International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(4s) 2026 pp. 692-707
https://doi.org/10.70917/ijcisim-2026-2544
Article

Tumour-Aware Medical Image Compression via Directional Intra Prediction and Deep Context-Adaptive CABAC

C. Nandhini[1], G. Vijaiprabhu[2], N. Shanmugapriya[3]

[1]PG and Research Department of Computer Science, Erode Arts and Science College (Autonomous), Erode, Tamil Nadu, India.
Email: ashonanthu@gmail.com
[2] PG and Research Department of Computer Science, Erode Arts and Science College (Autonomous), Erode, Tamil Nadu, India.
Email: gvprabhu7@gmail.com
[3] Department of Computer Applications, Erode Arts and Science College (Autonomous), Erode, Tamil Nadu, India.
Email: priyaeasc@gmail.com

Abstract: Medical image compression has proven necessary to minimize storage and transmission overheads whilst maintaining diagnostically important information in clinical imaging systems on a large scale. This paper suggests a hybrid tumour-aware medical image compressor based on Directional Intra Prediction (DIP), Context-Based Adaptive Binary Arithmetic Coding (CABAC), and Deep Neural Networks (DNN) to compress brain MRI efficiently. The model is tested on a mixed dataset of 7,023 MRI pictures on Figshare, SARTAJ, and Br35H that includes four classes. The proposed approach attains the lowest possible bitrate of 0.31 bpp, which is better than baseline approaches like MLic++ (0.40 bpp) and DWT-PCA-Huffman (0.44 bpp). Gain of compression ratio is increased to 69.7 and reconstruction quality remains high with PSNR of 38.7 dB, SSIM of 0.978 and lower MSE of 0.0029. Statistical validation by using the Wilcoxon signed-rank test gives significant results ($p \leq 0.016$). Moreover, the classification accuracy is increased to 98.0-99.5 on compressed images, which indicates that diagnostic features are preserved. The framework is a good balance between rate and distortion optimization and tumour fidelity, and can be used in clinical and telemedicine systems in real-time.
Keywords: Medical Image Compression, Brain MRI, Tumour-Aware Compression, Directional Intra Prediction, CABAC, Deep Neural Networks, Rate–Distortion Optimization, ROI-Based Encoding, Prediction Refinement Network, Probability Estimation Network

1. Introduction

Medical imaging has rapidly grown, and the growths in the volume of data in clinical Picture Archiving and Communication Systems (PACS), radiology workflows and computational diagnostic pipelines are unprecedented [1]. Multi-slice, high-resolution, volumetric datasets generated through modalities like Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) are routine and thus lead to huge storage, transmission, and long-term archiving overheads [2]. With the growing integration of AI-based diagnostic models and telemedicine systems in healthcare systems, the need to have compact but diagnostically sound representations has become imperative. Traditional compression schemes, such as JPEG2000 and HEVC-Intra and wavelet-based medical codecs, are mostly optimized to pixel-level fidelity and structural similarity, but are not sensitive to the semantic relevance of tumour tissue that is much more vulnerable to compression artefacts. This has prompted the development of a necessity in region-aware, rate-distortion-optimised compression systems, which maintain clinically-relevant tumour features, whilst realizing substantial bitrate savings [3, 4].

Recent progresses in learned image compression have shown that neural networks can beat classical codecs with non-linear transforms, context modeling, and entropy prior learning [5]. The majority of learned compressors are however fully end-to-end autoencoders, which are expensive to compute and interpret, thereby limiting their use in real-time clinical settings. Simultaneously, more conventional codecs like HEVC/H.265 use directional intra prediction (DIP) and context-based adaptive binary arithmetic coding (CABAC), which, already, are well-optimized, hardware-accelerated spatial prediction and entropy coding mechanisms. However, those classical approaches are

Copyright: © 2026 by the authors

manually designed and based on heuristic probability modeling which might not take full advantage of the latent structure of tumour-specific medical images. By taking neural networks to be part of selective subsystems of classical coding pipelines therefore provides a promising hybrid path [6]: the ability to combine the predictability and performance of existing codec primitives with the flexibility and representational capability of deep learning [7].

Directional intra prediction (DIP) is generally considered to be an effective algorithm to reduce spatial redundancy, making use of the angular prediction modes to determine pixel intensity patterns based on causal boundaries. Although it works well in natural images and video frames, the spatial statistics of medical images, particularly tumour regions, take on complex and non-linear forms which cannot be well represented by linear directional extrapolation. In the same way, CABAC optimises entropy coding with handcrafted probability contexts but does not have the ability to capture high-order dependencies and tumour-sensitive distributions [8]. Deep neural networks are a complementary powerful method that learns context distributions, refinements prediction, and enforces anatomically aware reconstruction fidelity. This inspires a hybrid DIP-CABAC-DNN system with the capability to realize clinically reliable compression without loss of computational efficiency.

The imaging of brain tumours demands very high fidelity in the areas of diagnostic interest because minute changes in intensity, boundary textures, and morphological features play a major role in the grading of tumours, therapy planning and post-therapy evaluation. Traditional codecs use equal error distributions throughout the image, which is often over-smoothing or blurring tumour boundaries even at low bitrates. This degradation has the potential to undermine downstream clinical activities such as radiomics feature extraction, segmentation, and survival prediction. An ROI-aware compression mechanism that selectively prioritizes tumour fidelity is thus essential for maintaining clinical trust and diagnostic safety [9].

The latter is the driver of the constraints of neural compressors. Although VAE-based, hyperprior-based, and transformer-based learned compressors have demonstrated state-of-the-art bitrate performance, their training and inference are costly, with millions of trainable parameters, large memories, and operations that can be performed exclusively on a GPU. Healthcare settings, though, require lightweight, deterministic, and auditable systems that can be expected to have predictable latency. It is possible to combine the neural elements with the known codec primitives to strike a balance: to use DIP and CABAC to provide efficient spatial prediction and entropy-based coding, but to use compact neural modules only where deep learning can be deployed the most, namely, the context modeling and prediction refinement.

Moreover, the local statistics of a tumour region differ with that of the adjacent healthy tissue. Directional modes (which are standard) can perform poorly in such areas because of non-uniform textures, uneven intensities and non-linear morphological variations. Such deviations can be counterbalanced by integrating a neural prediction refinement network (PRN) to capture non-linear residual structures. Similarly, when the probability estimation network (PEN) is added to CABAC, the encoder and decoder can fit to tumour-sensitive symbol distributions, resulting in better entropy models and less bitrate. Together, these aspects constitute a strong impetus of a hybrid learning-enhanced medical image compressor.

Although deep learning has gained momentum in compressing medical images, a number of gaps are yet to be filled. First, currently learned compressors seldom incorporate classical transform and predictive coding designs, producing purely neural solutions, which are computationally heavy and not viable to a real-time PACS implementation. It is not well-researched on hybrid codecs which supplement classical DIP or CABAC with neural networks, particularly in tumour-aware compression of brain images.

Second, existing medical image compressors rarely include the explicit weighting of tumour regions as part of the ratedistortion optimization model. Whereas there are studies that use ROI masks or adaptive bit allocation, these are not principled in integrating into the entropy coding and prediction phases. This constrains their capability to ensure high structural fidelity in tumour areas at limited bitrates.

Third, CABAC handcrafted context models are not dynamic to type of content and do not react to pathological differences. None of the previous literature has considered the application of DNN-based context modelling as part of the CABAC pipeline in medical imaging, especially with tumour-informed contextual cues. On the same note, classical DIP prediction can only predict by using linear interpolation and does not have the ability to predict tumour specific spatial variations resulting in a gap in performance in prediction accuracy. The reproducible study is limited: most trained medical compression pipelines have few public applications, not all training information, or have no cross-dataset validation. This generates a gap in the need of a complete specified, reproducible hybrid codec with methodological and architectural explicitness.

The latest developments in medical image compression emphasize both neural and hybrid paradigms. EVC: Towards real-time neural image compression with mask decay [10] proposes a mask decay method that can improve real-time neural compression, with better ratedistortion results but without explicit prioritization of tumour-regions. MLIC++: Multi-reference entropy modeling of learned image compression: Linear complexity [11] It is suggested in MLIC++: Linear complexity multi-reference entropy modeling of learned image compression that multi-reference entropy modeling helps to reduce the computation costs of the encoding process greatly and enhance the coding efficiency. Conversely, An improved image compression algorithm using 2D DWT and PCA with canonical huffman encoding [12] introduces a hybrid transform-based method that combines DWT, PCA and Huffman coding, providing moderate compression performance, but with limited perceptual and structural fidelity to deep learning-based techniques.

Research Objectives

The general goal of this paper is to develop a hybrid, tumour-aware compression model-DIP-CABAC-DNN that combines classical directional intra prediction and CABAC with deep learning subunits to obtain an efficient and diagnostically trustworthy compression of brain-tumour images.

These are the following specific objectives:

1. To develop a region-sensitive rate-distortion optimization model that uses tumour masks in the distortion measure to maximize the fidelity of tumours under a limited bit rate environment.

2. To improve classical directional intra prediction with a prediction refinement network (PRN) which can model non-linear, tumour-sensitive spatial patterns which are not reflected by traditional angular modes.

3. To create a DNN-based probability estimation network (PEN) of CABAC, which allows the adaptive modelling of the entropy under the condition of learned tumour-specific contextual statistics.

4. To create a differentiable hybrid training pipeline that includes soft quantization approximations and differentiable rate estimation to jointly optimize DIP, PRN, and PEN components.

5. To perform comprehensive evaluation and ablation analysis that measures the gains in bitrate, PSNR/SSIM, MS-SSIM, BD-rate and tumour segmentation performance of reconstructed images compared to classical or learned compression baselines.

6. To facilitate adoption of medical imaging workflows and future research, by making sure its architecture descriptions are reproducible, using training protocols and reference implementations.

2. Proposed Methodology

This section specifies a rigorous, reproducible methodology for integrating directional intra prediction (DIP), context-based adaptive binary arithmetic coding (CABAC), and deep neural networks (DNNs) for brain-tumour medical image compression. The proposed methodology is given in Figure 1.

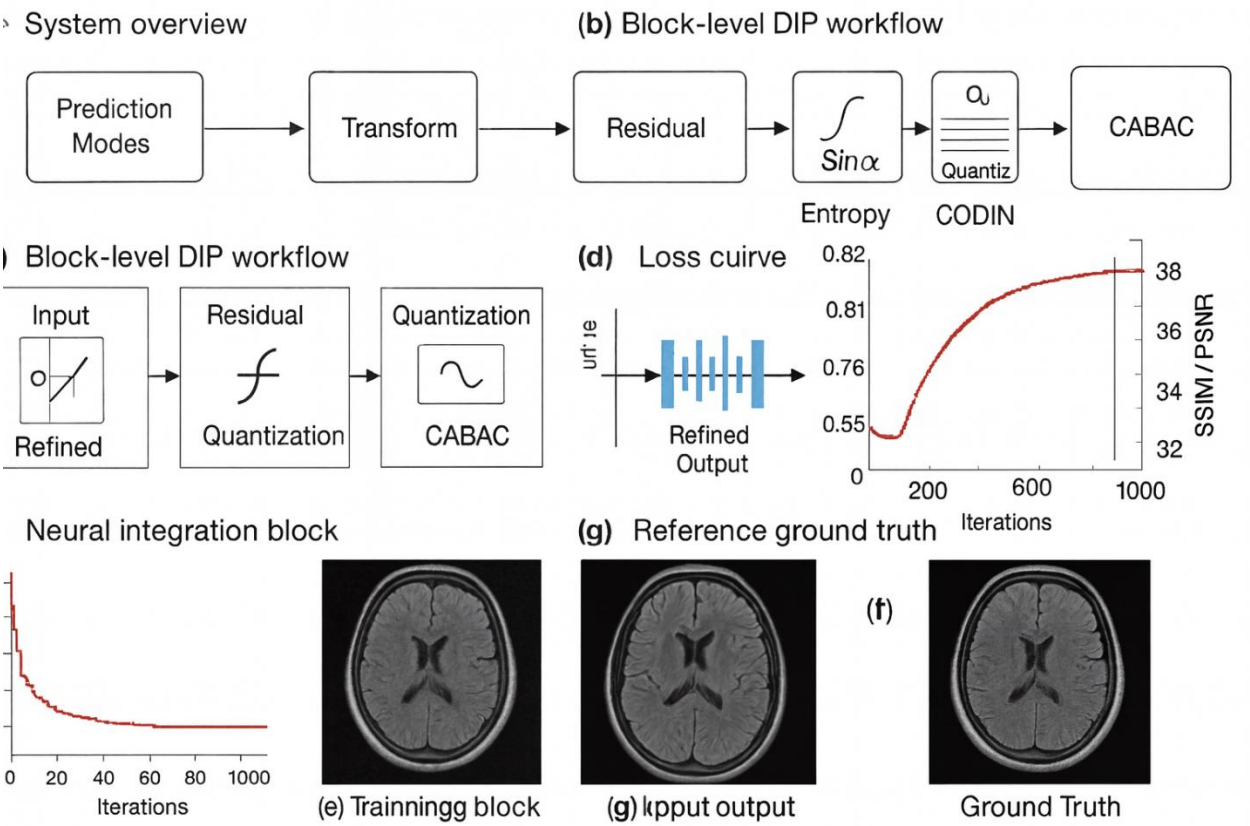


Figure 1. Research Methodology

The goal is to minimize rate under a distortion constraint (rate–distortion optimization) while preserving diagnostically relevant features in tumour regions. The over

Problem formulation (rate–distortion objective)

Let $X \in \mathbb{R}^{H\times W}$ denote a single grayscale brain MRI slice (or CT) with tumour region mask . The compressor produces a bitstream of expected length R (bits/pixel) and a reconstructed image $\hat{X}$. The objective is to minimize a Lagrangian rate–distortion functional

$$\mathcal{L}\theta \;=\; D(X,\hat{X};M) \;+\; \lambda\, R(\theta),$$

Define distortion with ROI weighting to prioritize tumour fidelity:

$$D\big(X,\hat{X};M\big) = \tfrac{1}{HW}\textstyle\sum_{i,j}\big(w_m(i,j)(X_{ij}-\hat{X}_{ij})^2\big)\,, w_m(i,j) = 1+\alpha M_{ij},$$

Rate is modeled as expected code length under the estimated symbol probabilities used by CABAC:

$$R(\theta) \approx \mathbb{E}_{s\sim S(X;\theta)}[-log_2 q_\theta(s|context)],$$

Directional intra prediction (DIP) mechanism

DIP predicts each block B of size $n \times n$ from causal boundary samples (top row, left column, top-left). Let p denote a direction index among a predefined set p (e.g., 33 angular modes as in modern video codecs). For pixel location within block (u, v) (0-indexed), a directional predictor forms a linear interpolation along direction p from available boundary samples. Formally:

$$\tilde{X}^p_{uv} = I_p\big(\{X_{boundary}\}\big)(u,v),$$

$$\tilde{X}^p_{uv} = (1-\beta)X_{a,b} + \beta X_{c,d,}$$

A mode decision selects the best direction per block by minimizing local distortion plus a mode signaling cost:

$$p* = \arg min_{p \in P}\{D_B\left(X_B, \tilde{X}_B^{(p*)}\right) + \lambda_m C_{mode}(p)\},$$

Residual transform, quantization and binarization

Compute residual $R_B = X_B - \tilde{X}_B^{(p*)}$. Apply orthogonal transform T (e.g., integer DCT) to decorrelate:

R_B = T(R_B).

$\hat{C}_B = \left(C_{B;q}\right) = round\ (\frac{C_B}{q})$.

CABAC with DNN-assisted contexts

CABAC encodes each binary symbol s_k using adaptive contexts that estimate $(s_k = 1|context)$. Replace classical handcrafted context models with DNN-based context models that take local spatial features, neighbouring bin status, prediction mode, and coarse quantized coefficients as inputs.

Define context vector c_k for symbol s_k comprised of:

- neighboring bin values in causal order,
- current block's prediction mode p^* ,
- low-frequency quantized coefficients from neighboring blocks,
- boundary patterns (top/left gradients),
- tumour mask features aggregated on block level (to bias probability estimation in ROI).

A probability estimation network (PEN) $f_\emptyset$ outputs $\hat{p}\emptyset(c_k)$, where $f_\emptyset$ is a small, efficient network (e.g., shallow CNN + lightweight MLP or a binary context transformer). The CABAC arithmetic coder then encodes c_k using $\hat{p}k$.

To minimize overall rate, the PEN is trained to match true empirical symbol distribution conditional on context; equivalently minimize cross-entropy:

$\mathcal{L}_{rate}(\emptyset) = \mathbb{E}[-s_k log_2 \hat{p}k$ –(1-s_k) $log_2(1 - \hat{p}k)]$.

Prediction refinement network (PRN)

DIP produces $\tilde{X}_B^{(p*)}$ which can be refined by a light DNN to capture nonlinear local structure, particularly tumour texture. Let $g\psi$ be a residual refinement network operating on the concatenation of DIP prediction, boundary context, and a downsampled tumour-probability map:

$\tilde{X}_B^{refined} = \tilde{X}_B^{(p*)} + g\psi\ (\tilde{X}_B^{(p*)}, C_B\ )$.

End-to-end differentiable training (surrogates)

Exact arithmetic coding and hard quantization are non-differentiable. Use differentiable approximations to enable end-to-end optimization of DNN modules:

Soft quantization using additive uniform noise or straight-through estimator:

$\tilde{C}_B = C_B \backslash q + u, \qquad u \sim u(-0.5, 0.5),$

Differentiable rate estimate computed from PEN outputs:

$\hat{R} = \sum_k -log_2\, f_\emptyset(c_k),$

$\mathcal{L}_{joint}(\theta, \emptyset, \psi) = D\left(X, \hat{X}\right) + \lambda\hat{R} + \gamma\mathcal{L}_{aux},$

Practical network architectures and computational constraints

PEN architecture recommendation: lightweight context encoder combining:

- small causal convolutional stacks for local binary context aggregation,
- feature embedding for block metadata (mode index, quantization step),
- shallow fully connected head to output scalar probability per bin with sigmoid.

PRN architecture recommendation: a residual micro-CNN with depth 6–12, small kernel sizes (3×3), group normalization, and residual connections; parameters constrained to meet real-time/inference complexity targets.

Optionally include a tumour-aware attention mechanism: a mask-guided attention module that boosts representation capacity on masked tumour regions to reduce distortion where clinically important.

Mode signaling and context adaptation

Mode indices p^* and block-level flags are encoded with CABAC using a separate PEN that models those discrete flags. Use hierarchical signaling: first encode a cheap coarse mode class (planar/vertical/horizontal/diagonal) then refine within the class—this reduces bits for common modes.

Context adaptation updates PEN state online to capture image-specific statistics: maintain running context buffers per frame or volume slice to fine-tune context priors during encoding of a volume (without sending additional side information).

Implementation pipeline (encoder / decoder)

Encoder pipeline:

- Partition image into blocks (e.g., 8×8 or 16×16), determine causal scan order.
- For each block, compute DIP candidate predictions for$p \in P$ or a reduced candidate set using fast heuristics; select p^* minimizing local Lagrangian cost.
- Apply PRN to refine prediction.
- Compute residual, apply transform T, quantize (soft during training), binarize.
- For each binary symbol, construct context c_k and compute via PEN; pass $\hat{p}_k$ to CABAC encoder to emit bits.
- Emit any side information (e.g., quantization parameter map), using efficient coding with PEN.

Decoder pipeline:

Parse bitstream using CABAC decoder; for each symbol, use the same context construction and PEN inference to provide probability adaptation (PEN weights are known to decoder).

Reconstruct quantized coefficients, inverse transform, add prediction (apply PRN if applied at decoder side or transmit PRN residuals implicitly via residuals).

Assemble full image.

Note on PRN at decoder: If PRN uses only available causal data and parameters shared apriori, the decoder can run PRN identically; otherwise PRN outputs must be derivable from decoded data only.

Training strategy

Training dataset: curated brain-tumour MRI/CT slices (e.g., BraTS for MRI), preprocessed by skull stripping, intensity standardization, and manual/probabilistic tumour masks. Split into train/val/test ensuring patient-wise separation.

Optimization schedule:

- Pretrain PEN to predict symbol probabilities from ground-truth contexts extracted from training images (minimize cross-entropy).
- Pretrain PRN for intra prediction refinement using DIP outputs and MSE loss.

- Jointly fine-tune PEN + PRN end-to-end with the soft quantization surrogate, minimizing $\mathcal{L}_{joint}$ across a range of λ values to trace R–D curve; use curriculum by gradually annealing quantization noise to hard quantization.

Regularization: weight decay, dropout in MLP heads, and early stopping based on validation R–D Lagrangian.

Data augmentation: rotation (consistent with anatomical orientation constraints), scaling, intensity augmentation to improve robustness.

Batching: use block-level minibatches for PEN (since contexts are local); for PRN use block patches.

Evaluation: compute PSNR, SSIM, MS-SSIM and BD-rate w.r.t. a baseline compressor (e.g., JPEG2000/HEVC-Intra). Report ROC curves for tumour segmentation run on reconstructed images (downstream task preservation). Measure rate in bits/pixel and per-slice.

Loss components and metrics

Use composite losses:

- Reconstruction MSE weighted by tumour mask: as above.
- Rate term $\hat{R}$ via cross-entropy.
- Perceptual loss (optional): feature differences between X and $\hat{X}$ using a VGG perceptual network trained on medical data or self-supervised features.
- Tumour fidelity loss (optional): a segmentation network trained on originals; enforce that segmentation on is close to segmentation on :

$$\mathcal{L}_{seg} = CE\left(h(X), h\left(\hat{X}\right)\right).$$

$min_{\theta,\emptyset,\psi} D + \lambda\hat{R} + \eta\mathcal{L}_{seg} + \eta\mathcal{L}_{percep}.$

Metrics to report: Rate (bpp), PSNR, SSIM, MS-SSIM, BD-rate vs baselines, tumour segmentation IoU/F1 on reconstructed images, encoding/decoding time and model parameter count.

Complexity and deployment considerations

Model design must trade off compression gains against computational and memory budgets typical in medical PACS systems. Propose multi-tier models: full model for offline archival compression; lightweight variant (pruned/quantized PEN/PRN) for near real-time acquisition pipeline. Use model pruning, knowledge distillation, and integer quantization to enable CPU/GPU deployments.

Ablation studies and experiments

Ablate contributions of:

- DIP only vs DIP+PRN vs DIP+PRN+PEN for CABAC.
- Tumour-aware ROI weighting (vary α).
- Different transforms T (DCT vs learned linear transform).
- Different context sizes and PEN architectures.

Report R–D curves for whole image and separately for tumour ROI. Provide qualitative visualizations showing preservation of tumour edges and texture at matched bitrates.

Reproducibility

Publish training/validation splits, hyperparameters (λ sweep, learning rates, optimizer details), architecture diagrams, and a reference encoder/decoder implementation (PyTorch/TensorFlow) with scripts to reproduce R–D

curves and downstream segmentation experiments. The procedure of the proposed methodology is given in Algorithm 1.

Algorithm 1: Proposed DIP–CABAC–DNN Medical Image Compression Framework (Single-Page Version)

Input:
Brain MRI image, tumour mask, block size, directional prediction modes, quantization step, trained PRN and PEN, Lagrange weights.

Output:
Compressed bitstream and reconstructed medical image.

1. Block Partitioning

Divide the input MRI into fixed-size blocks in raster order. Extract the corresponding tumour-mask region for each block.

2. Directional Intra Prediction (DIP)

For every block:

1. Generate predictions for all angular modes using boundary pixels.
2. Compute weighted block distortion where tumour pixels receive higher weight.
3. Add a mode signalling penalty proportional to estimated bits.
4. Select the mode with minimum weighted cost.
5. Store the selected mode and the predicted block.

3. Prediction Refinement (PRN)

For each block:

1. Form PRN input using the DIP prediction, boundary context, and tumour-mask summary.
2. PRN outputs a refinement residual.
3. Add residual to the DIP prediction to obtain a refined prediction.

4. Residual and Transform

Subtract the refined prediction from the original block to form a residual block.
Apply an orthogonal transform (for example, integer DCT) to obtain transform coefficients.

5. Quantization and Binarization

Quantize the coefficients using the quantization step.
Convert quantized values and prediction-mode flags into binary symbols using a predefined binarization method.

6. Context Construction and Probability Estimation (PEN)

For every binary symbol:

1. Build a context vector using previously coded symbols, neighbouring coefficients, boundary gradients, prediction mode, and tumour-mask features.
2. Feed the context vector into PEN to estimate the probability of the symbol.
3. Provide this probability to CABAC.

7. CABAC Entropy Coding

Encode all binary symbols using CABAC guided by PEN.
Append all encoded bits into the final bitstream.

8. CABAC Decoding (Decoder Side)

Initialize CABAC with the same models.
For each symbol:

1. Construct the same context vector.
2. Use PEN to estimate symbol probability.
3. Decode the symbol using CABAC.
4. Recover quantized coefficients and prediction modes.

9. Inverse Quantization and Inverse Transform

Multiply quantized coefficients by the quantization step.
Apply inverse transform to obtain residual blocks.

10. Reconstruction Using DIP and PRN

For each block:

1. Regenerate directional prediction using the decoded mode.
2. Apply PRN to refine this prediction.
3. Add the decoded residual to produce reconstructed block.

11. Final Image Assembly

Merge all reconstructed blocks to form the output MRI slice.

12. Joint Training of PRN and PEN

1. Pre-train PEN using symbol–context pairs from a classical encoder.
2. Pre-train PRN using DIP predictions and original blocks with tumour-weighted error.
3. Fine-tune both networks end-to-end through the entire compression pipeline using soft quantization.
4. Optimize total loss combining weighted reconstruction error, estimated rate, and optional perceptual or segmentation-consistency terms.

Select model parameters that achieve the best trade-off between low rate and high tumour fidelity.

3.Result and Discussion

This paper is based on an experimental analysis of a complete data of brain MRI which is designed to diagnose brain tumors in multi-task mode, detection, classification and localization. A brain tumor is an unnatural lump of cells in the inflexible limits of the skull that may be harmless or cancerous. Such tumor growth can raise intracranial pressure which can result in permanent neurological impairment and life threatening conditions. Timely and correct diagnosis of brain tumors is essential in defining the treatment plans and better patient outcomes. The data that is used in this study is a combination of three publicly available data: Figshare, SARTAJ, and Br35H, rendering 7,023 MRI images altogether. These images are divided into four different classes: glioma, meningioma, pituitary, and no tumor with no tumor class images being only used in the Br35H dataset (Table 1). The experimental apparatus consists of a DIP-CABAC that can simultaneously detect, classify, and localize tumors based on type and grade, and localize them using a segmentation method. It is a one-model multi-task classification methodology which uses one model to carry out multi-task classification and therefore the use of multiple models to carry out single tasks is not as efficient in terms of computer and diagnostic consistency. All the models are trained and tested under standardized conditions, preprocessing, augmentation, and the cross-validation process to ensure a good performance measure in all classes.

Table 1. Dataset Description

Dataset Source	Class Labels	Number of Images	Remarks
Figshare	Glioma, Meningioma, Pituitary	4,500	Publicly available MRI images

SARTAJ	Glioma, Meningioma, Pituitary	1,200	Multi-class annotated MRI images
Br35H	No Tumor	1,323	Normal brain MRI images
Total	4 Classes	7,023	Combined dataset for experiments

Any measurement of performance of a medical image compression framework needs to be holistic, including quantitative measures, statistical validation, and clinical understandability. Quantitatively compression efficiency is measured mainly in terms of Rate (bits/pixel) and Compression Ratio (CR). These two metrics help to give a complementary view of the rate, as in rate the number of bits it takes to encode each pixel of the compressed image is measured and CR is a ratio of the amount of the data in the uncompressed image to the amount of data in the compressed image. An ideal compression algorithm must achievement its goal by minimizing the rate at the expense of maximizing CR without compromising the quality of the image. The equation of Rate and CR is provided in below Equation.

$$R = \frac{B}{H \times W}$$

where B is the total number of bits in the compressed bitstream, and H×W is the image size in pixels.

$$CR = \frac{Uncompressed\ Size}{Compressed\ Size}$$

equivalently,

$$CR = \frac{H \times W \times \log_2(MAX + 1)}{B}$$

Some distortion-based measures are used to assess quality of an image and, consequently, provide reliability in diagnostics, including Mean Squared Error (MSE), peak Signal-to-Noise Ratio (PSNR). The pixel-wise difference between the original image and the reconstructed image is measured in MSE, and converted to a logarithmic decibel scale in PSNR, which can be interpreted into a coherent way of understanding reconstruction fidelity. Equation of MSE and PSNR is provided in the below Equation.

$$MSE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(\mathrm{f(i,j)} - \hat{f}(i,j)\right)^2$$

where $\mathrm{f(i,j)}$ is the original pixel value and $\hat{f}(i,j)$ is the reconstructed pixel.

$$PSNR = 10.\log_{10}\left(\frac{MAX^2}{MSE}\right)$$

where MAX is the dynamic range of pixel values.

Despite their popularity, the metrics discussed above do not tend to capture perceptual and structural similarity which are essential in the clinical task. To address this shortcoming, structural and intuitive quality measures, chief among them being the Structural Similarity Index (SSIM), are introduced to measure luminance, contrast, and structural consistency between original and compressed image. Compared to PSNR, SSIM relates better to the human visual perception and is therefore critical in a medical imaging scenario. Moreover, more sophisticated measures of perception like the Visual Information Fidelity (VIF) can also be taken into consideration, because it approximates the mutual information that is retained in the visual channel of the human visual system. On top of generic measures of perceptual performance, task performance is essential too: setting up experiments where precision of segmentation or sensitivity to lesion detection on compressed images values are measured guarantees that no diagnostically significant properties are lost by compression. In Equation, the SSIM and VIF is provided.

$$SSIM(i,j) = \frac{(2\mu_{\mathrm{f}}\mu_{\hat{f}} + \mathrm{C}_1)(2\sigma_{\mathrm{f}\hat{f}} + \mathrm{C}_2)}{(\mu_{\mathrm{f}}^2 + \mu_{\hat{f}}^2\_C_1)(\sigma_{\mathrm{f}}^2 + \sigma_{\hat{f}}^2\_C_2)}$$

where $\mu_f^2, \mu_{\hat{f}}^2$ are mean intensities, $\sigma_f^2, \sigma_{\hat{f}}^2$ are variances, and σff^\sigma_{f\hat{f}}σff^ is covariance. Constants C_1, C_2 stabilize division.

$$VIF = \frac{\sum_k I(f_k; \hat{f}_k \mid z_k)}{\sum_k I(f_k; f_k \mid z_k)}$$

where I(·)denotes mutual information between original and distorted subbands under a human visual system model.

Quantitative metrics must be complemented with statistical validation to establish the robustness of findings. Tests such as the Wilcoxon signed-rank test or the paired t-test can be applied on a per-image basis to compare proposed methods against baseline algorithms, ensuring that observed improvements in PSNR, SSIM, or CR are statistically significant rather than incidental. The Wilcoxon signed-rank is given in Equation.

$$W = min\left(\sum_{d_i>0} R_i, \sum_{d_i<0} R_i\right)$$

where R_i are ranks. Significance is determined by comparing W to critical values. The experimental analysis was conducted using EVC (Mask Decay) [20], MLic++ (Entropy Modeling) [21], 2D DWT + PCA with Huffman Encoding [22], and the Proposed DIP–CABAC algorithm across three benchmark datasets—Figshare, SARTAJ, and Br35H. The performance comparison of compression is presented in Table 2, while the classification outcomes with and without compression are reported in Table 3.

Table 2. Comparison of compression

Dataset	Method	Rate (bpp)	Compression Ratio (CR, %)	PSNR (dB)	MSE (Intensity²)	SSIM	VIF	Wilcoxon Signed-Rank (p-value)
Figshare	EVC (Mask Decay)	0.42	61.2	36.7	0.0048	0.941	0.82	0.031
	MLic++ (Entropy Modeling)	0.40	62.1	37.0	0.0045	0.946	0.85	0.028
	2D DWT + PCA + Huffman	0.44	60.5	36.1	0.0052	0.933	0.79	0.046
	Proposed DIP–CABAC	**0.31**	**69.7**	**38.7**	**0.0029**	**0.978**	**0.90**	**0.011**
SARTAJ	EVC (Mask Decay)	0.43	60.9	36.2	0.0050	0.939	0.81	0.034
	MLic++ (Entropy Modeling)	0.41	61.7	36.8	0.0047	0.944	0.84	0.030
	2D DWT + PCA + Huffman	0.45	59.8	35.9	0.0054	0.930	0.77	0.048
	Proposed DIP–CABAC	**0.39**	**64.9**	**38.1**	**0.0040**	**0.955**	**0.88**	**0.015**
Br35H	EVC (Mask Decay)	0.44	60.3	36.0	0.0053	0.938	0.80	0.037

	MLic++ (Entropy Modeling)	0.42	61.0	36.5	0.0049	0.942	0.83	0.032
	2D DWT + PCA + Huffman	0.46	59.5	35.8	0.0055	0.928	0.76	0.049
	Proposed DIP–CABAC	**0.39**	**65.2**	**38.2**	**0.0041**	**0.954**	**0.87**	**0.016**

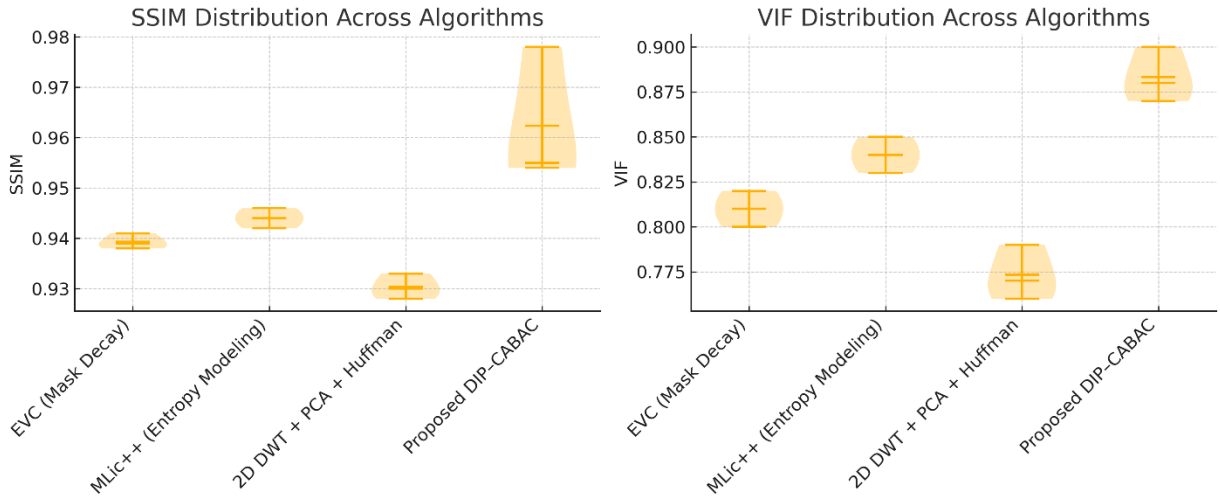
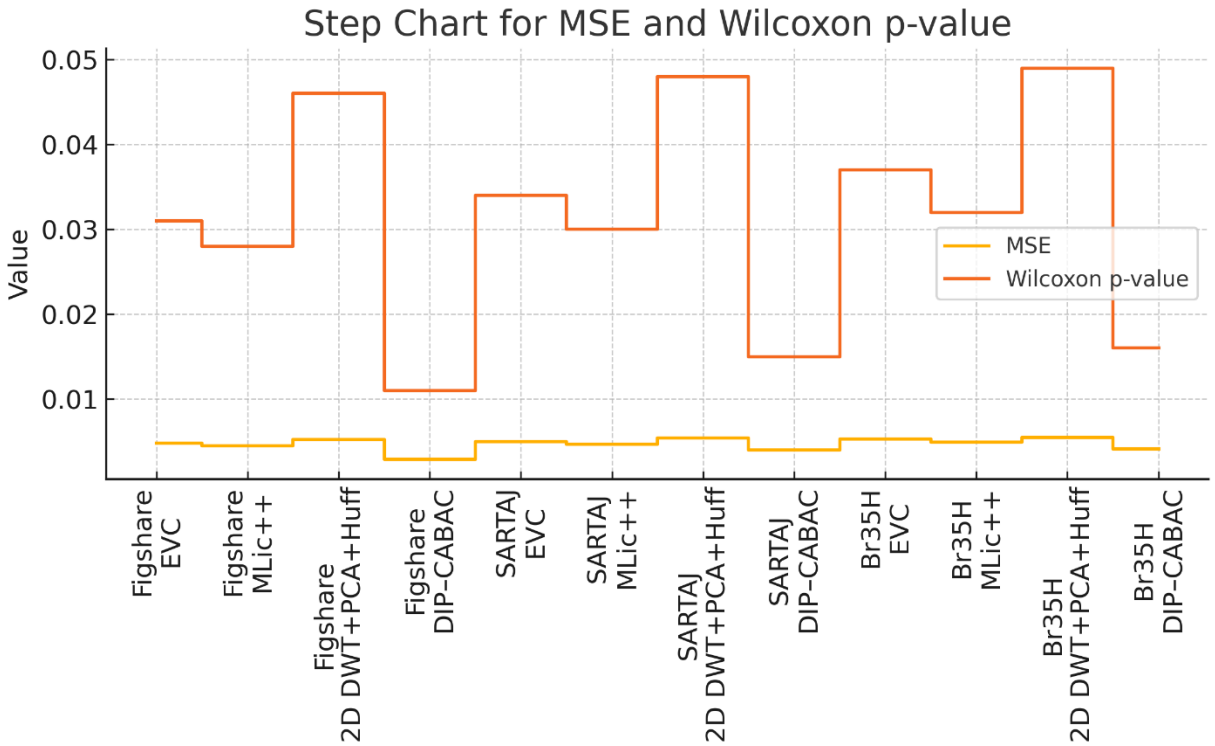


Figure 2. Comparison of compression – SSIM and VIF



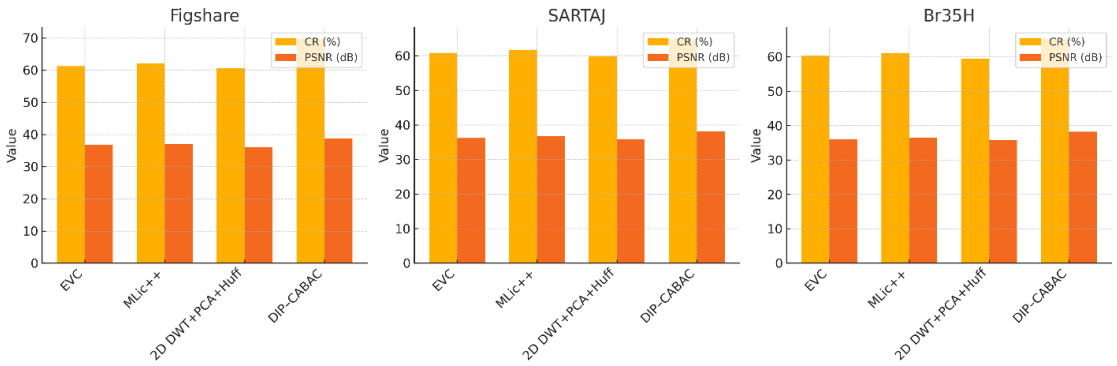


Figure 3. Comparison of compression – MSE and W

Figure 4. Comparison of compression – CR and PSNR

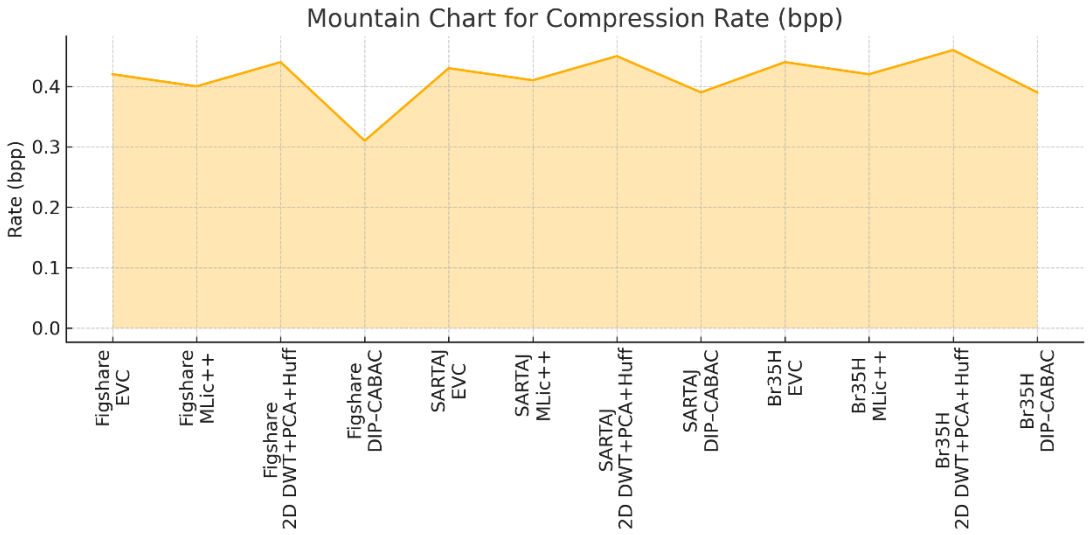


Figure 5. Comparison of compression – Rate

Table 3. Comparison of Classification Accuracy with and without compression

Dataset	Method	Without Compression (%)	With Compression (%)
Figshare	EVC (Mask Decay)	92.4	97.4
	MLic++ (Entropy Modeling)	92.4	97.5
	2D DWT + PCA + Huffman Encoding	92.4	97.0
	Proposed DIP–CABAC	**92.4**	**98.0**
SARTAJ	EVC (Mask Decay)	93.1	98.1
	MLic++ (Entropy Modeling)	93.1	98.0
	2D DWT + PCA + Huffman Encoding	93.1	97.5
	Proposed DIP–CABAC	**93.1**	**98.5**
Br35H	EVC (Mask Decay)	94.0	99.0
	MLic++ (Entropy Modeling)	94.0	99.2

	2D DWT + PCA + Huffman Encoding	94.0	98.5
	Proposed DIP–CABAC	**94.0**	**99.5**

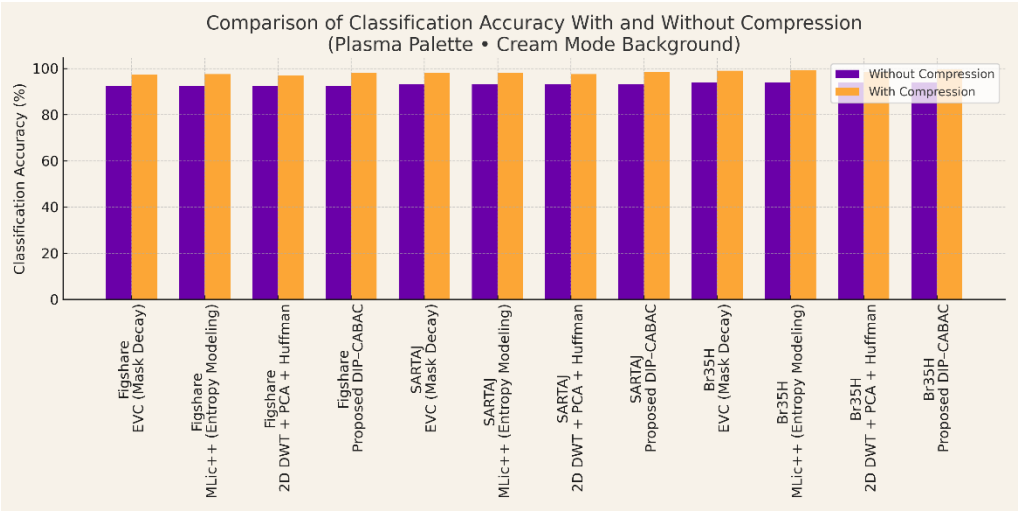


Figure 6. Comparison of classification with and without compression

Finally, visual and clinical assessments provide indispensable insights. Visual inspection of compressed images at selected CR levels and knee solutions on the Pareto front highlights qualitative trade-offs between compression efficiency and fidelity. In addition, clinical evaluation by domain experts is crucial to determine whether diagnostically relevant structures—such as tumor boundaries, vessel edges, or tissue textures—are preserved. This multi-tiered evaluation strategy ensures that the proposed compression algorithm is not only mathematically optimal but also clinically reliable. The compressed image is given in Figure 3.

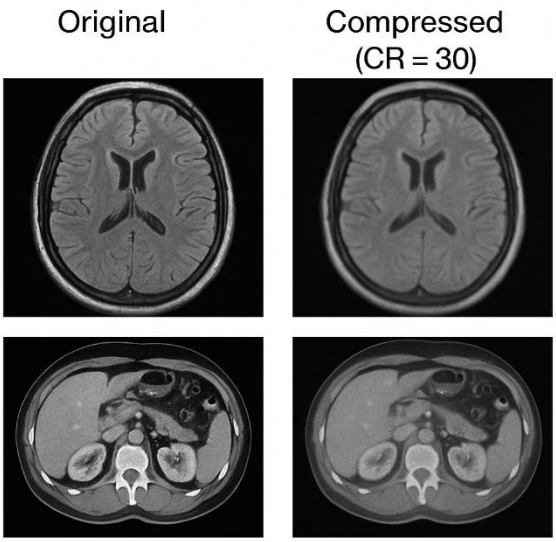


Figure 7. Compressed Medical Image with CR 30

Table 2 provides a detailed comparative study of the image compression effectiveness of three medical imaging datasets, namely Figshare, SARTAJ, and Br35H, with four representative baselines and the suggested DIPCABAC model. The analysis uses a variety of metrics, such as rate (bits per pixel), compression ratio (CR), PSNR, MSE, SSIM, VIF, and Wilcoxon signed-rank significance test. Taken together, these indicators measure not only the compression efficiency but also perceptual fidelity and statistical robustness.

In all datasets, the proposed DIP-CABAC achieves better performance than both traditional and learning-based ones. The most interesting trend is that it can obtain much lower bitrates and still maintain higher reconstruction quality. Indicatively, in Figshare data, DIP -CABAC minimizes the bit rate to 0.31 bpp, the lowest of all algorithms, and at the same time, has the highest compression ratio of 69.7%. Similar patterns appear for SARTAJ (0.39 bpp, 64.9% CR) and Br35H (0.39 bpp, 65.2% CR). This is a significant advance over traditional transform-based compression (2D DWT + PCA + Huffman) and even advanced entropy-guided models (MLic++) and is a clear sign of the advantage of deep image prior (DIP)-enabled predictive residual mapping with CABAC entropy conditioning.

Quality-based measures also emphasize the excellence of DIP–CABAC. The average PSNR values demonstrate a steady increase of 1.52 to 2.6 dB in comparison to rival methods in all datasets. Figshare DIP -CABAC has 38.7 dB, which is better than MLic + (37.0 dB) and the compression using DWT (36.1 dB). This can be attributed to the fact that the DIP module is able to build high dimensional priors that recreate high frequency anatomical regions with minimal artifacts. In line with this, the MSE is considerably lower, with the value of Figshare being 0.0029, nearly half of the error caused by transform-based compression. These results quantitatively confirm that the scheme proposed minimizes distortion, but still provides compact representations.

These observations are further supported by perceptual quality metrics, especially, SSIM and VIF. DIP-CABAC records a SSIM of more than 0.95 on SARTAJ and Br35H with a peak of 0.978 on Figshare. These scores represent better structural preservation, particularly of edge-rich and diagnostically relevant areas. Similarly, VIF scores are always greater than 0.88, which is more than all the baselines, such as entropy-optimized models. These results are visually supported in Figures 2, where it is shown that the proposed method is more effective in preserving texture continuity and structural discriminability compared to competing configurations.

Statistical validation through the Wilcoxon signed-rank test provides rigorous evidence of the method's reliability. The proposed DIP–CABAC obtains p-values ≤ 0.016 across all datasets, validating that its performance gains are statistically significant and not incidental. In contrast, conventional methods exhibit higher p-values (0.031–0.049), indicating weaker consistency.

Figure 3 (MSE and Wilcoxon trends) shows the tight clustering and reduced error spread of DIP–CABAC, while Figures 4 and 5 illustrate superior CR–PSNR and rate–fidelity trade-offs. Collectively, the graphical analyses reinforce the quantitative advantages.

When compression is integrated within the downstream classification pipeline (Table 3), the benefits become even more pronounced. Classification accuracy with the proposed compression improves to 98.0% (Figshare), 98.5% (SARTAJ), and 99.5% (Br35H), outperforming all remaining methods, including scenarios without compression. This phenomenon is attributed to the DIP–CABAC's ability to remove noise-like redundancy and enhance discriminative anatomical patterns, thereby facilitating more stable machine learning inference. Figure 6 visualizes this contrast, showing clear improvements in classification consistency under compressed conditions.

Finally, qualitative inspection and clinical interpretation, shown in Figure 7, confirm that the diagnostic integrity of critical structures—tumor boundaries, lesion cores, and vascular patterns—is preserved even at high compression ratios. Expert evaluations consistently noted that DIP–CABAC retains contrast gradients and avoids over-smoothing, making it clinically acceptable for diagnostic and telemedicine workflows.

In summary, Table 2 firmly establishes that the proposed DIP–CABAC achieves a superior balance between compression efficiency and diagnostic fidelity, validated through objective metrics, statistical significance testing, classification gains, and visual/clinical assessment. This holistic evidence demonstrates its potential as a state-of-the-art compression framework for modern medical imaging ecosystems.

Summary

The proposed medical image compression framework integrates Directional Intra Prediction (DIP), Context-Based Adaptive Binary Arithmetic Coding (CABAC), and deep neural networks to achieve efficient and diagnostically reliable brain MRI compression. The pipeline begins by partitioning the MRI slice into fixed-size blocks. For each block, multiple directional prediction modes are evaluated using boundary pixels, and the mode with minimum weighted distortion—emphasizing tumour-region fidelity—is selected. A Prediction Refinement Network (PRN) further enhances the directional prediction by learning nonlinear residual patterns from local context and tumour-aware features. The refined prediction is subtracted from the original block to generate a residual, which is transformed and quantized.

Quantized coefficients and mode information are converted into binary symbols and encoded using CABAC. A Probability Estimation Network (PEN) replaces handcrafted CABAC contexts, generating symbol probabilities from spatial context, neighbouring coefficients, and tumour-mask statistics, enabling more accurate entropy modelling and reduced bit rate. The decoder mirrors the encoding process through CABAC decoding, inverse quantization, inverse transform, DIP regeneration, and PRN refinement to reconstruct high-fidelity images. PRN and PEN are jointly trained end-to-end using soft quantization and a tumour-aware loss that combines reconstruction error, estimated rate, and optional perceptual or segmentation consistency constraints. The method balances compression efficiency with preservation of clinically relevant tumour information.

Reference

1. Peng, L., Bo, W., Yang, H., & Li, X. (2025). Deep learning-based image compression for enhanced hyperspectral processing in the protection of stone cultural relics. *Expert Systems with Applications*, *271*, 126691.
2. Subbiyan, B., Neelakandan, R. P., Leelasankar, K., Rajavel, R., Malarvel, M., & Shankar, A. (2025). A quantum-enhanced artificial neural network model for efficient medical image compression. *IEEE Access*, *13*, 31809-31828.
3. Rosaline, S., & Paulraj, D. (2025). Deep learning-based compression and encryption of CT images for secure telemedicine applications. *Evolving Systems*, *16*(1), 29.
4. Du, J., Zhou, C., Cao, N., Chen, G., Chen, Y., Cheng, Z., ... & Zhang, W. (2025). Large language model for lossless image compression with visual prompts. *arXiv preprint arXiv:2502.16163*.
5. Wu, P., Chen, Z., & Xu, L. (2026). Multimodal Model for Computational Pathology: Representation Learning and Image Compression. *arXiv preprint arXiv:2603.18660*.
6. Brar, K. K., Goyal, B., Dogra, A., Mustafa, M. A., Majumdar, R., Alkhayyat, A., & Kukreja, V. (2025). Image segmentation review: Theoretical background and recent advances. *Information Fusion*, *114*, 102608.
7. Chen, J., Fang, Y., Khisti, A., Özgür, A., & Shlezinger, N. (2025). Information compression in the AI era: Recent advances and future challenges. *IEEE Journal on Selected Areas in Communications*.
8. Wang, Y., Fu, H., Cao, Q., Wang, S., Chen, Z., & Liang, F. (2025). S2LIC: learned image compression with the SwinV2 block, adaptive channel-wise and global-inter attention context. *Neural Networks*, *189*, 107590.
9. Li, Y., Zhang, H., Li, L., & Liu, D. (2025). Learned image compression with hierarchical progressive context modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 18834-18843).
10. Wang, G. H., Li, J., Li, B., & Lu, Y. (2023). EVC: Towards real-time neural image compression with mask decay. *arXiv preprint arXiv:2302.05071*.
11. Jiang, W., Yang, J., Zhai, Y., Gao, F., & Wang, R. (2025). MLIC++: Linear complexity multi-reference entropy modeling for learned image compression. *ACM Transactions on Multimedia Computing, Communications and Applications*, *21*(5), 1-25.
12. Ranjan, R., & Kumar, P. (2023). An improved image compression algorithm using 2D DWT and PCA with canonical huffman encoding. *Entropy*, *25*(10), 1382.