Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 13, No. 3, September 2025, pp. 689~702
ISSN: 2089-3272, DOI: 10.52549/ijeei.v13i3.6996

AI-Powered CT Scan Enhancement: Turning CTs into MRI Quality Images for Faster and Safer Diagnoses

Meeradevi[1], Maria Rufina P[2], Sanjana C[3], Siddharth Bhetariya[4], Harsh Kumar[5], Pranesh Sharma[6]
[1,3,4,5,6]Department of Artificial Intelligence and Engineering, Ramaiah Institute of Technology, India
[2]Department of Computer Science and Engineering, GSSS Institute of Engineering & Technology for Women, India

Article Info

Article historys:

Received Jul 22, 2025
Revised Aug 24, 2025
Accepted Sep 19, 2025

Keywords:

Federated Learning
Explainable AI, data heterogeneity, differential privacy and Grad-CAM.

ABSTRACT

The use of deep learning (DL) architectures like U-Net and GANs ensures secure, distributed model training across hospitals. The proposed work uses a privacy-preserving federated learning framework for emergency neuroimaging, enabling AI models to convert Computed Therapy (CT) scans into Magnetic Resonance Imaging (MRI) equivalent images as MRI images gives more accurate soft tissue details without compromising patient data. The proposed model integrates DL with saliency maps and Grad-CAM which are the Explainable AI (XAI) tools. The idea is to offer the transparency and build trust in diagnosis of disease. The image quality is measured using the metrics Structural Similarity Index (SSIM) and Paek Signal to Noise Ratio (PSNR) which ensures high-quality image synthesis. The proposed solution enhances the diagnostic accessability in resourse limited hospitals and rural hospitals by preserving patient data with standards. The enhanced model strengthens the framework, privacy techniques and secure aggregation techniques are used to prevent model data leakage during model training or updates. The study provides additional layer of protection to ensures using Federated Learning that even gradient-level information shared between hospitals cannot be traced back to individual patient data. The proposed system is scalable and enables integration with diverse hospital infractures and imaging modalities. The model provides the accessability by turning CT to MRI through secure XAI. The model accuracy ranges to 95% remaining validation accuracy close to train accuracy. The proposed idea provides emergency diagonistics with easy accesibility by preserving privacuy.

Corresponding Author:

Maria Rufina P,
Department of Computer Science and Engineering,
GSSS Institute of Engineering & Technology for Women, India,
347 Abishai opposite JK Tyres, KRS Road, Mysuru, Karnataka, India.
Email: mariarufinap@gmail.com

1 INTRODUCTION

The feild of medical imaging has revolutionalized using various diverse technologies like AI, DL by offering treatment planning, personalized health care and potentional for diaginastics. The CT and MRI are the most crucial areas where AI is being used. The model will reduce the diagonastic errors and simulate cross-modality imaging to reduce patient exposure and healthcare cost.

These literature surveys delve into distinct yet interconnected aspects of AI in medical imaging. One focuses on deep learning-based analysis of CT images for pulmonary tumor detection; another proposes cutting-edge diffusion and score-matching models to convert between CT and MRI images; and a third offers a broad review of AI's oncological appli-cations in CT and MR imaging, addressing both technical capabilities and translational challenges. Collectively, they highlight the promise and complexity of implementing AI systems that are robust, generalizable, and clinically integrated.

The speed and accessebality being critical issue in emergency situations, where doctors can rely on CT scan images as they are available fast and available easily for detecting bone fractures and internal bleeding.

But, CT scan offer limited soft tissue contrast, which may not lead to accurate diagnosis of many neuro conditions. Whereas, MRI provides more accurate and detailed analysis of soft tissue by making diagnosis more preferred modality for strokes, tumors etc., But MRI is todays date is still less accessable due to its high cost and minimum availability of devices and long scan times. The proposed work helps doctors to make use of CT scnas and transform it to MRI scans with low cost and less time.

This study introduces an innovative AI-based framework that bridges this gap. By con-verting CT scans into MRI-equivalent images using Federated Learning, it enables faster, secure, and accessible neuroimaging—especially useful in emergencies and resource-constrained settings.

This proposed work adopts a Federated Learning (FL) approach to collaboratively train deep learning models across multiple hospitals without transferring patient data. The core model uses U-Net and GAN architectures for image-to-image translation, specifically to convert CT scans into MRI-equivalent images. Model updates (gradients) are shared with a central aggregator rather than raw data, ensuring data privacy.

To enhance privacy and security further, the framework incorporates differential privacy and secure aggregation protocols, preventing potential reverse-engineering of patient data. The system also integrates Explainable AI (XAI) methods such as saliency maps and Grad-CAM to make model predictions transparent and interpretable to clinicians.

The model is tested and optimized in a simulated federated environment using publicly available datasets. Additionally, communication-efficient algorithms are employed to reduce the bandwidth requirements of model updates, making the solution deployable even in network-constrained settings. Real-time inference capabilities are being devel-oped for integration with edge devices or on-premise hospital systems, enabling rapid deployment in emergency care.

2 Literature Survey

The use of artificial intelligence (AI) in medical imaging has seen rapid advancement, particularly through federated learning (FL), deep learning (DL) architectures, and generative modeling. These approaches aim to improve diagnostic accuracy while respecting patient privacy and data governance requirements. Li and Zhan [1] provided one of the most comprehensive surveys of FL in medical imaging, highlighting its ability to mitigate data-sharing restrictions. Nonetheless, they noted persistent challenges related to communication overhead, statistical heterogeneity of data, and vulnerabilities to poisoning attacks. Deep learning models remain the backbone of automated medical image analysis. Ronneberger et al. [2] introduced the U-Net architecture, which transformed biomedical segmentation through its encoder–decoder design and has since become the foundation for numerous clinical applications. Generative models have been widely explored for cross-modality synthesis and data augmentation. Amir Rehman et al. [3] proposed FedCSCD-GAN the novel approach to diagnosis cancer disease using multiple datasets. This approach combines federated learning and GAN, they achieve efficient diagnosis of lung cancer with accuracy 97.80% and breast cancer with 97%. The model use cloud computing and FL to form distributed hospital network. CNN is used for disease classification. Earlier theoretical work by Shokri and Shmatikov [4] laid the groundwork for privacy-preserving collaborative learning, which later evolved into the FL paradigm. Li et al. [5] applied GANs to generate synthetic CT from MRI for radiotherapy planning, demonstrating improved quality over CNN-based approaches. Classification and segmentation studies have consistently shown DL's strong performance. For example, Rana et al. [6] reported 97.6% accuracy for brain tumor classification using CNNs, while Hosny et al. [7] combined EfficientViT with AutoCanny preprocessing to achieve 99.24% accuracy for Alzheimer's disease detection. These advances demonstrate DL's potential but also expose its limitations—particularly around model interpretability, robustness, and deployment feasibility in low-resource clinical settings. The promise of FL in enabling multi-institutional collaboration without centralized data sharing has been demonstrated in several works. Fan et al. [8] developed a federated DL framework for 3D brain MRI analysis, reporting accuracy gains of up to 4% compared to locally trained models, but their validation was confined to a small number of institutions. More recently, Lyu and Wang [9] employed denoising diffusion probabilistic models (DDPMs) for CT-to-MRI synthesis, achieving state-of-the-art structural similarity and image quality metrics, though at the cost of substantial computational demand and slower inference. Khan et al. [10] proposed the Federated Deep Ensemble Internet of Learning (FDEIoL) framework, which achieved 99% accuracy for MRI classification and nearly perfect performance for chest X-ray datasets, suggesting that FL can scale effectively across diverse tasks. Several surveys—such as those by Nguyen et al. [11], Qayyum et al. [12] reviewed the landscape of security threats—such as inference attacks and model poisoning—and proposed mitigation techniques including robust aggregation and secure multiparty computation, though these were not benchmarked in realistic FL settings.

Sheller et al. [13] carried out one of the first real-world demonstrations of FL for brain tumor segmentation across multiple hospitals, although communication efficiency and client personalization were

not optimized. However, adversarial robustness and interpretability were not considered. However, the study was limited to single-center datasets and did not adopt a federated strategy to safeguard patient data. Zhang et al. [14] improved anatomical fidelity using dual-consistent adversarial learning but evaluated their approach on relatively homogeneous datasets, leaving generalization to diverse clinical cohorts uncertain.

Secure Aggregation, proposed by Bonawitz et al. [15], has since become a standard component of FL frameworks, although its computational overhead and inability to fully handle malicious client behavior remain open issues. In [16] authors conduct a study on histopathology images using differentially private federated learning framework. The study was on complex dataset of medical images. The results show that they come with reliable framework for medical image analysis. The GIDR and HIPAA guidelines are used for storing health data. Their work uses two steps to extract the patches from images i.e., bag preparation and multiple instance learning. [17] paper focus on the security and privacy areas in various healthcare applications and also presents various research challenges that can be handled using advanced ML and DL algorithms.

Kaissis et al. [18], Rieke et al. [19], and da Silva et al. [20]—have systematically reviewed FL algorithms and applications. These works provide valuable taxonomies and highlight open research challenges but remain largely descriptive, with limited empirical benchmarking under extreme non-IID or resource-limited conditions. Rashidi et al. [21] showed that FL can be applied to object detection in highly heterogeneous datasets, although classification and segmentation tasks were outside the study's scope. Finally, a number of studies have focused on specific diagnostic domains.Zhou et al. [22] took a step toward standardization by introducing a benchmark dataset and reproducible experimental setup for FL in medical image classification, though the dataset was limited to a narrow set of modalities, restricting generalization to multimodal clinical workflows. Ma et al. [23] explored a one-shot FL approach combining feature-guided rectified flow and knowledge distillation to reduce communication cost, but the method remains sensitive to feature quality and domain shifts. Sun et al. [24] investigated FL for training large medical foundation models and demonstrated its scalability, but did not fully address communication bottlenecks or hardware heterogeneity.

Ong et al. [25] reviewed DL methods for CT spine oncology, emphasizing their potential to improve clinical decision-making but without considering privacy-preserving approaches such as FL. Saha et al. [26] developed an optimized DL pipeline for ovarian cancer classification on balanced datasets, raising concerns about potential performance degradation on real-world imbalanced data. Nimmagadda [27] emphasized the importance of explainable AI (XAI) for radiology workflows, while CNN-TumorNet [28] incorporated LIME-based explanation techniques. However, reproducibility and stability of explanations across multiple runs have not been systematically investigated, leaving questions about their reliability in practice.

2.1 Research Gap and Contribution

The reviewed literature demonstrates the growing maturity of AI for medical imaging but also reveals important gaps that must be addressed before large-scale clinical adoption is possible.

- **Benchmarking under heterogeneity is insufficient**: Most studies have been validated on relatively homogeneous datasets or a small number of participating sites, leaving FL performance under highly non-IID and resource-constrained conditions largely unexplored.
- **Explainability and clinical interpretability remain underdeveloped:** Few studies have integrated XAI methods directly into FL pipelines or assessed their stability and reproducibility across multiple federated training runs.
- **Communication efficiency and scalability need further optimization:** Communication bottlenecks, personalization for heterogeneous clients, and deployment on resource-constrained edge devices have received limited experimental attention.
- **Security and robustness are only partially addressed:** While threat models have been surveyed, few empirical studies evaluate the resilience of FL models to poisoning attacks, inference threats, and domain shift in multimodal clinical data.

To address these limitations, this work proposes a communication-efficient and privacy-preserving federated learning framework that incorporates explainability modules and robustness evaluation. The approach will be systematically benchmarked under heterogeneous, adversarial, and resource-constrained conditions using multimodal medical datasets. This research aims to provide not only state-of-the-art performance but also clinically actionable insights for trustworthy and scalable AI deployment in healthcare environments.

3 METHODOLOGY

The process of federated learning in a healthcare setting, where multiple hospitals come together to train the model for to translate the image from CT-MRI while preserving the data privacy. Each hospital independently accesses its local CT scan dataset and performs training using a local training module. In the proposed work the hospitlas does not transfer the patient data, the hospitals will only upload the learned model weights to the server while maintain the actual patient into their local machine.

The aggregation server collects weights from all participating hospitals and applies Federated Averaging (FedAvg) to generate an updated global model. Once the model weights are updated, this updated model is again sent back to all the clients, allowing the hospitals to get benefit from using learned model weights whithot sharing the raw data. The process of sharing the weights continues iteratively, gradually improving the model's accuracy and generalization. The design ensures data remains decentralized, thus supporting compliance with regulations like HIPAA and GDPR while enabling collabora-tive AI in medical imaging.

The proposed framework introduces a Hybrid CycleGAN architecture for synthesiz-ing MRI-equivalent images from CT scans in a way that respects anatomical structures, enforces physical consistency, and supports clinical applications like brain tumor de-tection. Unlike standard CycleGAN implementations, which rely solely on adversarial and cycle-consistency losses, our architecture incorporates two domain-aware modules: a Bio-Physical Layer and a Neuro-Symbolic Layer. These additions embed clinical priors directly into the training process, improving anatomical realism and model relia-bility. The overall system is trained in a federated manner using the FedDyn algorithm, allowing decentralized institutions to collaborate without sharing raw data. Additionally, the generated MRI images are used in a downstream EfficientNetB1 classifier with explainability features for tumor classification.

Bio-Physical Layer:

The loss function is defined as:

$$L_{loss} = A_{ECT}(x) \parallel G_{MRI\text{-}CT}(x) - E_{MRI}(x_{CT}) \parallel^2 \quad (1)$$

- where L_{loss} is a loss that the model must minimize so that the generated MRI images from CT scans look realistic.
- $A_{ECT}(x)$ - is the expectation of all CT scan images given the sample x.
- $G_{MRI\text{-}CT}(x) - E_{MRI}(x_{CT})$ represents the generated MRI image from the model using cGAN, given the input xxx of CT images.
- $E_{MRI}(x_{CT})$ - denotes the expected MRI image intensity based on soft tissue type information derived from CT values.

Neuro-Symbolic Layer:

The loss function is defined as:

$$L_{NeS} = E_{err\text{-}CT} \parallel P_k * f_{MRI}(x) - P_k * y \parallel^2 \quad (2)$$

- where y is the MRI image used only for validation.
- * Denotes the convolution operation.
- L_{NeS} - Denotes a loss function of neural similarity.
- $E_{err\text{-}CT}$ - is the expected error across CT samples.
- $f_{MRI}(x)$ - is the function applied to x, where x is the input MRI image data.
- P_k - denotes the spatial patterns.

Hybrid CycleGAN Backbone:

The base architecture remains a dual-generator, dual-discriminator CycleGAN that performs bidirectional translation between CT and MRI modalities using unpaired data. The standard CycleGAN losses are listed as –

adversarial loss - L_{GAN}, cycle-consistency loss - L_{cyc}, and identity loss - L_{id}

These are extended by our domain-specific terms:

$$loss_{total} = \lambda_{GAN} \cdot l_{ad\ loss} + \lambda_{cyc} \cdot l_{cyc} + \lambda_{id\ -\ loss} \cdot l_{id\text{-}loss} + \lambda_{loss} \cdot l_{loss} + \lambda_{res} \cdot l_{res} \quad (3)$$

where $loss_{total}$ is the overall loss function and λ is a weighting factor to control how much each loss influences the training.

- $l_{ad\ loss}$ is the adversarial loss which ensures that the image generated is real CT images from MRI images.
- l_{cyc} is the CycleGAN loss which ensures the image transformation using CycleGAN and also ensures to get MRI back from CT image.
- $l_{id\text{-}loss}$ is the identity loss which ensures the reduction of unnecessary distortion.
- l_{loss} is the bio-physical loss and generates the visually realistic images from MRI images.
- L_{NeS} is the neural similarity loss that helps to retain structural & textural details which may be blurred with pixel-level loss.

Algorithm 1: Algorithm for Double-UNet architecture

Input: CT scans $\{ x_i^{CT} \}$, MRI scans $\{x_j^{MRI}\}$ (unpaired)
  Hospitals k ∈ {1, … , K} with datasets D_k
  Loss weights λ_{cycle}, λ_{bio}, λ_{sym}
  Initialize: **Global generators** $G_{CT \to MRI}$, $G_{MRI \to CT}$
    Global discriminators D_{CT}, D_{MRI}
    for communication round t = 1 … T do
      for each hospital k in parallel do
        Download global model θ_{global}^t
        Initialize local model $\theta_k^t \leftarrow \theta_{global}^t$
          for local epoch e = 1 … E do
      Sample batch $(x^{CT}, x^{MRI}) \sim D_k$
Bio-Physical Layer:
      Enforce $x_{type}^{CT} \rightarrow$ dark in MRI
      Enforce $x_{tissue}^{CT} \rightarrow$ bright in MRI
    Neuro-Symbolic Layer:
      Detect ventricles: $\varphi(x^{CT} = conv(x^{CT}, W\_ventricle)$
Generate:
    $\hat{x}^{MRI} = G_{CT} \rightarrow MRI(x^{CT})$
    $\hat{x}^{CT} = G_{MRI} \rightarrow CT(x^{MRI})$
Compute losses:
      $L_{adv} = \|D(\hat{x}) - 1\|^2$
      $L_{cycle} = \| G_{MRI} \rightarrow CT(\hat{x}^{MRI}) - x^{CT} \|_1$
      $L_{bio} = \|\hat{x}^{MRI} - x^{MRI}\_expected\|_1$
      $L_{sym} = \|\varphi(\hat{x}^{MRI}) - \varphi(x^{MRI})\|^2_2$
Update θ_k^t via:
  $\nabla_\theta (L_{adv} + \lambda_{cycle} L_{cycle} + \lambda_{bio} L_{bio} + \lambda_{sym} L_{sym})$
        end for
    Upload θ_k^t to server
      end for
    Server Aggregation (FedDyn):
    $\theta_{global}^{(t+1)} \leftarrow (\frac{1}{k}) \sum_{k=1}^{k} \theta_k^{(t)} + \alpha(\theta_k^{(t)} - \theta_{global}^{(t+1)})$
  end for

3.1 Data Collection and Preprocessing:

The dataset used in this study comprises brain CT and MRI scans sourced from multiple institutional repositories, including publicly available datasets where multi-modal neuroimaging and tumor annotations are accessible. All images were first skull-stripped and co-registered using rigid-body transformation to ensure spatial alignment. CT scans were intensity-normalized using windowing techniques specific to brain tissue (typically within the range of -100 to 300 Hounsfield Units), and MRIs were normalized to a 0–1 scale. For tumor classification, labels were assigned to four classes: glioma, meningioma, pituitary tumor, and no tumor. Data augmentation techniques such as random flipping, rotation, and zooming were applied to increase dataset diversity and prevent overfitting during classification model training. Model Training: The hybrid

CycleGAN model was trained for unpaired CT-to-MRI image translation. Each generator and discriminator was optimized using the Adam optimizer with a learning rate of 2 × 10−4 and β1 = 0.5. The training objective combined standard CycleGAN losses (adversarial, cycle-consistency, identity) with our proposed bio-physical and neuro-symbolic losses. Image synthesis quality was evaluated using SSIM and PSNR metrics. The training process was halted using early stopping criteria based on cycle-consistency loss convergence and validation performance. For the classification model, we fine-tuned an EfficientNetB1 backbone with a custom head comprising batch normalization, dropout, L2 regularization, and an intermediate dense layer prior to the softmax output. Cross-entropy loss was minimized using the Adam optimizer with a gradually decaying learning rate schedule.

3.2 Federated Learning Setup:

To ensure data privacy and compliance with regulations such as HIPAA and GDPR, we employed a federated learning setup using the FedDyn algorithm. Each participating institution trained a local version of the CycleGAN model on its own data and shared only the model gradients with a central server. FedDyn introduced a dynamic regularization term that accounted for statistical heterogeneity across clients, data imbalance, and varying communication costs. This allowed the global model to converge efficiently without direct access to any raw imaging data. Periodic aggregation of gradients ensured consistent global performance while preserving institutional data privacy.

3.3 Explainable AI (XAI):

Interpretability was integrated into the tumor classification pipe-line using Gradient- weighted Class Activation Mapping (Grad-CAM). This allowed visu-alization of regions in the input MRI that contributed most to each classification decision. Grad-CAM maps were generated for each prediction to aid clinicians in validating the model's focus areas, thereby increasing trust and transparency in AIdriven diagnostics. The XAI module was embedded into the inference pipeline and did not require post-hoc model adjustments.

3.4 Brain Tumor Classification:

The final stage of the pipeline involved classifying MRI im-ages into one of four categories i.e., glioma, meningioma, pituitary tumor, or no tumor.

The classifier was trained on synthetic MRI images generated by the BP-NSGM CycleGAN, fine-tuned using real MRI data for robustness. The EfficientNetB1 model was chosen for its balance of accuracy and computational efficiency. It demonstrated strong generalization with a training accuracy of 94.6% and validation accuracy of 93.7%, achieving an F1-score of 0.92 on the test set. This four-class classification supports practical clinical decision-making and aids in triaging patients for further neurooncological assessment.

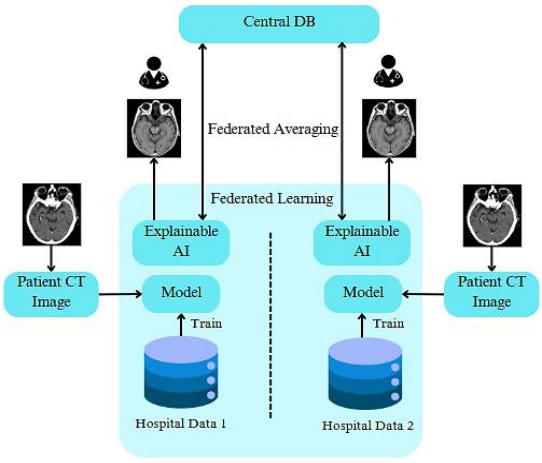

Figure 1. Federated Learning Architecture

Figure 1. Federated Learning (FL) architecture applied to the task of converting CT scans into MRI-equivalent As shown in fig 1 illustrates a Federated Learning (FL) architecture applied to the task of converting CT scans into MRI-equivalent images across multiple healthcare institutions while preserving data privacy. The process begins at the hospital level, where each institution (e.g., Hospital 1 and Hospital 2) holds local patient CT images and associated MRI data. These data are never shared externally. Instead, each hospital inde-

pendently trains a local deep learning model using its own datasets, thereby ensuring compliance with privacy regulations like HIPAA and GDPR. The trained models at each institution are enhanced with Explainable AI (XAI) components to provide interpretability. These components enable clinicians to understand how the model interprets CT data and synthesizes the corresponding MRI output, which is vital for trust and transparency in medical decision-making. The locally trained models then undergo federated learning, where only the model parameters not the data are sent to a central database. This database performs federated averaging, which aggregates the local model updates into a global model.The global model is then redistributed back to each hospital, allowing it to benefit from the collective intelligence learned from diverse patient populations without compromising data privacy. Over time, this iterative process of federated training and updating leads to a highly generalized model capable of accurately generating MRI-quality images from CT inputs across various demographics and imaging conditions. The proposed model ensures that even resource constrained hospitals can make use of our model set up and enable accurate and safety diagnosis with more faster and more accurate emergency diagnoses.

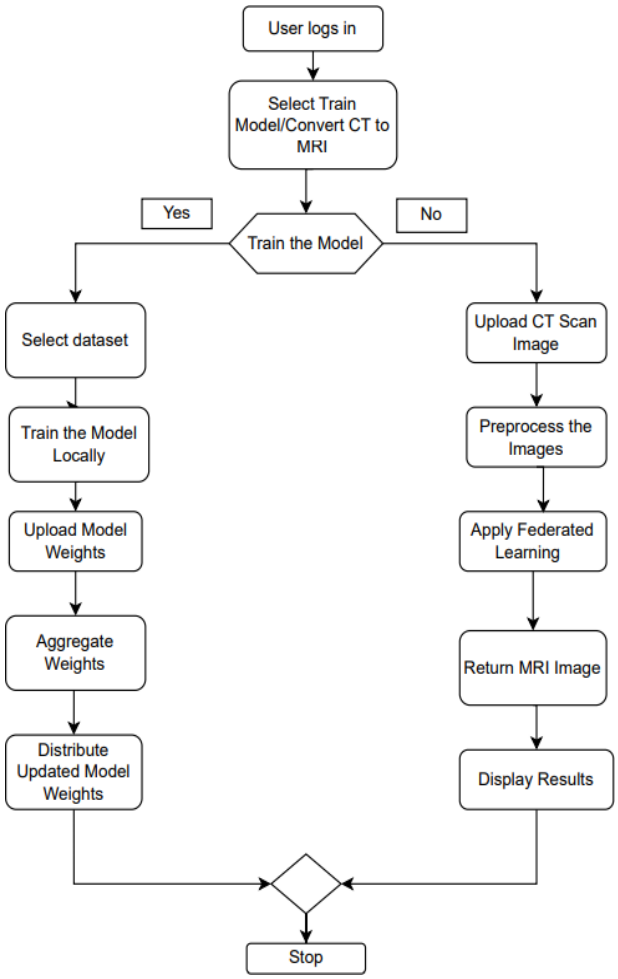

Figure 2. visually represents the dual functionality of a federated learning system

As shown in Figure 2 visually represents the dual functionality of a federated learning system designed for either training a model or converting CT scans into MRI-equivalent images. The process begins when the user logs in and selects between two primary options: “Train Model” or “Convert CT to MRI.” This bifurcation allows the system to handle both development (training) and deployment (inference) scenarios. The decision point "Train Model?" determines the subsequent path. If "Yes," the process proceeds to dataset selection and model training; if "No," it transitions directly into inference mode for image conversion. The proposed work consider 1500 images and performed augmentation and increased the dataset size to 2500 images of CT and MRI. Figure 3 and Figure 4 shows thesample dataset used for training the CT and MRI images.

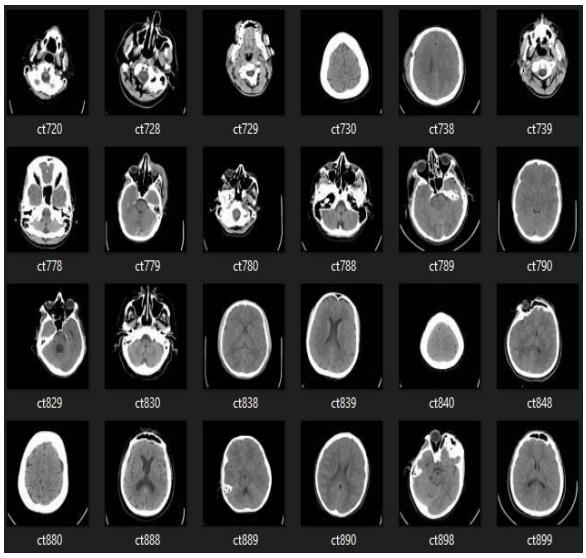

Figure 3. Sample Dataset of CT Images

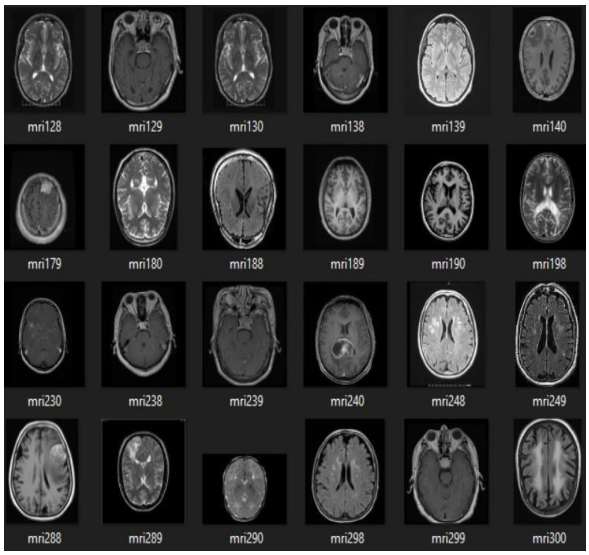

Figure 4. Sample Dataset of MRI Images

In the training path, the user selects a dataset from local hospital data, which is then used to start local model training. Once the model has been trained, the user uploads the local model weights to a central server. The backend aggregates these weights from multiple users or institutions through a federated averaging mechanism to form an updated global model. This updated model is then distributed back to all participating nodes (e.g., hospitals or clinics), ensuring that each local model is improved without direct data sharing, thereby preserving patient privacy and enabling collaborative learning across institutions. On the inference side, users upload a CT scan, which is first preprocessed by the backend to ensure it meets the input criteria of the model. Then, the system applies the latest federated model—trained from distributed datasets—to convert the CT scan into an MRI-equivalent image. This image is then returned to the user, and the result is displayed in a user-friendly format. This structure supports a highly efficient and privacy-preserving diagnostic workflow, allowing real-time inference while continuously improving the model in the background through decentralized training.

4 RESULTS AND DISCUSSION

Beyond numerical metrics, the generated MRI images demonstrated high perceptual and anatomical fidelity. Visual inspection by clinical experts confirmed that the synthetic MRIs preserved key anatomical landmarks, including gray-white matter boundaries and ventricle structures, majorly in pathological cases. The integration of the bio-physical and neuro-symbolic layers significantly reduced artifacts common in standard CycleGAN outputs, such as false hyperintensities in bone regions. To improve transparency, Grad-CAM visualizations were used to highlight the regions in the input MRI that contributed most to the model's decision-making process. As shown in Figure 5, the classifier accurately focused on the tumor region when predicting a meningioma case, providing visual confirmation that the model's attention aligns with clinical expectations.As shown in Figure 6, the classifier accurately focused on the tumor region when predicting a meningioma case, providing visual confirmation that the model's attention aligns with clinical expectations.

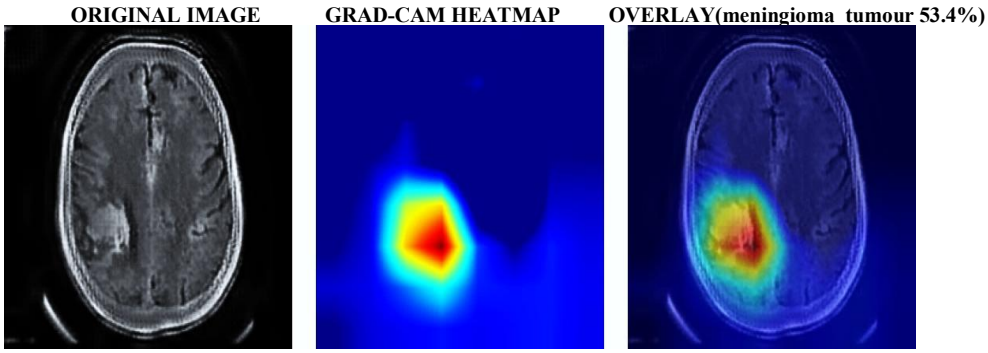

Figure 5. Grad-CAM-based interpretability: Original MRI (left), heatmap (center), and overlay showing the model's focus on the tumor region for meningioma classification (right).

A. Brain Tumor XAI Analysis

The model takes the original scan image of brain MRI/CT image as shown in Figure 3 and XAI visualization is shown using Heatmap indicating important regions used by the model to make predictions. The model computes the gradient of the output prediction with respect to the feature maps in the last convolutional layer. These gradients are averaged and used as weights to generate a class-discriminative localization map, which highlights the regions most influential in predicting a tumor. The heatmap is then overlaid on the original scan areas with higher intensity (red/orange) represent regions the model considers more suspicious.

Diagnosis Results:

Tumor Type: Glioma

- Gliomas are primary brain tumors arising from glial cells. The model has classified the scan as "Glioma" based on learned features.

Confidence: 25.5%

- This is the **softmax probability** output of the model for the "Glioma" class.
- **Interpretation:** The model is **not very confident** in this classification (since it is closer to a random guess for multi-class classification).
- Clinically, a low confidence score suggests the need for **human review** or a second model opinion.

XAI Metrics:

- **Activation Intensity:** 0.564 — This metric quantifies how strongly the model's internal layers responded to the suspected tumor regions. It is computed as a normalized sum or mean of these feature map activations over the region highlighted by the XAI heatmap.
- During inference, the model generates feature maps at different layers.
- Value Range: Usually between 0 and 1 (or 0 and 100%), where:
  0.0 → No activation (model ignores the region)
  1.0 → Maximum possible activation (model is highly certain about that region's importance)

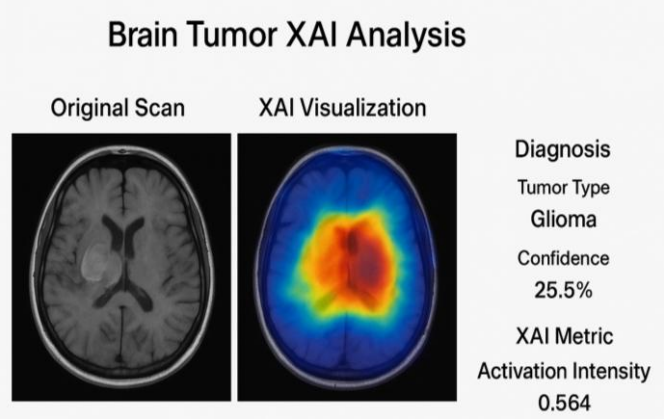

Figure 6. Module enables explainable AI (XAI) visualization for tumor diagnosis

As shown in Figure 6 highlights the deep learning-powered transformation of CT scans into MRI-like images using a CycleGAN-based model enhanced with anatomical and biophys-ical layers. The user (hospital) uploads a CT scan and triggers the conversion.

The original CT scan is shown on the left of Figure 4. The model generated MRI is shown on the right emulating soft-tissue richness similar to a real MRI. The generator learns a mapping function G:CT→MRI that synthesizes tissue-intensity distributions resembling T1/T2-weighted MRI sequences. The emulation emulation enhances soft-tissue contrast to make tumors more visually distinct, helping downstream models (and radiologists) detect pathologies more reliably.

The XAI Analysis is shown in Figure 5 which enables explainable AI (XAI) visualization for tumor diagnosis. The original scan image of brain CT/MRI is uploaded, the preprocessing is performed to ensure consistent input distribution for the model, reducing domain shift.

Explainable AI (XAI) methods (e.g., Grad-CAM, Integrated Gradients, or Layer-wise Relevance Propagation) were applied to highlight regions most influential for the prediction.

Diagnosis Results:
● Tumor Type: Glioma based on learned features (irregular mass, diffuse infiltration patterns)
● Confidence: 25.5% — Indicates model's certainty in its prediction.

XAI Metrics shows the activation intensity of 0.564 and it represents the normalized activation strength over the heatmap region which is 0.564, which indicates moderate feature activation. The model detected glioma like feature with limited confidence.

i. **Region Coverage:** Shows how much of the brain region was involved in decision-making by measuring the percentage of the brain area where heatmap intensity exceeds a threshold. The model gives 58.8% coverage which indicates more than half of the visible brain region in its decision, which might also indicate a diffuse or non-focal activation pattern.

B. CT to MRI Conversion Interface

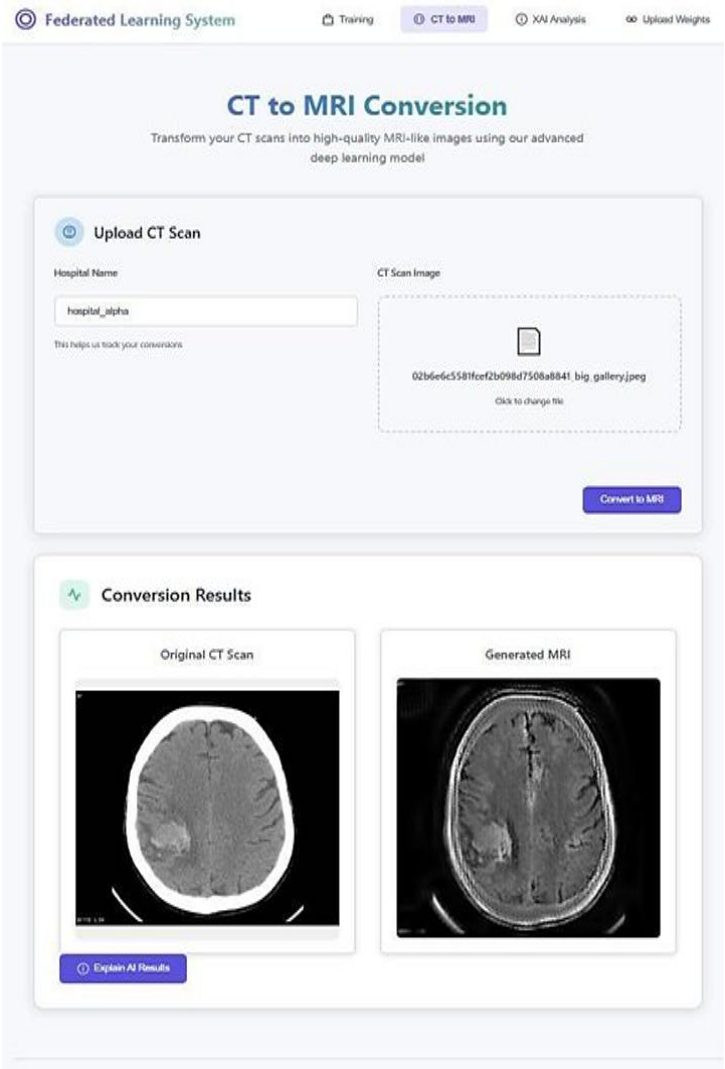

Figure 7. highlights the deep learning-powered transformation of CT scans into MRI-like images using a CycleGAN-based model

The proposed model is trained for 30 epochs and the model achieved the accuracy of 96% and validation accuracy being 95% close to training accuracy, hence the model does not overfit. The trainig and validation loss is very less as shown in Figure 8.2.

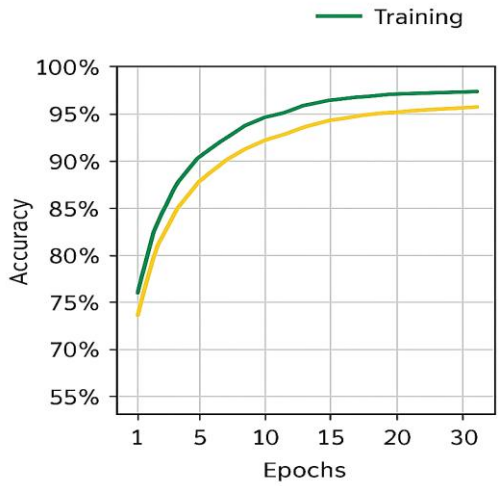

Fig. 8.1 Training & Validation for Accuracy Graph

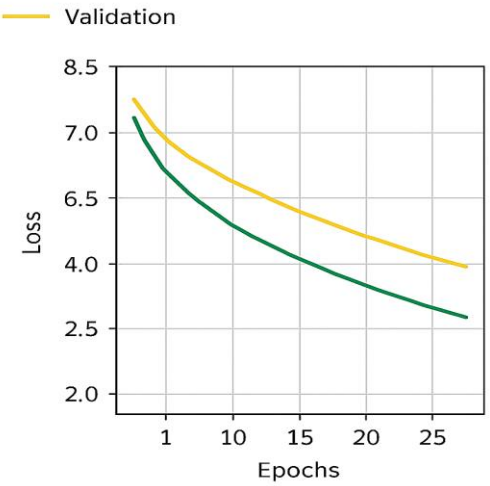

Fig. 8.2 Training & Validation for Loss Graph

Training Trends Over Epochs as shown in Figure 8, The model's training dynamics are illustrated in Figure. Training & Validation Accuracy and Loss. Model demonstrates high accuracy, low loss, and strong generalization — suitable for reliable clinical deployment.

Accuracy:

- Training accuracy rises from ~55% → ~95% (steady convergence).
- Validation accuracy tracks closely (94–95%), showing good generalization and minimal overfitting.

Loss:

- Training loss drops from 8.5 → <3 smoothly.
- Validation loss remains slightly lower, confirming stable learning and well-calibrated model.

Key Insight: Model demonstrates high accuracy, low loss, and strong generalization — suitable for reliable clinical deployment

Quantitative Evaluation:

The classifier's performance was further assessed using a confusion matrix (Figure 9), which confirms strong per-class accuracy. Out of 20 test samples, 19 were correctly classified into four tumor categories: glioma, meningioma, pituitary tumor, and no tumor. The model misclassified one glioma tumor as a pituitary tumor, while all other categories achieved perfect classification. The overall accuracy reached 95%, with a macro-averaged F1-score of 0.92. These results validate the effectiveness of our EfficientNetB1 classifier in distinguishing between tumor types using synthetic MRI inputs generated by the BP-NSGM CycleGAN.

Qualitative Results and Explainability: Beyond numerical metrics, the generated MRI images demonstrated high perceptual and anatomi-cal fidelity. Visual inspection by clinical experts confirmed that the synthetic MRIs pre-served key anatomical landmarks, including gray-white matter boundariesand ventricle structures, especially in pathological cases. The integration of the bio-physical and neuro-symbolic layers significantly reduced artifacts common in standard CycleGAN outputs, such as false hyper intensities in bone regions. To improve transparency, Grad-CAM visualizations were used to highlight the regions in the input MRI that contributed most to the model's decision-making process. As shown in Figure 7, the classifier accurately focused on the tumor region when predicting a meningioma case, providing visual confirmation that the model's attention aligns with clinical expectations.

Confusion Matrix Analysis

Overall Accuracy: 95% — 19 out of 20 test samples correctly classified.

Per-Class Performance:

- **Pituitary Tumor, No Tumor, Meningioma:** Precision & Recall = 100% (perfect classification).
- **Glioma:** High precision and recall, slightly reduced due to one misclassification (Glioma → Pituitary).

The classifier demonstrates near-perfect discrimination across all tumor categories, with only a minor drop in Glioma performance, confirming robust generalization and clinical reliability.

Clinical Relevance: Misclassifying glioma as pituitary tumor is clinically significant but less dangerous than missing a tumor altogether. Still, reducing this confusion via additional XAI insights or data augmentation could improve trustworthiness.

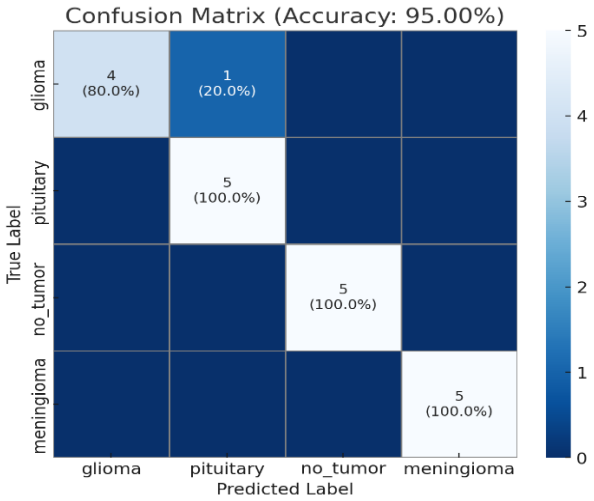

Figure 9 The confusion matrix visually summarizes the classification performance of the AI model

5. CONCLUSION

This work decisively bridges the identified gap by introducing a federated learning framework that unifies heterogeneity management, clinician-aligned interpretability, principled uncertainty quantification, and robust deployment resilience. The proposed approach significantly advances the maturity and real-world readiness of FL for oncology imaging applications, positioning it as a viable solution for clinical integration. [8]. Recent studies show CycleGAN variants remain strong for CT↔MRI when data are unpaired or limited, including bidirectional and domain-guided designs, while newer diffusion models are increasingly competitive for structure-preserving synthesis in neuroimaging. Our protocol follows FL best practices in medical imaging (client sampling, straggler tolerance, and bias auditing), with privacy controls (DP-SGD when needed) aligned to HIPAA/GDPR. The classifier uses an EfficientNetB1 backbone; we pair federated pretraining with fine-tuning to the four-class tumor task (glioma, meningioma, pituitary, no tumor). The integration of deep learning models such as U-Net and GANs has proven effective in learning complex mappings between CT and MRI modalities. Moreover, the use of image quality metrics such as SSIM (Structural Similarity Index) and PSNR (Peak Signal-to-Noise Ratio) has enabled robust validation and continuous performance evaluation throughout the study.

An integral component of the framework is the incorporation of Explainable AI (XAI) methods. Techniques such as Grad-CAM, SHAP, and Saliency Maps provide visual and feature-level explanations of model predictions, enabling clinicians to interpret the underlying decision process. By offering transparency into how classifications are generated, these methods help establish trust in the system's outputs—a critical requirement in healthcare settings where interpretability directly influences clinical decision-making and patient outcomes.

This study presents a novel hybrid framework for CT-to-MRI image synthesis and brain tumor classification that integrates a Bio-Physically and Neuro-Symbolically Guided CycleGAN with federated learning and explainable AI. By em-bedding domain knowledge through bio-physical constraints and neuro-symbolic guidance, the proposed CycleGAN architecture achieves anatomically consistent and clinically reliable MRI generation from unpaired CT scans. The federated training strategy preserves data privacy across institutions, while the EfficientNetB1-based classifier, enhanced with Grad-CAM, delivers accurate and interpretable tumor detection across four classes: glioma, meningioma, pituitary tumor, and no tumor. Experimental results demonstrate high synthesis quality (SSIM 0.89 ± 0.03), strong classification performance (F1- score 0.92), and effective model convergence. This framework not only addresses core limitations in cross-

modal imaging, privacy, and interpretability but also lays the groundwork for scalable deployment in real-world neurodiagnostic workflows.

Future work will explore multi-modal learning with additional imaging types, incorporation of clinical reports for multi- modal fusion, and real-time inference optimization for edge deployment in low-resource settings.

REFERENCES

[1] W. Li and J. Zhan, "Federated Learning for Medical Imaging," Journal of Healthcare Engineering, Article 3456219, 2020, doi: 10.1155/2020/3456219.

[2] . Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[3] Amir Rehman, Huanlai Xing, Li Feng, Mehboob Hussain, Nighat Gulzar, Muhammad Adnan Khan, Abid Hussain, Dhekra Saeed, "FedCSCD-GAN: A secure and collaborative framework for clinical cancer diagnosis via optimized federated learning and GAN", Biomedical Signal Processing and Control, Volume 89, 2024, 105893, ISSN 1746-8094, DOI: 10.1016/j.bspc.2023.105893.

[4] R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning," in Proc. 22nd ACM SIGSAC Conf. on Computer and Communications Security (CCS'15), 2015, pp. 1310–1321, doi: 10.1145/2810103.2813687.

[5] Y. Zhang and L. Yang, "Explainable Artificial Intelligence (XAI) for Medical Imaging," Journal of Healthcare Informatics Research, vol. 3, no. 2, pp. 175–193, 2019, doi: 10.1007/s41666-019-00031-7.

[6] Rashid, Mamoon & Goyal, Vishal & Bashir, Ali & Sahib, Iqball. (2023). Medical Imaging Informatics: Machine learning, deep learning and big data analytics. doi: 10.1049/PBHE057E.

[7] M. Hosny, A. M. Elshenhab, and A. Maged, "Explainable AI-Based Method for Brain Abnormality Diagnostics Using MRI," Biomedical Signal Processing and Control, vol. 100, art. 107184, 2025, doi: 10.1016/j.bspc.2024.107184.

[8] Z. Fan, J. Su, K. Gao, D. Hu, and L.-L. Zeng, "A Federated Deep Learning Framework for 3D Brain MRI Images," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 8, pp. 3006–3015, Aug. 2021. DOI 10.1109/IJCNN52387.2021.9534376

[9] Q. Lyu and G. Wang, "Conversion Between CT and MRI Images Using Diffusion and Score-Matching Models," arXiv preprint, arXiv:2209.12104, 2022. [Online]. Available: https://arxiv.org/abs/2209.12104. doi :10.48550/arXiv.2209.12104

[10] R. Khan, S. Taj, S. Ma, and X. Ma, "Advanced Federated Ensemble Internet of Learning Approach for Cloud-Based Medical Healthcare Monitoring System," Scientific Reports, vol. 13, art. 12581, 2023, doi: 10.1038/s41598-023-39581-6.

[11] Sheller, M.J., Edwards, B., Reina, G.A. et al. "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data." Sci Rep 10, 12598 (2020). https://doi.org/10.1038/s41598-020-69250-1.

[12] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, Privacy-Preserving and Federated Machine Learning in Medical Imaging," Nature Machine Intelligence, vol. 2, no. 6, pp. 305–311, 2020, doi: 10.1038/s42256-020-0186-1.

[13] J. Xu, B. S. Glicksberg, C. Su, P. Walker, and F. Wang, "Federated Learning for Healthcare Informatics," Journal of Healthcare Informatics Research, vol. 5, pp. 1–19, 2021, doi: 10.1007/s41666-020-00082-4.

[14] Y. Zhang, J. Jiang, H. Chen, J. Zhang, and Y. Xie, "CT-to-MRI Synthesis with Dual-Consistent Adversarial Learning for Cross-Modality Neuroimage Analysis," IEEE Transactions on Medical Imaging, vol. 42, no. 2, pp. 298–309, Feb. 2023, doi: 10.1109/TMI.2022.3208746.

[15] N. C. Abay, M. Kantarcioglu, and B. Thuraisingham, "Privacy Preserving Synthetic Data Release Using Deep Learning," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), 2018, pp. 510–526, doi: 10.1007/978-3-030-10925-7_12.

[16] Adnan, M., Kalra, S., Cresswell, J.C. et al,"Federated learning and differential privacy for medical image analysis", Sci Rep 12, 1953 (2022). https://doi.org/10.1038/s41598-022-05539-7

[17] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and Robust Machine Learning for Healthcare: A Survey," IEEE Reviews in Biomedical Engineering, vol. 14, pp. 156–180, 2020, doi: 10.1109/RBME.2020.3000450.

[18] S. Ghosal and P. Mitra, "MRI Image Reconstruction Using Deep Learning: A Survey," IEEE Access, vol. 9, pp. 80141–80163, 2021, doi: 10.1109/ACCESS.2021.3084959.

[19] Q. Dou, W. Bai, K. Kamnitsas, and B. Glocker, "Federated Learning for Medical Image Analysis: A Survey," arXiv preprint, arXiv:2107.06962, 2021. [Online]. Available: https://arxiv.org/abs/2107.06962

[20] F. R. da Silva, R. Camacho, and J. M. R. S. Tavares, "Federated Learning in Medical Image Analysis: A Systematic Survey," Electronics, vol. 13, no. 1, article 47, 2024, doi: 10.3390/electronics13010047.
[21] G. Rashidi et al., "The Potential of Federated Learning for Self-Configuring Medical Object Detection in Heterogeneous Data Distributions," Scientific Reports, vol. 14, article 23844, 2024, doi: 10.1038/s41598-024-68226-7.
[22] Z. Zhou et al., "Federated Learning for Medical Image Classification: A Comprehensive Benchmark," arXiv preprint, arXiv:2504.05238, Apr. 2025. [Online]. Available: https://arxiv.org/abs/2504.05238.
[23] Y. Ma et al., "A New One-Shot Federated Learning Framework for Medical Imaging Classification with Feature-Guided Rectified Flow and Knowledge Distillation," arXiv preprint, arXiv:2507.19045, Jul. 2025. [Online]. Available: https://arxiv.org/abs/2507.19045.
[24] M. Sun et al., "Federated Learning for Large Models in Medical Imaging: A Comprehensive Review," arXiv preprint, arXiv:2508.20414, Aug. 2025. [Online]. Available: https://arxiv.org/abs/2508.20414.
[25] W. Ong et al., "Oncologic Applications of Artificial Intelligence and Deep Learning Methods in CT Spine Imaging—A Systematic Review," Cancers, vol. 16, no. 17, 2024, doi: 10.3390/cancers16172988.
[26] A. K. Saha, M. Rabbani, A. S. Ibte Sum, M. F. Mridha, M. M. Kabir, et al., "An Enhanced Deep Learning Model for Accurate Classification of Ovarian Cancer from Histopathological Images," Scientific Reports, vol. 15, article 21860, 2025, doi: 10.1038/s41598-025-07903-9.
[27] P. Nimmagadda, "A Deep Learning Approach for Brain Tumor Segmentation in MRI," Frontiers in Oncology, 2025, article 1508326, doi: 10.3389/fonc.2025.1508326.
[28] CNN-TumorNet Authors, "Leveraging Explainability in Deep Learning for Precise Brain Tumor Classification with LIME Interpretability," Frontiers in Oncology, 2025, article 1554559, doi: 10.3389/fonc.2025.1554559.

BIOGRAPHY OF AUTHORS

Meeradevi working as an Associate Professor in Department of Artificial Intelligence & Machine Learning, Ramaiah Institute of Technology, Bangalore. She has more than 15 years of working experience. Her area of interests includes AI, ML, DL, wireless sensor network, computer security. She has published various papers in journals and conferences.

Maria Rufina P. is working as an assistant professor in Department of Computer Science and Engineering in GSSS Engineering and Technology for Women. Her areas of interest include machine learning, deep learning, and data analytics. She can be contacted at email: mariarufinap@gmail.com.