nature communications

Article

https://doi.org/10.1038/s41467-025-64459-y

Addressing data heterogeneity in distributed medical imaging with heterosync learning

Received: 1 November 2024

Accepted: 3 September 2025

Published online: 24 October 2025

Check for updates

Hang-Tong Hu[1,8,9], Ming-De Li[1,8], Xin-Xin Lin[1], Meng-Yao Cai[1], Shuai Liu[2], Shao-Hong Wu[1], Wen-Juan Tong[1], Feng-Yu Ye[2], Jin-Bo Hu[2], Wei-Ping Ke[1], Li-Da Chen[1], Hong Yang[3], Guang-Jian Liu[4], Hai-Bo Wang[5], Ming-De Lu[1,6], Qing-Hua Huang[7] ✉, Ming Kuang[1,6] ✉, Wei Wang[1] ✉ & Ultrasound Engineering Institute, Medical Industry Branch of China Association Plant Engineering (UE-MICAP)*

Data heterogeneity critically limits distributed artificial intelligence (AI) in medical imaging. We propose HeteroSync Learning (HSL), a privacy-preserving framework that addresses heterogeneity through: (1) Shared Anchor Task (SAT) for cross-node representation alignment, and (2) an Auxiliary Learning Architecture coordinating SAT with local primary tasks. Validated via large-scale simulations (feature/label/quantity/combined heterogeneity) and a real-world multi-center thyroid cancer study, HSL outperforms local learning, 12 benchmark methods (FedAvg, FedProx, SplitAVG, FedRCL, FedCOME, etc.), and foundation models (e.g., CLIP) by better stability and up to 40% in area under the curve (AUC), matching central learning performance. HSL achieves 0.846 AUC on the out-of-distribution pediatric thyroid cancer data (outperforming others by 5.1-28.2%), demonstrating superior generalization. Visualizations confirm HSL successfully homogenizes heterogeneous distributions. This work provides an effective solution for distributed medical AI, enabling equitable collaboration across institutions and advancing healthcare AI democratization.

Developing robust medical artificial intelligence (AI), particularly in medical imaging, aims to ensure performance across diverse patient populations and clinical settings[1]. Distributed learning technologies, such as Federated Learning (FL), have become essential in training AI on extensive multicenter datasets[2,3]. These methods enable collaborative model training within medical alliances with raw data distributed in their original physical locations (i.e., nodes). However, data heterogeneity across varied datasets poses a significant challenge to the effectiveness of collaborative modeling[4,5].

Data heterogeneity in medical imaging is characterized by disparities of three categories: feature distribution, label distribution, and data quantity[6]. For instance, feature distribution skew arises from diverse data sources, variations in disease stages, differences in data collection equipment, and distinct imaging protocols[7–9]. Label

[1]Department of Medical Ultrasonics, Institute of Diagnostic and Interventional Ultrasound, the First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. [2]School of Physics and Electronic Information, Guangxi Minzu University, Nanning, China. [3]Department of Medical Ultrasound, the First Affiliated Hospital of Guangxi Medical University, Nanning, China. [4]Department of Medical Ultrasonics, the Sixth Affiliated Hospital of Sun Yat-sen University (Guangdong Gastrointestinal Hospital), Guangzhou, China. [5]Research Center of Big Data and Artificial Intelligence for Medicine, the First Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China. [6]Center of Hepato-Pancreato-Biliary Surgery, the First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. [7]School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. [8]These authors contributed equally: Hang-Tong Hu, Ming-De Li. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: qhhuang@nwpu.edu.cn; kuangm@mail.sysu.edu.cn; wangw73@mail.sysu.edu.cn

distribution skew occurs when annotations are inconsistent or when certain labels are disproportionately represented in the dataset (e.g., varied disease prevalence)[10]. Quantity skew results from disparities in the number of patient records across different medical institutions (e.g., large-scale center vs. small clinic)[9]. Training models locally on heterogeneous data can lead to divergent updates in local models, and integrating these models, which may differ significantly, can disrupt the optimization process, thereby impairing the performance of the final model. Traditional methods like Federated Averaging (FedAvg) and emerging Swarm Learning fail to adequately address this issue[11,12].

Traditional approaches to heterogeneity mitigation often necessitate risky compromises: Data sharing methods (e.g., sharing subsets of raw data or feature maps) may improve model performance but violate privacy regulations[13–16]. For example, Zhao et al. [14] achieved a 30% accuracy increase on the CIFAR-10 dataset by sharing 5% of the original data globally. On the other hand, algorithm-centric methods like FedProx[17] and FedMoCo[18]) avoid data sharing but typically falter under severe heterogeneity[6]. Even when combining algorithm-centric approaches with data sharing, the performance improvements remain limited in highly heterogeneous datasets[15]. This highlights a fundamental challenge: existing methods treat heterogeneity and privacy as competing priorities, rather than interdependent challenges to be jointly addressed. Additionally, prior studies often focus on narrow skews or mild heterogeneity[15,19], and typically validate on small cohorts[10], limiting their clinical applicability.

To address this, we propose HeteroSync Learning (HSL), a privacy-aware distributed framework that mitigates data heterogeneity through collaborative representation alignment. HSL harmonizes two core components (Fig. 1A and Supplementary Fig. 1): (1) the Shared Anchor Task (SAT), a homogeneous reference task that establishes cross-node representation alignment, and (2) a customized Auxiliary Learning Architecture (MMoE, Multi-gate Mixture-of-Experts[20]) that coordinates the co-optimization of SAT with local primary tasks (e.g., cancer diagnosis). The SAT is a strategically designed that: (i) originates from public datasets (e.g., CIFAR-10, RSNA), (ii) maintains uniform distribution across nodes, (iii) locally co-trains with primary tasks, and (iv) globally synchronizes representation learning. To enhance the contribution of SAT to the primary tasks, we drew inspiration from knowledge distillation techniques and introduced a temperature parameter "T" for MMoE, designed to increase the information entropy of the SAT dataset. Through this dual-component framework, HSL achieves globally generalized model performance across distributed nodes while strictly preserving data of the primary task locally. We hypothesize posits that sharing a privacy-safe public dataset, combined with auxiliary learning, can jointly address both data heterogeneity and privacy concerns. We validate HSL through large-scale simulations and real-world multicenter studies, covering extreme clinical scenarios from small clinics to rare disease regions. By harmonizing heterogeneity and privacy, HSL promotes equitable participation in medical AI development, bridging disparities in imaging protocols, disease prevalence, and data resources. This work represents a step toward democratizing AI-driven healthcare, promoting advancements benefit both resource-rich and underserved populations.

The workflow of HSL in distributed learning is outlined as follows:

1. Local Training: Each node trains the MMoE model on its private primary task data and SAT dataset for a set number of iterations or epochs to generate local parameters.
2. Parameter Fusion: Each node aggregates the shared parameters from all nodes and continues training for additional iterations or epochs, updating local parameters.
3. Iterative Synchronization: Steps 1-2 repeat until convergence.

In this work, we introduce HSL, a privacy-preserving distributed learning framework designed to mitigate data heterogeneity in medical imaging through two core components: SAT for cross-node representation alignment and an auxiliary learning architecture that coordinates SAT with the primary task. The key contributions of this work include a harmonized learning framework that effectively aligns representations across heterogeneous nodes without sharing raw data, a dual-task coordination mechanism that improves model generalization and stability, extensive validation across both simulated and real-world data regimes demonstrating superiority over 12 federated and foundation model baselines, and the ability to achieve central learning-comparable performance while preserving privacy and supporting scalable collaboration across healthcare institutions.

Results

Simulation study: controlled validation of heterogeneity scenarios

The MURA dataset comprises 39,168 musculoskeletal radiographs for abnormality diagnosis of seven body sites: Finger, Hand, Elbow, Forearm, Humerus, Shoulder, and Wrist. We utilize it to highlight data heterogeneity across three dimensions: feature distribution, label distribution, data quantity, and their combination. Efficiency of HSL is compared with four classical methods for handling distributed data heterogeneity, including personalized learning[21], FedBN[22], FedProx[17], and SplitAVG[23]. (Fig. 1B)

1. Feature distribution skew: Radiographs from different anatomical regions (e.g., elbow vs. hand) are assigned to separate nodes, while label distribution (normal: abnormal = 1:1) and quantity (n1 = n2…=n7 = 1470) keeps consistency. Distributed learning is conducted among the 7 nodes for abnormality diagnosis (Fig. 2 A1). Figure 2 A2 (Supplementary Table 3) demonstrates that HSL consistently outperforms the other three learning methods across most nodes, except for being comparable to SplitAVG in nodes 3-6. HSL shows good performance stability (narrow bar length) compared with personalized learning in all nodes.
2. Label distribution skew: Radiographs from the same anatomical region (shoulder) are assigned to two distributed modes, with node 1 maintained a constant label distribution (1:1), while node 2 introduced gradients of change as 2:1, 4:1, 6:1, 8:1, 10:1, 20:1, 40:1, 60:1, 80:1, and 100:1. The quantity of both nodes are 3970 across all gradients (Fig. 2 B1). Figure 2 B2 demonstrates that as the label distribution skew increases, FedBN and FedProx experience a decline in performance (Supplementary Tables 4–5). HSL consistently outperformed FedBN, FedProx, and SplitAVG in terms of efficacy and stability. Personalized learning, which uses the backbone of HSL, achieved performance comparable to HSL.
3. Quantity skew: Radiographs from the same anatomical region (shoulder) are assigned to two distributed modes, with the data quantity ratio between them varied as 1:1, 2:1, 4:1, 6:1, 8:1, 10:1, 20:1, 40:1, 60:1, and 80:1, representing different degrees of skew. The total quantity of data across both nodes in each gradient summed up to 8,678. The label distribution was maintained at 1:1 in both nodes across all gradients (Fig. 2 C1). Figure 2 C2 (Supplementary Tables 6–7) indicates that HSL consistently exhibits the best performance across gradients in both nodes.
4. Combined heterogeneity: Five nodes are set mimicking a large-scale screening center, a large-scale disease-specialized hospital, two small clinics, and rare disease regions. Screening center stands for a healthcare facility specializing in population-wide disease screening, where the majority of examined cases are normal/benign, and only a small subset is abnormal. Rare disease regions have a prevalence of abnormal cases of less than 1 in 2000 according to European standards. Distributed learning is conducted among these five nodes. HSL is implemented to mitigate the influence of all three types of skews (Fig. 2 D1 and Supplementary Table 8). Figure 2 D2 demonstrates that HSL consistently outperforms all four classical methods in terms of

https://doi.org/10.1038/s41467-025-64459-y

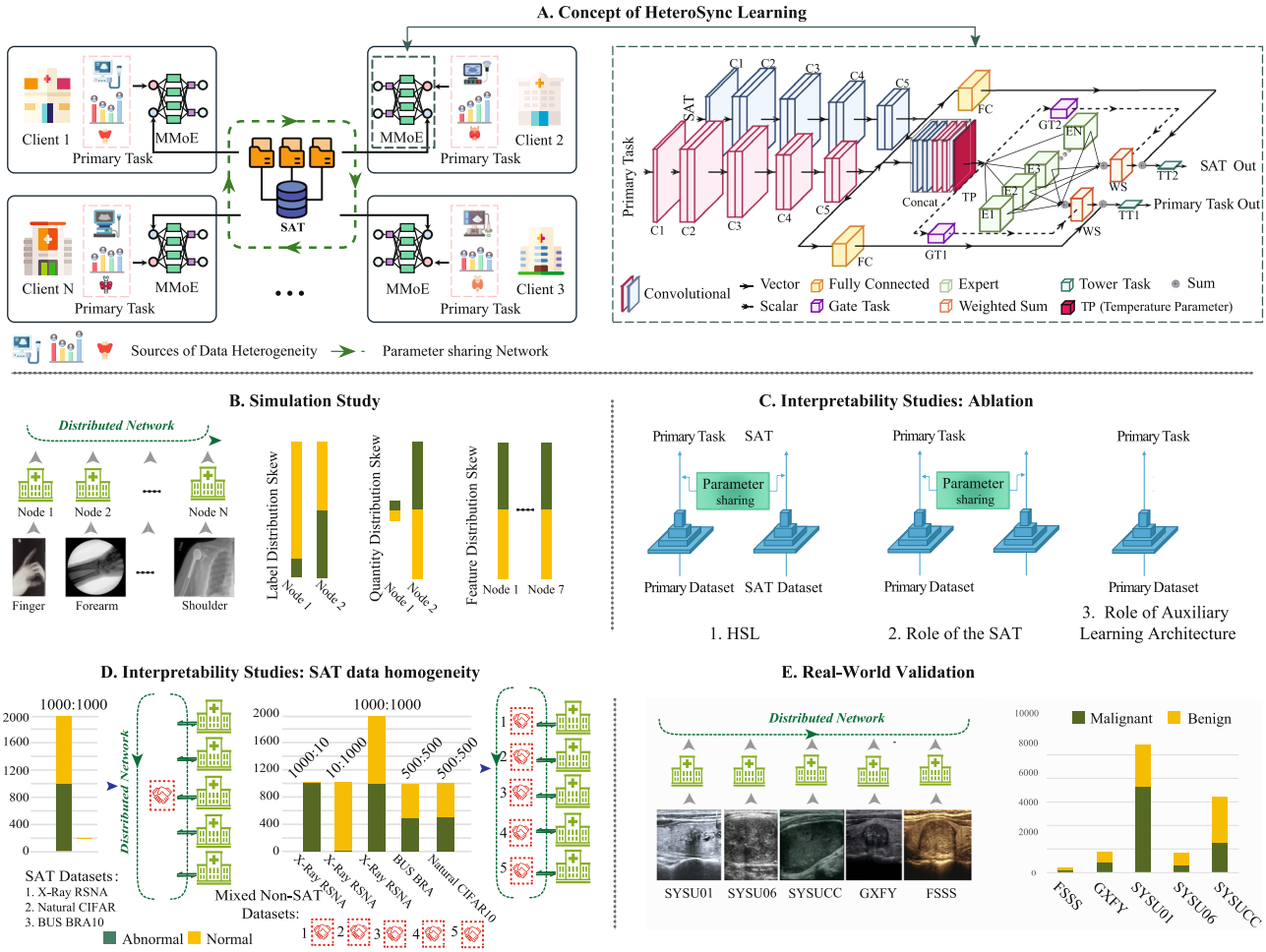

Fig. 1 | Overview of the study design. A Schematic and architecture of HSL: The four heterogeneous nodes of the primary task collectively utilize SAT for collaborative model training; Two separate ResNet18 branches are integrated using the MMOE architecture. **B** Schematic map of the simulation experiments. **C** Schematic map of the interpretability studies revealing role of SAT and auxiliary learning architecture. **D** Schematic map of the interpretability studies revealing the impact of SAT data homogeneity. **E** Schematic map of the Real-World Validation. Bar plots in **B**, **D**, **E** use colored bars (blue/yellow) to show both data volume (length) and label proportions (color segments), enabling direct comparison of dataset sizes and class distributions across nodes. Source data are provided as a Source data file. Figures are created using Adobe Illustrator 2023.

efficacy and stability. Specially, in the rare disease region node, classical methods show the poorest efficiency or stability, while HSL keeps good performance (Fig. 2 D2 and Supplementary Table 9).

Interpretability study: decoupling HSL components

In the combined heterogeneity scenario, HSL components are decoupled to test the contribution of the auxiliary learning architecture and SAT, and the importance of SAT data homogeneity. (Fig. 1C, D)

1. Role of SAT: Removing SAT while retaining the auxiliary learning architecture (Fig. 3A 2 No-SAT). Figure 3A (Supplementary Table 1) illustrates decreased model efficacy across most nodes, particularly in the rare disease region. Conversely, the performance in the two small clinic nodes remained unaffected.
2. Role of the Auxiliary Learning Architecture: Removing the auxiliary learning architecture (the SAT data plays no role in training) (Fig. 1C 3 NO-AL). Figure 3A (Supplementary Table 1) shows a pronounced drop in model efficacy across all nodes, with the rare disease region node experiencing the greatest decline. Additionally, model performance exhibited increased instability as indicated by higher variance (larger bar lengths) (Supplementary Table 1).
3. Importance of SAT data homogeneity: Fig. 3B shows that by replacing the homogenously distributed X-Ray RSNA SAT data with homogenously distributed CIFAR-10 or BUS-BRA SAT data, the performance keeps well and stable both in the repeated trials of a single node and trials across the five nodes. While by replacing the homogenously distributed SAT data with heterogeneously distributed multiple auxiliary datasets (mixed non-SAT datasets), performance drops and becomes unstable. (Supplementary Table 2).

We further visualize the data distribution to more intuitively demonstrate the effect of HSL-balanced distributed data heterogeneity, rather than just model performance. Figure 3B depicts the heterogeneous data distribution with varied shapes among the five nodes, as well as chaotic distribution within each single hospital. After training with HSL (Fig. 3C), the data distribution transforms into a similar shape, which indicates homogeneity. While by removing the auxiliary learning architecture (along with the SAT data) or removing SAT data only (Fig. 3C-2/3), the distribution among the five nodes remains heterogeneous with different shapes.

https://doi.org/10.1038/s41467-025-64459-y

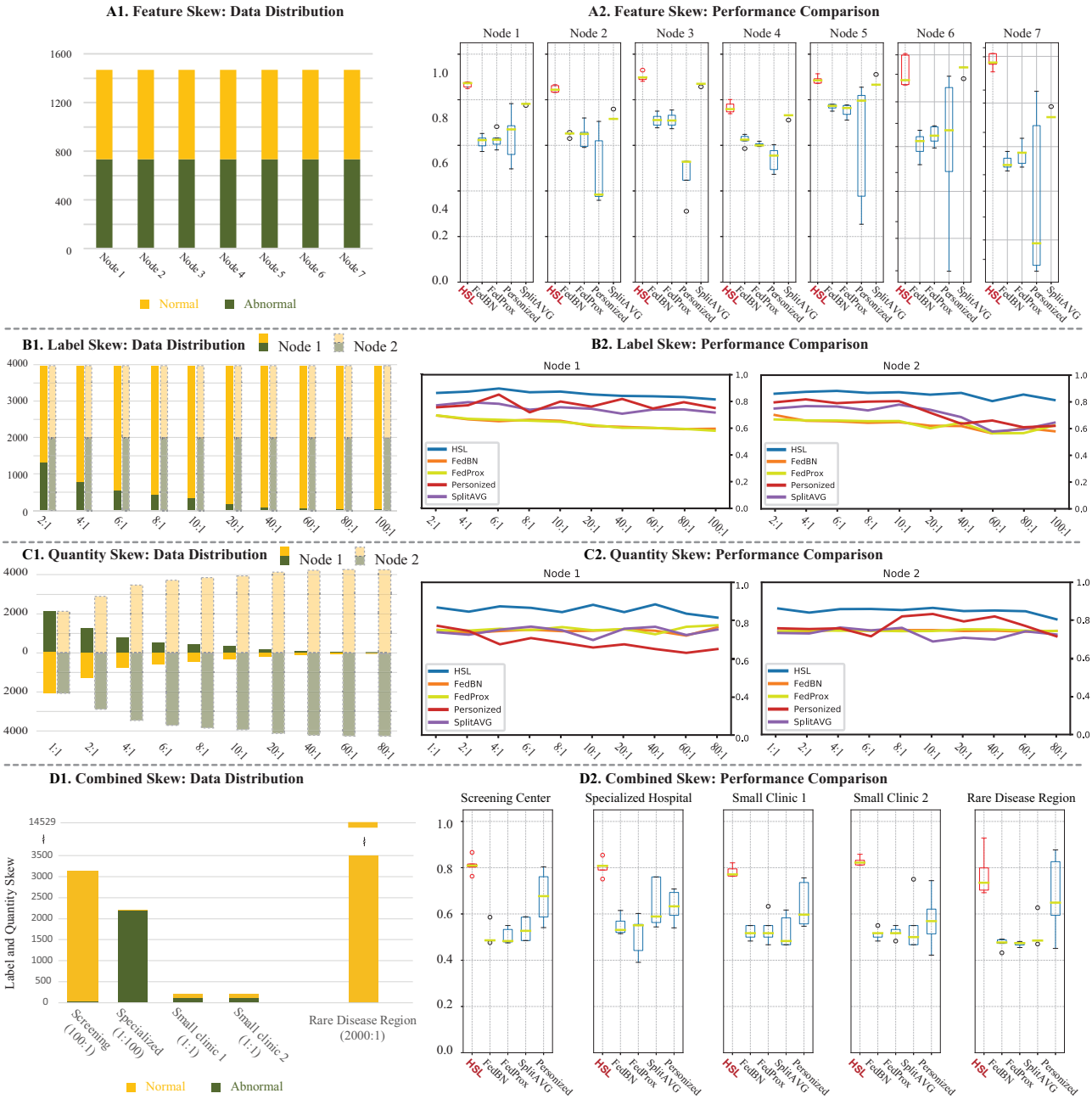

Fig. 2 | Efficacy of HSL vs. classical methods. A1 Data distribution across the 7 nodes for feature skew. **A2** AUC box plots of model testing efficacy on each node ($n = 440$ per testing set) for various learning methods for feature skew. **B1** Data distribution in label distribution gradients of the 2 nodes for label skew. **B2** AUC line charts of model testing efficacy of the 2 nodes ($n = 1191$ per testing set) for label skew. **C1** Data distribution in quantity distribution gradients of the 2 nodes for quantity skew. **C2** AUC line charts of model testing efficacy of the Node 1 (testing set sizes: $n = 1301$, 867, 520, 372, 289, 236, 123, 63, 42, and 31 per gradient) and Node 2 (testing set sizes: $n = 1301$, 1735, 2083, 2232, 2314, 2367, 2479, 2540, 2560, and 2571 per gradient) for quantity skew. **D1** Data distribution in the 5 nodes for combined heterogeneity. **D2** AUC box plots of model testing efficacy across nodes under combined heterogeneity (testing set sizes: $n = 938$, 665, 60, 60, and 4279). Box plots in **A2**, **D2** show the median (central line), the 25th and 75th percentiles (box bounds), and the minimum and maximum values excluding outliers (whiskers). Outliers are shown as individual points. Each bar represents the results of five repeated experiments for a learning method. Source data are provided as a Source data file. Figures aregenerated using Python and subsequently composited in Adobe Illustrator 2023.

Real-world validation: clinical efficacy and generalization

The THYROID and PTC (Pediatric Thyroid Cancer) datasets are an alliance-based multicenter dataset for thyroid cancer diagnosis, containing 20,043 ultrasound images collected from 5 hospitals across 3 municipalities in China. Models are trained within the alliance and tested in two aspects: intra-alliance efficiency and generalization to unseen populations. We conduct bootstrap for the 5 repeat-trail results to fit the ROC curves. Efficiency of HSL is compared with that of Central Learning, Local learning, four classical methods (personalized learning[21], FedBN[22], FedProx[17], and SplitAVG[23]), 8 newest published state-of-the-art (SOTA) methods in 2024 including: (1) Knowledge transfer-based methods: FedPPKT[24], FedKTL[25]; (2) Parameter personalization optimization methods: FedTP[26], AdaFedProx[27], FedRCL[28], FedCOME[29]; (3) Data sampling and selection-based methods: FedDpS[30], FedDyS[31]. To investigate alternative strategies for tackling data heterogeneity, we conducted comparative experiments using

https://doi.org/10.1038/s41467-025-64459-y

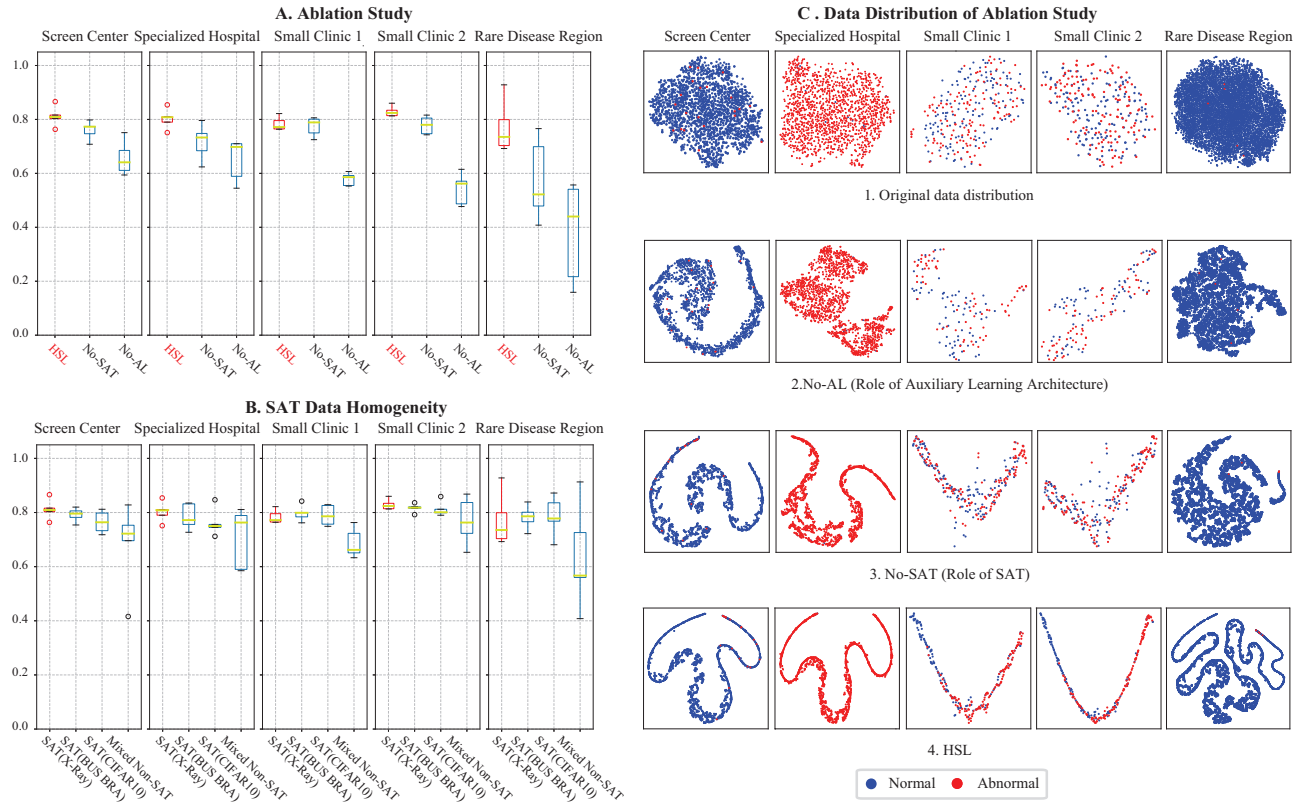

Fig. 3 | Interpretability study: decoupling HSL Components. A AUC box plots of the ablation study (testing set sizes: $n = 938$, 665, 60, 60, and 4279). No-AL: no auxiliary learning architecture. **B** AUC box plots of SAT data homogeneity on HSL (testing set sizes: $n = 938$, 665, 60, 60, and 4279). Box plots in **A**, **B** show the median (central line), the 25th and 75th percentiles (box bounds), and the minimum and maximum values excluding outliers (whiskers). Outliers are shown as individual points. Each bar represents the results of five repeated experiments for a learning method. Source data are provided as a Source data file. **C** Data distribution change before and after training is depicted for ablation study. Figures aregenerated using Python and subsequently composited in Adobe Illustrator 2023.

large foundation models. Specifically, we employed CLIP[32] for image feature extraction, followed by distributed learning. This approach took advantage of CLIP's pre-training on large-scale datasets, with the hypothesis that its highly generalizable feature representations could help alleviate heterogeneity challenges in federated settings. (Fig. 1E)

1. **Intra-alliance efficiency:** In the THYROID dataset, label skew (benign vs. malignant) is moderate, varying from 1.64:1 to 1:1.90. Quantity skew is significant, with sample sizes ranging from 422 to 11,571 (Fig. 1E). ROC curves (Fig. 4A) show that, across all hospitals, three methods stand out as the HSL, FedRCL and FedCOME (top-2 SOTA methods), which were comparable with central learning. Details indicate that AUCs of HSL is statistic significantly higher compared with the top-2 SOTA methods in the SYSU01 (0.931 vs. 0.894–0.921), SYSU06 (0.942 vs. 0.880–0.921), and SYSUCC (0.928 vs. 0.725–0.914), and comparable in GXFY (0.929 vs. 0.870-0.924) and FSSS (0.955 vs. 0.886–0.942) (Fig. 4A and Supplementary Table 10).
2. **Generalization to unseen populations:** In the PTC dataset, feature skew is notably pronounced due to the fundamental diversity of imaging representation in adult and pediatric thyroid cancer. The label skew (1:9.18) is moderate compared to the training Thyroid dataset (maximal skew of 1.64:1). However, quantity skew is significant, with the dataset containing 815 samples, which is considerably smaller than the largest training set, SYSU01, which has 11,571 samples. Figure 4A illustrates that HSL achieves the highest generalization (AUC = 0.846), which is statistically comparable with central learning (AUC = 0.822) and FedRCL (AUC = 0.837). HSL outperforms all other classical, CLIP, and SOTA methods considering AUC (AUC = 0.564–0.795) and performance stability (reduced AUC variance) (Fig. 4B and Supplementary Table 10).

Discussion

This study introduces HeteroSync Learning (HSL), a privacy-preserving distributed learning framework that effectively addresses data heterogeneity in medical imaging. By leveraging a shared public dataset (SAT) and a customed auxiliary learning architecture, HSL aligns representations across nodes without compromising data privacy. Our comprehensive evaluation demonstrates that HSL: (1) achieves performance comparable to central learning; (2) outperforms both classical methods (Personalized Learning, FedAvg, FedProx, SplitAVG), eight newly published SOTA approaches (FedRCL, FedCOME, etc.), and large foundation model (e.g., CLIP) by up to 40% of AUC improvement; and (3) exhibits superior generalization to unseen populations. Visualization analyses confirm that HSL transforms heterogeneous data distributions into harmonized representations, with the uniform SAT serving as a critical anchor for this alignment. These advantages persist across extreme clinical heterogeneity scenarios, from small clinics to rare disease populations.

The profound impact of data heterogeneity on the implementation of medical AI is unmistakable. Despite previous research affirming the potential of AI to enhance clinical practice[2,33–35], achieving consistent performance in subsequent tests proves challenging[36]. AI development necessitates diverse data, yet significant heterogeneity often results in the creation of biased models[4]. Moreover, "out of distribution" problem deteriorates the performance of a model substantially[37,38]. While swarm learning leverages blockchain for secure data sharing with stringent privacy measures, experiments conducted

https://doi.org/10.1038/s41467-025-64459-y

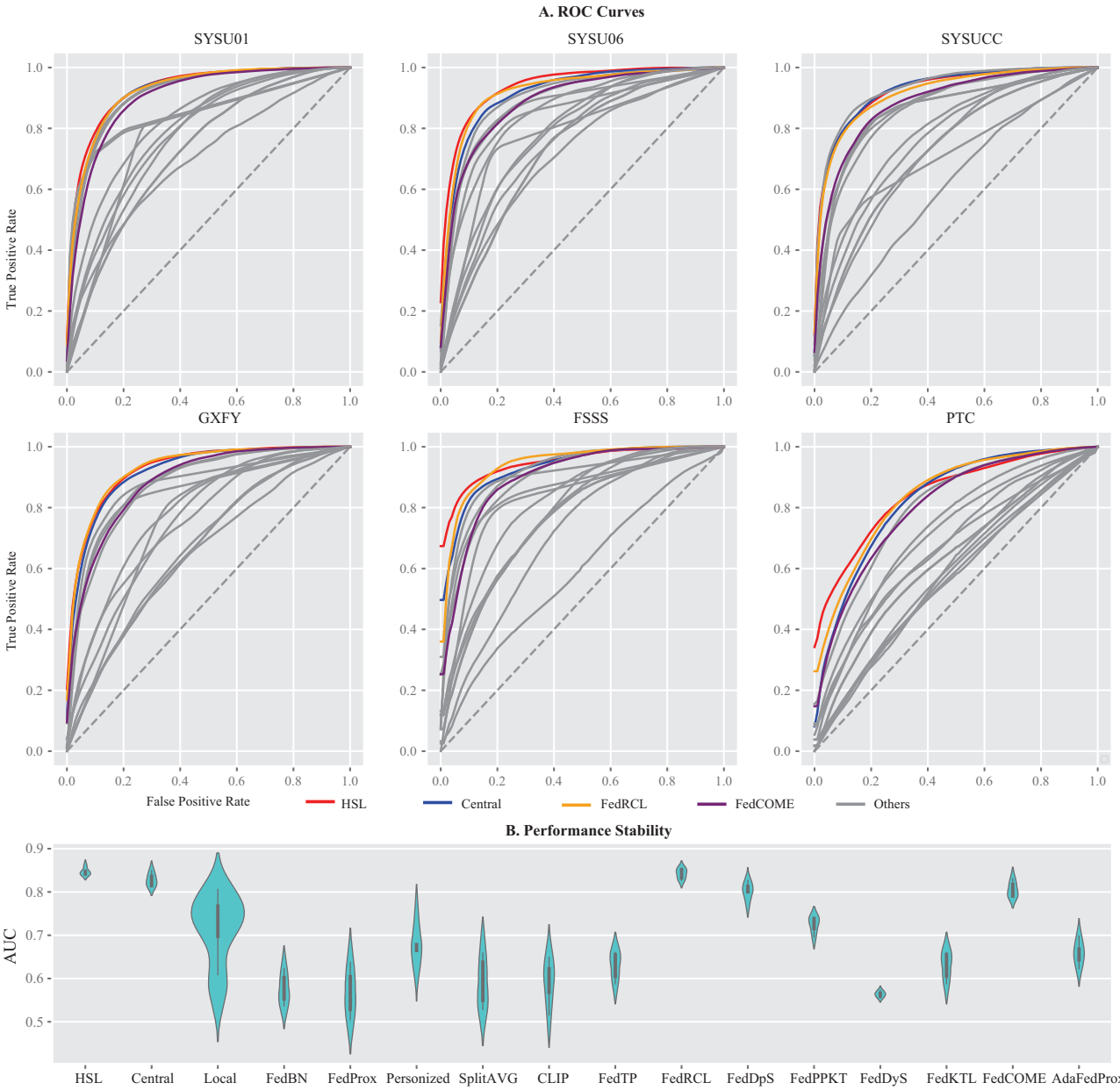

Fig. 4 | Real-world study. A ROC curves of the 16 learning methods' testing efficacy for each hospital. **B** Performance stability across 16 learning methods for the PTC dataset. Violin plots show the distribution of AUC scores derived from five independent experimental replicates ($n = 815$ for each replicate). Each violin displays the kernel density estimation of the distribution. The embedded box shows the interquartile range (25th–75th percentiles), and whiskers extend to the minimum and maximum values. Source data are provided as a Source data file. Figures are generated using Python and subsequently composited in Adobe Illustrator 2023.

on simulated heterogeneous datasets reveal a decline in model performance when tested on data with skewed distribution[12]. Real-world studies have demonstrated both improved and diminished model performance in heterogeneous datasets[11]. There are even cases of conflicting performance, with the model's odds ratio reversing in different datasets of heterogeneity[19]. To effectively mitigate the negative impact of data heterogeneity on distributed modeling is imperative for advancing AI applications, enabling adaptation to diverse heterogeneous scenarios in the clinical practice.

HSL mitigates heterogeneity by distributing SAT across all nodes, which is a shared, homogeneous public dataset that helps align representations and reduce data inconsistencies. It enables the model to learn generalizable representation, filtering out noise from local data discrepancies like label imbalances and varying imaging protocols[39]. Additionally, by training on both the primary task and SAT, HSL acts as a regularizer, preventing overfitting to local data and improving the model's ability to generalize to diverse and unseen datasets[20]. This design also improves computational efficiency and accelerates training through co-training using the MMoE architecture, which optimizes resource allocation across tasks. The performance of HSL critically depends on the proper configuration of SAT. The SAT data is selected basing on the following criteria and rationale: (1) Source of SAT data: The SAT data is sourced from publicly available datasets that are accessible and replaceable. (2) Uniform Distribution Requirement: The SAT data is uniformly distributed across all nodes, ensuring that each node uses the same SAT dataset with identical label ratios. For example, the RSNA dataset was divided into 1,000 normal and 1,000 abnormal images, maintaining consistency across all nodes in the training process. 3) Independence from Primary Task: The SAT data was not directly tied to the primary medical task, such as thyroid

cancer diagnosis. For instance, the CIFAR-10 dataset (natural images) was used alongside medical imaging tasks like ultrasound or X-ray. This independence ensured privacy and avoided overfitting to task-specific representations.

HSL demonstrates superior performance compared to methods reported in previous literature. While FedRCL[28] improves model convergence through contrastive learning, it struggles with slower convergence and feature collapse in highly heterogeneous settings due to intra-class attraction and insufficient representation diversity. FedCOME[29], using a consensus mechanism to reduce client-specific risk, relies on client sampling for gradient adjustment. While it performs well, it may not capture the full diversity of data distributions, limiting its generalization in clinical applications. Large foundation models (e.g., CLIP) can extract rich, task-agnostic representations from medical images. However, due to its reliance on large-scale datasets during pre-training, it may exhibit overfitting when applied to cross-domain and cross-modal medical imaging datasets, limiting its performance in heterogeneous data environments within the medical imaging field. In comparison, HSL's auxiliary learning framework, which aligns representations across nodes using a SAT dataset, directly addresses heterogeneity while preserving privacy. By decoupling feature extraction from task-specific training, HSL ensures robust performance and improved generalization across diverse clinical scenarios.

Our study presents a wide range of heterogeneity scenarios encountered in the clinical practice. In Tong et al.[19] and Qu et al.[15], the focus was solely on label distribution skew, with the prevalence of positive cases varied from 6% to <1% in Tong, with a maximal normal: abnormal ratio of 29:1 in Qu. Yan et al. [6] proposed a self-supervised FL framework method with significant advantage in situations with limited labeled data. However, it exhibited a notable performance drop in severe heterogeneity simulations and proved less efficient than certain other methods. HSL was tested not only across a range of simulated scenarios, including various clinical settings like specialized hospitals and rare disease regions with different degrees of heterogeneity, but also on a real-world heterogeneous thyroid cancer dataset from multiple medical centers. This extensive testing provides a thorough evaluation of HSL's robustness in both controlled simulations and real-world conditions. Further evaluation in additional clinical domains, such as breast cancer and epidemic diseases, will further validate HSL's generalization ability. It should be noted that the current framework has not been evaluated in dynamic clinical environments. Critical challenges for distributed learning, such as real-time data updates and incorporation of new nodes, would require corresponding model adaptation mechanisms. By incorporating continuous or incremental learning approaches, HSL could adapt to evolving data patterns or shifts in underlying distributions (e.g., changes in disease prevalence) without needing to retrain from scratch[40,41].

In conclusion, our study introduces HSL as a tailored solution for addressing the challenges of data heterogeneity in distributed medical imaging while preserving privacy. Extensive simulations and real-world evaluations demonstrate that HSL performs on par with central learning and outperforms existing state-of-the-art methods, even under extreme data heterogeneity. HSL demonstrates strong potential as a key technology to enhance the efficiency and prospects of distributed medical imaging modeling.

Methods

This work complies with all ethical regulations as approved by the Research Ethics Committee of the First Affiliated Hospital of Sun Yat-sen University ([2022]061). Informed consent was waived for retrospectively collected ultrasound images, which were de-identified before use. No data were excluded after initial preprocessing. All available data that met the inclusion criteria (e.g., image quality, label availability) were retained for model training, validation, and testing.

Participants

This study uses 137,585 images collected from four publicly available datasets and an alliance-based multicenter dataset. These include:

Primary tasks:

1. MURA dataset: The primary task MURA dataset comprises 39,168 musculoskeletal radiographs published by the StanfordML group (Stanford MURA Dataset, http://stanfordmlgroup.github.io/competitions/mura/). It covers seven body sites, including: Finger: 3256 normal and 2197 abnormal images; Hand: 4211 normal and 1661 abnormal images; Elbow: 3098 normal and 2223 abnormal images; Forearm: 1304 normal and 806 abnormal images; Humerus: 814 normal and 735 abnormal images; Shoulder: 4394 normal and 4340 abnormal images; Wrist: 5917 normal and 4211 abnormal images.
2. THYROID dataset: The Thyroid dataset comprises an alliance-based multicenter dataset of 20,043 thyroid nodule ultrasound images (one image per patient) collected from the Ultrasound Engineering Institute, Medical Industry Branch of China Association Plant Engineering (UE-MICAP)[42]. Ultrasound is the recommended imaging modality for screening various cancers, including thyroid cancer[43–45]. As being operator-dependent and patient-dependent, ultrasound exhibits a high degree of heterogeneity compared to other imaging modalities[46–48]. The details are as follows: Source: Collected from 5 hospitals across 3 cities, which include the First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China (SYSU01, benign: malignant = 3987:7584), the Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou, China (SYSU06, benign: malignant = 1066:650); Sun Yat-sen University Cancer Center (SYSUCC, benign: malignant = 3791:2,687), the First Affiliated Hospital of Guangxi Medical University, Nanning, China (GXFY, benign: malignant = 910:825), and Foshan San-shui District Hospital (FSSS, benign: malignant = 243:179); Ultrasound Device Types: 13.
3. PTC dataset: The PTC (Pediatric Thyroid Cancer) dataset offers a unique opportunity to explore the generalization capabilities of the HSL model to datasets that were not part of its initial training process. Although pediatric thyroid cancer is relatively rare, it tends to be highly aggressive, with a high incidence of neck invasion and distant metastases. Pediatric thyroid cancer presents distinct ultrasound characteristics compared to adult cases[49], and the relatively low prevalence of the disease makes it challenging to gather a sufficiently large dataset for comprehensive AI training. This dataset, which consists of 815 thyroid nodule ultrasound images-735 malignant and 80 benign cases-collected from multi-vendor ultrasound devices at SYSU01 and SYSUCC, has never been utilized in prior alliance-based model training. By leveraging this dataset, we aim to evaluate how well HSL generalizes to out-of-alliance data and to benchmark its performance against existing methods. This evaluation will provide insights into the model's robustness and adaptability across different patient demographics and equipment settings.

SAT Datasets:

1. RSNA dataset: The Radiological Society of North America (RSNA) dataset is a challenge dataset comprising 25,684 chest radiographs sourced from the NIH database (https://nihcc.app.box.com/v/ChestXray-NIHCC)[50]. The dataset was re-labeled by the RSNA and the Society of Thoracic Radiology (STR) (https://www.kaggle.com/c/rsna-pneumonia-detection-challenge)[50]. The dataset is categorized into three groups: Normal: Count: 8525 (33.2%); Abnormal with Lung Opacity: Count: 5659 (22.0%); Abnormal without Lung Opacity: Count: 11,500 (44.8%). This dataset is used as an auxiliary dataset.
2. BUS-BRA dataset: The BUS-BRA dataset comprises 1875 anonymized images from 1064 female patients acquired via four

ultrasound scanners during systematic studies at the National Institute of Cancer (Rio de Janeiro, Brazil)[51]. The dataset includes biopsy-proven tumors divided into 722 benign and 342 malignant cases. This dataset is used as an auxiliary dataset.

3. CIFAR-10 dataset: The CIFAR-10 dataset (Canadian Institute for Advanced Research, 10 classes) is a subset of the Tiny Images dataset and consists of 60,000 natural color images[52]. The 10 classes include airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck, with 6000 images per class. This dataset is used as an auxiliary dataset.

Model development

SAT is created by randomly selecting 1000 normal and 1000 abnormal lung opacity images from the RSNA dataset. We opted for MMOE (Multi-gate Mixture-of-Experts)[20] as our chosen auxiliary learning framework, and ResNet18 as the benchmark algorithms for feature extraction. In order to amplify the contribution of SAT to the primary task, we drew inspiration from knowledge distillation techniques, introducing a temperature parameter "T" for MMoE, designed to elevate the information entropy of SAT dataset (Fig. 1A). For parameter merging, we opted for the simplicity of parameter averaging:

$$\text{weight} = \frac{(weight_1 + weight_2 + \ldots + weight_i)}{i} \quad (1)$$

In the ablation study, as depicted in Fig. 1C, we: (1) remove the SAT data while keeping the auxiliary learning architecture intact; (2) remove the auxiliary learning architecture (the SAT data plays no role during training); (3) alternate SAT datasets or replace with non-SAT datasets (unevenly distributed multiple auxiliary datasets). SAT Datasets: These include X-Ray RSNA (1000:1000), BUS-BRA dataset (500:500), and natural CIFAR-10 dataset (500:500). The datasets maintain a 1:1 label ratio, and each node uses the same SAT dataset during implementation. Non-SAT Datasets: These include X-Ray RSNA (1000:1000, corresponding to Screening Center), X-Ray RSNA (1000:1, corresponding to Specialized Hospital), X-Ray RSNA (1:1000, corresponding to Small Clinic 1), BUS-BRA dataset (500:500, corresponding to Small Clinic 2), and natural CIFAR-10 dataset (500:500, corresponding to Rare Disease Region). Each node uses a different auxiliary dataset during implementation. (Fig. 1D)

The image input is resized to 224 × 224, and the model's output is the probability of the referred labels. The training process spans 100 epochs, and in every 5 epochs optimal local model parameters are generated for inter-node parameter fusion. Nodes synchronize all parameters, and the fusion takes place locally. Subsequent to parameter fusion, each node undergoes an additional 5 epochs of training based on the fused parameters. We use the ADAM optimizer with an initial learning rate of 0.001. Cosine annealing is utilized for dynamic learning rate adjustment, with a cosine period of 10 epochs. Weight decay is set to 0.0001. Batch size is determined based on local computing power.

In our federated learning framework, data from each node was divided into training (60%), validation (10%), and holdout testing (30%) sets using stratified random sampling to preserve label distributions. The validation set played a crucial role in preventing overfitting with early stopping (patience of 5 epochs), and selecting the model based on the highest validation AUC. Validation occurred locally every 5 epochs before parameter aggregation, ensuring both node-specific and global performance were monitored. The final model was chosen after achieving ≥3 consecutive AUC plateaus with less than 0.5% improvement, ensuring model stability. The test set was kept strictly separate for final evaluation, with results verified through five repeated runs to ensure reproducibility, while clinical relevance was ensured through AUC-based model selection.

Model training is conducted on GPUs using PyTorch (version 1.7.1). Different GPUs from various nodes, including 1 NVIDIA A100 with 40 GB of HBM2 memory, 2 RTX 4090 with 24 GB each, 1 RTX 3080 Laptop with 16 GB, and 3 RTX 1060 with 6 GB each, are utilized in the distributed learning process. We implemented node communication using the Paramiko library in conjunction with the encrypted SFTP protocol. User permissions are managed through SFTPGo, which strictly controls access and modification rights to ensure data security. Additionally, a fault-tolerance mechanism is in place to automatically retry or resume transfers in case of interruptions, maintaining robust and reliable data transmission.

Simulation study: controlled validation of heterogeneity scenarios

Radiographs from various anatomical sites exhibit diverse feature distributions, which can affect algorithm performance[53,54].

1. Feature distribution skew: To address feature distribution, we randomly select an equal number ($n = 1470$) of 1:1 (normal: abnormal) data from each of the seven anatomical sites and distribute them across seven nodes. Distributed learning is conducted among these nodes for abnormality diagnosis. (Fig. 2A1).
2. Label distribution skew: To address label distribution, we randomly selected data from the shoulder site and distributed them across two nodes. Within these two nodes, ten pairs of data were configured to simulate scenarios with varying label distributions (proportion of abnormal radiographs). Distributed learning was conducted between the two nodes for abnormality diagnosis. (Fig. 2B1).
3. Quantity skew: To address data quantity, we randomly selected data from the shoulder site and distributed them across two nodes. Within these two nodes, nine pairs of data were configured to simulate scenarios with varying data quantities (number of radiographs). Distributed learning was conducted for abnormality diagnosis between the two nodes. (Fig. 2C1).
4. Combined heterogeneity: The MURA dataset is used to simulate five distributed nodes, incorporating extreme clinical scenarios mimicking a large-scale screening center, a large-scale disease-specialized hospital, two small clinics, and rare disease regions (Fig. 2D1). Distributed learning is conducted among these five nodes. HSL is implemented to mitigate the influence of all three types of skews, promote generalization and enhance performance.

Interpretability study: decoupling HSL components

In the combined heterogeneity scenario, HSL components are decoupled to test the contribution of the auxiliary learning architecture and SAT, and the importance of SAT data homogeneity.

1. Role of the Auxiliary learning architecture: Removing SAT while retaining the auxiliary learning architecture (Fig. 1C 2 No-SAT).
2. Role of SAT: Removing the auxiliary learning architecture (AL) (the SAT data plays no role in training) (Fig. 1C 3 NO-AL).
3. Importance of SAT data homogeneity: replacing the homogenously distributed SAT data with homogenously distributed CIFAR-10 or BUS-BRA SAT data; replacing the homogenously distributed SAT data with heterogeneously distributed multiple auxiliary datasets (mixed non-SAT datasets) (Fig. 1D).

Real-world validation: clinical efficacy and generalization

1. Intra-alliance test: For each hospital, 30% of the data were held out for intra-alliance testing to assess model performance.
2. Generalization to unseen populations: The finalized model was further evaluated on the out-of-alliance PTC dataset to examine cross-population generalizability, which hadn't been seen during model development.

Statistical analysis

Our study did not include gender-based analysis. The research focused on data heterogeneity and does not apply specifically to one sex or gender, and does not include gender-based analysis or restrictions.

In each scenario simulation and in the real-world study, we conducted five permutations of repeated experiments, involving random dataset shuffling and reassignment. Model comparisons were conducted through non-parametric bootstrap resampling (1000 iterations) applied to predictions aggregated from five independent experimental replicates. For each method pair, we computed AUC (probability-based), Accuracy, Recall, and Precision using a fixed classification threshold of 0.5. Significance testing employed two-tailed percentile-based hypothesis testing. *P* values were calculated by assessing the proportion of bootstrap samples where performance differences reversed direction relative to the observed effect. All statistical inferences were evaluated against an alpha threshold of 0.05 without multiplicity adjustment. This approach provides robust estimation of effect sizes while accounting for variability across experimental repeats.

All computations are implemented in Python (version 3.8.8). Performance metrics, including AUC, ACC, recall, precision, are calculated using the "sklearn" Python package (version 0.24.1). ROC curves and boxplots are generated with the "matplotlib" Python package (version 3.1.1). Interpretability plots utilize t-distributed stochastic neighbor embedding (tSNE) visualization through the "sklearn.manifold.TSNE" Python package.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The RSNA, BUS-BRA datasets (https://www.kaggle.com/datasets/vuppalaadithyasairam/ultrasound-breast-images-for-breast-cancer) and CIFAR-10 dataset (https://www.kaggle.com/datasets/petitbonney/cifar10-image-recognition) are publicly available with no restricted access. The MURA dataset (http://stanfordmlgroup.github.io/competitions/mura/) requires an application on the official website (https://stanfordaimi.azurewebsites.net/datasets/3e00d84b-d86e-4fed-b2a4-bfe3effd661b) before it can be downloaded. The THYROID and PTC datasets are owned by the China MedAI Collaborative Research Alliance. Researchers from non-commercial entities can submit a request for cooperation by emailing the corresponding author at *wangw73@mail.sysu.edu.cn*. Requests will be reviewed within 10 business days and access will be granted to qualified researchers for academic use only. Source data are provided with this paper.

Code availability

The Python codes of HeteroSync Learning are available at Zenodo (https://doi.org/10.5281/zenodo.15869874)[55], with the source repository maintained at https://github.com/MedAI-UAIX/HeteroSync_Learning-HSL (Supplementary Software 1). The code is released under the Apache 2.0 License, an Open-Source Initiative-approved license, which permits reuse, modification, and distribution with appropriate credit and patent protection. No additional restrictions apply. Portions of the implementation adapt components from https://github.com/easezyc/Multitask-Recommendation-Library (MIT License), with copyright statements and license information retained in the corresponding source files.

References

1. Rajpurkar, P. & Lungren, M. P. The current and future state of AI interpretation of medical images. *N. Engl. J. Med.* **388**, 1981–1990 (2023).
2. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
3. Price, W. N. 2nd & Cohen, I. G. Privacy in the age of medical big data. *Nat. Med.* **25**, 37–43 (2019).
4. Rieke, N. et al. The future of digital health with federated learning. *NPJ Digit. Med.* **3**, 119 (2020).
5. Li, T., Sahu, A. K., Talwalkar, A. & Smith, V. Federated learning: challenges, methods, and future directions. *IEEE Signal Process. Mag.* **37**, 50–60 (2020).
6. Yan, R. et al. Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging. *IEEE Trans. Med. Imaging* **42**, 1932–1943 (2023).
7. Ding, W., Abdel-Basset, M., Hawash, H. & Pedrycz, W. MIC-Net: a deep network for cross-site segmentation of COVID-19 infection in the fog-assisted IoMT. *Inf. Sci.* **623**, 20–39 (2023).
8. Shen, Y., Zhou, Y. & Yu, L. CD2-pFed: cyclic distillation-guided channel decoupling for model personalization in federated learning. In *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 18–24 (IEEE, 2022).
9. Dayan, I. et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat. Med.* **27**, 1735–1743 (2021).
10. Gao, Z., Wu, F., Gao, W. & Zhuang, X. A new framework of swarm learning consolidating knowledge from multi-center non-IID data for medical image segmentation. *IEEE Trans. Med. Imaging* **42**, 2118–2129 (2023).
11. Ogier du Terrail, J. et al. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. *Nat. Med.* **29**, 135–146 (2023).
12. Warnat-Herresthal, S. et al. Swarm Learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 (2021).
13. Zhang, A., Xing, L., Zou, J. & Wu, J. C. Shifting machine learning for healthcare from development to deployment and from models to data. *Nat. Biomed. Eng.* **6**, 1330–1345 (2022).
14. Zhao, Y. et al. Federated learning with non-IID data. Preprint at: https://doi.org/10.48550/arXiv.1806.00582 (2018).
15. Qu, L., Balachandar, N., Zhang, M. & Rubin, D. Handling data heterogeneity with generative replay in collaborative learning for medical imaging. *Med. Image Anal.* **78**, 102424 (2022).
16. Vepakomma, P., Gupta, O., Swedish, T. & Raskar, R. Split learning for health: distributed deep learning without sharing raw patient data. Preprint at: https://doi.org/10.48550/arXiv.1812.00564 (2018).
17. Li, T. et al. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems* **2**, 429–450 (2020).
18. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2020).
19. Tong, J. et al. Distributed learning for heterogeneous clinical data with application to integrating COVID-19 data across 230 sites. *NPJ Digit. Med.* **5**, 76 (2022).
20. Ma, J. et al. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proc. 24th ACM SIGKDD Int Conf Knowl Discov Data Mining* 1930–1939 (Association for Computing Machinery, 2018).
21. Wang, J., Jin, Y., Stoyanov, D. & Wang, L. FedDP: dual personalization in federated medical image segmentation. In *Proc. IEEE Trans Med Imaging* 297–308 (IEEE, 2023).
22. Li, X. et al. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. In Proc. International Conference on Learning Representations (ICLR, 2021).
23. Torres-Lopez, V. M. et al. Development and validation of a model to identify critical brain injuries using natural language processing of text computed tomography reports. *JAMA Netw. Open* **5**, e2227109 (2022).

24. Wei, Q. et al. Enhancing privacy-utility tradeoff with few-round strategy in heterogeneous federated learning. In *Proc. 2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)* 1–6 (IEEE, 2024).
25. Zhang, J., Liu, Y., Hua, Y. & Cao, J. An upload-efficient scheme for transferring knowledge from a server-side pre-trained generator to clients in heterogeneous federated learning. In *Proc. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 12109–12119 (IEEE, 2024).
26. Li, H. et al. FedTP: federated learning by transformer personalization. *IEEE Trans. Neural Netw. Learn. Syst.* **35**, 13426–13440 (2024).
27. Sahoo, P. et al. AdaFedProx: a heterogeneity-aware federated deep reinforcement learning for medical image classification. In *Proc. IEEE Transactions on Consumer Electronics* 1–1 (IEEE, 2024).
28. Seo, S., Kim, J., Kim, G. & Han, B. Relaxed contrastive learning for federated learning. In *Proc. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 12279–12288 (IEEE, 2024).
29. Zheng, S., Ye, T., Li, X. & Gao, M. Federated learning via consensus mechanism on heterogeneous data: a new perspective on convergence. In *Proc. ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 7595–7599 (IEEE, 2024).
30. Ghosh, S., Kushwaha, A. & Singh, D. Tackling data heterogeneity in federated learning through global density estimation. In *Proc. 2024 IEEE International Conference on Big Data (BigData)* 7764–7773 (IEEE, 2024).
31. Kiyamousavi, S. E., Kraychev, B. & Koychev, I. FedDyS: enhancing federated learning efficiency with dynamic sample selection. In *Proc. 2024 IEEE Symposium on Computers and Communications (ISCC)* 1-8 (IEEE, 2024).
32. Radford, A. et al. Learning transferable visual models from natural language supervision. International conference on machine learning 8748–8763 (PmLR, 2021).
33. Tong, W. J. et al. Integration of artificial intelligence decision aids to reduce workload and enhance efficiency in thyroid nodule management. *JAMA Netw. Open* **6**, e2313674 (2023).
34. Huang, Q. et al. A novel image-to-knowledge inference approach for automatically diagnosing tumors. *Expert Syst. Appl.* **229**, 120450 (2023).
35. Luo, Y., Huang, Q. & Liu, L. Classification of tumor in one single ultrasound image via a novel multi-view learning strategy. *Pattern Recognit.* **143**, 109776 (2023).
36. Sohn, E. The reproducibility issues that haunt health-care AI. *Nature* **613**, 402–403 (2023).
37. Azizi, S. et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nat. Biomed. Eng.* **7**, 756–779 (2023).
38. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F. & Celi, L. A. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* **2**, e489–e492 (2020).
39. Zhao, Y., Wang, X., Che, T., Bao, G. & Li, S. Multi-task deep learning for medical image computing and analysis: a review. *Comput. Biol. Med.* **153**, 106496 (2023).
40. Pianykh, O. S. et al. Continuous learning AI in radiology: implementation principles and early applications. *Radiology* **297**, 6–14 (2020).
41. van de Ven, G. M., Tuytelaars, T. & Tolias, A. S. Three types of incremental learning. *Nat. Mach. Intell.* **4**, 1185–1197 (2022).
42. Peng, S. et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. *Lancet Digit. Health* **3**, e250–e259 (2021).
43. Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
44. Haugen, B. R. et al. 2015 American Thyroid Association Management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: The American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer. *Thyroid. Off. J. Am. Thyroid. Assoc.* **26**, 1–133 (2016).
45. Bevers, T. B. et al. NCCN guidelines® insights: breast cancer screening and diagnosis, version 1.2023. *J. Natl. Compr. Cancer Netw.* **21**, 900–909 (2023).
46. Todsen, T. et al. Reliable and valid assessment of point-of-care ultrasonography. *Ann. Surg.* **261**, 309–315 (2015).
47. Fetzer, D. T., Browning, T., Xi, Y., Yokoo, T. & Singal, A. G. Associations of ultrasound LI-RADS visualization score with examination, sonographer, and radiologist factors: retrospective assessment in over 10,000 examinations. *AJR Am. J. Roentgenol.* **218**, 1010–1020 (2022).
48. Li, M. D. et al. ADMNet: adaptive-weighting dual mapping for online tracking with respiratory motion estimation in contrast-enhanced ultrasound. *IEEE Trans. Image Process* **33**, 58–68 (2024).
49. Richman, D. M. et al. Thyroid nodules in pediatric patients: sonographic characteristics and likelihood of cancer. *Radiology* **288**, 591–599 (2018).
50. Stein, A. et al. RSNA Pneumonia Detection Challenge. https://kaggle.com/competitions/rsna-pneumonia-detection-challenge. Kaggle (2018).
51. Gómez-Flores, W., Gregorio-Calas, M. J. & Coelho de Albuquerque Pereira, W. BUS-BRA: a breast ultrasound dataset for assessing computer-aided diagnosis systems. *Med. Phys.* **51**, 3110–3123 (2024).
52. Krizhevsky, A., Nair, V. & Hinton, G. CIFAR-10: Canadian Institute for Advanced Research. http://www.cs.toronto.edu/~kriz/cifar.html (2010).
53. Murphy, Z. R., Venkatesh, K., Sulam, J. & Yi, P. H. Visual transformers and convolutional neural networks for disease classification on radiographs: a comparison of performance, sample efficiency, and hidden stratification. *Radiol. Artif. Intell.* **4**, e220012 (2022).
54. Kandel, I. & Castelli, M. Improving convolutional neural networks performance for image classification using test time augmentation: a case study using MURA dataset. *Health Inf. Sci. Syst.* **9**, 33 (2021).
55. Hang-Tong, Hu M.-D. L. et al. Addressing data heterogeneity in distributed medical imaging with heterosync learning (code repository). Zenodo. https://doi.org/10.5281/zenodo.15869874 (2025).

Acknowledgements

This study was supported by grants 82371983 (W.W.), 12326609 (Q.H.H.), 82102078 (Si-Min Ruan), 82171960 (W.W.), 82272076 (L.D.C.), 82102141 (H.T.H.), 82030047 (Q.H.H.) from the National Nature Science Foundation of China, grant NO 2023A04J2231 (H.T.H.) from the Guangzhou Science and Technology Project, grant NO 2021B1515120030 (L.D.C.) from the Guangdong Regional Joint Foundation. The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Author contributions

H.T.H., M.D.Li., and J.B.H.: acquisition and interpretation of data; AI development and statistical analysis; drafting of the manuscript. X.X.L., S.H.W., W.J.T., and F.Y.Y.: administrative and material support; curation of the data. L.D.C., H.Y., G.J.L., H.B.W., and M.D.Lu.: critical revision of the manuscript for important intellectual content. W.P.K.: hardware preparation, debugging, distributed network construction. Q.H.H., W.W., and M.K.: conceptual guidance, supervision; correspondence and requests for materials should be addressed to Wei Wang. M.D.Li, M.Y.C., and S.L.: AI development and manuscript support during revision. UEMICAP: Providing multi-center datasets.

Competing interests
The authors declare no competing interests.

Declaration of generative AI
During the preparation of this work, the authors used ChatGPT 3.5 for English language polishing. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-64459-y.

Correspondence and requests for materials should be addressed to Qing-Hua Huang, Ming Kuang or Wei Wang.

Peer review information *Nature Communications* thanks Bingsheng Huang, who co-reviewed with Bin Huang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025

Ultrasound Engineering Institute, Medical Industry Branch of China Association Plant Engineering (UE-MICAP)

Hang-Tong Hu[1,8,9], Ming-De Li[1,8], Xin-Xin Lin[1], Meng-Yao Cai[1], Shao-Hong Wu[1], Wen-Juan Tong[1], Wei-Ping Ke[1], Li-Da Chen[1], Hong Yang[3], Guang-Jian Liu[4], Ming-De Lu[1,6], Qing-Hua Huang[7] ✉, Ming Kuang[1,6] ✉ & Wei Wang[1] ✉