Experience with Single Domain Generalization in Real World Medical Imaging Deployments

Ayan Banerjee[1], **Komandoor Srivathsan**[2], **Sandeep K.S. Gupta**[1]

[1]Arizona State University,
[2]Mayo Clinic
{abanerj3, sandeep.gupta}@asu.edu, srivathsan.komandoor@mayo.edu

Abstract

A desirable property of any deployed artificial intelligence is generalization across domains. In medical imagining applications a coveted property for effective deployment is Single Domain Generalization (SDG),i.e., a model trained on a single domain should generalize well to unseen target domains. In multi-center studies, differences in scanners and imaging protocols introduce domain shifts that exacerbate variability in rare class characteristics. This paper presents our experience on SDG in real life deployment for two exemplary medical imaging case studies on seizure onset zone detection using fMRI data, and stress electrocardiogram based coronary artery detection. Utilizing the commonly used application of diabetic retinopathy, we first demonstrate that state-of-the-art SDG techniques fail to achieve generalized performance across data domains. We then develop a generic expert knowledge integrated deep learning technique DL+EKE and instantiate it for the DR application and show that DL+EKE outperforms SOTA SDG methods on DR. We then deploy instances of DL+EKE technique on the two real world examples of stress ECG and resting state (rs)-fMRI and discuss issues faced with SDG techniques.

Introduction

Domain generalization characterizes a model's capacity to maintain performance when exposed to data that diverges from the distribution it encountered during training (Stolte et al. 2023). This capability is especially vital in medical AI, where models must reliably interpret previously unseen patient cases originating from different clinical environments. In medical imaging, numerous patient and acquisition-specific variables, such as age, sex, scanner hardware, sedation protocols, and coexisting medical conditions, alter disease presentation across individuals (Bera and Biswas 2023)(Chen et al. 2023a)(Chen et al. 2023b)(Chen et al. 2023c)(Hu, Liao, and Xia 2023)(Parker et al. 2023). These sources of variation, collectively contributing to intraclass diversity (Che et al. 2023)(Liu et al. 2023), pose difficulties for deep learning (DL) systems, particularly when only limited examples of certain pathologies are available, leading to reduced sensitivity to acute abnormalities (Kim, Shin, and Hwang 2023). Such performance degradation frequently stems from dataset imbalance when models are often trained on overwhelmingly non-pathological cases, causing the rare but clinically essential pathological samples to be underrepresented. Although these rare instances contain the most informative signals, their scarcity limits the model's ability to learn robust representations. Consequently, identifying rare conditions necessitates specialized techniques that explicitly address their low prevalence, distinct visual markers, and nuanced patterns of positive and negative evidence, since purely data-driven statistical learning cannot fully capture the underlying distribution of rare classes from sparse observations (Abubakar et al. 2024). Rare class detection is especially important in scenarios like detection of stage 5 diabetes retinopathy (DR) or identifying the seizure onset zone (SOZ) from resting state (rs) fMRI (Che et al. 2023; Kamboj et al. 2025), or detecting coronary artery disease (CAD) from stress ECG and it remains a challenging task.

Researchers have proposed several strategies to mitigate domain generalization issues, including domain adaptation (DA) and multi-source domain generalization (MSDG), where training incorporates data originating from multiple external centers or source domains (Hu, Liao, and Xia 2023). Such approaches have demonstrated improved robustness when deep learning (DL) models are evaluated on datasets drawn from disparate clinical environments. Nevertheless, DA and MSDG frameworks typically require substantial effort to obtain data from those additional domains and often raise privacy concerns because they involve sharing or aggregating data across institutions and limit their real-world applicability in healthcare settings (Hu, Liao, and Xia 2023). Given these constraints, single-domain generalization (SDG) has emerged as a more practical alternative, focusing on achieving cross-domain robustness using only a single-source dataset (Vidit, Engilberge, and Salzmann 2023; Yan et al. 2023). Although SDG has been explored extensively for common image classification tasks, to the best of our knowledge, we find no existing work that tackles SDG specifically for rare-class scenarios, where the generalization problem is fundamentally more challenging.

Strategies enhancing SDG for rare categories can benefit from incorporating expert-derived insights, such as clinically recognized, domain-invariant cues that characterize disease signatures and help disentangle visually similar classes. These expert perspectives can also offer guidance on

handling intra-class variability. Yet, in clinical practice, such knowledge is often imprecise, subjective, and influenced by individual interpretation (Boerwinkle et al. 2017; Hunyadi et al. 2015). Moreover, medical datasets frequently contain measurement noise and acquisition artifacts, which may cause substantial overlap between disease presentations, further complicating the learning problem (Boerwinkle et al. 2017; Kamboj et al. 2024). As a result, many traditional knowledge-driven systems designed for healthcare applications exhibit considerable rates of false positives (FPs) and false negatives (FNs) (Banerjee et al. 2023; Galappaththige, Kuruppu, and Khan 2024; Kamboj et al. 2024; Kamboj, Banerjee, and Gupta 2024b; Nandakumar et al. 2023).

In this paper, we first share our motivating experience in deploying an AI technique for detection of CAD from stress ECG at Mayo Clinic that failed the SDG performance evaluation test (Banerjee et al. 2025). We then introduce a fundamentally different approach to SDG for rare classes. We utilize non-data driven, discriminative expert knowledge and integrate it with DL using pre-trained large vision model (LVM) and demonstrate superior SDG performance in rare class detection than SOTA on the commonly used SDG benchmark of diabetes retinopathy (DR). We then demonstrate the on-field performance of human expert integrated DL (DL+EKE) on two deplotment experiments on rs-fMRI based SOZ detection and CAD detection using stress ECG.

Our main contributions are: i) We introduce integration of domain invariant expert knowledge with DL as a viable solution for rare classes detection and SDG, using LVM and class-wise entropy, overcoming the challenge of limited and imbalanced data, as well as overlapping information gathered from expert knowledge about classes. ii) Demonstrate superior performance of DL+EKE on DR benchmarks , and iii) extensive deployment experiments on two example medical imaging applications (SOZ and CAD detection) and comparisons with SOTA show that DL models, when augmented with expert knowledge and designed to break information overlap, effectively detect rare classes from disparate sources, thereby demonstrating enhanced generalizability using single domain.

Related Work and Preliminaries

Domain Adaptation and Domain Generalization. Domain adaptation (DA) aligns a source distribution to a known target, while multi-source domain generalization (MSDG) uses multiple labeled sources to generalize to unseen targets. Both require either target or diverse source data, which is typically unavailable in medical settings (Hu, Liao, and Xia 2023). Single-source domain generalization (SDG) learns invariant features from one labeled domain (Vidit, Engilberge, and Salzmann 2023), but performs poorly under class imbalance and limited rare-class samples (Chen et al. 2023b; Kim, Shin, and Hwang 2023; Li et al. 2023; Yang et al. 2023). Image- and feature-level augmentations likewise underperform, as they sample from the same source distribution and fail to capture rare-class variability.

Knowledge and DL. Incorporating expert knowledge into DL has involved feature encoding (Banerjee et al. 2023) or domain-specific augmentations (Hu, Liao, and Xia 2023), but overlapping clinical concepts (e.g., similar power-spectra thresholds) limit discrimination (Boerwinkle et al. 2017). Physics-guided networks embed equations into the loss, yet falter when knowledge is vague. Knowledge-enhanced neural networks impose symbolic output constraints (Daniele and Serafini 2019), but cannot model intra-class variability and struggle with incomplete or inconsistent knowledge. Our method integrates knowledge with single-domain data to address rare-class imbalance and overlap without requiring target data.

Rare Class Properties "Rare classes are extremely infrequent classes whose characteristics make them or their consequences highly valuable. Such classes appear with extreme scarcity and are hard to predict, although they are expected eventually" (Sokolova et al. 2010). In this paper, we consider *rare class* to be a phenomenon which has four properties: a) **Discrimination**: observations of the phenomenon have distinctive characteristics than other observations, b) **Scarcity**: the phenomenon has less number of observations in the process, c) **Significance**: each observation of the phenomenon has much more information content than other observations of the process, d) **Overlap**: each observation of the phenomenon has several characteristic features that exhibit significant overlap or high similarity within the feature embedding space, independent of class labels. This means that the embedding space is not exposed to the specific classes of the domain in question.

Motivating Deployment Experience

The motivating problem was the development of an AI technique to detect CAD from a patient's 12-lead exercise stress ECG. The ground truth for the presence of obstructive CAD ($\geq$ 70% stenosis), is obtained using invasive coronary angiography (ICA) performed within 2 months of stress ECG. A visual transformer (ViT in Appendix (Banerjee 2026)) based architecture was trained with data from the year 2010 on 1200 patients (1000 train and 200 validation) with equal distribution of CAD positive and negative ICA outcomes. The ViT was then tested on an untouched data for 200 patients from the 2010 repository. The positive or negative predictive value (PPV or NPV) of this ViT model as shown in Table 1 are excellent on the test data with very little deviation from validation. However, when the ViT was tested on 92 patients from 2025, the PPV and NPV dropped drastically (Table 1).

Metric	Method	Validation (y2010)	Test (y2010)	Blind Test (y2025)	Reduction
PPV	ViT	80.4%	79.0%	46.0%	33%
NPV	ViT	83.0%	81.8%	49.0%	32.8%

Table 1: ViT performance under *single-domain* training: 1,000 patients; Validation: 200; Test: 200; Blind Test: 92.

Prior to 2012, if clinicians did not find S-T depression as a feature in the stress ECG image, then patients were not referred to ICA. However, stress ECG has other subtle positive CAD features manifested in the inter-relationship among time series from different leads which do not cause

any data distribution shift in individual leads and can be easily ignored by manual labeling. A change in triage policy in 2012 resulted in more patients being sent for ICA even if S-T depression was not found. Hence, in data from 2025, there were CAD positive stress ECG that had expert knowledge factors (inter lead relationship) affecting CAD diagnosis that were not present in data from 2010.

Methods

Our approach to solving the SDG rare class problem in Definition 1 (in Appendix) involves integrating expert knowledge with DL techniques. First, we consider the source domain Y_1 and isolate the rare class from the other available classes in the raw data, following the criteria outlined in Definition 2. For this purpose, we need a representation x_i of the observations $y_i \in Y_1$. For DL, we employ large vision model (LVM) CLIP to extract features from the raw data, independent of class information (Fig. 1) (Radford et al. 2021). For expert knowledge, we utilize symbolic AI based representations highlighted in Section "Expert Knowledge on rs-fMRI". This step is application dependent and requires specialized ML methods. We compute the similarity between the class agnostic feature embeddings of the rare class and each of the classes in C, referring the most similar class to the rare class as the overlap class $c_o = argmin(\theta_i)$. We use DL techniques to classify overlap class, and knowledge for rare class. We utilize a machine orchestration strategy to derive the best DL and knowledge based machine. This overall strategy is formulated in algorithm RareSaGe.

Machine orchestration strategy: We assume that there is a set of trained DL/ML classifiers $\mathcal{M}_{\mathcal{DL}}$ for the overlap and non-overlap classes' learning, and a set of trained knowledge based classifiers $\mathcal{M}_{\mathcal{K}}$ for the knowledge-based rare class and non-rare classes (normal classes) learning. Each classifier $M_d \in \mathcal{M}_{\mathcal{DL}}$ or $M_d \in \mathcal{M}_{\mathcal{K}}$, classifies instances of Y with a label from the set $S^{M_d} \subset 2^C$, such that each label in S^{M_d} meets the following rules: **a) Mutual exclusion:** no two labels in S^{M_d} share instance from the same original classes $c_i \in C$; **b) Class cover:** the union of class labels S^{M_d} includes all original classes; **c) Union rule:** the classes im S^{M_d} can be expressed as union of classes from C.

A classifier M_d represents the raw data $y_i \in Y$ in some latent representation space $x_i \in X$ using a discriminative feature function $\mathcal{F}_{M_d}$. The same function can be used to represent each raw data in the original class set C. We select the machine that reduces rare class entropy (concept described in our prior work (Kamboj, Banerjee, and Gupta 2024b)) the maximum amount. Following the entropy calculation in Eqn. 3 (Appendix) we can derive the entropy $\theta_r^{M_d}$ for rare class c_r in classifier M_d and select the classifier with the least entropy.

RARE class classification for Single domain Generalization (RareSaGe) overview: Algorithm 1 in Appendix (Banerjee 2026) proposed in (Kamboj, Banerjee, and Gupta 2024a,b) is used to design a classifier for rare class c_r.

Multiple Rare Classes: The Algorithm 1 can be applied iteratively on multiple rare classes one by one. We show this with three examples: single rare class CAD and SOZ detection, and multiple rare class DR grading.

Application of RareSaGe

We show RareSaGe application to SOZ detection problem. Stress ECG and DR grading examples are in appendix.

In standard pre-surgical screening, resting-state fMRI (rs-fMRI) is acquired, producing 4D spatio-temporal data. Independent component analysis (ICA) decomposes rs-fMRI into three classes of spatial–temporal independent components (ICs): (i) noise ICs capturing motion and measurement artifacts, (ii) resting-state network (RSN) ICs representing normal brain activity, and (iii) seizure onset zone (SOZ) ICs corresponding to epileptogenic activity. Each fMRI scan typically yields 100–200 ICs per patient, of which only a small subset ($\approx 5 - 10\%$) correspond to SOZ ICs, constituting a rare-class detection problem (Banerjee et al. 2023; Kamboj et al. 2024). Fig. 1 shows the architecture of expert knowledge integration with DL for rare class detection.

Deep Learning for noise DL techniques, including Vision transformers (ViT), language vision models (LVM)s (Radford et al. 2021), 2D-CNNs, and transfer learning using pretrained models are effective at capturing intricate spatial patterns and features within images (He et al. 2016). Given the prevalence of noise class instances in medical imaging datasets, often constituting over 50% of the data (Banerjee et al. 2023; Boerwinkle et al. 2017; Kamboj et al. 2024), DL can leverage its capability to detect subtle spatial patterns and features that distinguish noise from other classes.

Expert Knowledge on rs-fMRI We utilize two types of expert knowledge on rs-fMRI data:

a) Anatomical knowledge: This pertains to the spatial locations of anatomical brain parts crucial for SOZ recognition. These locations can be extracted employing established image processing algorithms. Locations are in Appendix.

b) Expert knowledge on specific rare class: This encompasses knowledge about rs-fMRI activation patterns observed for SOZ, compiled from the works of Hunyadi et al. (Hunyadi et al. 2015) and Boerwinkle et al. (Boerwinkle et al. 2017). SOZ specific knowledge is expressed as logical connectives of atomic propositions on the location of activation relative to brain anatomical regions (Appendix). The atomic proposition valuations are used in a support vector machine (SVM) based expert knowledge extractor (EKE).

Integration of expert knowledge with DL using Algorithm 1 Following Algorithm 1 we first identify rare class. From domain Y_A, the cross entropy using CLIP features for Noise was $\theta_{Noise}^{CLIP} = 0.004$, for RSN was $\theta_{RSN}^{CLIP} = 0.0046$, and for SOZ it is $\theta_{SOZ}^{CLIP} = 0.026$. We identify that SOZ is the rare class since θ_{SOZ}^{CLIP} satisfies Definition 2. Further, the cosine similarity of CLIP features between Noise and SOZ was 0.78 while that between RSN and SOZ was 0.74. Hence, the Noise class is determined to be the overlap class. We divide the dataset into $NOISE$ and $\neg NOISE$ class. Since we do not have knowledge based machines for Noise, we utilize DL techniques to classify $NOISE$ and $\neg NOISE$. The $\neg NOISE$ class is then used to determine SOZ. Here we have SOZ specific discriminative expert knowledge and we use the knowledge based machines to identify SOZ. We integrate expert knowledge with DL in the post processing

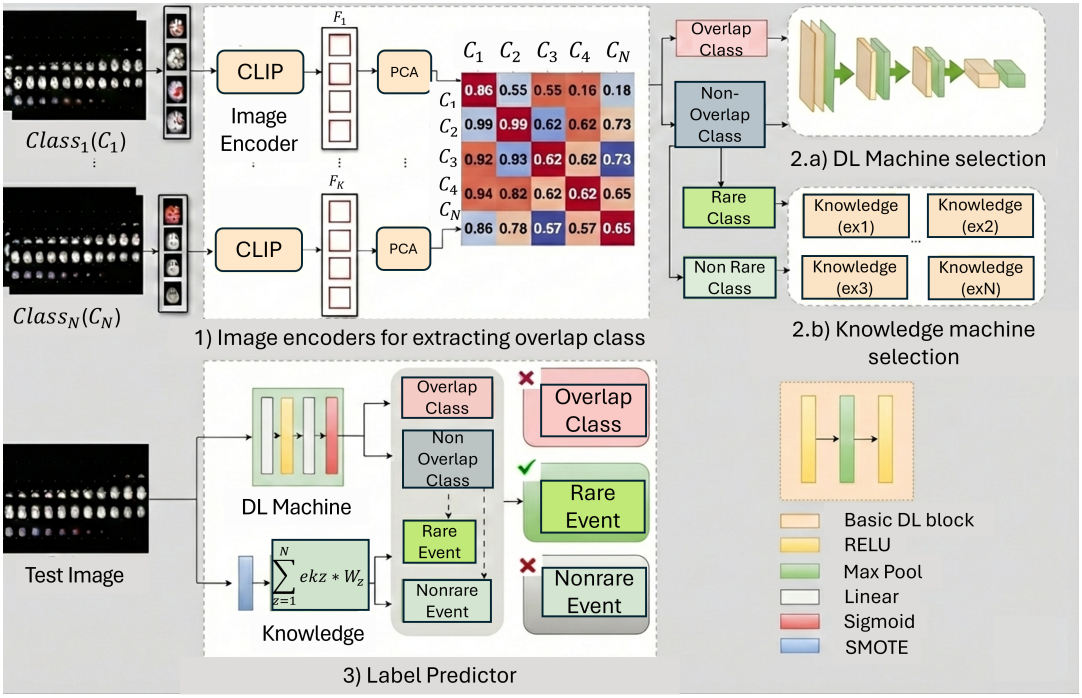

Figure 1: **RareSaGe - 1) Image Encoders:** CLIP identifies the class most similar to the rare class. **2.a) DL Machine** classifies overlap and non-overlap classes. **2.b) Knowledge Machine** extracts features from the rare class, and trains a quadratic optimization model. **3) Label Predictor:** Each test image is routed through DL classifier and knowledge machine for classification.

step (Fig 1, Label Predictor). The DL first classifies IC as Noise (overlap) or non-Noise (non-overlap) class. Simultaneously, the EKE classifies IC as SOZ (rare) and RSN (non-rare/normal). If DL classifies an IC as noise and EKE categorizes it as SOZ with a confidence score surpassing a selected threshold of 0.9, the IC is labeled as SOZ; otherwise, it retains noise label. For ICs labeled as non-noise by DL, the EKE label of RSN or SOZ is selected as the final label.

Experiments and Results

DR grading: benchmarks and baselines are in Appendix.

Deployment description

Deployment 1, SOZ detection: DL+EKE was trained using data from, Phoenix Children's Hospital (Center A) and deployed for testing in University of North Carolina (Center B), in compliance with IRB protocols and cross-university agreements. Center A includes 52 pediatric patients (23 Male, 29 Female, ages 3 months to 18 years) with 5,616 images (2,873 Noise, 2,427 RSN, 316 SOZ), acquired using a 3T Philips Ingenuity scanner. Test Center B includes 31 patients (14 Male, 17 Female, ages 2 months to 62 years) with 2,364 images (1,090 Noise, 1,072 RSN, 202 SOZ), acquired using a Siemens MAGNETOM Prisma FIT scanner.

SOZ baselines: There is no existing DL based baseline available for SOZ detection. We develop our own DL techniques for comparison including: pre-trained CNN using "VGG-16," pre-trained ViTs using "vit-small-patch16-224" and Google's "vit-base-patch16-224," ViTs trained from scratch with hyperparameters optimized using Optuna, LVM CLIP (Radford et al. 2021), and knowledge-based systems.

Deployment 2, CAD detection: The Collaborative Institutional Review Board (IRB) provides access to the Mayo Integrated Stress Center (MISC) database, containing over 100,000 Exercise Stress ECG (ESE) cases linked to invasive coronary angiography (ICA) from 2010. The dataset covers CAD patients from three centers (Banerjee et al. 2025).

Blind testing: Subsequently, the model was re-tested on a second validation cohort (n = 92) through blind testing where the labels of the stress ECG were not disclosed to the developers at ASU during inference. The labels were manually verified by the collaborators at Mayo Clinic. The distribution of CAD in this cohort was not predefined by the investigators. For this second cohort, random patients who underwent exercise stress ECG and coronary angiography within 3 months between February 2025 and May 2025 were included. Patients previously included in the first cohort, those with overlapping ECG waves or missing data, and patients who experienced acute coronary syndromes between the stress ECG and angiogram were excluded

Experiments

To evaluate generalizability, we perform two experiments motivated from (Chekroud et al. 2024):

A) Across trial validation: Here, $M_{DL(N)}$ and $M_{EKE(N)}$ is trained on data of center X with N patients and tested on dataset of the other centers $Y \neq X$ with P patients. This trial tests the SDG performance.

Method	Accuracy	Precision	Sensitivity	F1-score	Average F1-score	Time(min)	Ablation
Pre-trained ViT small Train A, Test B	64.5%	86.9%	71.4%	78.4%	77.2%	45(12)	DL only
Pre-trained ViT small Train B, Test A	61.5%	91.4%	65.3%	76.1%		42(13)	DL only
Knowledge based system, Train A, Test B	83.8%	89.6%	92.8%	91.2%	78.9%	13(6)	Knowledge only
Knowledge based system, Train B, Test A	50.0%	89.6%	53.0%	66.6%		12(6)	Knowledge only
DL+EKE Train A, Test B	**90.3%**	**90.3%**	**100%**	**94.9%**	**90.2%**	**35(10)**	**DL+EKE**
DL+EKE Train B, Test A	**75.0%**	**92.8%**	**79.5%**	**85.6%**		**38(9)**	**DL+EKE**

Table 2: Performance results of across-trial validation—single domain generalizability for rare class detection. Here we show only the best DL only approach. For all other baselines refer to Table 11 in Appendix

B) Aggregated trial validation: This is leave-one-domain-out method where data from all centers but one combined ($A \bigcup B$) was evaluated using 5-fold, 3-repeats validation.

Evaluation metrics, Results and Insight

Evaluation metrics: Trials are evaluated with standard metrics used in (Hunyadi et al. 2015; Kamboj, Banerjee, and Gupta 2024b) such as average accuracy of multi-class classification and F1 score for identifying rare class. For the SOZ detection example, we evaluate average execution time in detecting SOZ from each patient. This quantifies computational complexity of our technique and we compare it with SOTA. For CAD detection we present positive/ negative predictive value (PPV/NPV) which are clinically relevant.

DR grading results: Tables 5 through 8 in Appendix present systematic evaluation across all 12 source-target pairs for across trial validation SDG performance. DL+EKE consistently outperforms SOTA benchmarks. Table 9 in appendix presents fair comparisons where all methods are trained on three datasets and evaluated on the fourth for aggregate trial validation. DL+EKE demonstrates substantial improvements over SOTA domain generalization approaches. DL+EKE's superior performance stems from the integration of medical vision-language pre-training with domain conformal boundaries.

SOZ detection results: Table 2 presents the results of state-of-the-art DL techniques and our methodology on the across-trial validation test. All pre-trained models were fine-tuned on data from specific centers, with loss updated using class weights to address dataset imbalances due to rare classes, and early stopping strategies applied to prevent overfitting. For LVM, we conducted two evaluations: one fine-tuning with contrastive loss and another with cross-entropy loss, both using the "ViT-B/32" model. Our methodology utilized a 2D-CNN as a DL machine, which performed best for evaluating overlap (noise) and non-overlap (non-noise) classes (Fig. 1). In the across-trial experiment, our DL+EKE method achieved an F1 score of 90.2%, consistently maintaining F1 scores above 85.0% for both centers. This significantly outperforms other baseline comparators, highlighting our technique's generalizability to unseen data and its ability to learn domain-invariant expert knowledge without overfitting. Additionally, in the aggregated trial validation, where models were exposed to more data from both centers, Table 10 in Appendix demonstrates that our technique achieves the best results, with an average F1 score improvement of 4.5%.

SOZ detection insights: Our across-trial experiments reveal

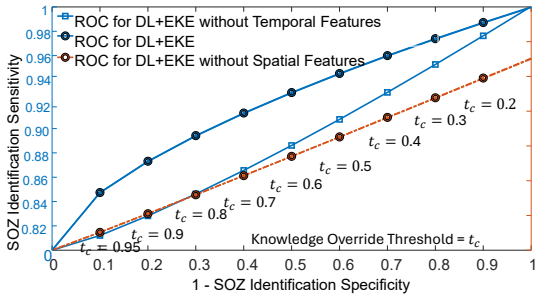

Figure 2: ROC of DL+EKE for different knowledge override thresholds and ablation analysis of knowledge.

that state-of-the-art DL techniques struggled to differentiate between NOISE and SOZ classes, as indicated by CLIP similarity results, leading to suboptimal performance in rare class detection. A knowledge-based system trained on Center A accurately identified SOZ characteristics in Center B with high precision. However, when trained on Center B and tested on Center A, precision remained stable, but sensitivity dropped significantly, highlighting an increase in FNs due to patient variability. Integrating DL for overlap class separation with knowledge-based methods improved both FPs and FNs, resulting in a more generalized performance. The aggregate trial achieved higher F1 score of 94.7%.

Ablation studies: Table 2 shows that DL+EKE performs better than DL only and knowledge only approaches. Fig. 2 further explores ablation of knowledge types. It shows that temporal knowledge components are not as important as spatial components since removing them makes the DL+EKE technique as good as random choice.

Sensitivity to knowledge override threshold t_c: The receiver operating characteristics (ROC) curve shows the variance of sensitivity and specificity of DL+EKE as the knowledge override threshold is varied from 0.1 to 0.95. It shows determining this threshold is not straightforward and reduction of this threshold may hamper performance.

CAD detection results: We test four different configurations of DL+EKE, the best performing K1 and subsequent ablations: *a) 5-Lead Transformer at Max METs (K1 - comparator):* This primary architecture integrates the highest expert knowledge level, utilizing data from all maximum MET levels and restricting the lead count(L=5). *b) 5-Lead Transformer Across All METs (K2):* Maintaining a 5-lead constraint while including data from all MET levels. *c) 12-Lead Transformer at Max METs (K3):* Expanding the lead

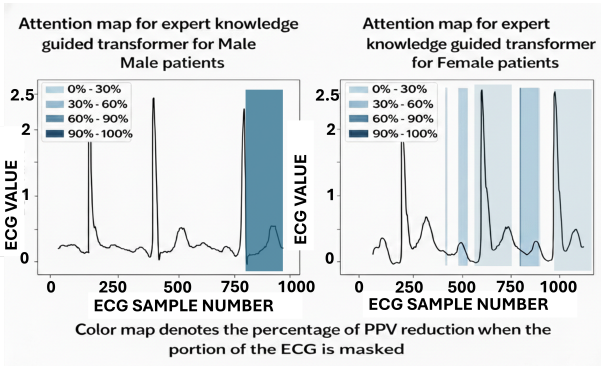

Figure 3: LIME attention maps of DL+EKE focuses on ST segment for males, but on QRS and P waves for females.

coverage to all 12 leads (L=12) focuses solely on data from maximum MET levels. *d) 12-Lead Transformer Across All METs (K4):* It uses all leads and all speeds.

Metrics	Method	Validation	Test Data (2010)	Blind Test (2025)
PPV	ViT	80.4%	79%	46.0%
	K1	91.2% (3.3)	91.2%	75.0%
	K2	87.2% (3.0)	88% (0.04)	74.0%
	K3	83.3% (4.3)	82.4% (0.001)	72.0%
	K4	83.4% (3.0)	82.3% (0.003)	71.5%
NPV	ViT	83%	81.8%	49.0%
	K1	93.0% (1.5)	93%	76.0%
	K2	85% (22.3)	87% (0.07)	73.0%
	K3	90% (22.4)	81% (0.045)	74.0%
	K4	83% (17.2)	79% (0.006)	72.0%

Table 3: Comparison with SOTA unguided-ViT models, and DL+EKE models on validation and test performance.

CAD detection insights: Among the four expert-guided configurations, K1, incorporating five expert-selected leads and maximum METs, achieved the best performance. Relative to K1, using all METs reduced PPV by 4% and NPV by 8%, while using all leads with maximum METs reduced PPV by 7.9% and NPV by 3%. Using all leads and all METs further reduced PPV by 7.8% and NPV by 10%. For unseen test data, K1 achieved a PPV of 91.2% and an NPV of 93% (Table. 3), representing a substantial improvement over the state-of-the-art PPV of 77%. Additionally, expert-guided AI outperformed unguided approaches, yielding average gains of 12.2% in PPV and 11.2% in NPV. The 5-fold cross-validation ROC AUC was 92.2 ($\pm$ 1.1), and knowledge guided methods demonstrated strong SDG performance.

Discussion on Deployment Experience

The **main** observation is that state-of-the-art single-domain generalization methods fail to generalize across both deployment applications and the DR benchmark, and **do not achieve clinically significant performance**.

Incorporating expert knowledge can improve generalization, but comes at the **cost of substantial manual effort** to identify and encode domain expertise. Moreover, many inter-center variations are difficult to anticipate without a collaborative development environment that enables continuous interaction between clinical and engineering teams. The absence of such interaction can significantly delay deployment-ready AI. Clinicians often require clinically meaningful parameters that are not exposed by standard AI/ML pipelines; retrofitting explainability after development can necessitate extensive re-engineering and must therefore be addressed early in the design phase.

Post deployment interaction at Mayo Clinic: After the blind testing on 92 patients, the research team presented the technique and results to team of experts in internal medicine and cardiology. Interaction with experts resulted in the following questions and suggestions:

Q1. Can the attention map explain why a positive manual stress ECG interpretation was a false positive (not confirmed in ICA)? DL+EKE has the capacity to show the part of a signal that has a domnant effect on the outcome. However, this component is not thoroughly validated.

Q2. Can DL+EKE grade data quality and qualify the recognition result with a confidence? DL+EKE assumes good quality stress ECG images where the leads are not overlapped. The performance of DL+EKE with noisy stress ECG is yet to be quantified.

Q3. Can the attention map educate a trainee about signs of a positive stress ECG result that is confirmed by ICA? The question indicates an interest in using DL+EKE as an AI tutor for ECG readers. While this is theoretically possible, DL+EKE was not originally designed for this purpose.

What signal characteristics did DL+EKE use to determine CAD positive results for women? LIME (Simsar et al. 2024) based attention maps of the transformer is shown in Fig. 3. The question shows strong interest in reducing the disparity between men and women. This delves into ethical use of AI where AI can potentially reduce health disparities.

Conclusions

SDG is an essential property of deployment ready AI. SOTA DG techniques may show improvements in benchmarks but are nowhere near clinically relevant performance that can be readily deployed in patient facing applications. Expert knowledge can improve generalized performance however, knowledge is vague and often conflicting between experts. Hence, knowledge refinement is essential which can only occur through iterative dialogue early in the development process. There needs to be a strong theoretical exploration on the reasons for failure of SOTA DG methods. A formalism of the core reason of causal changes across domains may result in more efficient and better SDG methods. This is essential for deployable AI in the medical field.

Acknowledgments

We thank NSF FDTBiotech (2436801), NIH R21 (1R21HL175632), Mayo Cardiovascular Research, the Mayo–ASU Seed Grant, and the Arizona New Economy Initiative for partially funding, Dr. Boerwinkle for SOZ expertise, and Payal Kamboj, Riya Salian, Kuntal Prakash Thakur, and Midhat Urooj for compiling results.

References

Abubakar, Y. I.; Othmani, A.; Siarry, P.; and Sabri, A. Q. M. 2024. A Systematic Review of Rare Events Detection Across Modalities using Machine Learning and Deep Learning. *IEEE Access*.

Banerjee, A. 2026. Appendix for Experience with Single Domain Generalization in Real World Medical Imaging Deployments. https://www.dropbox.com/scl/fi/abwwx23ox5pdkdzjb6vih/IAAIPaper_V2.pdf?rlkey=zzklj7fwcuc2a6a4rsb7evoee&st=pu2p94jp&dl=0.

Banerjee, A.; Kamboj, P.; Wyckoff, S. N.; Sussman, B. L.; Gupta, S. K.; and Boerwinkle, V. L. 2023. Automated Seizure Onset Zone Locator from Resting-State Functional MRI in Drug-Resistant Epilepsy. *Frontiers in Neuroimaging*, 1(Brain Imaging Methods).

Banerjee, A.; Salian, R. S.; Vemulapalli, H. S.; Sriramoju, A. K.; Prajapati, P.; Rodriguez-Riascos, J. F.; Muthu, P.; Iyengar, S. K.; Shen, W.; Gupta, S. K.; et al. 2025. Enhancement of Stress ECG Performance with Machine Learning: A Single-Center Study. *JACC: Advances*, 4(10_Part_2): 102141.

Bera, S.; and Biswas, P. 2023. Noise Conditioned Weight Modulation for Robust and Generalizable Low Dose CT Denoising. In Greenspan, H.; et al., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14229 of *Lecture Notes in Computer Science*. Springer, Cham.

Boerwinkle, V. L.; Mohanty, D.; Foldes, S. T.; Guffey, D.; Minard, C. G.; Vedantam, A.; et al. 2017. Correlating Resting-State Functional Magnetic Resonance Imaging Connectivity by Independent Component Analysis-Based Epileptogenic Zones with Intracranial Electroencephalogram Localized Seizure Onset Zones and Surgical Outcomes in Prospective Pediatric Intractable Epilepsy Study. *Brain Connectivity*, 7: 424–442.

Che, H.; Cheng, Y.; Jin, H.; and Chen, H. 2023. Towards Generalizable Diabetic Retinopathy Grading in Unseen Domains. In Greenspan, H.; et al., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14224 of *Lecture Notes in Computer Science*. Springer, Cham.

Chekroud, A. M.; Hawrilenko, M.; Hieronimus, L.; et al. 2024. Illusory Generalizability of Clinical Prediction Models. *Science*, 383(6679): 164–167.

Chen, M.; Jiang, M.; Dou, Q.; Wang, Z.; and Li, X. 2023a. FedSoup: Improving Generalization and Personalization in Federated Learning via Selective Model Interpolation. In Greenspan, H.; et al., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14221 of *Lecture Notes in Computer Science*. Springer, Cham.

Chen, S.; et al. 2023b. Federated Condition Generalization on Low-dose CT Reconstruction via Cross-domain Learning. In Greenspan, H.; et al., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14222 of *Lecture Notes in Computer Science*. Springer, Cham.

Chen, Z.; Pan, Y.; Ye, Y.; Cui, H.; and Xia, Y. 2023c. Treasure in Distribution: A Domain Randomization Based Multi-source Domain Generalization for 2D Medical Image Segmentation. In Greenspan, H.; et al., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14223 of *Lecture Notes in Computer Science*. Springer, Cham.

Daniele, A.; and Serafini, L. 2019. Knowledge Enhanced Neural Networks. In *Pacific Rim International Conference on Artificial Intelligence*.

Galappaththige, C. J.; Kuruppu, G.; and Khan, M. H. 2024. Generalizing to unseen domains in diabetic retinopathy classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 7685–7695.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hu, S.; Liao, Z.; and Xia, Y. 2023. Devil is in Channels: Contrastive Single Domain Generalization for Medical Image Segmentation. In Greenspan, H.; et al., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14223 of *Lecture Notes in Computer Science*. Springer, Cham.

Hunyadi, B.; Tousseyn, S.; Dupont, P.; Van Huffel, S.; De Vos, M.; and Van Paesschen, W. 2015. A Prospective fMRI-Based Technique for Localising the Epileptogenic Zone in Presurgical Evaluation of Epilepsy. *Neuroimage*, 113: 329–339. Epub 2015 Mar 14.

Kamboj, P.; Banerjee, A.; Boerwinkle, V. L.; and Gupta, S. K. 2024. The Expert's Knowledge Combined with AI Outperforms AI Alone in Seizure Onset Zone Localization Using Resting State fMRI. *Frontiers in Neurology*, 14(Artificial Intelligence in Neurology).

Kamboj, P.; Banerjee, A.; and Gupta, S. K. 2024a. STORM: Strategic Orchestration of Modalities for Rare Event Classification. In *2024 58th Asilomar Conference on Signals, Systems, and Computers*, 554–558.

Kamboj, P.; Banerjee, A.; and Gupta, S. K. S. 2024b. Expert Knowledge Driven Human-AI Collaboration for Medical Imaging: A Study on Epileptic Seizure Onset Zone Identification. *IEEE Transactions on Artificial Intelligence*, 5(10): 5352–5368.

Kamboj, P.; Banerjee, A.; Xu, B.; and Gupta, S. 2025. Generating Customized Prompts for Zero-Shot Rare Event Medical Image Classification Using LLM. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Kim, H.; Shin, Y.; and Hwang, D. 2023. DiMix: Disentangle-and-Mix Based Domain Generalizable Medical Image Segmentation. In Greenspan, H.; et al., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14222 of *Lecture Notes in Computer Science*. Springer, Cham.

Li, H.; et al. 2023. Frequency-Mixed Single-Source Domain Generalization for Medical Image Segmentation. In Greenspan, H.; et al., eds., *Medical Image Computing and*

Computer Assisted Intervention – MICCAI 2023, volume 14225 of *Lecture Notes in Computer Science*. Springer, Cham.

Liu, Y.; Liu, M.; Zhang, Y.; and Shen, D. 2023. Development and Fast Transferring of General Connectivity-Based Diagnosis Model to New Brain Disorders with Adaptive Graph Meta-Learner. In Greenspan, H.; et al., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14227 of *Lecture Notes in Computer Science*. Springer, Cham.

Nandakumar, N.; Hsu, D.; Ahmed, R.; and Venkataraman, A. 2023. DeepEZ: A Graph Convolutional Network for Automated Epileptogenic Zone Localization From Resting-State fMRI Connectivity. *IEEE Transactions on Biomedical Engineering*, 70(1): 216–227.

Parker, L.; Moreno-Garijo, A.; Chilet-Rosell, E.; Lorente, F.; and Lumbreras, B. 2023. Gender Differences in the Impact of Recommendations on Diagnostic Imaging Tests: A Retrospective Study 2007–2021. *Life*, 13: 289.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Simsar, E.; Tonioni, A.; Xian, Y.; Hofmann, T.; and Tombari, F. 2024. LIME: Localized Image Editing via Attention Regularization in Diffusion Models. arXiv:2312.09256.

Sokolova, M.; El Emam, K.; Chowdhury, S.; Neri, E.; Rose, S.; and Jonker, E. 2010. Evaluation of rare event detection. In *Advances in Artificial Intelligence: 23rd Canadian Conference on Artificial Intelligence, Canadian AI 2010, Ottawa, Canada, May 31–June 2, 2010. Proceedings 23*, 379–383. Springer.

Stolte, S.; et al. 2023. DOMINO++: Domain-Aware Loss Regularization for Deep Learning Generalizability. In Greenspan, H.; et al., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14223 of *Lecture Notes in Computer Science*. Springer, Cham.

Vidit, V.; Engilberge, M.; and Salzmann, M. 2023. CLIP the Gap: A Single Domain Generalization Approach for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3219–3229.

Yan, S.; et al. 2023. EPVT: Environment-Aware Prompt Vision Transformer for Domain Generalization in Skin Lesion Recognition. In Greenspan, H.; et al., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14226 of *Lecture Notes in Computer Science*. Springer, Cham.

Yang, X.; Chin, B.; Silosky, M.; Litwiller, D.; Ghosh, D.; and Xing, F. 2023. Learning with Synthesized Data for Generalizable Lesion Detection in Real PET Images. In Greenspan, H.; et al., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14224 of *Lecture Notes in Computer Science*. Springer, Cham.