Multimodal Fusion for Domain-Incremental Continual Learning in 3D Medical Imaging

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Scaling deep learning models across evolving clinical environments requires robust Continual Learning (CL) mechanisms to overcome severe domain shifts. However, standard visual CL approaches in medical imaging frequently suffer from domain overfitting and shortcut learning, where models memorize spurious physical artifacts rather than genuine biological pathology. Furthermore, visual-only dynamic routing often exacerbates this by causing negative transfer in out-of-distribution clinical scenarios. To address the stability-plasticity dilemma under severe clinical covariate shift, we propose a Multimodal-Guided Continual Adaptation framework. Our approach decouples 3D visual representations from tabular clinical metadata using a dual-stream architecture. To enable Domain-Incremental Learning (Domain-IL), we freeze a generalized visual backbone and sequentially train domain-specific Deep Latent Adapters. Crucially, the framework leverages the tabular modality to compute a demographic sparsity metric. This multimodal gating dynamically routes common patient profiles to specialized visual adapters, while deploying a cautious cross-domain ensemble to prevent negative transfer for atypical clinical profiles. Additionally, an entropy-weighted prior correction mechanism utilizes the tabular stream to mathematically subtract hardware-specific priors directly from the visual predictions. Evaluations on a multi-hospital 3D CT continual learning benchmark demonstrate that our multimodal fusion strategy effectively preserves representation stability, ensures target plasticity, and neutralizes domain-specific biases across sequential clinical deployments.

Keywords: Continual Learning · Multimodal Vision · Domain Adaptation · Domain-Incremental Learning · Medical Imaging.

1 Introduction

The transition of medical deep learning models to heterogeneous, multi-hospital deployment is bottlenecked by the challenge of Domain-Incremental Learning (Domain-IL) [23]. As models process data from new clinical environments, varying imaging protocols, scanner manufacturers, and local demographic distributions induce significant covariate shifts [2]. In these sequential deployment scenarios, deep neural networks frequently succumb to shortcut learning, a phenomenon where models bypass complex biological reasoning to memorize spurious physical correlations [12]. Instead of learning genuine pathology, the network overfits to

domain-specific hospital priors, such as varying slice thicknesses, reconstruction algorithms, or scanner noise profiles [33]. When deployed across sequential clinical settings, these models risk misdiagnosing patients by inappropriately relying on learned hardware artifacts rather than true anatomical evidence [8].

Existing Continual Learning (CL) frameworks attempt to mitigate these domain shifts by sequentially fine-tuning on newly acquired hospital data [27]. However, standard approaches face a severe stability-plasticity dilemma, often triggering catastrophic forgetting that erases historical knowledge as network weights are overwritten [21, 7]. While Parameter-Efficient Fine-Tuning (PEFT) [14] and dynamic routing architectures, such as Mixture-of-Experts (MoE) [32], offer pathways to modular learning, their standard gating mechanisms operate as black boxes reliant purely on entangled visual features [6]. In medical imaging, purely visual gating fails to separate a patient's anatomical presentation from their demographic and hardware context. Consequently, these models frequently misroute out-of-distribution (OOD) patients who exhibit atypical traits for their specific pathology, leading to negative transfer and systemic diagnostic failures during sequential deployment [30, 13].

To resolve the intersection of catastrophic forgetting, domain shift, and hardware induced shortcut learning, we propose a Multimodal-Guided Continual Adaptation framework. We argue that robust continual learning in 3D medical imaging requires integrating non-visual contextual data to safely guide the visual adaptation process. We establish a decoupled dual-stream architecture that separates 3D visual representations from tabular clinical metadata. To enable robust Domain-IL, the framework freezes a generalized spatial backbone and injects parameter-efficient 3D Deep Latent Adapters optimized via a dual-objective training strategy. Crucially, instead of relying on opaque visual gating, we introduce a multimodal routing mechanism directed exclusively by decoupled tabular covariates. This design aims to protect atypical patient profiles with high demographic sparsity by using a targeted, similarity-weighted cross-domain ensemble, dynamically mitigating negative transfer.

Furthermore, to encourage the final diagnosis to rely on genuine biological evidence across domains, our framework utilizes an entropy-guided prior correction mechanism. By utilizing an independent tabular stream to model the hardware and population priors, we mathematically quantify inherent hospital bias and adjust the 3D spatial predictions.

In summary, the main contributions of our work are:

- **Multimodal Covariate-Guided Routing**: We replace purely visual dynamic gating with a formal routing logic driven by tabular metadata (Demographic Sparsity Risk). This novel decoupling transitions atypical patients into a similarity-weighted cross-domain ensemble, mitigating negative transfer and overconfident misclassification in OOD scenarios.
- **Parameter-Efficient Domain-IL**: We present a dual-objective (Focal and Supervised Contrastive) continual learning strategy utilizing 3D Deep Latent Adapters injected into a frozen backbone. This approach maps anatomically similar patients to proximate latent coordinates to support system-level representation stability.

– **Entropy-Weighted Hardware Prior Correction**: We introduce a post-hoc logit adjustment mechanism that quantifies the uncertainty of hardware protocol and population priors via the tabular stream. This dynamically subtracts scanner bias from the 3D visual predictions without artificially suppressing inherently confident diagnostic signals.

2 Related Work

Parameter-Efficient Continual Learning: The deployment of deep learning models across evolving clinical environments requires the ability to learn new data distributions without overwriting historical knowledge, formalized as Continual Learning (CL) [7]. Traditional CL approaches typically utilize regularization [21] or rehearsal-based methods [29]. However, in medical imaging, rehearsal strategies are heavily restricted by patient privacy regulations (e.g., HIPAA), and regularization penalties often restrict the model's plasticity when adapting to radically different hospital hardware [9]. Parameter-Efficient Fine-Tuning (PEFT) and adapter modules have emerged as viable alternatives, allowing networks to adapt to new tasks by training a small number of injected parameters while keeping the backbone frozen [14, 5]. Our framework advances this by optimizing 3D Deep Latent Adapters with a dual-objective strategy (Focal and Supervised Contrastive Loss). This minimizes catastrophic forgetting while ensuring that the latent space remains geometrically separable across sequential domains, conditioning the features for downstream multi-modal retrieval.

Shortcut Learning and Multimodal Fusion: Deep neural networks are notoriously susceptible to shortcut learning, achieving high performance by memorizing spurious environmental correlations rather than learning the underlying biological pathology [12]. Under Domain-IL, models frequently overfit to domain priors, such as scanner-specific noise profiles or specific reconstruction algorithms [33, 8]. Existing debiasing techniques often rely on adversarial training to strip domain information from the latent space [10], which can inadvertently destroy subtle pathological features. Our framework minimizes the risks of latent unlearning by introducing an independent Tabular Stream. By explicitly training a Multi-Layer Perceptron (MLP) [11] to predict diagnostic distributions based solely on clinical covariates (hardware protocols and demographics), we mathematically isolate the data shortcut bias. This secondary modality is then used to dynamically scale logit adjustments, ensuring hardware and demographic priors are corrected without suppressing confident biological signals.

Dynamic Routing and Negative Transfer: To handle diverse, heterogeneous data distributions during continual learning, researchers frequently employ Mixture-of-Experts (MoE) architectures, which utilize a gating network to route inputs to specialized sub-networks [32]. However, standard routing mechanisms operate as black boxes heavily reliant on entangled visual features [6]. Consequently, these visual gating networks frequently trigger negative transfer when presented with OOD clinical scans. Because the gating mechanism merges anatomical pathology with domain-specific hardware noise, atypical patients are often misrouted with high confidence [13]. Our framework redesigns adaptive

routing by decoupling the gating mechanism from the visual stream. By utilizing a mathematically formal logic governed exclusively by tabular covariates, we dynamically identify OOD profiles and transition them to a cautious cross-domain ensemble, explicitly mitigating negative transfer during inference.

3 Methodology

Our multimodal continual learning architecture operates via a decoupled dual-stream design to seamlessly adapt to sequential hospital domains, mitigating both catastrophic forgetting and hardware shortcut learning. The process operates in four sequential mathematical phases:

1. Parameter-Efficient Domain-IL via Deep Latent Adapters: To mitigate the intersection of representation degradation and modality entanglement, we isolate the extraction of biological spatial features from the modeling of clinical metadata.

Visual Stream: Let $\mathcal{F}_\theta$ represent a 3D Backbone Encoder parameterized by θ_{shared}. During an initial pre-training phase on a source domain, $\mathcal{F}_\theta$ learns generalized anatomical representations, after which θ_{shared} is strictly and permanently frozen [21]. To acquire knowledge from sequential target clinical domains ($\mathcal{D}_t$), lightweight Spatial Deep Attention Adapter modules parameterized by ϕ_t are injected into the intermediate latent layers of $\mathcal{F}_\theta$ [14]. For a given intermediate feature map x, the adapted output z is formalized as:

$$z = x + s \cdot \text{Dropout}\left(\text{SE}(\text{LN}(x)) \odot \mathcal{W}_{up}(\sigma(\mathcal{W}_{down}(\text{LN}(x))))\right)$$

where LN denotes Layer Normalization, $\mathcal{W}_{down}$ and $\mathcal{W}_{up}$ are spatial convolutions forming a volumetric bottleneck, σ is the GELU activation, SE is a Squeeze-and-Excitation spatial attention gate [15], $\odot$ represents channel-wise multiplication, and s is a learnable scalar.

During continual training on new domains, gradients strictly update the adapter parameters ϕ_t. To ensure the latent representations are highly separable, preventing majority-class collapse, the adapters are optimized using a dual-objective loss: $\mathcal{L}_{total} = \mathcal{L}_{Focal} + \lambda \mathcal{L}_{SupCon}$ [24, 20]. This combined objective dynamically penalizes misclassification of rare volumes while explicitly forcing the network to cluster patients based on anatomical pathology rather than local hospital noise [34, 1].

Tabular Stream: Simultaneously, an independent Tabular Multi-Layer Perceptron (MLP) [25] is trained on vectorized metadata $m_{train} \in \mathbb{R}^f$ (demographics, hardware, imaging protocols) from the global training corpus. Optimized via a class-weighted Cross-Entropy Loss to counteract demographic imbalances, this MLP learns to predict the diagnosis based solely on non-visual covariates. Because this stream has absolutely no access to the 3D CT volumes, the trained MLP mathematically encapsulates the population and domain prior distributions, learning exactly how specific scanning protocols or patient demographics correlate with a positive diagnosis independent of true biological pathology [8].

2. Hierarchical Prototype Memory Bank: To establish domain similarities purely via spatial geometry, the extracted generalist visual features ($v_{base} \in \mathbb{R}^d$) from the frozen backbone are compressed into canonical prototypes [4]. For each

hospital domain and diagnostic class, K-Means clustering ($K = 3$) [17] is applied to partition the latent space. The visual centroids (c_k) are extracted alongside their corresponding tabular covariate medoids (m_k), forming a unified multimodal prototype $P_k = (c_k, m_k)$. The visual centroids are L_2-normalized ($\hat{c}_k = \frac{c_k}{\|c_k\|_2}$). During inference, a test patient's visual vector $\hat{v}_{test}$ queries a FAISS index [18] using cosine similarity: $S(\hat{v}_{test}, \hat{c}_k) = \hat{v}_{test}^T \hat{c}_k$. The system retrieves the Top-3 unique visual domains, establishing their spatial configurations as $\mathcal{P}_{top} = \{P_1, P_2, P_3\}$.

3. Multimodal Covariate-Guided Routing: We integrate the parallel data streams during inference by replacing opaque visual MoE gating [32, 6] with a formal mechanism governed by the decoupled covariates. The test patient's metadata m_{test} [16] is processed by the Tabular MLP to output a prior probability distribution $\pi = f_{MLP}(m_{test})$, quantifying the metadata prior for this profile. To prevent negative transfer in OOD scenarios, the routing path is dictated by two dynamically calculated limits:

A. ***Demographic Sparsity Risk:*** This metric measures the Euclidean divergence of the test patient from the global training mean ($\bar{m}$), bounded to ensure gradient stability:

$$R(m_{test}) = \min\left(\frac{\|m_{test} - \bar{m}\|_2}{\|\bar{m}\|_2 + \epsilon}, 1.0\right)$$

B. ***Dynamic Tolerance:*** A visual safety hurdle (τ_{gap}) based on a calibrated base margin (β) dynamically expands as demographic sparsity increases to protect atypical patient records: $\tau_{gap} = \beta \cdot (1 + R(m_{test}))$.

The final spatial inference logits ($L_{spatial}$) are defined by a piece-wise routing function based on the actual spatial Visual Gap ($\Delta = |P_1 - P_2|$). Let $\mathcal{F}_{\phi_i}(x)$ denote the 3D volume processed through adapter i:

$$L_{spatial} = \begin{cases} \mathcal{F}_{\phi_1}(x) & \text{if } R(m_{test}) < R_{ceiling} \text{ and } \Delta > \tau_{gap} \\ \sum_{i=1}^{3} W_i \mathcal{F}_{\phi_i}(x) & \text{otherwise} \end{cases}$$

If the patient triggers the cautious soft fallback condition, predictions are aggregated using an unscaled similarity-weighted distribution derived directly from the FAISS retrieval distances. Let $d_i = \|\hat{v}_{test} - \hat{c}_i\|_2$ represent the L_2 distance to the retrieved centroid i. The aggregation weights are formalized as $W_i = \frac{\exp(-d_i)}{\sum_{j=1}^{3} \exp(-d_j)}$, organically suppressing severe spatial outliers.

4. Entropy-Weighted Hardware Prior Correction: While the domain adapters map localized biological features, the aggregated spatial logits ($L_{spatial}$) inherently contain residual shortcut biases learned from specific acquisition hardware. Our framework clarifies this hospital prior to support pathology isolation. Because applying a static debiasing penalty uniformly across all domains risks suppressing inherently confident, unbiased predictions, we utilize Information Entropy-Guided Penalty Scaling, leveraging predictive entropy as a formalized measure of uncertainty [19]. The uncertainty of the tabular prior (π) is quantified using Shannon Entropy [31]:

$$H(\pi) = -\sum_{c \in \{0,1\}} \pi_c \log(\pi_c + \epsilon)$$

This absolute entropy is normalized into a multiplier ranging from 0 to 1: $C_{bias} = \max\left(0, 1 - \frac{H(\pi)}{\log 2}\right)$. An optimal base penalty (α_{base}), discovered via a Youden's J statistic grid search [26], is dynamically scaled per-patient ($\alpha_{dynamic} = \alpha_{base} \cdot C_{bias}$). A post-hoc logit adjustment mathematically subtracts this scaled prior from the raw 3D spatial representations, yielding the corrected diagnostic distribution: $L_{adjusted} = L_{spatial} - \alpha_{dynamic} \log(\pi + \epsilon)$.

4 Dataset and Experimental Setup

We utilize a Domain-IL benchmark of 3D chest CT scans across five distinct hospital networks (H_1–H_5) [22]. Each volume is diagnostically annotated as COVID-19 positive or Normal. To pioneer a multimodal evaluation, we extract standardized metadata directly from the DICOM headers. Categorical features (e.g., Scanner Manufacturer) are one-hot encoded, while continuous features (Age, Slice Thickness, Voltage, Exposure) are min-max normalized to $[0, 1]$ before computing the Euclidean distance for Demographic Sparsity Risk.

To enforce a strict sequential deployment scenario, the datasets are functionally partitioned. The Source Domain (H_1; $N = 1,699$) contains a diagnostically balanced distribution (51.0% COVID-19) acquired predominantly via Siemens scanners (slice thickness 2.1 ± 1.9 mm). This domain is used exclusively to train the foundational base encoder. The Target Domains (H_2–H_5) sequentially optimize the Deep Latent Adapters and introduce radical covariate shifts. These target hospitals feature heavily imbalanced disease prevalences (ranging from 7.9% to 11.6% COVID-19) and alternate hardware, predominantly utilizing Philips and Toshiba scanners with extreme slice thickness variations (ranging from 1.0 ± 0.1 mm to 5.3 ± 1.6 mm). Within this sequence, H_4 serves as the "shortcut learning" stress-test: comprising 2,569 patients with a massive imbalance toward Normal cases (11.6%) acquired via Toshiba hardware, standard models risk collapsing into the majority class by memorizing scanner noise.

To evaluate the framework across differing architectural inductive biases, the visual backbone was implemented using four distinct 3D networks: ResNet-3D-18, DenseNet-121, Swin-UNETR, and ViT-3D. All validation and testing splits were strictly partitioned at the patient level (80/10/10) to prevent intra-patient data leakage. Adapter optimization was performed using the AdamW optimizer (learning rate 10^{-4}, batch size 8). The Supervised Contrastive Loss temperature (τ) was set to 0.07. Models were built using PyTorch [28] and the MONAI framework [3]. Due to the extreme long-tailed target distributions, diagnostic viability was evaluated strictly through class-balanced metrics: ROC-AUC (AUC), Macro F1-Score (F1), PR-AUC (PR), and Balanced Accuracy (BAcc), alongside Accuracy (Acc), Sensitivity (Sens), and Specificity (Spec). Finally, Domain-IL performance was tracked across three dimensions: Target Plasticity (adaptation to H_2–H_5), Source Stability (knowledge retention on H_1), and Fallback Plasticity (routing robustness in OOD simulations).

5 Results and Discussion

Evaluating Target Plasticity and Source Stability: An analytical requirement of any Domain-IL system is acquiring target plasticity while tracking repre-

Table 1. Comprehensive Continual Learning Evaluation. **(A)** Target Plasticity demonstrating adaptation without catastrophic forgetting. **(B)** Routing Robustness under Out-Of-Distribution (OOD) scenarios. Notice how baseline routing methods frequently collapse to 0.00% Sensitivity (predicting only the majority class), whereas our proposed framework successfully recovers and balances minority-class detection.

(A) Continual Learning Target Plasticity

Architecture	Method	Target H_2 (%)				Target H_3 (%)				Target H_4 (%)				Target H_5 (%)			
		F1	BAcc	AUC	PR	F1	BAcc	AUC	PR	F1	BAcc	AUC	PR	F1	BAcc	AUC	PR
ResNet-3D-18	Zero-Shot (H_1)	79.14	78.50	90.47	71.46	68.65	70.77	73.77	45.93	76.81	76.59	90.08	67.78	69.35	67.67	69.77	57.02
	Fine-Tuned	86.59	85.78	95.18	81.05	60.81	59.68	76.06	42.28	88.94	91.50	96.40	87.65	76.89	70.00	78.60	66.99
	Deep Adapter	**81.22**	**87.94**	**94.24**	**79.93**	**25.12**	**58.45**	**76.58**	**61.20**	**85.32**	**87.34**	**96.01**	**82.28**	**76.89**	**70.00**	**76.74**	**58.30**
Swin-UNETR	Zero-Shot (H_1)	59.48	61.13	64.44	25.70	74.22	72.89	74.12	54.46	54.50	54.19	64.10	19.63	46.67	48.84	68.84	24.32
	Fine-Tuned	73.62	72.84	85.73	61.15	70.66	67.34	69.01	50.53	70.33	70.17	81.04	43.99	74.33	76.51	85.12	67.11
	Deep Adapter	**60.85**	**71.68**	**77.94**	**32.50**	**31.02**	**41.90**	**59.15**	**55.02**	**72.33**	**78.70**	**83.60**	**40.93**	**77.67**	**77.67**	**80.93**	**65.56**

(B) Multimodal Routing Robustness

Architecture	Routing	Known Domains (Include $H_1 ... H_5$)									OOD Simulation (Exclude H_4)								
		Source Stability (H_1) (%)				Target Plasticity (H_4) (%)					Source Stability (H_1) (%)				Fallback Plasticity (H_4) (%)				
		Acc	AUC	Sens	Spec	Acc	AUC	Sens	Spec	H/S Split	Acc	AUC	Sens	Spec	Acc	AUC	Sens	Spec	H/S Split
ResNet-3D-18	Base Hard	76.34	98.21	60.71	100.0	88.39	96.11	1.89	100.0	100/0	76.34	98.21	60.71	100.0	N/A	N/A	N/A	N/A	N/A
	Base Soft	44.44	97.14	7.74	100.0	88.62	91.14	3.77	100.0	0/100	49.82	96.99	16.67	100.0	89.29	88.92	9.43	100.0	0/100
	Visual Only	87.81	94.11	93.45	79.28	73.66	81.54	73.58	73.67	66/34	77.42	89.09	66.07	94.59	72.54	77.34	69.81	72.91	71/29
	Proposed	**72.23**	**81.57**	**69.78**	**77.61**	**81.19**	**80.85**	**71.70**	**82.62**	77/23	**66.99**	**79.04**	**79.28**	**59.62**	**60.59**	**64.39**	**80.57**	**63.05**	77/23
Swin-UNETR	Base Hard	39.78	41.38	0.00	100.0	94.87	90.01	69.81	98.23	100/0	42.29	46.20	27.98	63.96	N/A	N/A	N/A	N/A	N/A
	Base Soft	39.78	41.38	0.00	100.0	89.96	89.29	20.75	99.24	0/100	42.29	46.20	27.98	63.96	88.17	53.17	0.00	100.0	0/100
	Visual Only	73.48	80.86	66.67	83.78	81.92	86.38	83.02	81.77	62/38	77.78	84.83	76.79	79.28	78.79	67.65	56.60	79.49	57/43
	Proposed	**67.96**	**76.33**	**61.15**	**82.09**	**75.26**	**78.29**	**69.81**	**76.07**	67/33	**73.79**	**80.09**	**76.98**	**67.16**	**59.65**	**67.73**	**66.04**	**58.69**	66/34

Table 2. Ablation of Multimodal Fusion Components during Domain-IL.

Backbone	CL Framework Components				Metrics (H_4) (%)		
	Adapters (ϕ_t)	Visual (S_d)	Tabular (R_m)	Debias (α_{dyn})	F1	BAcc	AUC
Swin-UNETR	–	–	–	–	54.50	54.19	64.10
	✓	–	–	–	72.33	78.70	83.60
	✓	✓	–	–	42.07	82.40	86.38
	✓	✓	✓	–	37.50	69.29	77.14
	✓	✓	✓	✓	**52.53**	**72.94**	**78.29**

sentation stability. Table 1A evaluates adaptation performance across target partitions H_2–H_5. Standard zero-shot inference exhibits severe degradation across all inductive biases, indicating that unadapted visual features are susceptible to scanner and protocol variations. For example, the zero-shot Swin-UNETR model degrades to a PR-AUC of 19.63% on the H_4 domain. Full fine-tuning establishes an empirical upper bound for local performance but triggers substantial system-level forgetting, overwriting historical representations.

Integrating parameter-efficient Multi-Layer Deep Latent Adapters maintains high target plasticity. As detailed in Table 1A, on the ResNet-3D-18 backbone, the adapters achieve an AUC of 94.24% on H_2 and 96.01% on H_4, performing comparably to unconstrained fine-tuning. Because the backbone weights remain frozen, target adaptation avoids backpropagation interference with historical representations, though inference stability across the system depends heavily on downstream routing configurations.

Multimodal Routing Prevents Negative Transfer: Standard baseline routing architectures can struggle to balance source stability and target plasticity under severe domain variations. As shown in Table 1B, "Base Hard" routing rigidly allocates inputs to a single expert domain; while effective for clear in-

distribution samples, it fails to capture cross-domain uncertainty and drops to near-zero sensitivity on highly skewed OOD partitions like H_4. Conversely, "Base Soft" routing unconditionally averages prediction logits across all available experts, introducing out-of-domain feature noise that degrades source stability, as seen in the drop in Swin-UNETR source accuracy.

Our Multimodal routing mechanism balances this stability-plasticity tradeoff by regulating adapter activation splits through the Demographic Sparsity risk metric. In OOD deployment simulations, the framework dynamically triggers a conservative ensembling strategy (allocating a 77/23% hard/soft split for ResNet and a 66/34% split for Swin-UNETR). This regularized gating structure supports source stability (maintaining a ResNet AUC of 79.04%) while enabling fallback target plasticity (64.39% AUC). Decoupling the safety hurdle via non-visual parameters helps minimize overconfident visual misrouting, mitigating negative transfer during cross-hospital inference.

Multimodal Component Ablation and Limitations: To analyze the structural interaction of our framework components, a progressive ablation was conducted on the challenging H_4 partition presented in Table 2). The adapter-only baseline (Row 2) yields a high raw global F1 score (72.33%) and an AUC of 83.60%. Adding visual-only prototype routing (S_d) expands the AUC tracking to 86.38% but reduces the global F1 score to 42.07%. Incorporating tabular covariate gating (R_m) further adjusts the operational metrics (F1 37.50%, AUC 77.14%). The fully integrated framework (Row 5) converges at an F1 score of 52.53%, balanced accuracy of 72.94%, and an AUC of 78.29%.

This non-monotonic behavior highlights a core architectural trade-off. The isolated adapter baseline maximizes raw global classification metrics by exploiting domain-specific prior distributions and dataset shortcuts in the imbalanced target environment. Introducing cross-modal routing and post-hoc debiasing explicitly penalizes these tabular priors and enforces expanded safety hurdles. While this regularization suppresses shortcut exploitation and reduces the absolute raw global F1 score, it serves to constrain domain overfitting, correcting for uncalibrated hardware priors to build balanced cross-domain stability.

6 Conclusion

Scaling deep learning models across clinical environments requires Continual Learning (CL) mechanisms robust to severe hardware covariate shifts. Standard Domain-IL architectures relying on purely visual routing often trigger shortcut learning and severe negative transfer for OOD or minority profiles. To resolve this stability-plasticity dilemma, our Multimodal-Guided Continual Adaptation framework decouples 3D spatial features from tabular clinical covariates. By injecting parameter-efficient Deep Latent Adapters into a permanently frozen backbone, we achieve target adaptation while preserving representation stability. Additionally, utilizing tabular metadata to assess domain sparsity dynamically routes atypical patients into a cautious cross-domain ensemble, bypassing opaque visual misrouting. Finally, our Information Entropy-Weighted post-hoc adjustment mathematically corrects for inherent hardware biases directly within the predictive logits. This multimodal approach establishes a stable, highly plastic foundation for the continuous deployment of 3D medical vision models.

References

1. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al.: Big self-supervised models advance medical image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3478–3488 (2021)
2. Bayasi, N., Hamarneh, G., Garbi, R.: GC2: Generalizable continual classification of medical images. IEEE Transactions on Medical Imaging (TMI) **43**(11) (2024)
3. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, I., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
4. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. In: Advances in Neural Information Processing Systems. vol. 32 (2019)
5. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. In: Advances in Neural Information Processing Systems. vol. 35, pp. 16664–16678 (2022)
6. Chi, Z., Dong, L., Huang, S., Dai, D., Ma, S., Singhal, B., Bajaj, P., Song, X., Wei, F.: On the representation collapse of mixture of experts. arXiv preprint (2022)
7. De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Alesis, A., Slump, C.H., Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks. IEEE transactions on pattern analysis and machine intelligence **44**(7) (2021)
8. DeGrave, A.J., Janizek, J.D., Lee, S.I.: Ai for radiographic covid-19 detection selects shortcuts over pathology. Nature Machine Intelligence **3**(7), 610–619 (2021)
9. Derakhshani, M.M., Najafi, M., Hossny, M., Omidshafiei, S., et al.: Continual learning in medical image analysis: A comprehensive review. Medical Image Analysis **88**, 102875 (2023)
10. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research **17**(1), 2096–2030 (2016)
11. Gardner, M., Dorling, S.: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric Environment **32**(14), 2627–2636 (1998)
12. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
13. Ghassemi, M., Oakden-Rayner, L., Kohane, I.S.: The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health **3**(11), e745–e750 (2021)
14. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)
15. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
16. Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P.: Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digital Medicine **3**(1), 136 (2020)
17. Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., Heming, J.: K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences **622**, 178–210 (2023)

18. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. IEEE Transactions on Big Data **7**(3), 535–547 (2019)
19. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in neural information processing systems (2017)
20. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
21. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
22. Kollias, D., Arsenos, A., Kollias, S.: Domain adaptation, explainability fairness in ai for medical image analysis: Diagnosis of covid-19 based on 3-d chest ct-scans. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 4907–4914 (2024)
23. Kumari, P., Chauhan, J., Bozorgpour, A., Huang, B., Azad, R., Merhof, D.: Continual learning in medical image analysis: A comprehensive review of recent advancements and future prospects (2024)
24. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (2017)
25. Mao, A., Mohri, M., Zhong, Y.: Cross-entropy loss functions: theoretical analysis and applications. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
26. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. In: International Conference on Learning Representations (2021)
27. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks **113**, 54–71 (2019)
28. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
29. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
30. Seyyed-Kalantari, L., Liu, H., McDermott, M.B., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine **27**(12), 2176–2182 (2021)
31. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)
32. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (2017)
33. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine **15**(11) (2018)
34. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 10795–10816 (2023)