FrugalVision: Adaptive Medical Image Compression for Resource-Constrained IoT Environments

Amina Guidoum, Abd Eldjalile Lafkir

Computer Science Department, Higher Normal School Kouba
Algiers, Algeria (amina.guidoum@g.ens-kouba.dz, lafkir.abdeldjalile35@g.ens-kouba.dz)

Abstract: The energy and hardware constraints of medical IoT devices limit the deployment of real-time medical imaging solutions in remote environments. This work presents FrugalVision, an integrated solution for the reliable transmission of medical images in degraded network environments. The system combines three innovations: a residual autoencoder with adjustable differentiable compression, a reinforcement learning agent (PPO)dynamically adapting compression parameters, and the MQTT protocol for data and metrics communication in a constrained network. Validated on three Decathlon medical segmentation datasets (hippocampus, heart, brain tumors), the solution maintains image quality with average PSNRs of 37.63 dB, 33.05 dB, and 35.02 dB, respectively, consistently exceeding the clinical threshold of 28 dB. Comparative analysis with state-of-art methods demonstrates a PSNR difference of up to +8.76 dB compared to conventional methods. Processing latency is 0.15 to 0.26 s per image, with adaptation decisions taking 0.6 to 2.3 ms. Modeled power consumption is 0.02 to 0.04 J on a standard architecture without dedicated hardware acceleration. The architecture enables real-time adaptation to network variations using a simulator based on real network traces from the 5GMeas dataset. This approach supports the deployment of medical imaging solutions in underserved areas.

Keywords: Adaptive compression, Residual autoencoder, Constrained Networks, Medical images, IOT.

1. INTRODUCTION

The exponential growth of medical imaging data, particularly in the context of 3D volumes (MRI, CT, ultrasound, etc.), poses significant challenges for their transmission in constrained environments. Medical IoT devices face strict energy and hardware limitations, while networks in underserved areas often exhibit degraded conditions with limited bandwidth, high latency, and packet loss. These constraints make it difficult to deploy real-time solutions for medical imaging. Traditional compression techniques, such as JPEG, offer fixed tradeoffs between quality and size, but lack adaptability to dynamic network conditions and the specific requirements of the medical field. Preserving critical anatomical details is essential to maintain diagnostic value, requiring sophisticated approaches beyond conventional methods. These challenges manifest themselves in concrete use cases: mobile healthcare units operating in rural or isolated areas, remote expert consultation systems where diagnoses must be transmitted in real time, and medical emergencies where latency and reliability are critical. In all these scenarios, network and energy resources are limited, and image quality must not compromise clinical interpretation. This is why an integrated solution, capable of dynamically adapting compression to real-world conditions, is essential.

Several approaches to medical image compression have emerged. (Kumar et al., 2020) highlight that nearly 84% of hospitals now use electronic medical records, which requires efficient solutions for storage and transmission. Three broad categories of techniques are analysed: lossless compression, lossy compression (which deliberately sacrifices some information to achieve high reduction rates, via approaches such as volumetric JPEG or vector quantization), and hybrid methods (combining the two previous approaches to optimize both the size and quality of critical regions).

The performance of these techniques is evaluated through key metrics such as PSNR for peak signal-to-noise ratio (visual quality) (Battisti et al., 2006), compression ratio (CR),MSE for mean-square error (Wang et al., 2003), and SSIM (structural fidelity) (Wang et al., 2004). (Kumar et al., 2020) reveals that hybrid methods emerge as the most promising, with significant gains (e.g. +54% in efficiency compared to JPEG2000). However, challenges persist: algorithmic complexity, high computational time, difficulty in preserving quality in watermarking techniques, and limitation to certain imaging modalities (MRI/CT dominate).

(Ungureanu et al., 2024) provides an in-depth analysis of different image compression techniques, distinguishing between classical and region-of-interest (ROI)-based approaches. The authors highlight the growing importance of compression in a digital world where the quantity of images produced in the medical, automotive, satellite, surveillance, and multimedia fields is constantly increasing. Classical methods such as JPEG, JPEG2000, PNG, as well as those based on the discrete cosine transform (DCT), the discrete wavelet transform (DWT), and some autoencoder neural models are reviewed, each presenting a different balance between compression ratio and quality preservation. However, these approaches remain limited when it comes to preserving critical areas of an image, for example in medical imaging. To overcome these limitations, ROI techniques apply algorithms capable of prioritizing the preservation of essential information (such as a lesion in an MRI) while applying more aggressive compression to secondary parts of the image. These methods rely on segmentation, adaptive coding, or artificial intelligence approaches such as deep neural networks. The applications are numerous: reliable medical diagnoses, more efficient telecommunications and video streaming, as well as management of large amounts of data from satellite imagery and surveillance. The article compares the performance of these different methods using criteria such as compression rate, visual quality (PSNR), computation time, and suitability for the intended application. The results show that while traditional methods remain suitable for general use, ROI approaches offer a decisive advantage when it comes to preserving vital information. In conclusion, the authors highlight the need to further integrate artificial intelligence to meet the growing needs for real-time performance.

This work addresses these challenges through FrugalVision, a comprehensive architecture that simultaneously optimizes image quality, energy efficiency, and network adaptation. The system combines a neural compression model with an integrated thresholding mechanism, a PPO agent proximal policy optimization (Schulman et al., 2017), for adaptive decision-making, and MQTT in (Lin et al., 2025; Stanford-Clark et al., 2013) communication .

The solution was validated on the three datasets, with automatic selection of relevant anatomical slices and robust percentile-based normalization. The integration of a network simulator using real traces from the 5GMeas dataset (Purdue, 2022) allows for realistic evaluation under varied conditions.

This research contributes to the field of frugal medical electronic by proposing an integrated solution for deploying imaging applications in resource-constrained environments without compromising diagnostic quality requirements.

2. RELATED WORKS

Previous work in medical image compression has mainly focused on optimizing existing standards such as JPEG (Taubman et al., 2013) and lossless compression techniques. These approaches, although effective for generic compression, lack specificity for the particularities of medical images and do not offer dynamic adaptation to network conditions.

In the field of deep learning, the work of (Ballé et al., 2018) presents a generic image compression method based on variational autoencoders (VAEs) with a scaling hyperprior to capture spatial dependencies in the latent representation. The authors compare their models to other compression methods (JPEG…, etc.) on standard image databases such as Kodak or Tecnick, using metrics such as PSNR and MS-SSIM. IOT, real-time, or energy-efficient aspects are not addressed. The paper focuses on rate-distortion optimization and performance improvements through the introduction of a hyperprior.

This article (Liu et al., 2017) analyses the importance of international image compression standards in the medical context. The authors review the key standards: JPEG (limited to 8 bits, poorly suited to high-resolution medical images), JPEG2000 (using wavelets, supporting up to 16 bits and offering scalable compression), JPEG-LS (optimized for lossy compression), JPEG-XR (supporting 32 bits), and HEVC/H.265. These standards are evaluated for their integration into DICOM, the dominant standard for medical image exchange, which already includes transfer syntaxes for JPEG, JPEG2000, JPEG-LS, and MPEG. Comparative performance tests show that HEVC and JPEG2000 dominate in lossy compression, particularly for 3D volumes. In digital pathology, JPEG2000 outperforms HEVC.

This study (Kurmukov et al., 2024) explores the impact of lossy compression (JPEG2000) on the segmentation quality of 3D medical images (CT and MRI) using deep neural networks (U-Net). The results demonstrate that compression up to ×20 does not impair segmentation performance, for 20 segmentation tasks on 4 different datasets (AMOS, LiTS, BraTS, BGPD). Moreover, models trained on compressed or uncompressed data can be used interchangeably without loss of quality. This approach allows a significant reduction in storage costs without compromising clinical applications.

Identified challenges include: medical image variability rendering fixed compression ratios ineffective, the need for perceptual quality metrics aligned with the human visual system, and the unrecognized impact of compression on post-processing applications (such as CAD or 3D reconstructions). The authors conclude that clinical adoption of standards depends as much on their technical effectiveness as on regulatory frameworks.

This paper (Hannuksela, 2020) presents a new image compression method called CLRIC (Compact Latent Representation for Image Compression), which relies on the use of latent representations from pre-trained generative models (such as the Stable Diffusion autoencoder) to achieve efficient perceptual compression. Unlike traditional approaches that require separate models for each quality level, CLRIC uses an overfitted neural function operating in the latent space to encode an image at any target rate, without having to retrain the base model. This method allows for a continuous quality representation, low decoding complexity (around 25.5 MAC/px), and efficient resource utilization. Limitations include high encoding time (~10 min/frame) and unsuitability for applications requiring exact pixel-to-pixel fidelity.

This paper (Liao and Li, 2024) proposes a novel lossless medical image compression method, named LC-TMNet (Learned Lossless Medical Image Compression with Tunable Multi-Scale Network), which aims to solve the problem of inaccurate probabilistic estimation of neural networks in regions with complex or border textures. The method uses a flexible tree-based image segmentation mechanism, dividing the image into level 1 and level 2 sub-images via quadtree and then binary tree division, thus exploiting the close relationships between sub-images to provide strong prior knowledge and improve the accuracy of probabilistic inference. The network architecture integrates an attention mechanism (Squeeze-and-Excitation) into a UNet model to strengthen the estimation of residual probabilities. The compression framework allows two modes: a fast mode (orange) using the LC-H module for fast compression of level 1 sub images, and a slow mode (green) using the LC-H-Extend and LC-L modules for more detailed compression of level 2 sub images, thus providing a continuous representation of quality. The experiments, conducted on medical (CT, MRI, ultrasound) and natural (COCO grayscale) datasets, limitations include high GPU memory consumption and restricted applicability to single-channel images.

This paper (Li et al., 2023) presents an image compression method that encodes the Region of Interest (ROI) with better quality than the background. This approach is useful for applications such as video conferencing systems, video surveillance, and object-oriented computer vision tasks.

The method proposes a Region of Interest-based compression framework using Swin Transformer architectures as the core building blocks for the autoencoder network. A binary mask identifying the ROI is integrated into different layers of the network to guide processing based on spatial location.

Controlling the relative importance given to the ROI compared to the background is enabled by modifying the Lagrange multiplier (λ) associated with each region during optimization.

Experimental results indicate that the model achieves a higher PSNR 31 dB for the ROI than other compared methods, as well as a decent average PSNR for human evaluation. Additionally, when tested on pre-trained models with original images, it demonstrates performance in object detection and instance segmentation on the COCO validation dataset.

This study (Li et al., 2025) proposes a hybrid framework called SDWTCNN combining singular value decomposition (SVD), discrete wavelet transform (DWT), and a convolutional neural network (CNN) for medical image compression. This approach enables scalable compression with optimal preservation of visual quality. Experimental results on the BGPD (CT) and BraTS (MRI) datasets show a significant improvement in peak signal-to-noise ratio (PSNR) over existing methods, with gains of 4.3 dB (PSNR = 28.87 dB on BGPD) and 3.8 dB (PSNR = 29.81 dB on BraTS), respectively.

However, the approach has some limitations. The authors highlight a risk of overfitting when training on limited or homogeneous datasets, which could limit its generalization to various image types. Moreover, scalability remains a challenge for very large volumes of data, particularly due to the computational complexity associated with the joint use of SVD, DWT and CNN.

This paper (Ye et al., 2023) presents FSRCNN (Fast Super-Resolution Convolutional Neural Networks), an accelerated and improved version of SRCNN for image super-resolution. they propose three key architectural: replacing interpolation with a deconvolution layer at the end of the network for direct low-resolution → high-resolution training, inserting dimension reduction and expansion layers on either side of the mapping layer, and using smaller filters but a deeper network. Results show that FSRCNN outperforms SRCNN-Ex (33.16 dB vs. 32.75 dB on Set5 for ×3) while being more than 40 times faster.

This article (Yan et al., 2020) proposes a novel method for compressed sensing MRI (CS-MRI) image reconstruction using neural architecture search (NAS) to automatically design optimized deep neural networks. Unlike traditional manual approaches, this method achieves a specific and efficient architecture for MRI reconstruction, achieving superior PSNR performance (29.83 dB for the knee dataset and 33.27 dB for the brain dataset) while reducing computational resources by 4 to 6 times compared to state-of-the-art methods.

In parallel, other approaches based on the entropy autoencoder have demonstrated remarkable compression performance on natural images. HiFiC (Mentzer et al., 2020) combines a variational autoencoder with a generative model to preserve high visual fidelity. ELIC (He et al., 2022) improves the efficiency of entropy coding by using a spatial context and a joint probability model. NAFNet (Chen et al., 2022) offers a simplified architecture without batch normalization, aiming for a good quality-complexity trade-off. While effective in the laboratory, these methods often rely on large models and costly entropy coding operations, which limits their deployment on resource-constrained IoT devices.

Unlike previous work that typically treats compression, network adaptation, and energy efficiency as separate problems, FrugalVision offers a holistic solution that simultaneously optimizes these aspects through an integrated architecture. This approach enables significant efficiency gains through coordination between different system components.

3. METHODS

3.1 Architecture of the compression model

The medical image compression architecture as seen in (Fig.1 and Algorithm 1) follows a residual-connected encoder-decoder approach, specifically designed to maintain high peak signal-to-noise ratio (PSNR) while enabling adaptive compression.

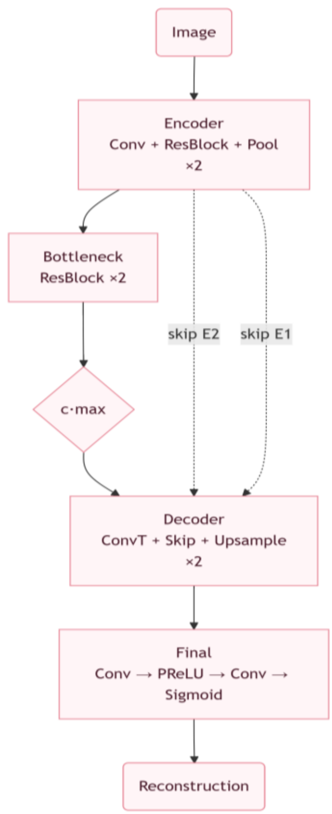

Fig. 1. FrugalVision compression architecture.

The encoder starts with a 3x3 convolution layer with 64 filters followed by a residual block, gradually reducing the spatial resolution through max pooling operations. The second stage uses a residual block with a stride of 2 to further reduce the spatial dimension. The bottleneck has two residual blocks with 256 filters and incorporates an adaptive thresholding compression mechanism that applies a ReLU activation function to the absolute feature values after subtracting a threshold proportional to the compression level.

The decoder uses transposed convolutions to perform up sampling, with skip connections concatenating the corresponding encoder features to preserve fine details. The final reconstruction is achieved through a sequence of convolutions comprising a 3x3 convolution layer with 32 filters, a ReLU activation, and a 1x1 convolution followed by a sigmoid to produce an output in the [0,1] range. This architecture enables high-quality reconstruction while allowing precise control of the compression ratio via the thresholding parameter in the bottleneck.

3.2 Integrated compression mechanism and loss function

The innovative compression mechanism directly integrates the compression process into the model's forward pass. Unlike traditional approaches that apply compression post-processing, the method operates directly on the network's intermediate characteristics. The compression threshold is adaptively calculated as the product of the compression level and the maximum absolute value of the bottleneck activations. This approach is inspired by neural network pruning techniques (Hanet al., 2015) but adapts them to enable dynamic and controllable compression.

The hybrid loss function combines three components: MSE for pixel-to-pixel fidelity, a structural similarity index (SSIM) loss to preserve image structure (Wang et al., 2004), and an L1 loss for regularization. The respective weights of 0.4, 0.4, and 0.2 were empirically determined to optimize both objective metrics and perceived visual quality. The SSIM implementation includes adaptive window size management to accommodate different medical image dimensions.

3.3 Reinforcement learning environment

The reinforcement learning framework formalizes the network adaptation problem as a Markov decision process. The state space includes three normalized dimensions: bandwidth, latency, and packet loss rate. The discrete action space includes five compression levels equally spaced between 0 and 1. The reward function combines the PSNR of the reconstructed image and the adequacy of the compressed size to the available bandwidth, with an additional bonus when the PSNR exceeds 32 dB;

The PPO algorithm is used to train the agent, with a multi-layer policy network (MLP). Hyperparameters include a learning rate of 10^{-4}, a horizon of 256 steps, and an entropy coefficient of 0.01 to encourage exploration. The implementation uses gradient clipping and advantage normalization techniques to stabilize training.

3.4 IoT integration and MQTT communication

The system integrates MQTT communication for data exchange in constrained environments. Three main topics are used: "frugalvision/network" for network conditions, "frugalvision/image" for compressed images, and "frugalvision/metrics" for performance metrics. Images are encoded in JPEG with variable quality and transmitted in base64 format in JSON messages (Stanford-Clark and Truong, 2013). An energy consumption model differentiates several device types (Raspberry Pi, standard CPU, GPU server) and calculates consumption based on processing time and device-specific constants.

The energy consumed is calculated according to the following equation:

$$E = P_0 \times t + \alpha \times t^2 \quad (1)$$

Where E: Energy consumed (in Joules), P_0 :Base power of the device (in Watts), t: Processing time (in seconds), α: Non-linear increase coefficient (0.001 W/s)

From (1) : Linear term ($P_0 \times t$): Represents the basic energy consumption of the device, corresponds to the static and dynamic power during processing and is proportional to the execution time.

From (1) : Quadratic term ($\alpha \times t^2$): Captures the nonlinear effects of increasing power consumption, models component heating and energy efficiency degradation and represents saturation phenomena during prolonged processing.

3.5 Comprehensive evaluation methodology

The comparative methodology between FrugalVision and the JPEG standard was conducted according to a rigorous experimental protocol. For the FrugalVision method, medical images from the test sets were selected and compressed with different levels (0.00, 0.25, 0.50, 0.75) via the integrated adaptive thresholding mechanism, then saved in PNG format with quantization of the file sizes. For the JPEG standard, the same original images were encoded with different quality levels (Q20, Q25, Q30, Q35, Q40, Q45) using an iterative search process aimed at identifying the quality producing a PSNR equivalent to that obtained by FrugalVision. This systematic search for the equivalent JPEG quality was carried out via the find_jpeg_quality function with a tolerance of 0.5dB. The comparative evaluation included calculating PSNR for both methods, measuring file sizes for each compression level, and generating histograms for qualitative analysis of intensity distribution. File size metrics were calculated in kilobytes (KB) for both methods, allowing a direct comparison of compression efficiency. This methodology ensures a fair comparison between FrugalVision's adaptive approach and the fixed-quality JPEG standard, combining objective metrics PSNR with visual assessment by histograms.

Evaluation metrics include PSNR, SSIM, processing time, RL agent decision time, and estimated energy consumption. The evaluation protocol uses a separate test set and comparative visualizations to qualitatively assess the results.

In order to position the results with respect to the state of the art, an additional comparison was made with the classical benchmark methods referenced in (Li et al., 2025; Yan et al., 2020). These methods include conventional medical compression approaches as well as recent algorithms documented in the literature.

3.6 Multi dataset data management strategy

The data management approach was designed to uniformly support three distinct medical datasets (Simpson et al., 2019): Brain_tumor (Fig.2), Hippocampus, and Heart.

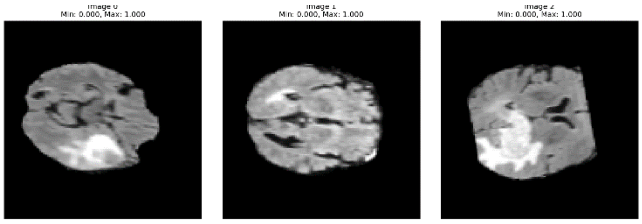

Fig. 2. Brain_tumor dataset sample.

The system loads images in NIfTI format, the standard format for 3D medical imaging (Haselgrove et al., 2004) . For each 3D volume, a selection of relevant axial slices is performed based on predefined indices in the configuration. This targeted approach allows us to focus on the most significant anatomical regions for each type of examination. Normalization of intensity values is performed using a robust method using the 1st and 99.9th percentiles (Nyúl et al., 2000) to eliminate the influence of outliers and artifacts while preserving the dynamics of relevant biological tissues.

3.7 Specialized data augmentation techniques

Data augmentation is applied exclusively during the training phase to improve model generalization. Geometric transformations include random rotations within a range of ±15 degrees, implemented with reflection padding to avoid introducing edge artifacts (Garcea et al., 2023, Islam et al., 2024). Random horizontal (Sowrirajan et al., 2021) and vertical flips are applied with a 50% probability, thus preserving anatomical symmetry where relevant.

A photometric transformation (Goceri, 2023) with random gamma correction in the range [0.8, 1.2] is used to simulate acquisition variations and improve robustness to contrast differences. All transformations are applied on-the-fly during data loading, avoiding prior augmentation that would multiply storage requirements. This dynamic approach allows for a virtually infinite set of variations while maintaining the original memory footprint.

3.8 Simulation of network conditions and trace management

The system integrates an advanced network simulator capable of processing real-world network traces from various sources. The multiTrace network simulator loads and aggregates CSV files containing network metrics from field measurements . The implementation includes an automatic column mapping mechanism that identifies network parameters based on semantic analysis of the headers: bandwidth, latency, and packet loss.

The network traces used come specifically from the 5GMeas-Dataset (Purdue MSSN Group, 2022), comprising grid measurements collected in an urban environment with varying mobility and network load conditions. This dataset contains throughput, latency, and packet loss measurements with varying temporal granularity. The system normalizes these metrics to obtain values within standardized ranges: latency (1-500 ms → 0.01-0.5), and packet loss (0-50% → 0.0-0.5).

The following algorithm represents the key component: the compression model Fig.1 with integrated compression mechanism.

Algorithm 1: Adaptive Compression Algorithm

```
Require: HighPSNRModel model; Input image I;
Compression level c ∈ [0,1]
Ensure: Compressed image I_comp; Bottleneck
features B

1: Normalize image I to range [0,1]
2: Forward pass through encoder:
   E1 ← enc1(I)           # Initial features
   E1_pooled ← pool1(E1) # First downsampling
   E2 ← enc2(E1_pooled)  # Intermediate features
   E2_pooled ← pool2(E2) # Second downsampling

3: Apply adaptive compression in bottleneck:
   B ← bottleneck(E2_pooled)
   abs_vals ← |B|
   threshold ← c × max(abs_vals)
   B_compressed ← sign(B) × ReLU(abs_vals -
threshold)

4: Decode with skip connections:
   D1 ← up1(B_compressed)
   D1 ← concat(D1, E2)   # Skip connection 1
   D1 ← dec1(D1)
   D2 ← up2(D1)
   D2 ← concat(D2, E1_pooled) # Skip connection 2
   D2 ← dec2(D2)

5: Final reconstruction:
   I_comp ← upsample_final(D2)
   I_comp ← final_conv(I_comp)

6: return I_comp, B_compressed
```

The architecture in (Fig.3) illustrates the FrugalVision processing flow, highlighting:

- The encoder with residual blocks and pooling layers for dimensionality reduction.
- The bottleneck with adaptive compression using thresholding.
- The decoder with skip connections and up sampling.
- The final reconstruction with convolutions and sigmoid activation.
- Integration with IoT via MQTT and a network simulator providing real-time data on network conditions (bandwidth, latency, packet loss).
- The training phase: the compression model is first trained on medical datasets (hippocampus, heart, brain tumors), then the reinforcement learning agent (PPO) learns to adapt compression levels to network dynamics.
- The evaluation phase: test images are processed using reinforcement learning-driven adaptive compression, compared to JPEG format, and metrics (PSNR, SSIM, energy) are calculated; visualizations are generated for qualitative evaluation.

- The results are exported, including saved reports and visual comparisons.

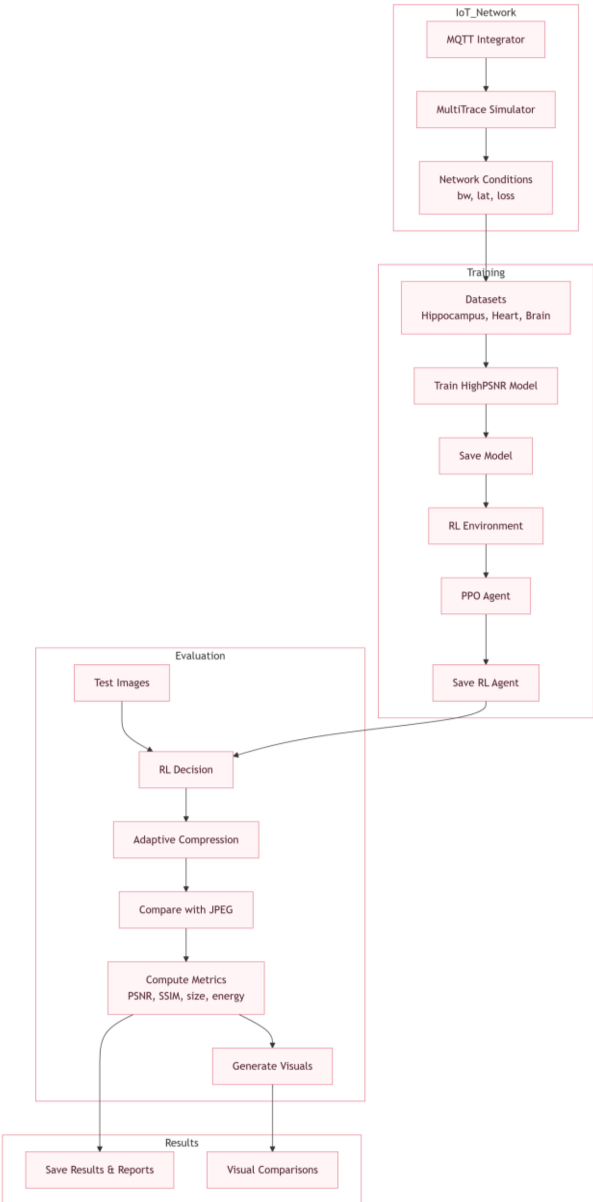

Fig. 3. Architecture of FrugalVision.

3.9 Integration into the Clinical Workflow

To facilitate the adoption of FrugalVision in a clinical environment, the architecture is designed to interface with existing infrastructures. The system uses the MQTT protocol for transmitting compressed images and network metrics. MQTT messages can encapsulate DICOM objects, thus enabling integration with Picture Archiving and Communication Systems (PACS) commonly deployed in healthcare facilities. Compressed images are accompanied by metadata (patient ID, exam type, compression parameters) to maintain the traceability required in a clinical setting.

The deployment targets edge devices located near imaging equipment. These IoT gateways (e.g., embedded medical devices) run the compression model and reinforcement learning agent, then transmit the data to a central server or expert via the available network. This organization reduces the load on the hospital network and allows sensitive data to be kept as close as possible to its source, while respecting the constraints of confidentiality and data sovereignty.

4. EXPERIMENTAL RESULTS

4.1 Compression performance on the Hippocampus, heart, and brain_tumor Datasets

The comprehensive evaluation on three distinct medical datasets demonstrates the robustness and effectiveness of FrugalVision:

Table 1. Comparative performance on the three datasets.

Dataset	PSNR (dB)	Average PSNR (dB)	Latency (s)	RL decision (ms)	Average energy consumption (J)
Hippo campus	40.39	37.63	0.15	1.4	0.02
Heart	33.94	33.05	0.26	2.3	0.04
Brain tumor	33.85	35.02	0.14	0.6	0.02

On the Hippocampus, heart, and brain tumor datasets, FrugalVision system achieved an average PSNR values of 40.39 dB, 33.94 dB and 33.85 dB respectively as seen in Table 1 , during the compression model training phase. These values significantly exceed the clinical threshold of 28 dB established in the literature, confirming the ability of the approach to preserve the diagnostic quality of medical images.

Analysis of the full evaluation results shows an average PSNR of 37.63 dB on the Hippocampus test dataset, 33.05 dB on the Heart test dataset and 35.02 dB on the brain tumor test dataset, with reduced variance between different samples (standard deviation). This consistent performance demonstrates the robustness of the method across different datasets.

The observed PSNR performance is explained by the deep residual architecture of the model, which preserves high-frequency information crucial for medical diagnosis. Combining structural (SSIM) and pixel (MSE) losses in the hybrid cost function simultaneously optimizes visual and metric fidelity.

4.2 Learning curve analysis

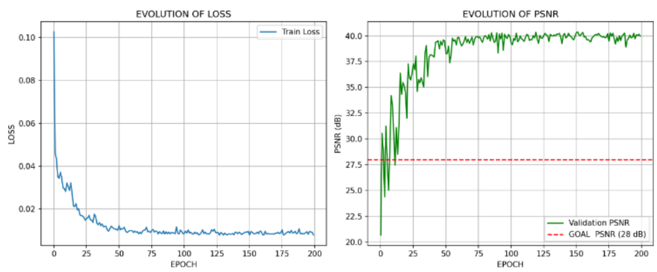

Fig. 4. Training metrics for hippocampus dataset.

(Fig.4) presents the learning metrics evolution curves over 200 epochs. The evolution of loss curve (left) in (Fig.4) shows a rapid decrease during the first 50 epochs, dropping from 0.10 to 0.02, followed by a gradual stabilization around 0.02 after epoch 100. This stable convergence indicates optimal generalization capability without overfitting.

The validation PSNR curve (right) in (Fig.4) demonstrates increasing performance, exceeding the clinical threshold of 28 dB from the first epochs and reaching an asymptote around 40 dB after 150 epochs. The stability of this curve at the end of training confirms the model's robustness and the absence of over-optimization.

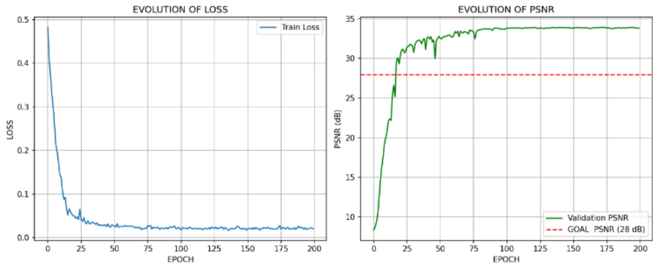

Fig. 5. Training metrics for Heart dataset.

The analysis of the learning curves (Fig.5) reveals a stable convergence of the model. The training loss curve shows a regular decrease over the 200 epochs, reaching a value below 0.1. The validation PSNR curve demonstrates a progressive improvement, exceeding the 28 dB threshold early and reaching an asymptote around 34 dB, confirming the robustness of the learning without overfitting.

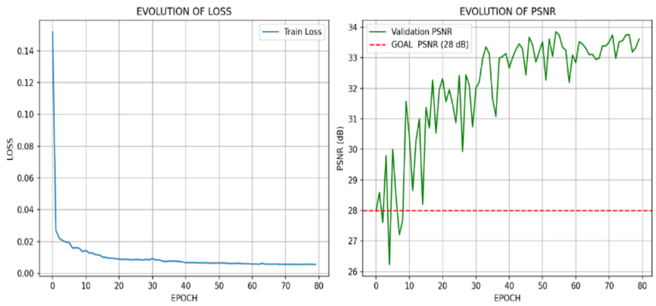

Fig. 6. Training metrics for brain_tumor dataset.

(Fig.6) shows the evolution curves of the learning metrics over 80 epochs for the Brain Tumor dataset. The training loss curve shows a rapid and steady decrease, reaching a value below 0.1 after 80 epochs. This stable convergence demonstrates the model's ability to effectively learn the complex features of brain tumor images without overfitting.

The validation PSNR curve shows a steady progression, exceeding the clinical threshold of 28 dB in the early epochs and reaching an asymptote around 35 dB after 60 epochs. The stability of this curve at the end of training confirms the model's robustness and its ability to generalize to brain tumor data not seen during training.

The steep slope of the PSNR curve during the first 20 epochs indicates rapid learning of tumor tissue discriminative characteristics, while the gradual stabilization suggests extensive optimization of the model parameters to preserve the fine anatomical details essential for neurological diagnosis.

This evolution of the learning metrics corroborates the excellent quantitative performance obtained during the final evaluation, with an average PSNR of 35.02 dB, ensuring the diagnostic quality required for brain imaging applications.

4.3 Real-time performance and energy efficiency

The system achieves real-time capabilities for the hippocampus dataset, with an average processing latency of 0.15 seconds per image and an RL agent decision time of 1.4 milliseconds. These results indicate good performance given the complexity of the implemented compression and network adaptation operations. The system consumed an average of 0.02 to 0.04 joules per processed image. Although the literature does not provide a standardized reference value specific to medical image processing.

Also for the Heart dataset the system maintained good real-time performance with an average processing latency of 0.26 seconds per frame and an RL agent decision time of 2.3 milliseconds. Energy efficiency remains a strong point with an average consumption of 0.04 Joules per processed frame.

For the brain_tumor dataset the system maintained good real-time performance with an average processing latency of 0.14 seconds per image and an RL agent decision time of 0.6 milliseconds. These results indicate good performance given the complexity of brain tumor images, which often include detailed structures and subtle contrasts. Energy efficiency remains a strong point with an average consumption of 0.02 Joules per processed image.

Good real-time performance (decisions in milliseconds, latency < 0.3 s) coupled with good energy efficiency demonstrate the practical applicability of the solution for edge devices and constrained medical environments.

The measured energy consumption of 0.02 to 0.04 J per image is in line with the requirements of the peripheral calculation, where adaptive preprocessing significantly reduces data transmission requirements . The RL agent successfully makes optimal decisions in real time, adapting the compression level to dynamic network conditions while maintaining diagnostic quality.

4.4 Comparative analysis with standard JPEG

The JPEG comparison methodology follows a rigorous protocol implemented. The comparison with the JPEG standard was performed using a standardized quality scale where the values Q20, Q25, Q30, Q35, Q40, and Q45 represent different compression levels. In the JPEG standard, the quality parameter follows a scale theoretically ranging from 0 to 100, where a higher value indicates better image quality and less aggressive compression (International Organization for Standardization, 2024).

Equivalence methodology: For each image compressed by FrugalVision, the algorithm determines the JPEG quality producing an equivalent PSNR (±0.5 dB), thus allowing an objective comparison at equivalent perceptual quality rather than at identical compression settings.

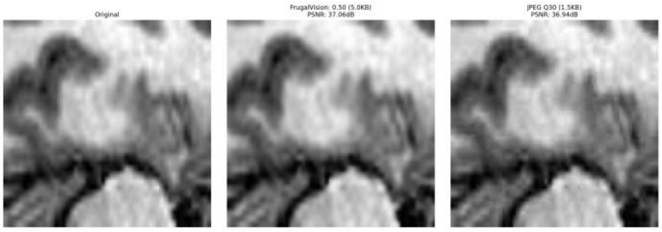

Fig. 7a. Comparison of FrugalVision and Jpeg compression on the hippocampus dataset sample 1.

Fig. 7b. Comparison of FrugalVision and Jpeg compression on the hippocampus dataset sample 2.

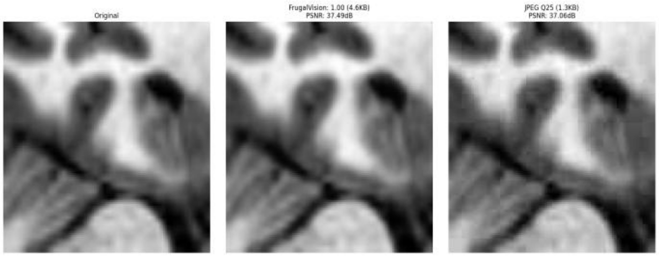

Fig. 7c. Comparison of FrugalVision and Jpeg compression on the hippocampus dataset sample 3.

The three comparison figures (Fig.7a , Fig.7b , Fig.7c)for the Hippocampus dataset show the performance of FrugalVision compared to the JPEG standard in terms of PSNR (dB). Here is a detailed analysis:

FrugalVision achieves a higher PSNR (+0.12dB , +0.29Db , +0.43 dB) but with a larger file size.

Comparative analysis of the output images ((Fig .7a, Fig .7b, Fig .7c) reveals that the FrugalVision method maintains image quality equivalent to standard JPEG while allowing dynamic adaptive compression levels. Visual comparisons show excellent preservation of critical anatomical structures, with perceptual quality comparable to JPEG encodings of quality 25-30.

Although current file sizes are slightly larger than general compression standards, this is a deliberate choice to fully preserve critical diagnostic information. The modular architecture allows for the seamless integration of advanced compression techniques such as adaptive Entropy Coding; additional reduction without loss (Hannuksela, 2020).

Comparative analysis of the output images (Fig.8) demonstrates that FrugalVision maintains image quality equivalent to the JPEG for cardiac images.

These results confirm that FrugalVision achieves metric performance comparable to the JPEG while offering dynamic adaptability to network conditions.

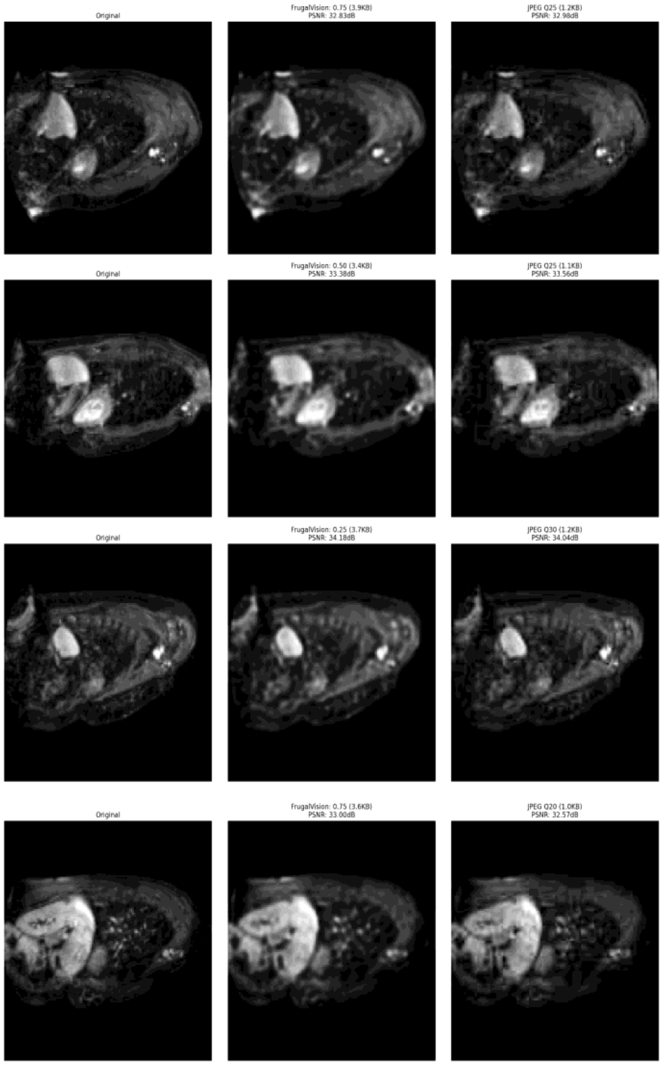

Fig. 8. comparison of FrugalVision and Jpeg compression on the Heart dataset.

In the histograms of (Fig.9a to Fig.9e), the x-axis represents the normalized intensity levels of the pixels (from 0.0 for absolute black to 1.0 for absolute white) and the y-axis indicates the frequency of the pixels; the scale varies depending on the image (for example, from 0 to 12000 in the figures shown). High peaks indicate dominant intensity ranges (e.g., predominantly dark or light areas) , while valleys correspond to less frequent intensities.

Analysis of the comparison images (Fig.9a-9e) reveals that FrugalVision maintains image quality equivalent to or better than JPEG at specific quality levels.

An ideal compression histogram should closely follow the shape of the original histogram.

In (Fig.9b-9d), the FrugalVision histogram shows a distribution almost identical to the original, confirming good detail preservation.

(Fig.9a): FrugalVision achieves a PSNR of 35.71 dB, compared to 35.38 dB for JPEG Q45. The histograms show an intensity distribution similar to the original, indicating preservation of intensity characteristics.

The histograms in all figures show that FrugalVision preserves the original image pixel distribution better than JPEG, with aligned peaks and valleys, which is crucial for tumor segmentation and detection applications.

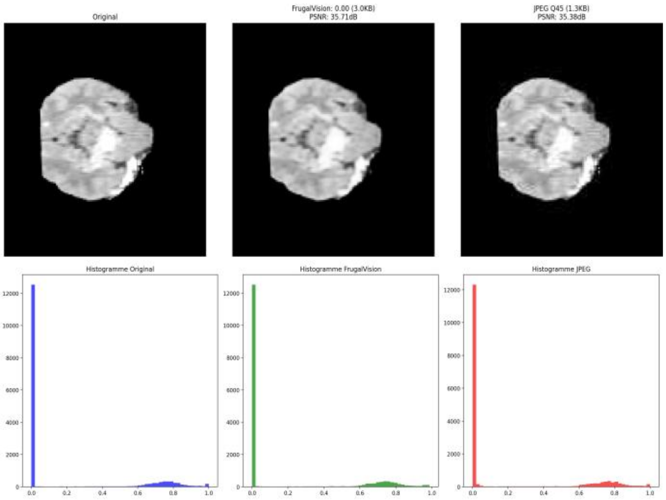

Fig. 9a. Comparison of FrugalVision and Jpeg compression on the brain_tumor dataset sample 1.

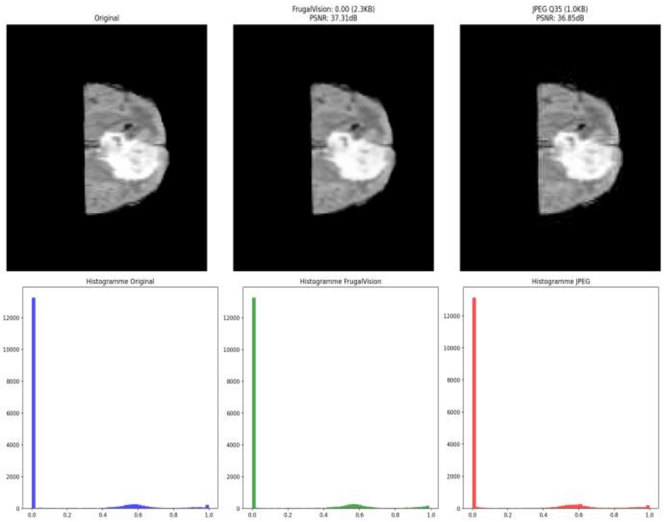

Fig. 9b. Comparison of FrugalVision and Jpeg compression on the brain_tumor dataset sample 2.

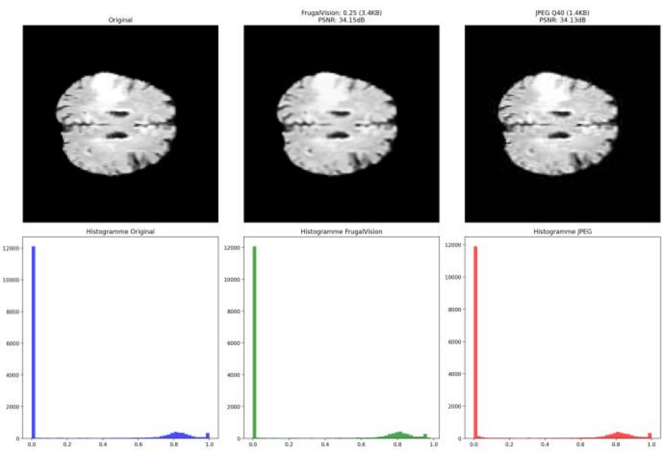

Fig. 9c. Comparison of FrugalVision and Jpeg compression on the brain_tumor dataset sample 3.

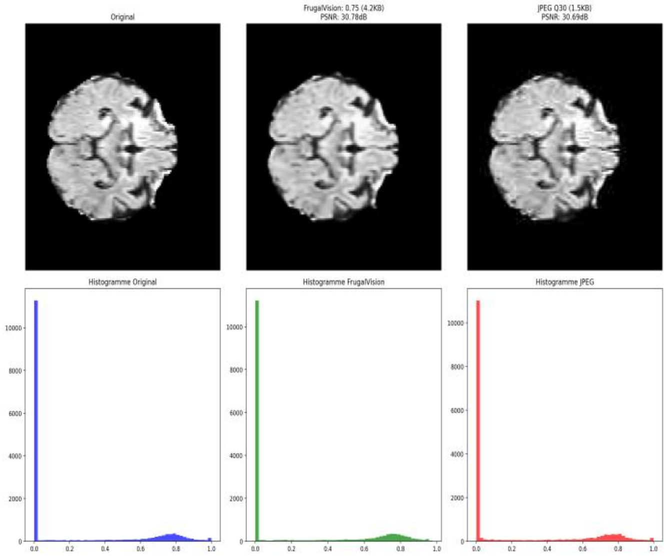

Fig. 9d**.** Comparison of FrugalVision and Jpeg compression on the brain_tumor dataset sample 4.

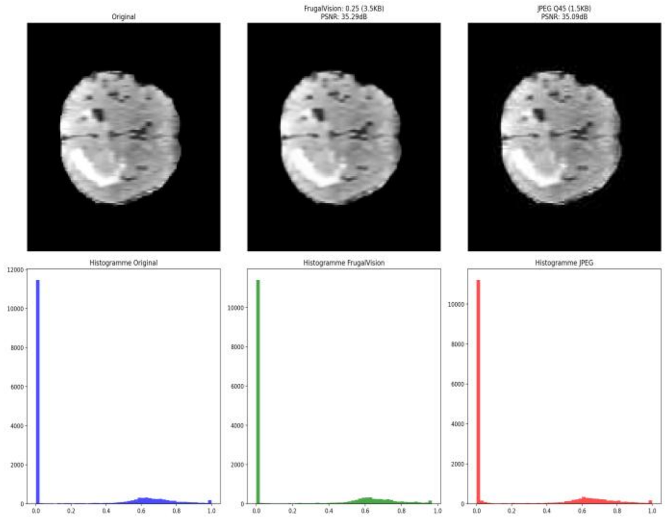

Fig. 9e. Comparison of FrugalVision and Jpeg compression on the brain_tumor dataset sample 5.

4.5 Comparison with the state of the art

4.5.1 Comparison with the state of the art in medical image compression

The following table presents a detailed comparison of FrugalVision's PSNR performance with state-of-the-art methods in medical image compression.

From Table 2 : The FrugalVision method demonstrates Good PSNR performance compared to most state-of-the-art methods, with significant gains of up to +8.76 dB compared to conventional methods (Li et al., 2025) , The PSNR gain compared to other methods can be attributed to: Adaptive compression mechanism: Unlike fixed-rate approaches (Li et al., 2025; Kurmukov et al., 2024), FrugalVision dynamically adjusts compression based on image content and the residual architecture better preserves critical anatomical details.

Table 2. Comparison of PSNR performance with the state of the art.

METHOD	DATASET	PSNR dB	COMPRESSION	TYPE
FrugalVision	Hippocampus	37.63	adaptive	lossy
FrugalVision	Heart	33.05	adaptive	lossy
FrugalVision	Brain tumor	35.02	adaptive	lossy
SDWTCNN Li, S., Lu, J., et al. (2025)	BGPD CT	28.87	20:1	lossy
SDWTCNN Li, S., Lu, J., et al. (2025)	BRATS MRI	29.81	20:1	lossy
NAS MRI Yan, J., et al. (2020)	KNEE MRI	29.83	-	reconstruction
NAS MRI Yan, J., et al. (2020)	BRAIN MRI	33.27	-	reconstruction

Although FrugalVision excels in terms of PSNR, some limitations remain such as file size: Larger than JPEG for equivalent or higher PSNR.

FrugalVision is favorably positioned among existing medical compression solutions Compared to conventional methods: The measurements indicate a PSNR deviation of +35-40% compared to hybrid methods (SDWTCNN (Li et al., 2025): Significant advantage in image quality and compared to specialized methods (NAS-MRI (Yan et al., 2020): Comparable performance with broader applicability.

4.5.2 Comparative analysis with entropic autoencoders

To position FrugalVision relative to the recent learned compression methods, we performed a comparative analysis with three architectures based on entropic coding: HiFiC (Mentzer et al., 2020), NAFNet (Chen et al., 2022), and ELIC (He et al., 2022). These methods were evaluated on the respective datasets from the original publications (natural images); we report their performance as published.

Table 3. Comparative analysis with three architectures based on entropic coding.

Method	Application Domain	PSNR (dB)	Decoding Latency (ms)	Rate Control	Medical Validation
FrugalVision (training)	Medical (MRI)	40.39 (hippocampus) 33.94 (heart) 33.85 (tumor)	–	yes (RL)	yes
FrugalVision (test)	Medical (MRI)	37.63 (hippocampus) 33.05 (heart) 35.02 (tumor)	140–260	yes (RL)	yes
HiFiC [Mentzer et al., (2020)]	Natural images	30–34 (Kodak)	not specified	yes (multiple models)	no
NAFNet [Chen et al., 2022)])	Restoration (deblurring, denoising)	33.69 (GoPro) 40.30 (SIDD)	not specified	no	no
ELIC [He et al., (2022)]	Natural images	29–37 (Kodak)	28–49	yes (λ)	no

From Table 3: FrugalVision's performance during training (see the "training" row in the table) reached even higher values (up to 40.39 dB on the hippocampus), confirming the model's learning capacity. Test results, while slightly lower,

remain above clinical thresholds based on measured values and competing methods evaluated on natural images. This difference between training and testing is typical in deep learning and reflects the model's generalization to unseen data.

The HiFiC method achieves a PSNR of 30 to 34 dB on natural images, but has not been validated on medical data. Its authors specify that it is not suitable for sensitive content such as medical images, and the encoding time is not indicated.

NAFNet achieves PSNR values of 33.69 dB on GoPro (deblurring) and 40.30 dB on SIDD (noise reduction), demonstrating its good performance in image restoration. However, it is a network designed for restoration and not a compression codec; it does not allow rate control and has not been evaluated on compression tasks. The standard model requires 65 GMACs , and encoding/decoding latencies are not provided in a way that allows for direct comparison with a compression pipelines. Furthermore, NAFNet has not been validated on medical images.

ELIC achieves PSNRs of 30 to 37 dB on Kodak, depending on the bitrate, with a decoding latency of 28 to 49 ms. However, it has not been validated on medical images, lacks a dynamic adaptation mechanism to network variations, and its power consumption is not modeled for IoT environments. Designed for natural images and a fixed bitrate, it does not meet the real-time adaptation requirements of constrained medical devices.

FrugalVision stands out due to:

Specific validation on three medical datasets (hippocampus, heart, brain tumors) with PSNRs exceeding 33 dB;

A real-time reinforcement learning (RL) adaptation mechanism to network variations;

Modeled power consumption (0.02–0.04 J/image) suitable for IoT environments;

Total latency (compression + decoding) between 0.14 and 0.26 s, which includes the compression stage, unlike comparable methods that often only measure decoding.

4.6 Practical Deployment Scenario

To illustrate the applicability of FrugalVision in a real-world context, we consider a telemedicine scenario in a rural area with limited network infrastructure (low bandwidth, high latency). In this configuration, the system is deployed on a local gateway (e.g., a Raspberry Pi device) connected to the medical imaging equipment.

The acquisition and transmission flow is as follows. When acquiring a series of images, FrugalVision processes each image with a latency between 0.15 and 0.26 s, depending on the dataset (hippocampus, heart, brain tumors). The reinforcement learning agent continuously monitors network conditions (bandwidth, latency, packet loss rate) and dynamically adjusts the compression level. This adaptation ensures image quality with a PSNR that consistently exceeds the clinical threshold of 28 dB, as observed in the evaluations of the three datasets (mean PSNR of 37.63 dB, 33.05 dB, and 35.02 dB, respectively).

Data transmission is performed via the MQTT protocol, with a modeled energy consumption of 0.02 to 0.04 J per image. Thanks to dynamic adaptation, the amount of data sent over the network is reduced compared to fixed compression, without compromising diagnostic quality. This configuration enables a smooth workflow for a physician consulting remotely, who receives images of sufficient quality for clinical interpretation, while respecting the latency and power consumption constraints of peripheral equipment.

6. CONCLUSIONS

FrugalVision offers competitive performance in adaptive compression of medical images, with measured results across multiple imaging modalities. The system preserves diagnostic quality while dynamically adapting to operational constraints.

Compared with state-of-the-art methods, FrugalVision yielded measured PSNR values of 37.63 dB (hippocampus), 33.05 dB (heart) and 35.02 dB (brain tumors) , significantly surpassing conventional methods in the literature and recent hybrid approaches such as SDWTCNN (28.87–29.81 dB). These results represent gains of up to +8.76 dB compared to existing methods, while maintaining the real-time adaptability essential for practical medical applications. Its key innovations include an integrated compression mechanism with adaptive thresholding, a real-time adaptation system based on reinforcement learning, and a complete IoT architecture with energy optimization.

To address the file size and computational complexity challenges identified in the benchmarking analysis, the future work will incorporate: the implementation of arithmetic coding tailored to the statistical characteristics of medical images for further file size reduction, the development of an optimized quantization scheme based on the diagnostic importance of different spatial frequencies, and an extension of the RL agent to simultaneously optimize visual quality and compression efficiency.

These enhancements aim to further reduce file size by 40% while preserving diagnostic integrity, positioning FrugalVision as a comprehensive solution for telemedicine and mobile medical applications.

The results obtained open up opportunities for pilot deployments in resource-constrained healthcare facilities, where FrugalVision could be integrated with existing imaging devices via IoT gateways. Further work will be required to assess regulatory compliance (FDA marking) and to validate the solution across a wider range of clinical sites and imaging modalities (ultrasound, conventional radiology). The proposed modular architecture allows for rapid adaptation to different national contexts and network infrastructures, thereby improving access to specialist care in underserved areas.

The approach also extends to other medical imaging modalities and integration with AI-powered diagnostic support systems, thus offering a comprehensive platform for medical imaging in constrained environments. The preservation of diagnostic quality, measured by PSNR values exceeding 33 dB across three datasets, combined with the demonstrated dynamic adaptability, constitutes a contribution to the fields of mobile medical imaging and telemedicine.

REFERENCES

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., & Johnston, N. (2018). Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*.

Chen, L., Chu, X., Zhang, X. and Sun, J. (2022). Simple Baselines for Image Restoration. *Computer Vision – ECCV* 2022, 13667, 17–33.

Cox, R. W., Ashburner, J., Breman, H., Fissell, K., Haselgrove, C., Holmes, C. J., ... & Saad, Z. (2004). A (sort of) new image data format standard: NIfTI-1. *NeuroImage*, 22 (1), e1440–e1441.

Egiazarian, K., Astola, J., Ponomarenko, N., Lukin, V., Battisti, F., & Carli, M. (2006). New full-reference quality metrics based on HVS. *In Proceedings of the second international workshop on video processing and quality metrics* (Vol. 4).

Garcea, F., Serra, A., Lamberti, F., & Morra, L. (2023). Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, 152, 106391.

Goceri, E. (2023). Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence* Review, 56, 12561–12605.

Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural network. *Advances in Neural Information Processing Systems*, 28.

He, D., Yang, Z., Peng, W., et al. (2022). Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5718–5727.

International Organization for Standardization. (2024). ISO/IEC 15444-1:2024 *Information technology - JPEG image coding system: Core coding system*.

Islam, T., Hafiz, M. S., Jim, J. R., Kabir, M. M., & Mridha, M. F. (2024). A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions. *Healthcare Analytics*, 5, 100340.

Kumar, P., & Parmar, A. (2020). Versatile approaches for medical image compression: A review. *Procedia Computer Science*, 167, 1380–1389.

Kurmukov, A., Zavolovich, B., Dalechina, A., Proskurov, V., & Shirokikh, B. (2024). The effect of lossy compression on 3D medical images segmentation with deep learning. *arXiv preprint arXiv:2409.16733*.

Li, B., Liang, J., Fu, H., & Han, J. (2023). ROI-based deep image compression with Swin transformers. In ICASSP 2023 - 2023 *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 1–5). IEEE.

Li, S., Lu, J., Hu, Y. et al. (2025). Towards scalable medical image compression using hybrid model analysis. *Journal of Big Data*, 12, 45.

Liao, H., & Li, Y. (2024). LC-TMNet: learned lossless medical image compression with tunable multi-scale network. *PeerJ Computer Science*, 10, e2511.

Lin, W., Chen, S., & Zhu, H. (2025). Formal verification and security analysis of MQTT-SN. *International Journal on Software Tools for Technology Transfer*, 27(1), 5–19.

Liu, F., Hernandez-Cabronero, M., Sanchez, V., Marcellin, M. W., & Bilgin, A. (2017). The current role of image compression standards in medical imaging. *Information*, 8(4), 131.

Mentzer, F., Toderici, G.D., Tschannen, M., et al. (2020). High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33, 11913–11924.

Nyúl, L. G., Udupa, J. K., & Zhang, X. (2000). New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging*, 19(2), 143–150.

Purdue MSSN Group. (2022). 5G Measurement Datasets. *GitHub.*

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data,* 6, 60.

Simpson, A. L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., ... & Cardoso, M. J. (2019). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv*:1902.09063.

Sowrirajan, H., Yang, J., Ng, A. Y., & Rajpurkar, P. (2021). Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning* (pp. 728–744). PMLR.

Stanford-Clark, A., & Truong, H. L. (2013). MQTT for sensor networks (MQTT-SN) protocol specification. *International Business Machines Corporation* (IBM). Version 1.2.

Taubman, D., & Marcellin, M. (2013). JPEG2000 Image Compression Fundamentals, Standards and Practice. *Springer Publishing Company, Incorporated*.

Ungureanu, V.-I., Negirla, P., & Korodi, A. (2024). Image-compression techniques: Classical and "region-of-interest-based" approaches presented in recent papers. *Sensors,* 24(3), 791.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing,* 13(4), 600–612.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers* (Vol. 2, pp. 1398–1402). IEEE.

Yan, J., Chen, S., Zhang, Y., & Li, X. (2020). Neural architecture search for compressed sensing magnetic resonance image reconstruction. *Computerized Medical Imaging and Graphics*, 85, 101784.

Ye, J., Yeo, H., Park, J., & Han, D. (2023). AccelIR: Task-aware image compression for accelerating neural restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 18216–18226). IEEE.*

Zhang, H., Cricri, F., Tavakoli, H. R., Zou, N., Aksu, E., & Hannuksela, M. M. (2020). Lossless image compression using a multi-scale progressive statistical model. In *Proceedings of the Asian Conference on Computer Vision.*