or system-specific artifacts rather than the biological patterns of the disease itself.

Due to these factors, the model tends to learn superficial, domain-specific correlations—so-called "shortcuts"—that exist only in the data of a particular hospital, rather than learning the essential generalizable features. For example, a medication frequently used only at a specific hospital or the prescribing habits of a particular physician could be incorrectly learned as a

significant predictor of the disease. When such a model is applied to another hospital where these shortcuts no longer exist, a sharp decline in performance is an inevitable consequence.

To address this fundamental challenge of shortcut learning and enhance model robustness, this study moves beyond conventional regularization techniques. While methods like standard Mixup[9] and Dropout[10] are effective, they operate on the principle of random perturbation and are inherently domain-agnostic, failing to leverage the rich structural information and expert knowledge embedded within clinical data.

We begin with the hypothesis that systematically integrating the Clinical Knowledge of medical experts into the data augmentation process can serve as a powerful, domain-aware regularizer. Moving beyond the conventional "black-box" approaches, we aim to suppress shortcut learning and guide the model toward learning the essential patterns of disease by directly injecting domain expertise into the model training process.

To this end, this study proposes a Dual Knowledge-Guided Data Augmentation Framework composed of the following two core strategies:

Similarity-guided Mixup. This first strategy aims to generate more plausible virtual patient data. Unlike the conventional random Mixup[9] method, this study selectively interpolates clinically similar patients. For instance, the probability of a data pair being selected for Mixup is increased if their key indicators of kidney disease, such as protein-to-creatinine ratio (UPCR) or estimated glomerular filtration rate (eGFR), are similar. This approach generates clinically plausible data, helping the model learn smoother and more robust decision boundaries.

Group-based Masking. The second strategy acts as a structured regularizer to prevent the model from overfitting to spurious features. Through discussions with medical experts, we identify and group features that are clinically related or likely to be missing together. The training process then involves stochastically masking these entire feature groups. This process prevents the model from becoming overly reliant on any single feature and compels it to learn alternative reasoning paths based on the remaining biological relationships, thereby enhancing its robustness.

To rigorously evaluate the efficacy of our proposed framework as a regularization method, we tested it in one of the most challenging scenarios: Single-Source Domain Generalization (SSDG). Using only the EHR data from a single institution in the KNOW-pedCKD[11] cohort (Seoul National University Hospital), we trained a predictive model based on our framework. We then validated its generalization performance on data8 from three external institutions that were never used for training (Asan Medical Center, Kyungpook National University Hospital, and Severance Hospital).

Experimental results show that our framework not only slightly improves the overall predictive accuracy (AUROC) but also significantly enhances the most clinically crucial metric, Recall—the ability to correctly identify patients who actually kidney function decline. This outcome validates that our knowledge-guided regularization strategy effectively improves a model's ability to generalize under severe domain shifts.

In conclusion, this research presents a novel regularization strategy that systematically injects clinical knowledge into the data augmentation process. This approach offers a pathway to develop AI models that can operate reliably across different medical environments using data from only a single institution. This will serve as a vital cornerstone for enabling medical AI technology to move beyond the laboratory and achieve successful implementation in real-world clinical practice, surmounting the barriers of data sharing.

Related Works

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[9] 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[19] applies interpolation within the hidden layers.

Another popular approach, CutMix[20], 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[21] leverages saliency detection to ensure the cropped region is informative, while ResizeMix[22] simplifies the process by resizing a source image into a small patch.

However, these methods has a fundamental problem: it combines patient data randomly without any consideration for the clinical context[9,19] or leverage characteristics of vision domain[20–22]. 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[23], 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.

Problem setting

The goal of domain generalization is to train a model on a source domain that can generalize to an unseen target domain. We formally define the setting as follows. We denote the source domain dataset as $\mathscr{D}_s$ and the target domain dataset as $\mathscr{D}_t$. The core challenge arises because the source and target domains, $\mathscr{D}_s$ and $\mathscr{D}_t$, are drawn 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 $\mathscr{D}_t$.

Similarity-Guided Mixup

A prior study, CMixup[24], improved upon the original Mixup by leveraging label information to encourage interpolation between similar classes. However, its methodology was specifically tailored for regression tasks, limiting its applicability to classification problems. Our proposed method overcomes this limitation by defining sample similarity based on a clinically meaningful feature set, making it effective for classification tasks. We propose Similarity-guided Mixup, a method that first identifies a subset of clinically significant features based on prior domain knowledge. The Mixup augmentation is then performed exclusively between patient samples that exhibit high similarity within this predefined 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 $S(\mathbf{x}_i, \mathbf{x}_j)$:

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

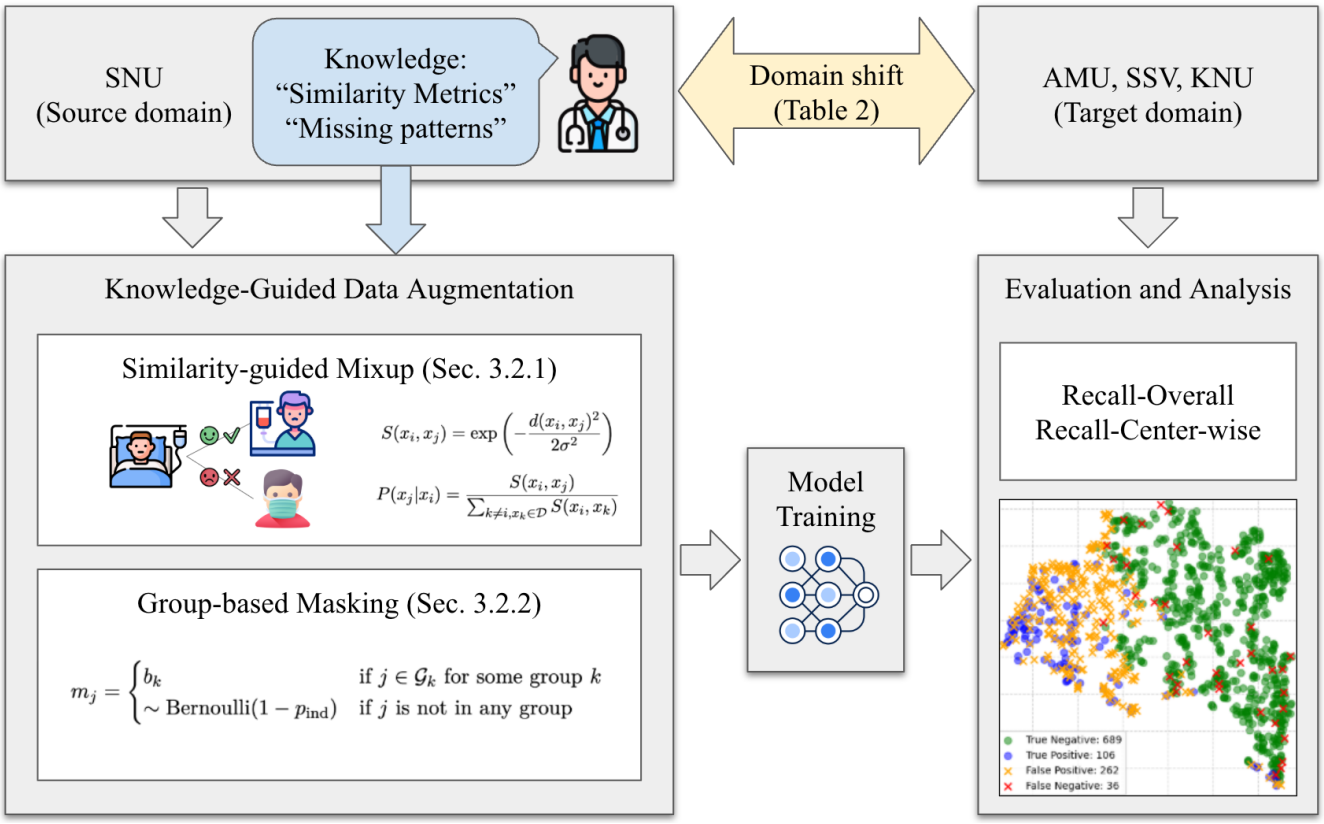

Figure 1. Overview of our framework. We presented a benchmark for evaluating model generalization performance under domain shift conditions. In this setting, we evaluated and analyzed the performance of models trained using clinical knowledge-guided data augmentation.

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

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 $\mathscr{G} = \{\mathscr{G}_1, \mathscr{G}_2, \dots\}$.

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 $\{\mathscr{G}_1, \dots, \mathscr{G}_K\}$ defined by clinical knowledge. For each group G_k, a single binary variable $b_k \sim \text{Bernoulli}(1 -$

p_{group}) is sampled.

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

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

Table 1. Demographics and clinical characteristics

Characteristic	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%)

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.

The Seoul National University Hospital (SNUH) dataset, which had the largest number of patients, was selected as the source domain ($\mathcal{D}_s$). The next three largest hospitals—Asan Medical Center (AMC), Kyungpook National University Hospital (KNUH), and Severance Hospital (SSV)—were selected as the target domains ($\mathcal{D}_t$) to construct a single-source domain generalization scenario. Each data record was converted to a person-visit pair, following previous work[6]. The 18 features used for training and the statistical characteristics for each hospital are available in Table 1. Furthermore, the results of the t-tests assessing the distribution shift from the source domain to the target domains are shown in Table 2. Considering the p-values in Table 2, a significant distribution difference was observed between the source and target domains for all features except for Sex.

The primary metric used in this experiment was Recall. In this study, recall measures how accurately the model predicts patients whose condition actually deteriorates[25–27]. For selecting the evaluation model during training, the source dataset was

Table 2. T-test p-values of source domain (A) to target domains. T-test p-values are reported for continuous variables, while 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

split into train (80%) and validation (20%). The model was trained on the training set, and the model achieving the highest Area Under the Receiver Operating Characteristic curve (AUROC) performance on the validation set was selected for target domain evaluation. We report both the performance measured across all combined datasets (Overall) and the average performance across each center (Center-wise).

Baselines

We compared the performance of our proposed method against several baselines: For non-augmetnation-based baseline, Empirical Risk Minimization (ERM), Invariant Risk Minimization (IRM[28]). 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 SNUH source data, thereby simulating a multi-source environment for training. For augmentation based methods as non-knowledge guided baselines, we select Mixup[9], Manifold Mixup[19], and Input Masking.

Implement details

All experiments were conducted using Multi-Layer Perceptron (MLP) with single hidden layer. For Similarity-guided Mixup, we defined clinical similarity using features identified as critical in the widely-used Pediatric Estimated Time to Kidney Replacement Therapy (KRT) Calculator. This clinically meaningful feature set included [UPCR, eGFR, Hemoglobin, Systolic BP, Diastolic BP].

For Group-based Masking, we established predefined feature groups based on pairs of features that are often concurrently missing in clinical practice. The groups utilized were [("Diastolic BP", "Systolic BP"), ("Diastolic BP", "UPCR"), and ("Systolic BP", "UPCR")].

Results

Overall Performance Comparison. Table 3 presents the performance results. We observe that the proposed method significantly improves performance in identifying positive cases across domain shifts. Our method achieved a mean Recall of **0.7879**, substantially outperforming the best baseline (Mixup + Masking, 0.7277) by 6.02 percentage points (pp). This indicates that our approach is particularly effective at minimizing False Negatives (FNs)—i.e., accurately identifying patients truly at risk of kidney function failure across unseen domains. Consistent performance gains were also confirmed across

Table 3. Performance comparison on the KNOW-pedCKD dataset under the single-source domain generalization setting in terms of Recall. Center-wise performance across individual centers (B, C, D) is shown, alongside the mean and standard deviation across all centers.

MLP	Method	B	C	D	Mean ± 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	**0.7879 ± 0.1473**

all individual centers (B, C, and D), ranging from a minimum of 4.94 to a maximum of 6.67 percentage points (pp). This consistency across different centers highlights the robustness of our approach in handling diverse domain shifts in a clinically critical metric like Recall.

Ablation of Proposed Components able 3 also presents an ablation study on each component of our proposed method. While the existing conventional methods (Input Masking and Mixup) demonstrated performance improvements due to their inherent regularization effects, Group-based Masking and Similarity-guided Mixup achieved further gains compared to their respective base methods (8.76, 1.52 percentage points). Furthermore, a synergistic effect, consistent with that observed in combinations of conventional methods, is evident within the Mixup + Masking categories. This confirms that each component contributes independently to the overall performance enhancement and that their combination yields a significant synergistic effect, validating the design of our proposed architecture.

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 2. In the case of the baseline ERM and Manifold Mixup, which yielded the lowest performance, the learned feature spaces are largely unstructured, with positive (blue, red) and negative (green) samples being heavily intermingled. Notably, the False Negatives (red 'x' markers) are scattered deep within the negative class cluster, visually explaining the model's poor Recall score of 0.3451. While Input Masking and Mixup show a slight trend towards separation, their feature spaces remain largely disorganized. In stark contrast, our proposed method (Ours) learns a highly structured and linearly separable feature space. The positive and negative classes form distinct clusters, and most importantly, the number of False Negatives is dramatically reduced. Furthermore, the few remaining False Negatives are primarily located near the decision boundary between the two clusters. This visualization demonstrates that our method does more than just improve a single metric; it learns a robust and meaningful feature space that better represents the underlying structure of the data. This well-separated representation is the key reason for achieving a significantly higher Recall score of 0.7465.

Center-wise analysis. When examining the center-specific recall scores 3, we observed that performance generally rose across all hospitals, in line with the overall performance improvement. Notably, C achieved a substantial gain when Mixup was incorporated, suggesti ng that Mixup's data augmentation, by increasing the number of samples in the minority class, was a contributing factor.

Arhcitecture Robustness. To validate the robustness of our proposed method across different model architectures, we conducted additional experiments using TabPFN[30], a state-of-the-art foundation model for tabular data. The results, presented in Table 4, demonstrate that our method consistently outperforms all baselines in terms of Recall, achieving an Overall Recall of 0.7042 and a Center-wise Recall of 0.6514. This further confirms the effectiveness and generalizability of our approach across different model architectures.

Discussion

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 on unseen target domains. We proposed two novel data augmentation techniques that leverage prior clinical knowledge: Similarity-guided Mixup and Knowledge-guided Input Masking. Experimental results

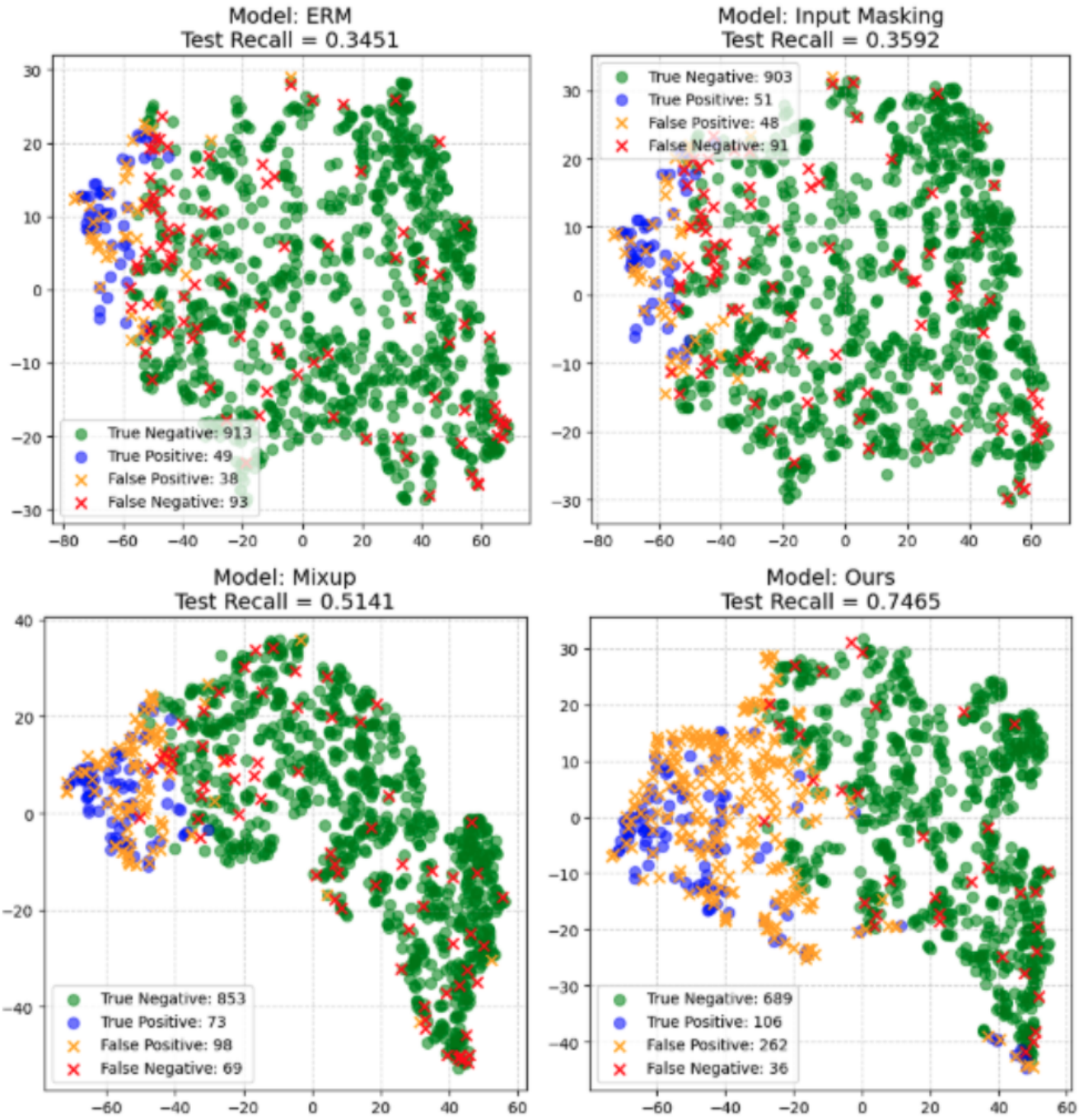

Figure 2. Feature Space Analysis via t-SNE Visualization. Our method yielded a structured feature space with robust separation between positive and negative classes, and the resulting decision boundary was appropriately positioned to effectively identify True Positives (TPs).

demonstrated that our proposed method significantly outperformed existing baselines, including ERM, IRM, Mixup, and Input Masking, particularly in terms of recall, thereby validating its efficacy.

The primary factor contributing to the success of our method was the active integration of clinical context into the data augmentation process. Whereas conventional Mixup generated unrealistic noise and degraded performance by indiscriminately interpolating data from clinically disparate patients, our Similarity-guided Mixup created more plausible and meaningful augmented samples. It achieved this by selectively mixing data only from similar patient groups within a clinically significant feature space. This suggests that our approach acted as an effective regularization strategy, enhancing data diversity without

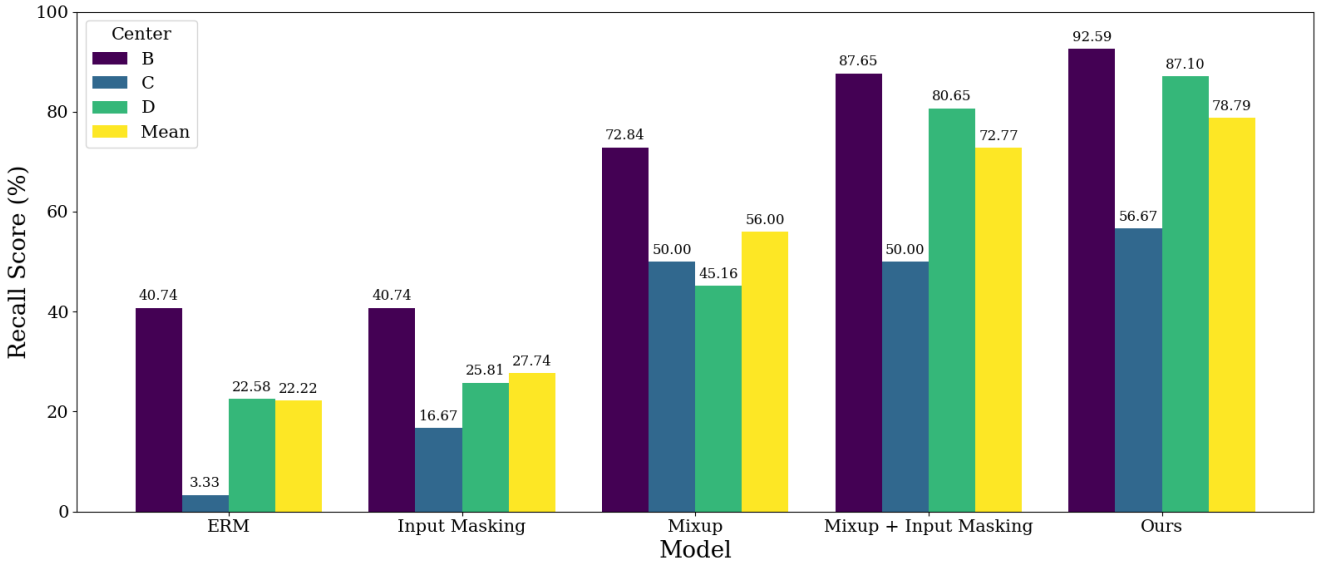

Figure 3. Center-wise recall analysis we present center-wise recall score.

corrupting the original data distribution.

Similarly, our Knowledge-guided Input Masking technique forced the model to learn from more realistic imperfections by simulating authentic missing data patterns observed in clinical practice (e.g., the concurrent absence of related lab values). This method proved more effective at improving model robustness than conventional masking, which applies dropout uniformly across all features. It better prepared the model for the various missing data scenarios it would encounter in real-world target domains.

The limited performance of the baseline models can also be clearly interpreted. For IRM, the strategy of creating pseudo-domains based on patient information within a single source failed to adequately capture the complex domain shifts that exist between different institutions. These shifts can be attributed to a variety of factors, including differences in measurement equipment, patient demographics, and data recording protocols. This finding underscores that a central challenge in domain generalization is the ability to realistically simulate the true disparities between domains.

This study offers several significant contributions. First, it empirically demonstrates the critical importance of integrating domain knowledge in the medical field, especially for tabular data where standard augmentation techniques often fall short. Second, it presents a concrete methodology for building a highly generalizable model using only single-source data, offering a practical solution to the challenges of limited data access in medical AI research.

Nevertheless, this study has several limitations.First, our validation was confined to a specific pediatric chronic kidney disease dataset (KNOW-pedCKD). Further research is required to determine if the proposed methods yield similar benefits in other disease domains or with adult patient data.Second, the process of selecting "clinically significant features" and defining feature groups relied on expert knowledge, which may introduce subjectivity. Future work could explore data-driven methods for automatically identifying these crucial features and groupings.Third, this study was conducted using only a single source domain (SNUH). This was due to the insufficient sample sizes of the other hospital datasets, which made them unsuitable for use as independent source domains for training. Consequently, the impact of varying the source domain remains unexplored.

Finally, it is important to clarify that this study's primary contribution is the proposal of a novel, knowledge-guided regularization framework rather than a direct 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 our methodology by applying it to a broader range of clinical datasets and on developing techniques to automate the feature selection process. We also plan to extend this work to a Multi-Source Domain Generalization setting to investigate effective ways of fusing domain knowledge from multiple institutions. In conclusion, this study demonstrates that combining clinical knowledge with advanced data augmentation is a promising direction for dramatically improving the generalization performance of medical AI models.

Table 4. Performance comparison on the KNOW-pedCKD dataset under the SSDG setting (TabPFN Model). Overall performance is measured on the aggregated (unified) target datasets, whereas Center-wise performance represents the calculated statistics (e.g., mean and standard deviation) across each individual center. The row with the highest recall is highlighted in bold.

TabPFN	Method	Recall		AUROC	
		Overall	Center-wise	Overall	Center-wise
W/o Aug.	ERM	0.5070	0.4603 ± 0.0989	0.8161	0.8068 ± 0.0120
	IRM	0.5141	0.4644 ± 0.1044	0.8166	0.8079 ± 0.0123
Masking only	Input Masking	0.5352	0.4837 ± 0.1128	0.8173	0.8091 ± 0.0113
	Group-based Masking	0.5423	0.4878 ± 0.1182	0.8177	0.8097 ± 0.0108
Mixup only	Mixup	0.5775	0.5290 ± 0.1142	0.8152	0.8059 ± 0.0115
	Manifold Mixup	0.5775	0.5220 ± 0.1202	0.8137	0.8036 ± 0.0148
	Similarity-guided Mixup	0.6338	0.5756 ± 0.1314	0.8145	0.8047 ± 0.0136
Mixup + Masking	Mixup + Masking	0.6690	0.6172 ± 0.1456	0.8178	0.8098 ± 0.0096
	Manifold Mixup + Masking	0.5775	0.5220 ± 0.1202	0.8137	0.8036 ± 0.0148
	Ours	**0.7042**	$\mathbf{0.6514 \pm 0.1473}$	**0.8184**	$\mathbf{0.8088 \pm 0.0064}$

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. Islam, M. A., Majumder, M. Z. H. & Hussein, M. A. Chronic kidney disease prediction based on machine learning algorithms. *J. Pathol. Informatics* (2023).
5. Ghosh, S. K. & Khandoker, A. H. Investigation on explainable machine learning models to predict chronic kidney diseases. *Sci. Reports* (2024).
6. Moon, S. *et al.* Development of a prediction tool for kidney function decline in children with chronic kidney disease. *Kidney Res. Clin. Pract.* (2025).
7. 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).
8. 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).
9. Zhang, H., Cisse, M., Dauphin, Y. N., Lopez-Paz, D. *et al.* mixup: Beyond empirical risk minimization. In *ICLR* (2018).
10. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* (2014).
11. Kang, H. G. *et al.* Know-ped ckd (korean cohort study for outcomes in patients with pediatric ckd): design and methods. *BMC nephrology* (2016).
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. Verma, V. *et al.* Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning* (2019).

20. Yun, S. *et al.* Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV* (2019).

21. 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).

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

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

24. Yao, H., Wang, Y., Zhang, L., Zou, J. Y. & Finn, C. C-mixup: Improving generalization in regression. In *Advances in neural information processing systems* (2022).

25. Haque, M. E. *et al.* Improving chronic kidney disease detection efficiency: Fine tuned catboost and nature-inspired algorithms with explainable ai. In *CSNT* (2025).

26. Priyadharshini, M. *et al.* A population based optimization of convolutional neural networks for chronic kidney disease prediction. *Sci. Reports* (2025).

27. Supriana, I. W., Pramartha, C., Putra, I. G., Raharja, M. & Wiguna, P. Modified k-nearest neighbor optimization with genetic algorithm in chronic kidney disease classification. In *ICAMSAC 2023* (2024).

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

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

30. Hollmann, N. *et al.* Accurate predictions on small data with a tabular foundation model. *Nature* (2025).

For data citations of datasets uploaded to e.g. *figshare*, please use the `howpublished` option in the bib entry to specify the platform and the link, as in the `Hao:gidmaps:2014` example in the sample bibliography file.

Acknowledgements (not compulsory)

Acknowledgements should be brief, and should not include thanks to anonymous referees and editors, or effusive comments. Grant or contribution numbers may be acknowledged.

Author contributions statement

Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Additional information

The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.