Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

FedCKD: Cluster-Aware Knowledge Distillation for Heterogeneous Medical Federated Learning

Yi Xu

xuyi1023@126.com

Kun Chen
Haoyu Luo
Xiao Liu

Research Article

Keywords: Federated Learning, Medical Information Systems, Knowledge Distillation, Data Heterogeneity, Knowledge Clustering

Posted Date: August 22nd, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7250878/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

FedCKD: Cluster-Aware Knowledge Distillation for Heterogeneous Medical Federated Learning

Yi Xu[1*], Kun Chen[1], Haoyu Luo[2], Xiao Liu[3]

[1*]School of Computer Science and Technology, Anhui University, , Hefei, 230601, China.

[2]College of Mathematics and Informatics, South China Agricultural University, , Guangzhou, China.

[3]School of Information Technology, Deakin University, , Geelong, Australia.

*Corresponding author(s). E-mail(s): xuyi1023@126.com;

Contributing authors: e23301221@stu.ahu.edu.cn; haoyuluo@scau.edu.cn; xiao.liu@deakin.edu.au;

Abstract

In medical knowledge systems, federated learning provides a promising paradigm for collaborative knowledge extraction while preserving data privacy. However, inherent heterogeneity in medical information—stemming from variations in disease distribution, imaging protocols, and patient demographics—severely degrades the performance of traditional federated frameworks. To address this challenge, we propose **FedCKD**, a knowledge-driven federated framework tailored for heterogeneous medical information systems. FedCKD introduces three key innovations: (1) a label-driven knowledge clustering mechanism that partitions medical nodes based on disease-specific knowledge representations, ensuring intra-cluster semantic consistency; (2) a two-stage adaptive aggregation strategy for knowledge-oriented model fusion within each cluster, balancing local specialization and cluster-level consistency; (3) a cross-cluster knowledge distillation protocol that enables privacy-preserving transfer of complementary knowledge across specialized medical domains via weighted teacher ensembles. By simulating interoperability in distributed medical systems, FedCKD achieves cross-domain knowledge integration while respecting statistical heterogeneity. Comprehensive experiments on multiple datasets demonstrate that FedCKD significantly outperforms state-of-the-art methods, establishing it as an effective solution for

knowledge extraction and integration in privacy-sensitive, heterogeneous medical ecosystems.

Keywords: Federated Learning, Medical Information Systems, Knowledge Distillation, Data Heterogeneity, Knowledge Clustering

1 Introduction

Deep learning models in medical imaging have demonstrated remarkable efficacy in diagnostic knowledge extraction, driving significant advancements in healthcare information systems. Sheller et al.'s multi institutional brain tumor segmentation framework verifies the feasibility and superiority of federated learning in heterogeneous medical images[1].However, their reliance on large-scale, high-quality annotated datasets presents substantial challenges in real-world medical settings. Stringent privacy regulations (e.g., HIPAA, GDPR) and the high cost of expert annotations restrict data centralization, leading to pronounced heterogeneity in both data distribution and knowledge representation across healthcare institutions [1, 2].In addition, Xu et al. made a systematic review of the privacy and security challenges of federal learning in the medical scenario, providing more practical guidance for regulatory compliance[3]. This fragmentation hinders the development of unified diagnostic knowledge models, as each institution accumulates specialized expertise in its focus areas—such as oncology, cardiology, or pulmonology—with limited cross-domain knowledge coverage.

Federated learning (FL) has emerged as a privacy-preserving paradigm for collaborative knowledge extraction in distributed environments. By enabling local model updates to be aggregated without raw data sharing, FL addresses privacy concerns while fostering collaborative knowledge building. However, medical FL faces unique challenges from non-IID knowledge distributions, which arise from variations in disease prevalence, imaging protocols, and patient demographics [4, 5]. For instance, a pulmonary hospital's dataset dominated by lung nodule images differs drastically from a cardiovascular center's collection of echocardiograms, making global model aggregation via traditional methods prone to knowledge drift and performance degradation [6].

This heterogeneity manifests in three core challenges for medical knowledge systems:

- **Statistical Heterogeneity**: Divergent knowledge distributions across institutions cause model drift and suboptimal convergence in knowledge extraction[2].
- **Knowledge Isolation**: Specialized clinical knowledge remains siloed, preventing cross-institutional expertise sharing and knowledge integration.
- **Personalization Gaps**: A one-size-fits-all global model cannot adapt to the unique diagnostic knowledge needs of specialized institutions.

1.1 Advanced Federated Learning Paradigms

Recent research has developed specialized FL frameworks to address these challenges, which can be categorized into four main strands with distinct knowledge management approaches:

Classical Federated Learning.

Early FL approaches like FedAvg [7] and FedProx [8] form global knowledge models by averaging local updates. FedAvg, the canonical FL algorithm, iteratively aggregates weighted model parameters from all clients to form a global model, assuming knowledge homogeneity. FedProx introduces a proximal term to regularize local updates, mitigating the impact of slight distribution shifts. However, both methods fail to account for severe label distribution shifts in medical knowledge systems, leading to model drift when aggregating updates from institutions with skewed disease distributions.In order to solve the problem of global update deviation, SCAFFOLD uses control variables to correct local updates[9]; FedNova alleviates the unbalanced computing load by unifying and normalizing the local gradient step size[10].

Personalized Federated Learning.

Personalization techniques aim to balance global knowledge generalization with local knowledge adaptation. FedALA [11] introduces an adaptive local aggregation mechanism that weights updates based on each institution's knowledge characteristics, using an attention mechanism to prioritize task-relevant features. Meta-learning-based approaches like pFedMe [12] optimize model initialization to facilitate rapid local adaptation, learning a global meta-model that can be fine-tuned efficiently for specific institutions.Another idea is to add a personalization layer to the global model. FedPer implements personalization by fixing the global layer and only updating the local layer[13] ; Ditto adopts dual model mutual learning, and multi model parallel optimization of global and personalized goals[14]. While these methods excel in local accuracy, they often neglect cross-institutional knowledge sharing, leading to fragmented knowledge landscapes where complementary diagnostic insights remain untapped.

Clustered Federated Learning.

Cluster-based FL groups institutions with similar knowledge distributions to enhance intra-cluster collaboration. FedClusterNet [15] employs an attention-guided mechanism to weight cluster contributions, first deriving semantic representations of each client's knowledge vector and then using K-means to form homogeneous clusters. Within each cluster, models are aggregated to preserve domain-specific knowledge, significantly improving intra-cluster performance.IFCA proposes to automatically form clusters according to model deviation degree, and jointly optimize the model in the cluster[16]; Sattler et al. enhanced inter cluster robustness in non IID settings to resist malicious node interference[17]. However, these methods often lack mechanisms

for inter-cluster knowledge transfer, leading to isolated knowledge silos that hinder cross-domain diagnostic capabilities.

Federated Knowledge Distillation.

Knowledge distillation techniques enable privacy-preserving knowledge transfer via soft predictions. FedHKD [18] introduces a hierarchical distillation strategy, first transferring knowledge within clusters to reinforce shared features and then across clusters to promote complementary knowledge exchange. It uses a teacher-student framework where cluster-level models act as teachers to guide students in other clusters, using a combination of logit-based and feature-based distillation losses.FedMD implements soft label distillation through public proxy data, supporting heterogeneous model structure[19]; FedGKT will migrate the student teacher paradigm to the client server scenario to achieve efficient knowledge transfer[20]. Despite these advances, existing methods often overlook cluster-specific knowledge nuances in multi-disease scenarios, applying uniform distillation strategies that fail to capture specialized diagnostic patterns.

1.2 Our Approach and Contributions

Motivated by the limitations of existing methods, we propose FedCKD (Federated Cross-Cluster Knowledge Distillation), a hierarchical FL framework for heterogeneous medical knowledge systems, featuring:

- **Label-Driven Knowledge Clustering**: Grouping institutions by disease label distributions to form semantically coherent knowledge clusters, using K-means on label distribution vectors to ensure intra-cluster knowledge similarity.
- **Two-Phase Adaptive Aggregation**: Dynamically fusing local and cluster models through a two-step process: first aggregating intra-cluster knowledge to capture specialized features, then integrating global knowledge to maintain generalization.
- **Cross-Cluster Knowledge Distillation**: Enabling privacy-preserving transfer of complementary knowledge across clusters via a weighted teacher ensemble, combining logit-based distillation with feature alignment to preserve both global patterns and local nuances.

Our contributions include a novel knowledge clustering mechanism aligned with clinical knowledge domains, an adaptive aggregation strategy balancing specialization and generalization, and a hierarchical distillation protocol. Experiments validate that FedCKD achieves significant accuracy improvements over state-of-the-art methods in extreme non-IID settings, with performance gains verified through rigorous comparative analysis.

2 Motivation and Application Scenario

In contemporary healthcare knowledge ecosystems, hospitals function as specialized knowledge nodes, each generating copious data in core domains—such as oncology,

cardiology, or pulmonology—while retaining a smaller volume of out-of-domain cases. This results in highly non-IID distributions of medical imaging knowledge across institutions [1, 2, 5]. Centralizing these sensitive images in a single repository is neither operationally feasible nor compliant with regulations (e.g., HIPAA, GDPR) [3, 21]. Consequently, two critical challenges emerge:

1. **Performance degradation under non-IID knowledge distribution.** Aggregating model updates from institutions with divergent pathological focuses leads to knowledge parameter drift, slow convergence, and reduced diagnostic accuracy compared to locally optimized models [6, 22]. For example, a global model trained on a mix of oncology and cardiology data may underperform on both domains due to conflicting optimization directions, as the model struggles to balance the specialized features of each disease category.
2. **Lack of personalized cross-domain knowledge adaptation.** A single global model cannot be finely tuned to each institution's specialty nor effectively leverage complementary imaging features from other domains [12, 23, 24]. This limitation restricts its ability to diagnose complex cases that require integrating knowledge from multiple specialties, such as a patient with both cardiovascular and pulmonary complications, where insights from both domains are essential for accurate diagnosis.

These challenges underscore the need for a federated knowledge computing framework that:

- Preserves patient privacy by maintaining data locality under regulatory constraints [1, 21], ensuring that sensitive medical data remains within each institution's infrastructure.
- Adapts to each institution's local disease knowledge distribution through adaptive aggregation [11, 23], enabling the framework to learn specialized diagnostic patterns while maintaining some level of global consistency.
- Enables privacy-preserving knowledge sharing across specialty clusters to leverage complementary expertise without exposing raw data, allowing institutions to benefit from each other's specialized knowledge while protecting patient confidentiality.

Motivated by these needs, we developed FedCKD to cluster institutions by disease-label knowledge distributions, perform adaptive intra-cluster knowledge aggregation, and apply cross-cluster knowledge distillation—thereby achieving both local knowledge specialization and global knowledge generalization with full data privacy preservation.

3 Methodology

3.1 Framework Overview

In a federated healthcare knowledge system comprising N hospitals, each institution acts as a knowledge node holding sensitive and heterogeneous medical imaging

data. To collaboratively enhance diagnostic knowledge extraction while preserving privacy, we propose **FedCKD** (Federated Cross-Cluster Knowledge Distillation). As illustrated in Figure 1, this framework integrates three core components: **knowledge-node clustering**, **two-phase adaptive knowledge aggregation**, and **cross-cluster knowledge distillation**. These components work in tandem to address knowledge heterogeneity and enable efficient knowledge transfer in distributed medical information systems.

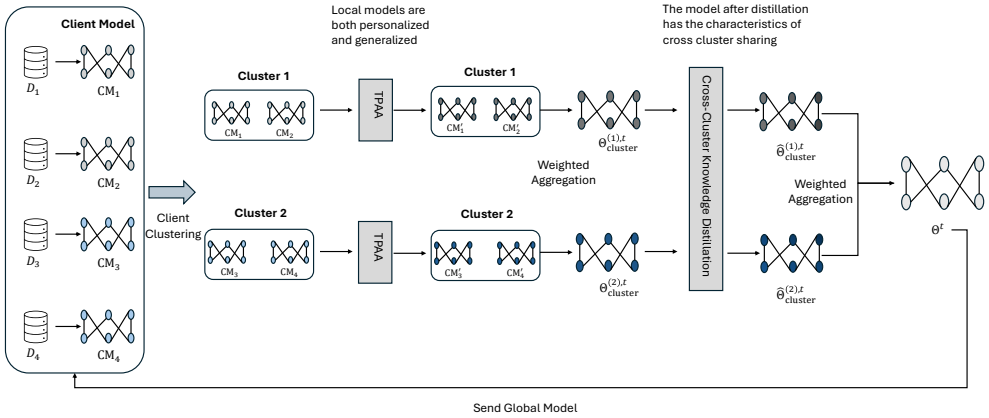


Fig. 1: Overview of the FedCKD framework: (1) Clustering knowledge nodes by disease label distributions; (2) Intra-cluster knowledge aggregation via TPAA; (3) Cross-cluster knowledge distillation (CCKD); (4) Global knowledge model aggregation.

3.2 Knowledge-node Clustering Based on Label Distributions

Hospitals typically accumulate imaging data dominated by specific disease categories, leading to distinct knowledge profiles. To align local knowledge characteristics with collaborative learning objectives, we cluster hospitals into groups with similar label distributions.

For each hospital i, its disease label distribution is represented as a knowledge vector $\mathbf{v}_i \in \mathbb{R}^K$, where K denotes the number of disease categories. The elements of $\mathbf{v}_i$ represent the proportion of each disease category in the hospital's dataset. Using these vectors, we apply K-means clustering to partition hospitals into K knowledge clusters $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$ by minimizing:

$$\min_{S} \sum_{i=1}^{N} \left\| \mathbf{v}_i - \boldsymbol{\mu}_{S(i)} \right\|^2, \tag{1}$$

where $\boldsymbol{\mu}_{S(i)}$ is the centroid of the cluster assigned to hospital i. This clustering step groups institutions with similar diagnostic knowledge focuses, ensuring that knowledge

sharing and aggregation within each cluster are more targeted and effective for the specific disease categories dominant in that cluster.

3.3 Two-Phase Adaptive Knowledge Aggregation

Direct global aggregation of knowledge models often overlooks heterogeneity in local knowledge distributions. To address this, we propose a two-phase adaptive aggregation (TPAA) strategy that integrates cluster-level and global-level knowledge, as depicted in Figure 2 (left: intra-cluster aggregation; right: global integration).

3.3.1 Cluster-Level Knowledge Aggregation

Within each knowledge cluster $\mathcal{C}_k$, we aggregate local models to capture shared diagnostic knowledge specific to the cluster's dominant disease categories. For any hospital $i \in \mathcal{C}_k$, the intra-cluster aggregated knowledge model $\Theta_{\text{intra}}^{((k,i),t-1)}$ is computed as:

$$\Theta_{\text{intra}}^{((k,i),t-1)} = \frac{\sum_{j \in \mathcal{C}_k,\, j \neq i} |\mathcal{D}_j| \cdot \Theta_j^{(t-1)}}{\sum_{j \in \mathcal{C}_k,\, j \neq i} |\mathcal{D}_j|}, \tag{2}$$

where $|\mathcal{D}_j|$ denotes the size of the dataset at hospital j. To balance cluster-level knowledge with local relevance, we adaptively fuse the hospital's previous model $\Theta_i^{(t-1)}$ with the intra-cluster model via:

$$\widetilde{\Theta}_i^t = \text{ALA_Intra}\left(\Theta_i^{(t-1)}, \Theta_{\text{intra}}^{((k,i),t-1)}\right), \tag{3}$$

where $\text{ALA_Intra}(a, b) = \alpha \cdot a + (1 - \alpha) \cdot b$ defines the adaptive fusion function. The weight $\alpha = \exp(-\text{Sim}(\mathbf{v}_i, \boldsymbol{\mu}_k))$ quantifies the similarity between the hospital's knowledge vector $\mathbf{v}_i$ and cluster centroid $\boldsymbol{\mu}_k$ (using cosine similarity), ensuring hospitals near the centroid prioritize the intra-cluster model, while peripheral nodes retain more local knowledge.

3.3.2 Global-Level Knowledge Integration

Following intra-cluster aggregation, the central server broadcasts the global knowledge model $\Theta^{(t-1)}$ to all nodes. Each hospital then adaptively fuses its intra-cluster model $\widetilde{\Theta}_i^t$ with the global model:

$$\widehat{\Theta}_i^t = \text{ALA_Global}\left(\widetilde{\Theta}_i^t, \Theta^{(t-1)}\right), \tag{4}$$

where $\text{ALA_Global}(a, b) = \beta \cdot a + (1-\beta) \cdot b$ with $\beta = \text{Perf}(\widetilde{\Theta}_i^t)/(\text{Perf}(\widetilde{\Theta}_i^t) + \text{Perf}(\Theta^{(t-1)}))$. Here, $\text{Perf}(\cdot)$ denotes the validation accuracy on a local held-out set, ensuring weights align with model performance across clusters.

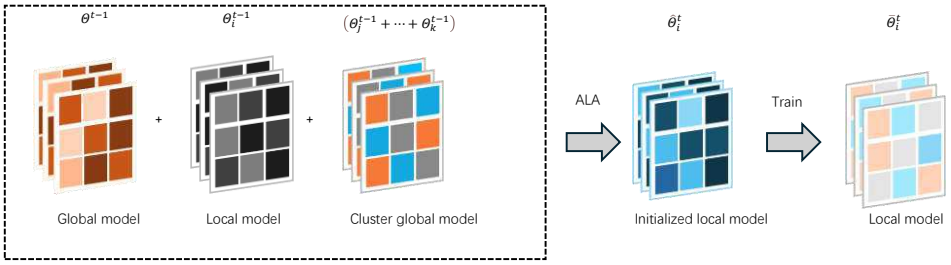


Fig. 2: Workflow of Two-Phase Adaptive Knowledge Aggregation (TPAA): (1) Intra-cluster fusion via ALA_Intra; (2) Global integration via ALA_Global.

3.4 Cross-Cluster Knowledge Distillation

While clustering enhances intra-cluster collaboration, it may isolate inter-cluster knowledge. We address this via cross-cluster knowledge distillation (CCKD), enabling transfer of complementary features across clusters (Figure 3).

3.4.1 Cluster Knowledge Model Generation

At the end of each training round, the server aggregates intra-cluster fused models to generate a cluster-level knowledge model for $\mathcal{C}_k$:

$$\Theta_{\text{cluster}}^{(k),t} = \frac{\sum_{i \in \mathcal{C}_k} |\mathcal{D}_i| \cdot \widetilde{\Theta}_i^t}{\sum_{i \in \mathcal{C}_k} |\mathcal{D}_i|}, \tag{5}$$

where $\widetilde{\Theta}_i^t$ is the intra-cluster fused model of hospital i (from Eq. 3), capturing collective diagnostic patterns of the cluster's dominant diseases.

3.4.2 Inter-Cluster Knowledge Distillation

Each cluster model $\Theta_{\text{cluster}}^{(k),t}$ acts as a student, learning from models of other clusters $\{\Theta_{\text{cluster}}^{(j),t} \mid j \neq k\}$ (teachers). Using distillation samples $\mathcal{D}_{\text{distill}}^{(k)}$ (5

$$p_{\text{teacher}} = \frac{\sum_{j \neq k} w_{t_j} \cdot p_{t_j}}{\sum_{j \neq k} w_{t_j}}, \tag{6}$$

where $w_{t_j} = \text{Acc}(\Theta_{\text{cluster}}^{(j),t}, \mathcal{D}_{\text{distill}}^{(k)})$ denotes teacher j's accuracy on $\mathcal{D}_{\text{distill}}^{(k)}$.
The student learns via a hybrid distillation loss:

$$\mathcal{L}_{\text{distill}} = \lambda_1 \cdot \tau^2 \cdot \text{KL}(p_s \parallel p_{\text{teacher}}) + \lambda_2 \cdot \|\phi_s - \phi_{\text{teacher}}\|^2, \tag{7}$$

where $\tau = 3$ (temperature), $\lambda_1 = 0.5$, $\lambda_2 = 0.5$ (hyperparameters validated in Sec. 4.3), $\phi_s/\phi_{\text{teacher}}$ are intermediate features, ensuring both output probabilities and feature representations are transferred.

3.4.3 Global Knowledge Model Update

Post-distillation, the global model is updated by aggregating enhanced cluster models:

$$\Theta^t = \frac{\sum_{k=1}^{K} |\mathcal{C}_k| \cdot \Theta_{\text{cluster}}^{(k),t}}{\sum_{k=1}^{K} |\mathcal{C}_k|}, \tag{8}$$

incorporating intra-cluster specialization and cross-cluster complementarity. This model is broadcast to all nodes for the next iteration, completing the cycle.

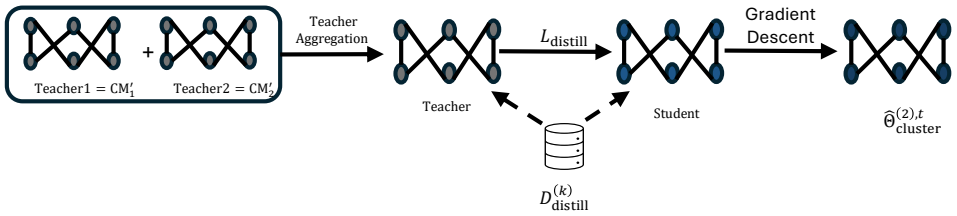


Fig. 3: Workflow of Cross-Cluster Knowledge Distillation (CCKD): (1) Cluster model generation; (2) Teacher ensemble; (3) Knowledge transfer via distillation loss.

4 Experiments

4.1 Experimental Setup

We evaluate the effectiveness of the proposed FedCKD in highly heterogeneous medical imaging scenarios and compare it against several baseline federated learning methods. These include conventional approaches such as FedAvg [7] and FedProx [8], as well as personalized methods like Per-FedAvg [23], FedRep [25], pFedMe [12], Ditto [1], FedAMP [22], FedPHP [26], FedFomo [27], APPLE [24], PartialFed [28], FedCAC [29], and FedALA [11]. This comparison highlights the advantages of FedCKD in handling heterogeneous knowledge representations, protecting privacy, and enhancing personalized diagnostic performance.

We focus on computer vision tasks using MNIST, CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets (100,000 images over 200 classes). To study the impact of neural network architectures on knowledge extraction, we test two backbones on Tiny-ImageNet: a 4-layer CNN and ResNet-18.

All experiments are performed on a uniform GPU server environment to ensure fair comparisons. The local training batch size is fixed at 10, and each client performs one step of SGD per round. Learning rates are 0.005 for the 4-layer CNN, 0.1 for MNIST tasks, and 0.1 for ResNet-18. There are 1000 global iterations with 20 clients, and all clients participate in each round ($\rho = 1$), consistent with FedALA [11].

To simulate Non-IID scenarios, we use two data partitioning strategies. The *pathological heterogeneous setting* assigns each client 2–10 nonoverlapping classes (for MNIST,

CIFAR-10, and CIFAR-100), creating fragmented knowledge distributions. The *practical heterogeneous setting* uses a Dirichlet distribution, Dir(β), with $\beta = 0.1$ to simulate extreme heterogeneity [30].

For evaluation, we report the best global model accuracy on the test set for conventional methods, while for personalized methods, the average accuracy across all local test sets is computed. Each client reserves 25% of its data for testing, with 75% for training. Hyperparameters for FedCKD are kept identical to baseline methods to ensure fairness. Further, we analyze the impact of key parameters (e.g., temperature coefficient, cluster count, and distillation sample size) and conduct ablation studies to assess the contributions of each module.

4.2 Performance Comparison and Analysis

4.2.1 Experimental Evaluation on Medical Dataset: Kvasir

To assess the practical applicability of our method in medical knowledge systems, we conducted experiments on the Kvasir dataset, a widely used endoscopic image collection for gastrointestinal disease classification[31]. The dataset exhibits a high degree of heterogeneity, making it suitable for evaluating federated learning methods under non-IID knowledge distribution settings.

As illustrated in Figure 4, FedCKD consistently outperforms existing baselines, including FedAvg, FedProx, Ditto, and SCAFFOLD, in terms of test accuracy and convergence stability. Notably,FedCKD achieves near-optimal accuracy within the first 100 communication rounds and maintains performance above 90% throughout the training process. In contrast, methods like SCAFFOLD struggle in highly skewed data scenarios, with performance fluctuating around 60%.

These results further validate the effectiveness of our proposed cross-cluster distillation and intra-group aggregation mechanisms in handling real-world non-IID distributions, particularly in medical image analysis tasks where knowledge extraction from heterogeneous data sources is critical.

4.2.2 Experimental Evaluation on Medical Dataset: ISIC

To further evaluate the robustness and generalizability of FedCKD in real-world medical scenarios, we conducted additional experiments on the ISIC 2018 dataset, a widely adopted dermoscopic image collection for skin lesion classification [32]. The dataset contains diverse lesion categories, such as melanoma, nevus, and basal cell carcinoma, presenting significant inter-class visual heterogeneity.

Following the same experimental configuration as in the Kvasir setting, we simulate non-IID conditions by partitioning data across clients based on lesion types, ensuring each client receives only a subset of classes. This setup mimics real-world deployment where dermatology centers may specialize in diagnosing specific lesion types.

As illustrated in Figure 5, FedCKD exhibits a clear advantage over other baseline methods, including FedAvg, FedProx, Ditto, and SCAFFOLD. It demonstrates

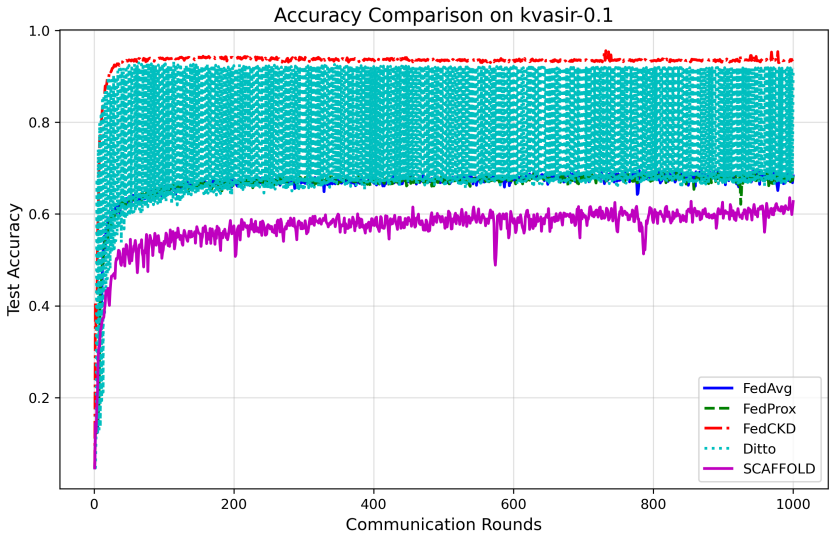


Fig. 4: Accuracy Comparison on Kvasir-0.1 under Non-IID Setting: FedCKD demonstrates superior convergence speed and stable knowledge extraction performance compared to baseline methods.

consistently faster convergence and higher final test accuracy. Notably, FedCKD surpasses 70% accuracy within the first 200 communication rounds, while other methods converge more slowly and plateau at lower levels, particularly under severe data imbalance. These results reaffirm the efficacy of FedCKD's hierarchical aggregation and cross-cluster distillation mechanisms in managing skewed data distributions across heterogeneous medical clients.

4.2.3 Heterogeneous Setting

In both pathological and practical heterogeneous settings, knowledge distributions among service nodes (hospitals) are highly fragmented, either due to disease-specific categories or imbalanced local data. In the pathological heterogeneity scenario, nodes are partitioned into mutually exclusive category groups—each node holds only a few disease categories—severely undermining the generalization of the global knowledge model. Existing methods like FedALA [11] focus on adaptive aggregation at the client level but primarily adjust locally without effective cross-node knowledge collaboration.

In contrast, FedCKD combines intra-group aggregation and cross-cluster service knowledge distillation. Intra-cluster aggregation reinforces learning of dominant disease features within each cluster, while cross-cluster distillation enables clusters to share complementary knowledge features without exposing raw images. As shown in Table 1, FedCKD consistently outperforms other methods on MNIST, CIFAR-10, CIFAR-100 (pathological partitioning) and on CIFAR-10, CIFAR-100, Tiny-ImageNet (Dirichlet $\alpha = 0.1$), demonstrating superior robustness in knowledge extraction under extreme heterogeneity.

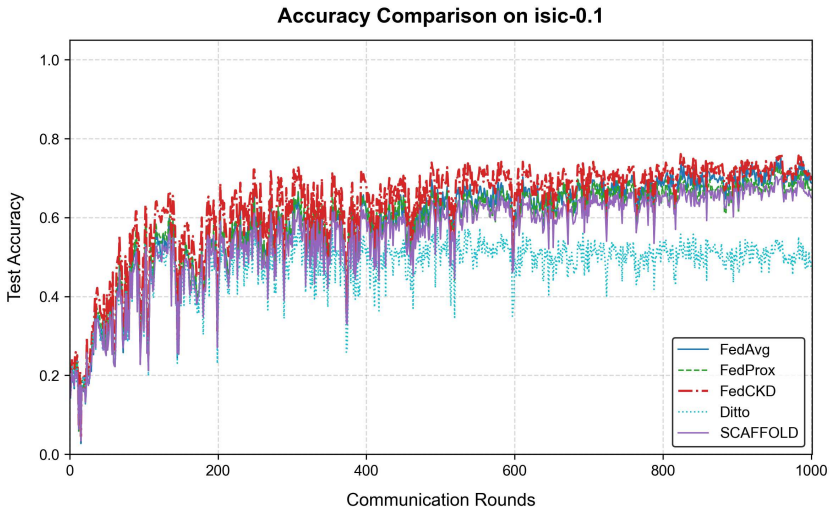


Fig. 5: Accuracy Comparison on ISIC-0.1 under Non-IID Setting: FedCKD consistently outperforms baseline methods such as FedAvg, FedProx, Ditto, and SCAFFOLD in convergence speed and final accuracy, demonstrating robustness against severe label distribution shift in skin lesion classification.

Table 1: Global Accuracy (%) under Pathological and Practical Heterogeneity

Method	Pathological			Practical ($\alpha = 0.1$)		
	MNIST	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100	Tiny-ImageNet
FedAvg	97.90	55.45	25.28	59.22	31.88	19.43
FedAmp	99.72	90.12	63.68	88.65	47.59	27.89
FedProx	98.04	55.18	25.18	59.18	31.95	19.39
PerAvg	99.60	88.94	55.93	87.80	44.23	25.02
MOON	96.69	56.33	35.65	59.50	34.58	30.43
Ditto	99.79	91.28	67.20	90.49	52.77	32.36
APPLE	99.70	90.36	65.97	89.87	53.52	35.01
FedALA	99.86	91.29	65.96	90.20	55.56	40.51
FedCAC	99.77	89.04	66.49	89.43	47.52	33.95
FedCKD	**99.89**	**92.39**	**68.86**	**91.76**	**56.93**	**44.01**

4.2.4 Local Accuracy in Pathological/Practical Heterogeneity

Table 2 reports the peak local accuracy of each method under both pathological partitioning and practical Dirichlet heterogeneity ($\alpha = 0.1$). FedCKD achieves the best performance in both scenarios across all benchmarks, demonstrating its robustness and adaptability to highly skewed knowledge distributions.

Under the pathological split, on MNIST FedCKD attains 99.89% accuracy, slightly improving over FedCAC (99.77%) and FedALA (99.86%). On CIFAR-10, the gap

is more pronounced: FedCKD reaches 92.39%, outperforming FedCAC (91.80%), APPLE (90.36%), and MOON (56.63%). For CIFAR-100 the advantage is again clear, with FedCKD at 68.86%, compared to FedCAC's 66.49% and FedALA's 65.96%.

In the practical Dirichlet setting ($\alpha = 0.1$), FedCKD similarly leads: achieving 91.76% on CIFAR-10 (next best FedCAC 89.43%), 56.93% on CIFAR-100 (next best FedCAC 47.62%), and 44.01% on Tiny-ImageNet (next best FedCAC 33.95%). Notably, on the hardest benchmark, Tiny-ImageNet, FedCKD's 44.01% represents a more than 10-point improvement over APPLE (30.51%) and a full 10% gain over FedAvg (19.43%).

These results confirm that the combination of disease-profile clustering, intra-cluster aggregation and cross-cluster knowledge distillation in FedCKD effectively mitigates severe non-IID effects, yielding significant accuracy gains in knowledge extraction under both extreme and realistic heterogeneity.

Table 2: Local Accuracy (%) under Pathological and Practical Heterogeneity

Methods	Pathological			Practical ($\alpha = 0.1$)		
	MNIST	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100	Tiny-ImageNet
FedAvg	97.90	55.45	25.28	59.22	31.88	19.43
FedAmp	99.72	90.12	63.68	88.65	47.59	27.89
FedProx	98.04	55.18	25.18	59.18	31.95	19.39
PerAvg	99.60	88.94	55.93	87.80	44.23	25.02
MOON	96.69	56.33	35.65	59.50	34.58	30.43
Ditto	99.79	91.28	67.20	90.49	52.77	32.36
APPLE	99.70	90.36	65.97	89.87	53.52	35.01
FedALA	99.86	91.29	65.96	90.20	55.56	40.51
FedCAC	99.77	89.04	66.49	89.43	47.52	33.95
FedCKD	**99.89**	**92.39**	**68.86**	**91.76**	**56.93**	**44.01**

4.2.5 Heterogeneity Analysis

To evaluate the performance of various federated learning methods under different levels of knowledge heterogeneity, we conducted experiments on the Tiny-ImageNet dataset using the Dirichlet distribution parameter β. When β is small (e.g., $\beta = 0.01$), the knowledge distribution among clients exhibits high heterogeneity, causing most personalized federated learning methods to outperform traditional approaches due to enhanced local adaptation. However, some methods that overly focus on personalization and neglect the importance of global shared knowledge experience performance degradation when the heterogeneity decreases to a moderate level (e.g., $\beta = 0.5$), where traditional methods like FedAvg yield more stable results. In contrast, FedCKD consistently demonstrates superior performance across the full spectrum of heterogeneity, from highly heterogeneous to moderately heterogeneous knowledge distributions. This underscores the effectiveness of its clustering strategy and cross-cluster distillation mechanism in balancing local knowledge personalization with global knowledge sharing.

4.2.6 Scalability Analysis

To assess the robustness of federated learning methods in large-scale knowledge systems, we increased the number of clients from 50 to 100 while keeping the total dataset size constant. This configuration introduces a significant small-sample challenge, as the amount of data available to each client decreases. While most personalized federated learning methods experience noticeable declines in accuracy as the number of clients increases, FedCKD exhibits exceptional scalability, with its performance degrading by less than 1%. This robustness is primarily attributed to its cross-cluster knowledge distillation mechanism, which successfully integrates global feature sharing with local adaptation. Consequently, FedCKD maintains strong generalization and robustness even in scenarios with sparse data and high heterogeneity. The detailed results for heterogeneity and scalability are summarized in Table 3, where FedCKD consistently achieves the best performance across all settings.

Table 3: Performance comparison of federated learning methods

Methods	**Heterogeneity (Tiny-ImageNet)**		**Scalability (CIFAR-100)**	
	Dir(0.01)	Dir(0.5)	50 Clients	100 Clients
FedAvg	15.70	21.14	31.90	31.95
FedProx	15.66	21.22	31.94	31.97
FedAvg-C	49.88	16.21	49.82	47.90
FedProx-C	49.84	16.36	49.79	48.02
Per-FedAvg	39.39	16.36	44.31	36.07
FedRep	55.43	16.74	47.41	44.61
pFedMe	41.45	17.48	48.36	46.45
Ditto	50.62	18.98	54.22	52.89
FedAMP	48.42	12.48	44.39	40.43
FedPHP	48.63	21.09	52.44	49.70
FedFomo	46.36	11.59	42.56	38.91
APPLE	48.04	24.28	55.06	52.81
PartialFed	49.38	24.20	48.95	39.31
FedALA	55.75	27.85	55.61	54.68
FedCAC	50.84	25.39	50.72	49.85
FedCKD	**57.81**	**28.87**	**58.30**	**57.68**

4.3 Effect of Hyperparameters

4.3.1 Distillation Temperature T

To evaluate the impact of distillation temperature on knowledge transfer accuracy, we conducted experiments on the CIFAR-100 dataset. As shown in Figure 6a, increasing T from 1 to 3 led to a steady improvement in accuracy, from 65.32% to 68.05%. This improvement can be attributed to the smoothing effect of higher temperatures, which enhances the student's ability to learn inter-cluster knowledge relationships from the

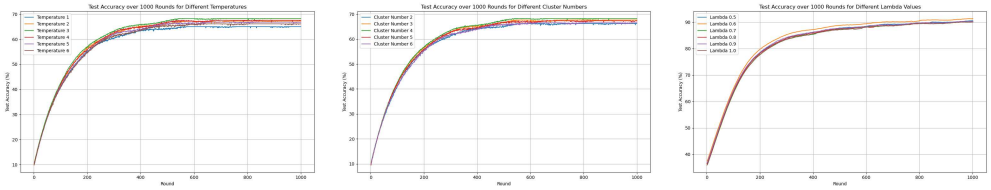
teacher's outputs. However, further increases in T to 4, 5, and 6 resulted in declining accuracy, suggesting that excessively high temperatures reduce the discriminative power of the teacher's output distribution. Based on these findings, we adopt $T = 3$ as the optimal distillation temperature for subsequent experiments.

4.3.2 Number of Clusters K

We analyzed the effect of varying the number of clusters K on the CIFAR-100 dataset. As shown in Figure 6b, increasing K from 2 to 4 improved accuracy, with the highest accuracy of 68.55% observed at $K = 4$. This indicates that increasing K up to a moderate value improves intra-cluster similarity, thereby enhancing the quality of knowledge distillation. However, further increasing K to 5 or 6 reduced accuracy, likely due to insufficient data in each cluster for effective knowledge extraction. Therefore, we set $K = 4$ as the default value for clustering in subsequent experiments.

4.3.3 Feature Loss Weight λ

To investigate the effect of the feature loss weight λ on knowledge transfer, we conducted experiments on the CIFAR-10 dataset. As shown in Figure 6c, increasing λ from 0.5 to 0.6 improved accuracy from 91.08% to 91.88%, with the peak performance at $\lambda = 0.6$. This suggests that a moderate weighting of the feature alignment loss enables the student model to better capture the teacher's knowledge representations. However, further increases in λ to 0.7 and above caused minor fluctuations and failed to surpass the performance at $\lambda = 0.6$, likely due to reduced adaptability to local knowledge features. Consequently, we select $\lambda = 0.6$ as the optimal weight for feature loss in our experiments.



(a) Accuracy vs. distillation temperature T. Optimal at $T = 3$.

(b) Accuracy vs. number of clusters K. Best at $K = 4$.

(c) Accuracy vs. feature loss weight λ. Optimal at $\lambda = 0.6$.

Fig. 6: Effects of hyperparameters on knowledge extraction accuracy: (a) distillation temperature, (b) cluster number, and (c) feature loss weight.

4.4 Ablation Study

To quantify the contributions of clustering and cross-cluster knowledge distillation, we perform ablation experiments on four datasets (MNIST, CIFAR-10, CIFAR-100, Tiny-ImageNet), evaluating both global and local model accuracy under four configurations: **Base** (no clustering or distillation), **Clustering Only** (only client-side label

clustering), **Clustering + Distillation*** (clustering plus distillation, without aggregating the distilled cluster models into a global model; distilled models sent directly to clients), and **Clustering + Distillation** (clustering plus distillation, with the distilled cluster models first aggregated into a global model and then distributed to clients).

Table 4: Global Model Accuracy (%) on Four Datasets

Configuration	MNIST	CIFAR-10	CIFAR-100	Tiny-ImageNet
Base	99.86	90.27	55.85	40.54
+ Clustering Only	99.88	91.18	56.55	43.68
+ Clustering + Distillation*	98.35	82.76	30.82	28.68
+ Clustering + Distillation	**99.90**	**91.55**	**56.95**	**44.02**

Table 5: Local Model Accuracy (%) on Four Datasets

Configuration	MNIST	CIFAR-10	CIFAR-100	Tiny-ImageNet
Base	99.81	91.10	55.98	40.33
+ Clustering Only	99.84	91.77	56.83	43.49
+ Clustering + Distillation*	98.03	82.76	30.76	28.35
+ Clustering + Distillation	**99.88**	**91.93**	**57.29**	**44.08**

The ablation results in Tables 4 and 5 reveal two main insights. First, client-side clustering yields immediate gains in both global and local accuracy, especially on challenging benchmarks (e.g., +3.14% globally and +3.16% locally on Tiny-ImageNet), confirming that intra-cluster aggregation mitigates statistical heterogeneity in knowledge representation. Second, adding cross-cluster distillation and aggregating the distilled models further boosts accuracy (up to +0.37–1.34 points globally and +0.16–0.78 points locally), demonstrating the value of inter-cluster knowledge sharing. Overall, the combination of adaptive clustering and distillation forms a robust framework for both knowledge personalization and generalization under non-IID data distributions.

5 Conclusion

The comprehensive experiments demonstrate that FedCKD consistently outperforms both conventional and state-of-the-art personalized federated learning methods across diverse non-IID scenarios and datasets. In extreme pathological and practical heterogeneity settings, FedCKD's combination of client-side clustering and cross-cluster distillation yields substantial gains in both global and local accuracy for knowledge extraction tasks. The ablation study further confirms that clustering alone already mitigates statistical heterogeneity by adapting aggregation to label distributions, while

the addition of distillation—especially when the distilled cluster models are aggregated into a global model—provides complementary knowledge sharing that refines performance even further.

FedCKD addresses the critical challenges in federated medical knowledge systems by: (1) effectively handling non-IID knowledge distributions through label-driven clustering; (2) balancing local specialization with global generalization through two-phase adaptive aggregation; and (3) enabling privacy-preserving knowledge transfer across specialized medical domains via cross-cluster distillation. These capabilities make FedCKD particularly well-suited for knowledge-intensive applications in healthcare information systems, where both privacy preservation and knowledge integration across heterogeneous data sources are paramount.

Future work will focus on expanding the framework to adapt to the evolution of dynamic knowledge in medical systems, where disease patterns and institutional models may change over time. In addition, we plan to explore more sophisticated data heterogeneous types to enhance the adaptability of this method to various medical scenarios, which is a key requirement for clinical application of the medical federated learning system.

References

[1] Sheller, M.J., Reina, G.A., Edwards, B., *et al.*: Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. Medical Image Analysis **53**, 136–148 (2019)

[2] Zhang, F., Li, Y., Liu, X.: Recent advances in federated learning for medical imaging: A survey. Artificial Intelligence in Medicine **144**, 102258 (2024) https://doi.org/10.1016/j.artmed.2023.102258

[3] Xu, H., Wang, L., Zhang, Y.: Federated learning meets edge computing: A hierarchical framework for smart healthcare systems. IEEE Transactions on Services Computing **16**(1), 102–115 (2023) https://doi.org/10.1109/TSC.2022.3145678

[4] Kairouz, P., McMahan, B., Avent, B., *et al.*: Advances and open problems in federated learning. Foundations and Trends® in Machine Learning **14**(1–2), 1–210 (2021)

[5] Sheller, M.J., Reina, G.A., Edwards, B., *et al.*: Federated learning in medical imaging: Part i: Toward multicentral health care ecosystems. Academic Radiology **29**(3), 456–467 (2022)

[6] Zhao, Y., Li, M., Lai, L., et al.: Federated learning with non-iid data. arXiv preprint arXiv:1806.00582 (2018)

[7] McMahan, B., Moore, E., Ramage, D., *et al.*: Communication-efficient learning of deep networks from decentralized data. In: AISTATS (2017)

[8] Li, T., Sahu, A.K., Zaheer, M., *et al.*: Federated optimization in heterogeneous networks. In: Proceedings of MLSys (2020)

[9] Karimireddy, S.P., Kale, S., Stich, S.U., Jaggi, M.: Scaffold: Stochastic controlled averaging for federated learning. In: Proceedings of the 37th International Conference on Machine Learning (ICML), pp. 5132–5143 (2020). http://proceedings.mlr.press/v119/karimireddy20a.html

[10] Wang, J.D., Joshi, G.: Tackling the objective inconsistency problem in heterogeneous federated optimization. In: Proceedings of NeurIPS 2020, pp. 7611–7623 (2020). https://proceedings.neurips.cc/paper/2020/file/8ceb3363c96bd7c2f277843bf6b392ca-Paper.pdf

[11] Zhang, J., Hua, Y., Wang, H., Song, T., Xue, Z., Ma, R., Guan, H.: Fedala: Adaptive local aggregation for personalized federated learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 11237–11244 (2023)

[12] Dinh, T., Tran, N., Nguyen, T.: Personalized federated learning with moreau envelopes. In: NeurIPS (2020)

[13] Arivazhagan, M., Aggarwal, A., Singh, A., Balan, P.: Federated learning with personalization layers. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 32, pp. 12226–12237 (2019). https://proceedings.neurips.cc/paper/2019/file/1a7f0c8803274393fd2c7de9cf618a0f-Paper.pdf

[14] Li, T., He, J., Song, S., Gong, Q., Liu, X.: Ditto: Fair and robust federated learning through personalization. In: International Conference on Learning Representations (ICLR) (2021). https://openreview.net/forum?id=7Byf7M4YwH

[15] Chen, Z., Liu, Y., Wu, X.: Fedclusternet: Federated learning with attention-guided cluster knowledge transfer. Knowledge-Based Systems **281**, 110092 (2024) https://doi.org/10.1016/j.knosys.2023.110092

[16] Blum, J., Lee, K., Li, Y., Kanza, Y., Nagaraja, S.: Federated learning through a model personalization lens. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 32, pp. 13880–13892 (2019). https://proceedings.neurips.cc/paper/2019/file/3c49c7d5d50d72cfdafb75c6a76b4569-Paper.pdf

[17] Sattler, F., Wiedemann, S., Müller, K.-R., Samek, W.: Clustered federated learning: Model-agnostic distributed multi-task optimization under privacy constraints. In: International Conference on Machine Learning (ICML), pp. 879–889 (2020). http://proceedings.mlr.press/v119/sattler20a.html

[18] Zhang, Y., Li, X., Wang, L.: Hierarchical knowledge distillation for federated learning in heterogeneous medical systems. IEEE Transactions on Medical Imaging **43**(5), 1325–1336 (2024) https://doi.org/10.1109/TMI.2023.3325411

[19] Li, D., Wang, J.: Fedmd: Heterogeneous federated learning via model distillation. arXiv preprint arXiv:1910.03581 (2019)

[20] He, J., Goodfellow, I., Hou, X.: Group knowledge transfer: Federated learning of large cnns at the edge. In: International Conference on Learning Representations (ICLR) (2020). https://openreview.net/forum?id=rkgMmuEKPB

[21] Liang, F., Chen, M., Zhao, R.: A decentralized federated learning framework for privacy-preserving medical image analysis. arXiv preprint arXiv:2206.12345 (2022)

[22] Huang, Y., Chu, L., Zhou, Z., *et al.*: Personalized cross-silo federated learning on non-iid data. In: AAAI (2021)

[23] Fallah, A., Mokhtari, A., Ozdaglar, A.: Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. Advances in neural information processing systems **33**, 3557–3568 (2020)

[24] Luo, J., Wu, S.: Adapt to adaptation: Learning personalization for cross-silo federated learning. arXiv preprint arXiv:2110.08394 (2021)

[25] Wang, H., Yurochkin, M., Sun, Y., et al.: Federated learning with matched averaging. arXiv preprint arXiv:2002.06440 (2020)

[26] Li, X.-C., Zhan, D.-C., Shao, Y., Li, B., Song, S.: Fedphp: Federated personalization with inherited private models. In: ECML PKDD (2021)

[27] Zhang, M., Sapra, K., Fidler, S., Yeung, S., Alvarez, J.M.: Personalized federated learning with first order model optimization. In: ICLR (2021)

[28] Sun, B., Huo, H., Yang, Y., Bai, B.: Partialfed: Cross-domain personalized federated learning via partial initialization. In: NeurIPS (2021)

[29] Wu, X., Liu, X., Niu, J., Zhu, G., Tang, S.: Bold but cautious: Unlocking the potential of personalized federated learning through cautiously aggressive collaboration. In: ICCV (2023)

[30] Zhang, J., Liu, Y., Hua, Y., Wang, H., Song, T., Xue, Z., Ma, R., Cao, J.: Pfllib: A beginner-friendly and comprehensive personalized federated learning library and benchmark. arXiv preprint arXiv:2312.04992 (2023)

[31] Pogorelov, K., Randel, K.R., Griwodz, C., Eskeland, S.L., Lange, T., Johansen, D., Schmidt, P.T., Haugen, H., Lux, M., Riegler, M.A., Halvorsen, P.: Kvasir: A multi-class image dataset for computer aided gastrointestinal disease

detection. In: Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys), pp. 164–169 (2017). https://doi.org/10.1145/3083187.3083212 . https://doi.org/10.1145/3083187.3083212

[32] Codella, N.C., Rotemberg, V., Tschandl, P., al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)