# Robust Federated Anomaly Detection Using Dual-Signal Autoencoders: Application to Kidney Stone Identification in Ureteroscopy

Ivan Reyes-Amezcua[1], Francisco Lopez-Tiro[2,3], Clément Larose[4], Christian Daul[3], Andres Mendez-Vazquez[1], and Gilberto Ochoa-Ruiz[2]

[1] CINVESTAV, Guadalajara, Mexico
[2] Tecnologico de Monterrey, School of Engineering and Sciences, Mexico
[3] CRAN UMR 7039, Université de Lorraine and CNRS, Nancy, France
[4] CHRU de Nancy-Brabois, Service d'urologie, Vandœuvre-les-Nancy, France

**Abstract.** This work introduces Federated Adaptive Gain via Dual Signal Trust (FedAgain), a novel federated learning algorithm designed to enhance anomaly detection in medical imaging under decentralized and heterogeneous conditions. Focusing on the task of kidney stone classification, FedAgain addresses the common challenge of corrupted or low-quality client data in real-world clinical environments by implementing a dual-signal trust mechanism based on reconstruction error and model divergence. This mechanism enables the central server to dynamically down-weight updates from untrustworthy clients without accessing their raw data, thereby preserving both model integrity and data privacy. FedAgain employs deep convolutional autoencoders trained in two diverse kidney stone datasets and is evaluated in 16 types of endoscopy-specific corruption at five severity levels. Extensive experiments demonstrate that FedAgain effectively suppresses "expert forger" clients, enhances robustness to image corruptions, and offers a privacy-preserving solution for collaborative medical anomaly detection. Compared to traditional FedAvg, FedAgain achieves clear improvements in all 16 types of corruption, with precision gains of up to +14.49% and F1 score improvements of up to +10.20%, highlighting its robustness and effectiveness in challenging imaging scenarios.

**Keywords:** Federated Learning · Anomaly Detection · Kidney Stone Identification.

## 1 Introduction

Deep learning has significantly advanced medical image analysis, notably enhancing tasks such as kidney stone classification [1, 8]. However, conventional methods typically rely on local training approaches, limiting data diversity due to the process and/or instrument for the acquisition, and exposing models to performance degradation under perturbations like noise, blur, illumination changes, and camera instability. Such perturbations, collectively termed image corruptions, frequently arise in clinical environments, underscoring the necessity for

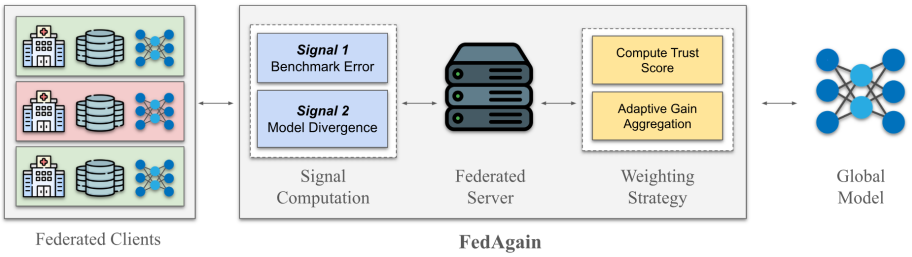


Fig. 1: Overview of the FedAgain framework. Federated clients (left) train local models and send updates to a central server. Clean clients are shown in green, while corrupted clients are highlighted in red. The server computes two trust signals: benchmark error and model divergence, which are used to assign trust scores. These scores guide the adaptive gain aggregation process, reducing the impact of unreliable clients. The aggregated result is used to update the global model.

robust and reliable models that ensure diagnostic certainty in medical practice [9].

Federated Learning (FL) provides a strategic response to these limitations by enabling multiple medical centers to collaboratively train models without directly sharing sensitive patient data [15]. Each client independently trains a local model and communicates only model parameters to a central server for aggregation, effectively addressing privacy concerns while incorporating diverse and decentralized data [18, 10]. Nevertheless, federated training is vulnerable to corrupted datasets affected by artifacts like blur, noise, or poor lighting, which can degrade global model performance.

To effectively identify and mitigate these image corruptions in a federated, multiclient environment, autoencoder-based anomaly detection methods have emerged as promising solutions. Autoencoders detect deviations using reconstruction errors, successfully identifying medical anomalies such as tumors and imaging artifacts [10, 2]. Unlike standard convolutional neural networks (CNNs), autoencoders excel by learning intrinsic data representations, making them particularly effective at capturing subtle anomalies. Integrating autoencoders into federated learning frameworks enables the collaborative, privacy-preserving detection of anomalies, significantly improving the robustness and generalization of models across varied clinical settings [19].

Federated autoencoder-based anomaly detection is still underexplored in kidney stone imaging, mainly due to limited annotated data. This study addresses the gap using a specialized dataset with endoscopic images and synthetic corruptions, enabling systematic evaluation of model robustness in federated settings. A key challenge addressed in this work is the *expert forger* problem, which arises when a client consistently submits corrupted data (e.g., from persistently blurry endoscopic cameras), inadvertently training its local model to specialize in these corruptions. This specialization compromises the generalizability of the global

model when aggregated via conventional Federated Averaging (FedAvg) [14]. To mitigate this issue, a novel federated autoencoder framework, *FedAgain* (Federated Adaptive Gain via Dual Signal Trust) is introduced. FedAgain incorporates a dual-signal trust mechanism to dynamically evaluate and down-weight untrustworthy clients, ensuring enhanced resilience, robustness, ultimately, greater diagnostic certainty (Figure 1).

This paper introduces the first federated autoencoder framework specifically tailored for robust endoscopic kidney stone imaging. Using six specialized autoencoder models, the approach is rigorously evaluated across 16 different corruption types, employing various performance metrics to ensure comprehensive analysis.

The rest of the paper is organized as follows: Section 2 details the methodology, including the FL framework, autoencoder architecture, and the trust mechanism employed by FedAgain. Section 3 describes the datasets and preprocessing. Section 4 covers corruption simulation and evaluation metrics. Section 5 presents results and client behavior analysis. Section 6 concludes with key findings and implications.

## 2 Related Work

FL first gained traction with FedAvg, which averages local updates instead of sharing raw data, but it falters when even a few clients are corrupted [14]. Aggregation-based defenses such as FLTrust attach every round to a small, server-held "root" dataset, sharply curbing Byzantine influence [3], while Bayesian ensembling in FedBE samples and averages multiple global models to stay stable on highly non-IID (independent and identically distributed) data [4]. However, a recent systematic survey notes that these strategies still leave long-term, client-specific artifacts largely unaddressed [20, 7]. Autoencoder anomaly detectors fill part of this gap: autoencoders minimize bandwidth and spot outliers at the edge [16], and global summary-statistics thresholds further boost detection under skewed client distributions [10]. Nevertheless, most autoencoder work targets generic noise or blur and rarely tackles the persistent, domain-specific corruptions common in clinical images, an issue repeatedly highlighted in medical-imaging FL reviews [9, 7].

Kidney stone ureteroscopic data highlight challenges such as glare, blur, and color drift across hospitals [12, 13]. A corruption-aware FL system combines a pre-trained CNN with two stages: Learning Parameter Optimization and Federated Robustness Validation on corrupted data [18]. These results support FL pipelines that integrate robust aggregation and client-side anomaly detection for reliable, privacy-preserving diagnosis.

## 3 Federated Adaptive Gain via Dual-Signal Trust

FL allows clients (e.g., hospitals) to collaboratively train a shared model without sharing raw data, preserving privacy and ensuring compliance (e.g., HIPAA,

GDPR). Clients send only model updates, reducing communication costs and enabling scalable, privacy-preserving training on distributed medical data.

FL offers a principled solution by training models where the data reside, but standard FL algorithms remain vulnerable to persistent data-quality heterogeneity: A handful of *noisy* or corrupted clients can bias model updates and degrade performance [11]. The problem is especially pronounced in kidney stone imaging, where differences in endoscopic and laboratory acquisition pipelines routinely produce specular reflections, blur, sensor noise, and lens artifacts that vary from one clinical site to another. This leads to a critical vulnerability referred to as the *expert forger* problem. A client with a persistent, local data corruption (e.g., a perpetually blurry camera) will train its local model to become an *expert* at reconstructing that specific corruption. In conventional FedAvg, the parameter updates from such an *expert-forger* client are incorporated into the global weight average, contaminating the aggregated model and substantially degrading its ability to recognize the corresponding defect as an anomaly.

FedAgain is a strategy that builds robust anomaly detection models by weighting client contributions based on trust scores, replacing uniform averaging with a two-phase process: client-side dual-signal generation (reconstruction error and model divergence), followed by server-side trust assessment and adaptive aggregation, see Figure 1.

### 3.1 FedAgain Algorithm

Each step in FedAgain comprises two phases: A client one (Dual-Signal generation) followed by a server phase (Trust assessment and adaptive aggregation), see Figure 1.

**Client phase:** For every selected client $C_k$:

1. *Benchmark signal:* It evaluates incoming global models $w_t$ on the local dataset $\mathcal{D}_k$ by using:

$$E_k = \frac{1}{|\mathcal{D}_k|} \sum_{(x_i, y_i) \in \mathcal{D}_k} \ell\big(f_{w_t}(x_i),\, y_i\big). \tag{1}$$

   where $\ell$ is a loss function that measures the discrepancy between the prediction of the model $f_{w_t}(x_i)$ and the true label $y_i$. The resulting scalar $E_k$ is sent back to the server. A high value of $E_k$ indicates that the client's data distribution $\mathcal{D}_k$ differs significantly from the global distribution.
2. *Local adaptation:* It performs $e$ epochs of stochastic gradient descent on $\mathcal{D}_k$ to obtain an updated model $w_{t+1}^{(k)}$. Finally, it transmits $w_{t+1}^{(k)}$ to the server.

**Server phase:** For each responding client:

1. *Divergence signal:* It measures the results with the Euclidean or cosine distance by defining
$$D_k = \left\| w_{t+1}^{(k)} - w_t \right\|, \tag{2}$$
A large $D_k$ implies that substantial parameter changes are required to fit the local data.
2. *Dual-signal trust score:* It is calculated with a small constant $\varepsilon$ to ensure numerical stability by
$$T_k \;=\; \frac{1}{E_k\, D_k + \varepsilon}. \tag{3}$$
The score sharply discounts clients exhibiting high error, high divergence, or both.
3. *Adaptive gain aggregation:* The next global model is the trust-weighted average (Eq. 4)
$$w_{t+1} \;=\; \frac{\sum_k T_k\, w_{t+1}^{(k)}}{\sum_k T_k}. \tag{4}$$

Clients with high error and divergence receive low aggregation weight, while reliable clients contribute more. Repeated trust-weighted updates yield a robust, privacy-preserving global model.

### 3.2 Autoencoder Architecture

A deep convolutional autoencoder, trained on clean data, encodes kidney stone images into a 128-dimensional latent space and detects anomalies via reconstruction errors. It serves as both the global model and client evaluator, with the server distributing the current global autoencoder $w_t$ each round to represent the clean image distribution. Each client $C_k$ first passes its local dataset through $w_t$ to obtain a mean-squared reconstruction error $E_k$ (***Signal 1***). Next, the client then fine-tunes the autoencoder on its data, producing an updated model $w_{t+1}^{(k)}$. The server measures the parameter divergence (Eq. 2) (***Signal 2***) and derives a trust weight (Eq. 3). These weights are used to compute the next global model via a trust-weighted average (Eq. 4).

Thus, the autoencoder functions simultaneously as benchmarking tool, trainable local model, divergence source, and aggregation target, enabling FedAgain to down-weight corrupted clients and iteratively refine a robust, privacy-preserving global detector.

## 4 Datasets and Case Study

Kidney stone formation is a significant public health concern, especially in industrialized countries, affecting up to 10% of the population and showing recurrence rates of approximately 40% in the United States. Therefore, accurate

identification of stone types is crucial for personalized treatment and effective prevention strategies [5]. The current standard for kidney stone classification, Morpho-Constitutional Analysis (MCA), integrates visual examination and biochemical evaluation. However, MCA is time-consuming, reliant on expert interpretation, and unsuitable for real-time clinical use. An alternative, Endoscopic Stone Recognition (ESR), performed during ureteroscopy, remains subjective and is further complicated by variable lighting conditions, motion artifacts, and image noise [6].

To address these challenges, deep learning models have emerged as a promising solution to automate stone classification from endoscopic and non-endoscopic images [6, 13]. However, existing models are typically trained on local datasets, each presenting distinct challenges for classification and segmentation tasks. Consequently, there is a critical need for robust models capable of maintaining consistent performance in the presence of anomalies and artifacts, while also facilitating effective collaboration between multiple institutions.

Integrating this task within a FL framework offers a practical solution, enabling multiple institutions to collaboratively train models without sharing sensitive image data directly. Furthermore, FL provides an effective platform to assess and enhance model robustness against the variety of real-world artifacts encountered during clinical practice.

### 4.1 Datasets

In this contribution, two ex-vivo (extracted) kidney stone datasets were used for the evaluation of FedAgain. The key characteristics of these datasets are detailed as follows:

Dataset A [5] includes 366 CCD images (Charge-Coupled Device) of kidney stone fragments. Dataset A images were acquired under controlled illumination conditions, minimizing reflections and camera motion, with high spatial resolution (4288×2848 pixels) and image quality, and a uniform background.

Dataset B [6] contains 409 endoscopic images acquired with a flexible ureteroscope in an environment that simulates real acquisition conditions. Dataset B is close in a simulated clinical environment. Images in dataset B are visually similar to those captured during in-vivo utereroscopy, including changes in lighting, blur caused by camera movement, and limited spatial resolution (1920×1080 pixels).

All images were manually segmented under expert supervision. Preprocessing follows the protocol of [12], ensuring normalization and balancing of stone classes while excluding tissue background. Stone patches of fixed size 256×256 pixels are extracted to support robust training and evaluation splits of 80% for training and 20% for testing, with strict measures to prevent data leakage [13].

### 4.2 Endoscopy Corruptions

In this work, the existing framework provided by the *endoscopy corruptions* [17] library is leveraged to simulate clinically realistic image degradations in the kidney stone datasets. This library offers a comprehensive suite of 16 corruption

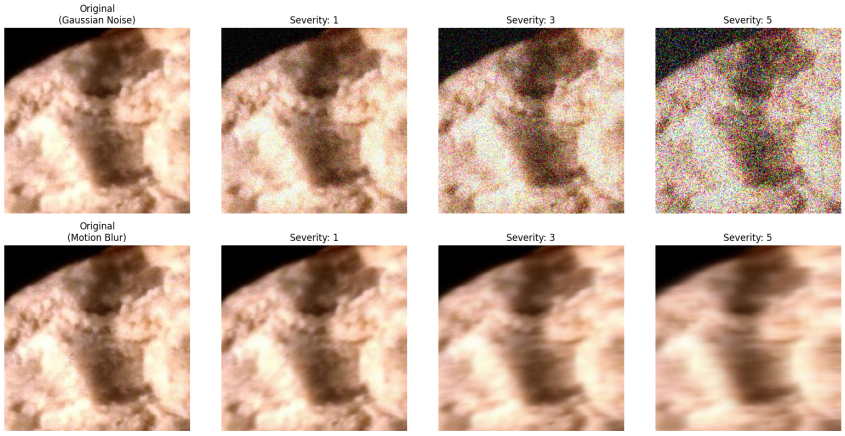


Fig. 2: Visualization of four representative image corruptions applied to a kidney stone endoscopic image. Each row shows the original image followed by increasing levels of corruption severity (1, 3, and 5).

functions, each designed to replicate visual artifacts commonly observed in endoscopic imaging, such as specular reflections, motion blur, sensor noise, and lens distortions.

Each corruption function accepts a clean input image and a severity level (1 to 5), see Figure 2, enabling a fine-grained evaluation of model robustness. By applying these perturbations to the dataset used in this study, a controlled experimental setup is created to assess the sensitivity of the proposed federated autoencoder to diverse and increasingly severe real-world image corruptions.

### 4.3 Configuration Setup

A federated learning setup with 20 clients is simulated, selecting 10 per round over 20 rounds. Each client trains a convolutional autoencoder locally for 3 epochs. While class labels are IID, 30% of clients receive uniquely corrupted images, introducing non-IID data quality. The global model is evaluated on 16 corruption types at 5 severity levels using the *endoscopycorruptions* library, with performance assessed by classification metrics.

## 5 Experiments and Evaluation

The proposed method was evaluated on two kidney stone datasets with various realistic corruptions, revealing its ability to handle decentralized, heterogeneous data and distinguish between trustworthy and unreliable clients. Table 1 presents the classification accuracy for 16 corruption types at severity level 4, evaluated across datasets A and B and three imaging views: mixed, section, and surface. The results reveal that both the dataset domain and the imaging perspective

Table 1: Accuracy performance across 16 corruption types at severity level 4 for different dataset variants. Column headers follow the format `<Dataset>-<View>`, where `A` and `B` denote the datasets, and `MIX`, `SEC`, and `SUR` indicate mixed, section, and surface views, respectively. Highest values per row are in bold.

| **Corruption** | **A-MIX** | **A-SEC** | **A-SUR** | **B-MIX** | **B-SEC** | **B-SUR** |
|---|---|---|---|---|---|---|
| brightness | 0.686 | **0.717** | 0.537 | 0.525 | 0.536 | 0.565 |
| color_changes | 0.615 | 0.645 | 0.565 | 0.707 | 0.718 | **0.750** |
| contrast | 0.500 | 0.500 | **0.500** | 0.500 | 0.500 | 0.500 |
| darkness | 0.500 | 0.500 | 0.500 | 0.516 | **0.528** | 0.503 |
| defocus_blur | **0.501** | 0.500 | 0.500 | 0.500 | 0.500 | 0.500 |
| fog | 0.648 | 0.698 | 0.574 | 0.679 | **0.728** | 0.719 |
| gaussian_noise | 0.972 | 0.968 | 0.986 | **1.000** | **1.000** | **1.000** |
| glass_blur | 0.516 | 0.521 | 0.503 | 0.516 | **0.535** | 0.531 |
| impulse_noise | 0.967 | 0.987 | 0.981 | **1.000** | **1.000** | **1.000** |
| iso_noise | 0.958 | 0.968 | 0.973 | **1.000** | **1.000** | **1.000** |
| lens_distortion | 0.762 | 0.839 | 0.753 | 0.949 | **0.950** | 0.944 |
| motion_blur | 0.500 | 0.500 | **0.500** | 0.500 | 0.500 | 0.500 |
| resolution_change | 0.500 | 0.500 | **0.500** | 0.500 | 0.500 | 0.500 |
| shot_noise | 0.990 | 0.987 | 0.991 | 0.999 | **1.000** | 0.999 |
| specular_reflection | 0.670 | 0.735 | 0.648 | **0.828** | 0.804 | 0.820 |
| zoom_blur | 0.500 | 0.500 | **0.500** | 0.500 | 0.500 | 0.500 |

influence how well the model handles degradation. Some corruptions, such as Gaussian and ISO noise, are consistently easier to detect, while others like resolution loss prove more challenging.

Performance gains become clearer when comparing against the standard FedAvg approach. As shown in Table 2, the A-MIX configuration of FedAgain achieves substantial improvements, with precision rising by up to +14.49%, recall by +7.78%, and F1 score by +10.20%. These gains reflect stronger anomaly detection and fewer false positives. Configurations A-SEC and A-SUR also show moderate benefits, while B-MIX and B-SUR yield smaller or even negative changes. Overall, the results point to the advantage of FedAgain, particularly when trained on diverse data distributions.

Table 2: Average percentage improvement over FedAvg [14] across multiple metrics and dataset variants. Highest values per row are highlighted in bold.

| **Metric** | **A-MIX** | **A-SEC** | **A-SUR** | **B-MIX** | **B-SEC** | **B-SUR** |
|---|---|---|---|---|---|---|
| AUC | **2.66%** | 1.77% | 0.13% | 0.89% | 1.34% | 0.62% |
| Accuracy | 0.80% | 0.65% | 0.13% | 0.39% | **0.98%** | 0.48% |
| Recall | **7.78%** | 3.90% | 2.06% | 0.56% | 6.12% | -0.27% |
| Precision | **14.49%** | 3.45% | 4.20% | -2.85% | 5.99% | -2.05% |
| F1 | **10.20%** | 3.65% | 2.74% | -0.72% | 5.31% | -0.42% |

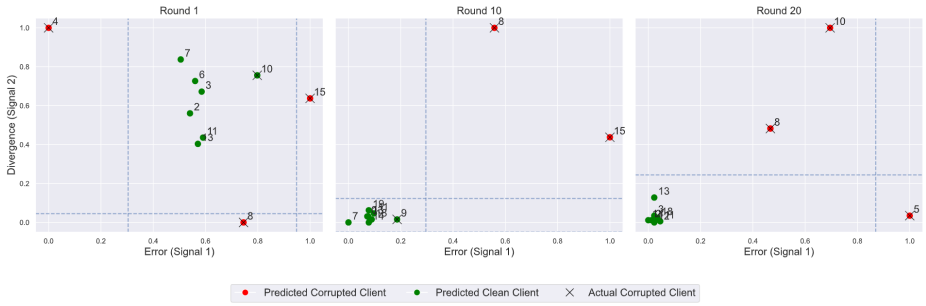


Fig. 3: Visualization of client behavior at rounds 1, 10, and 20 using FedAgain on Dataset A-SEC. The plot shows increasing separation between trustworthy and untrustworthy clients over time.

Figure 3 illustrates the evolution of client trust across rounds 1, 10, and 20. Initially, clean and corrupted clients are indistinguishable. By round 10, clean clients cluster in the low-error, low-divergence region, while corrupted ones begin to drift. By round 20, separation is clear, showing how FedAgain progressively identifies and down-weights unreliable clients, enhancing robustness and preserving privacy.

## 6 Conclusion

This work proposes FedAgain, a federated learning framework for robust anomaly detection in kidney stone imaging. By combining reconstruction error and model divergence as trust signals, FedAgain mitigates the *expert forger* problem caused by corrupted clients. Six specialized auto-encoders were trained in various configurations, and simulations with persistent corruptions showed effective anomaly detection and model robustness. FedAgain offers a scalable, privacy-preserving solution for medical imaging in federated settings.

## Acknowledgments

The authors wish to acknowledge the Mexican Secretaría de Ciencia, Humanidades, Tecnología e Innovación (Secihti) and CINVESTAV for their support in terms of postgraduate scholarships in this project, and the Data Science Hub at Tecnologico de Monterrey for their support on this project. This work has been supported by Azure Sponsorship credits granted by Microsoft's AI for Good Research Lab through the AI for Health program. The project was also supported by the French-Mexican ANUIES CONAHCYT Ecos Nord grant 322537. We also gratefully acknowledge the support from the Google Explore Computer Science Research (CSR) Program for partially funding this project through the LATAM Undergraduate Research Program.

## References


1. Bi, W.L., Hosny, A., Schabath, M.B., Giger, M.L., Birkbak, N.J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I.F., et al.: Artificial intelligence in cancer imaging: clinical challenges and applications. CA: a cancer journal for clinicians **69**, 127–157 (2019). https://doi.org/10.3322/caac.21552
2. Cai, Y., Chen, H., Cheng, K.T.: Rethinking autoencoders for medical anomaly detection from a theoretical perspective. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 544–554. Springer (2024)
3. Cao, X., Fang, M., Liu, J., Gong, N.Z.: Fltrust: Byzantine-robust federated learning via trust bootstrapping. arXiv preprint arXiv:2012.13995 (2020)
4. Chen, H.Y., Chao, W.L.: Fedbe: Making bayesian model ensemble applicable to federated learning. arXiv preprint arXiv:2009.01974 (2020)
5. Corrales, M., Doizi, S., Barghouthy, Y., Traxer, O., Daudon, M.: Classification of stones according to michel daudon: a narrative review. European Urology Focus **7**(1), 13–21 (2021)
6. El Beze, J., Mazeaud, C., Daul, C., Ochoa-Ruiz, G., Daudon, M., Eschwège, P., Hubert, J.: Evaluation and understanding of automated urinary stone recognition methods. BJU international (2022)
7. Guan, H., Yap, P.T., Bozoki, A., Liu, M.: Federated learning for medical image analysis: A survey. Pattern Recognition p. 110424 (2024)
8. Hamdi, M., Bourouis, S., Rastislav, K., Mohmed, F.: Evaluation of neuro images for the diagnosis of alzheimer's disease using deep learning neural network. Frontiers in Public Health **10**, 834032 (2022). https://doi.org/10.3389/fpubh.2022.834032
9. Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F.: Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence **2**(6), 305–311 (2020)
10. Laridi, S., Palmer, G., Tam, K.M.M.: Enhanced federated anomaly detection through autoencoders using summary statistics-based thresholding. Scientific Reports **14**(1), 26704 (2024)
11. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: Challenges, methods, and future directions. IEEE signal processing magazine **37**(3), 50–60 (2020)
12. Lopez, F.e.a.: Assessing deep learning methods for the identification of kidney stones in endoscopic images. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 2778–2781. IEEE (2021)
13. Lopez-Tiro, F., Estrade, V., Hubert, J., Flores-Araiza, D., Gonzalez-Mendoza, M., Ochoa, G., Daul, C.: On the in vivo recognition of kidney stones using machine learning. IEEE Access **12**, 10736–10759 (2024)
14. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
15. Ng, D.e.a.: Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. Quantitative Imaging in Medicine and Surgery **11**, 852 (2021)
16. Novoa-Paradela, D., Fontenla-Romero, O., Guijarro-Berdiñas, B.: Fast deep autoencoder for federated learning. Pattern Recognition **143**, 109805 (2023)

17. Reyes-Amezcua, I., Espinosa, R., Daul, C., Ochoa-Ruiz, G., Mendez-Vazquez, A.: Endodepth: A benchmark for assessing robustness in endoscopic depth prediction. In: MICCAI Workshop on Data Engineering in Medical Imaging. pp. 84–94. Springer (2024)
18. Reyes-Amezcua, I., Rojas-Ruiz, M., Ochoa-Ruiz, G., Mendez-Vazquez, A., Daul, C.: Leveraging pre-trained models for robust federated learning for kidney stone type recognition. In: Mexican International Conference on Artificial Intelligence. pp. 168–181. Springer (2024)
19. Shvetsova, N., Bakker, B., Fedulova, I., Schulz, H., Dylov, D.V.: Anomaly detection in medical imaging with deep perceptual autoencoders. IEEE Access **9**, 118571–118583 (2021)
20. Uddin, M.P., Xiang, Y., Hasan, M., Bai, J., Zhao, Y., Gao, L.: A systematic literature review of robust federated learning: Issues, solutions, and future research directions. ACM Computing Surveys **57**(10), 1–62 (2025)