

BRIQA: Balanced Reweighting in Image Quality Assessment of Pediatric Brain MRI

Alya Almsouti*, Ainur Khamitova*,
Darya Taratynova*, and Mohammad Yaqub

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE

Abstract. Assessing the severity of artifacts in pediatric brain Magnetic Resonance Imaging (MRI) is critical for diagnostic accuracy, especially in low-field systems where the signal-to-noise ratio is reduced. Manual quality assessment is time-consuming and subjective, motivating the need for robust automated solutions. In this work, we propose BRIQA (Balanced Reweighting in Image Quality Assessment), which addresses class imbalance in artifact severity levels. BRIQA uses gradient-based loss reweighting to dynamically adjust per-class contributions and employs a rotating batching scheme to ensure consistent exposure to underrepresented classes. Through experiments, no single architecture performs best across all artifact types, emphasizing the importance of architectural diversity. The rotating batching configuration improves performance across metrics by promoting balanced learning when combined with cross-entropy loss. BRIQA improves average macro F1 score from 0.659 to 0.706, with notable gains in Noise (0.430), Zipper (0.098), Positioning (0.097), Contrast (0.217), Motion (0.022), and Banding (0.012) artifact severity classification. The code is available at <https://github.com/BioMedIA-MBZUAI/BRIQA>.

Keywords: Low-field MRI · Quality Assessment · MRI Artifacts · Gradient-Based Reweighting · Class Imbalance

1 Introduction

Brain Magnetic Resonance Imaging (MRI) is an essential imaging modality to study pediatric brain development. In the early postnatal period, the human brain undergoes rapid growth and structural development; therefore, capturing these changes is important for improving our understanding of brain maturation and allowing the early detection of neurodevelopmental conditions [3,4,8,13]. While MRI is considered safe due to the absence of ionizing radiation [3], high-field systems produce loud noise and require children to remain still in enclosed spaces for extended periods, often needing sedation, which is not ideal. In addition, these systems’ high cost and maintenance requirements limit their accessibility in low- and middle-income countries.

* Equal contribution.

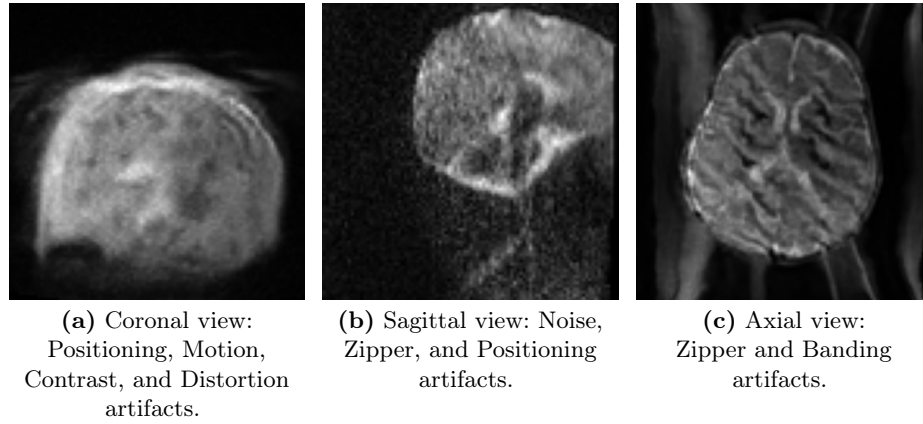


Fig. 1. Scans from multiple patients obtained with the 0.064T Hyperfine SWOP system, showing severe artifacts in different anatomical planes.

To address this, low-field MRIs offer an alternative solution with portable, point-of-care systems, reduced cost, quieter scans, and open designs, eliminating the need for sedation. However, the decreased signal-to-noise ratio in low field MRI poses limitations on the acquired image quality [1], and introduces artifacts, as shown in Figure 1. This makes image quality assessment essential to ensure that images meet specific standards and support diagnostic reliability, and given that manual Image Quality Assessment (IQA) is time-consuming and costly, automated solutions are crucial. This motivates Task 1 of the LISA Challenge 2025, the automatic assessment of the quality of the MRI scan in seven artifact classes.

Previous efforts have been made in brain MRI IQA, including the machine learning approach by Sanchez et al. [9], which extracts image quality metrics from fetal brain MRI for automatic quality assessment. Deep learning approaches include Zhang et al. [15], who proposed jointly segmenting the brain and assessing quality in fetal MRI slices, while Lou et al. [5] developed a contrastive learning method to enhance feature extraction, leveraging both spatial and frequency representations for quality assessment. The previous LISA 2024 Challenge [4] featured Kim et al. [2], who predicted scan orientation alongside quality assessment, Sundaresan et al. [11], who suggested synthesizing artifacts, and Zhu et al. [16], who developed a multi-label model combining gated CNNs and an ML-Decoder.

However, these prior studies rely on a single model architecture for all artifact types. In practice, performance can vary depending on the medical application; for example, ResNet may outperform DenseNet in some scenarios and vice versa [12,6]. Moreover, larger models do not necessarily perform better, particularly on small datasets [14]. Therefore, it is beneficial to leverage diverse architectures of varying sizes, as different models may excel at identifying different artifact types based on their distinct visual patterns.

In this work, we introduce BRIQA, a method for the automatic assessment of artifact severity of MRI scans. BRIQA features a tailored model architecture for each artifact type, along with a gradient-weighting strategy and a custom batching technique to address class imbalance. The remainder of this paper is organized as follows: Section 2 describes the dataset used and BRIQA framework including gradient-based reweighting and rotating batching, Section 3 and 4 presents experimental results with discussion, followed by conclusion at Section 5.

2 Methods

2.1 Dataset

In the LISA 2025 Challenge, quality assessment involves scoring the presence of seven common artifacts on pediatric brain magnetic resonance images: Banding, Contrast, Motion, Distortion, Noise, Positioning, and Zipper. Each artifact is rated on a three-point severity scale: 0 for no artifact, 1 for moderate, and 2 for severe. The dataset provided by the challenge organizers consists of 532 brain magnetic resonance images acquired at a low magnetic field strength of 0.064T, representing 244 unique pediatric subjects. Each subject had up to three scans acquired in different orientations: axial, coronal, and sagittal. The severity of artifacts varied across scans.

As illustrated in Table 1, scans containing artifacts are underrepresented. A solution proposed by the first-place winner [11] involved increasing the proportion of scans with artifacts through simulation. Following this approach, we applied artifact simulation using TorchIO, adopting the same parameters as in previous work for all artifact types except motion. Specifically, for moderate motion (level 1), we increased the rotation severity from three to five degrees, and for severe motion (level 2), from seven to ten degrees. These adjustments resulted in more visually distinguishable motion artifacts, ensuring clearer degradation corresponding to the assigned severity level. The distribution of artifacts before and after simulation is shown in Table 1. It is worth noting that although the number of class 1 and 2 instances increased, the overall distribution remains imbalanced. For training, scans were resized to $128 \times 128 \times 128$, followed by augmentations such as normalization, center spatial cropping, and random rotation.

2.2 Model Description

To predict the severity of artifacts from MRI scans, we employ a multitask learning framework. As demonstrated by [2], incorporating scan plane classification as an auxiliary task enhances quality assessment, as the appearance of artifacts can vary with anatomical orientation.

In BRIQA, each input scan $\mathbf{x}$ is processed by an encoder $f_\theta(\cdot)$, which branches into two heads: one for the classification of the severity of the artifact and the other for the classification of the scan plane (axis). To mitigate the effects of class

Table 1. Distribution of artifact severity before and after simulation.

Artifact	Before			After		
	Class 0	Class 1	Class 2	Class 0	Class 1	Class 2
Noise	426	60	46	734	122	97
Zipper	398	105	29	686	201	66
Positioning	470	47	15	810	106	37
Banding	504	15	13	871	52	30
Motion	384	78	70	672	147	134
Contrast	375	134	23	637	265	51
Distortion	435	56	41	782	101	70

imbalance between severity levels $c' \in 0, 1, 2$, BRIQA adopts a gradient-based loss reweighting strategy.

For each class c , BRIQA calculates how much that class contributes to the training signal by measuring the size of the gradients it produces. Specifically, BRIQA computes the ℓ_2 norm of the gradient of the classification loss $\mathcal{L}_{\text{cls}}^{(c)}$ when considering only the samples that belong to class c . This gradient is taken with respect to the parameters of the classification head, denoted as θ_{cls} . The result is a scalar value ϕ_c , which reflects the overall magnitude of the update that class c would induce on the classification head if it were trained in isolation:

$$\phi_c = \left\| \nabla_{\theta_{\text{cls}}} \mathcal{L}_{\text{cls}}^{(c)} \right\|_2. \quad (1)$$

To rebalance the contributions from each class, BRIQA normalizes the gradients by the smallest observed norm ℓ_2 :

$$\alpha_c = \frac{\min_{c'} \phi_{c'}}{\phi_c}. \quad (2)$$

These weights are then used to compute a weighted classification loss:

$$\mathcal{L}_{\text{cls}} = \sum_{c \in c'} \alpha_c \cdot \mathcal{L}_{\text{cls}}^{(c)}. \quad (3)$$

Finally, the total loss for each batch is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{axis}}. \quad (4)$$

where $\mathcal{L}_{\text{axis}}$ encourages orientation-sensitive representations.

Batch Configurations. We experiment with two different configurations to form training batches to handle class imbalance and improve learning stability.

Standard Batching. The first configuration employs standard random sampling, where training batches are formed by shuffling the dataset without enforcing any class distribution constraints.

Rotating Batching. This is a custom configuration that introduces epoch-wise variation in the selection of class 0 samples while maintaining a fixed class

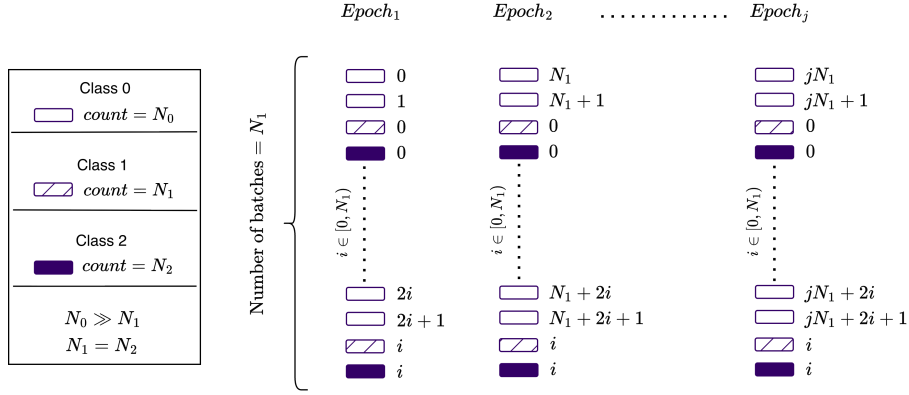


Fig. 2. Rotating batch configuration. Samples from class 0 are drawn using a rotating buffer mechanism, where their indices are cyclically shifted across epochs according to a modular offset. Note that for each sample, we take the remainder of the calculated index with respect to N_0 to achieve the cycling effect; this operation is not shown in the figure.

ratio within each batch. Each batch of size 4 contains two samples from class 0, one from class 1, and one from class 2. To address the scarcity of class 2, we apply random upsampling to match the number of class 1 samples. Unlike standard random sampling, class 0 examples are drawn using a rotating buffer strategy: their indices are cyclically shifted across epochs using a modular offset. This strategy is illustrated in Figure 2. This ensures uniform usage of all class 0 samples over time, improving training diversity. Importantly, batches are constructed before data augmentations, enabling consistent per-batch composition while introducing controlled epoch-level variation. To our knowledge, this rotating buffer mechanism is a novel contribution.

2.3 Experimental Setup

To evaluate BRIQA, we experimented with a range of encoder backbones, including DenseNet and ResNet variants, as well as MedNeXtS [7]. In addition to backbone comparisons, we compare our method against [16], which considers the problem as multilabel classification. We also compare with the incorporation of a frequency-based encoder from [5]. Specifically, we applied the Discrete Fourier Transform (DFT) to the input MRI and passed the transformed image through a separate encoder. The output of the DFT-based encoder was concatenated with the features of the spatial MRI encoder before classification to capture complementary information in the frequency domain.

To analyze the impact of orientation-aware training, we compared models trained without rotation augmentation to those trained with random rotation up to 180 degrees. Additionally, we evaluated the effect of different loss functions,

Table 2. Weighted F1-scores across different MRI artifact categories using various encoder backbones and MambaOut variants. The best scores per artifact are highlighted in **bold**.

Architecture	Noise	Zipper	Positioning	Banding	Motion	Contrast	Distortion	Mean
Encoder								
Resnet10	0.844	0.836	0.826	0.947	0.687	0.813	0.765	0.817
Resnet18	0.805	0.822	0.830	0.947	0.736	0.822	0.754	0.817
Resnet50	0.827	0.803	0.846	0.947	0.727	0.800	0.753	0.814
Resnet101	0.815	0.807	0.856	0.926	0.730	0.788	0.753	0.811
DenseNet169	0.844	0.853	0.872	0.926	0.716	0.788	0.776	0.825
DenseNet264	0.698	0.853	0.870	0.942	0.731	0.795	0.767	0.808
MedNeXtS [7]	0.807	0.761	0.789	0.881	0.486	0.746	0.601	0.725
MLMambaOut								
MambaOut tiny [16]	0.807	0.845	0.826	0.924	0.722	0.744	0.763	0.805
MambaOut small [16]	0.809	0.838	0.801	0.946	0.690	0.789	0.821	0.813
MambaOut base [16]	0.811	0.814	0.834	0.947	0.698	0.767	0.790	0.809

including Cross-Entropy (CE), Weighted Cross-Entropy, Focal Loss, and Ordinal Loss[10].

The dataset was split into training and internal validation sets at the patient level, with 80% of patients used for training and 20% for validation. All models were trained for 150 epochs on an NVIDIA A100-SXM4-40GB GPU, using the Adam optimizer with a learning rate of 1×10^{-5} , and a Cosine Annealing learning rate scheduler.

3 Results

Table 2 presents weighted F1 scores for detecting seven MRI artifacts in various encoders with classification heads, showing that no single architecture performs best in all types of artifacts. For example, DenseNet169 excels in detecting Zipper, Positioning, and Noise artifacts (0.844, 0.853, and 0.872, respectively), while Resnet18 achieves the highest scores on Banding, Motion and Contrast (0.947, 0.736, and 0.822, respectively). Simpler models like Resnet10 also perform competitively, particularly on Banding and Noise, outperforming deeper models in some cases. Meanwhile, MedNeXtS [7] struggles with Motion and Distortion. The MLMambaOut [16] architecture, which uses a single backbone for multi-label classification, demonstrates limited effectiveness across several artifact types regardless of the size of the model.

After selecting the best-performing backbone for each artifact, we conducted experiments under different training configurations. Table 3 compares the baseline setup, where reweighting is not applied, with BRIQA. In the baseline, we experiment with various loss functions, regularization techniques, and a fusion model that combines spatial MRI features with Discrete Fourier Transform (DFT) representations.

When comparing loss functions in the baseline setting, we observe that weighted cross-entropy and ordinal loss outperform standard cross-entropy, achiev-

Table 3. Performance metrics for different methods. $\circ$ – No rotation (0°), $\cup$ – Rotation 180° . Best results are highlighted in **bold**, and second-best results are underlined.

Method	Weighted					Macro					Micro					Mean
	Prec.	Rec.	F1	F2	Acc.	Prec.	Rec.	F1	F2	Acc.	Prec.	Rec.	F1	F2	Acc.	
Baseline																
Standard Batch: Loss Variations																
CE $\circ$	0.800	0.846	0.818	0.834	0.846	0.587	0.552	0.560	0.554	0.552	0.846	0.846	0.846	0.846	0.846	0.745
CE $\cup$	0.817	0.834	0.818	0.829	0.839	0.708	0.576	0.619	0.590	0.576	0.839	0.840	0.840	0.840	0.839	0.761
Ordinal Loss $\circ$	0.828	0.841	0.821	0.831	0.841	0.741	0.572	0.607	0.582	0.572	0.841	0.841	0.841	0.841	0.841	0.763
Weighted CE $\circ$	0.834	0.857	<u>0.842</u>	<u>0.851</u>	0.857	0.661	0.620	0.629	0.622	0.620	<u>0.857</u>	<u>0.857</u>	<u>0.857</u>	<u>0.857</u>	<u>0.857</u>	0.779
Standard Batch: Regularization																
CE $\circ$	0.835	0.860	0.842	0.852	0.860	0.660	0.621	0.628	0.623	0.621	0.860	0.860	0.860	0.860	0.860	<u>0.780</u>
CE $\cup$	0.825	0.815	0.810	0.812	0.815	0.670	0.607	0.608	0.605	0.607	0.815	0.815	0.815	0.815	0.815	0.750
Standard Batch: DFT Fusion																
CE $\circ$	0.829	0.845	0.834	0.840	0.845	0.701	0.631	0.659	0.641	0.631	0.845	0.845	0.845	0.845	0.845	0.779
BRIQA																
Standard Batch: Loss Variations																
CE $\circ$	<u>0.840</u>	<u>0.853</u>	0.844	0.849	0.853	<u>0.732</u>	0.657	0.688	0.668	0.657	0.853	0.853	0.853	0.853	0.853	0.794
CE $\cup$	0.789	0.743	0.763	0.750	0.743	0.576	0.612	0.592	0.607	0.619	0.743	0.743	0.743	0.743	0.743	0.701
Ordinal Loss $\circ$	0.838	0.786	0.801	0.789	0.786	0.653	0.649	0.621	0.625	0.649	0.786	0.786	0.786	0.786	0.786	0.742
Focal loss $\circ$	0.818	0.818	0.818	0.818	0.818	0.642	0.648	0.645	0.647	0.648	0.818	0.818	0.818	0.818	0.818	0.760
Standard Batch: DFT Fusion																
CE $\circ$	0.821	0.827	0.823	0.825	0.827	0.665	0.683	0.671	<u>0.677</u>	0.683	0.827	0.827	0.827	0.827	0.827	0.776
Rotating Batch: BRIQA																
CE $\circ$	0.843	0.849	0.846	0.848	0.849	0.724	0.690	0.706	0.696	0.690	0.849	0.849	0.849	0.849	0.849	0.799

ing mean scores of 0.779 and 0.763, respectively, compared to 0.745 without rotation. Interestingly, applying rotation in the cross-entropy setup leads to a 0.016 improvement in the mean score. However, rotation does not consistently yield better performance. For instance, in the regularization setting, the no-rotation variant achieves a mean score of 0.780, the second highest overall, and obtains the best micro-averaged scores, outperforming the rotated version which averages 0.750. Finally, the fusion experiment, which incorporates spectral information via the DFT, shows performance close to that of the second-best configuration, suggesting that frequency-domain features may offer complementary benefits for artifact detection.

Across all configurations, BRIQA improves performance. For cross-entropy without rotation, the mean score increases from 0.745 to the highest overall score of 0.799 with rotating batching. This setup also achieves the best macro scores among all experiments while maintaining weighted and micro scores within 0.01 of the second-best results.

To better understand where these gains are most impactful, Table 4 presents a detailed breakdown of performance improvements in the seven types of MRI artifacts. The model demonstrates consistent improvements in most artifacts, particularly in Noise, Zipper, and Distortion, where all metrics show notable gains. For example, noise and distortion exhibit substantial increases in macro F1 and F2 scores, indicating enhanced sensitivity to rare or harder-to-classify severity levels. Although Banding achieved the highest weighted and micro F1 scores (0.919 and 0.905, respectively), its macro performance was slightly lower

Table 4. Performance metrics of the best-performing model across artifact types. ($\uparrow$) indicates improvement over the CE without gradient reweighting, and ($\downarrow$) denotes a performance decrease.

Metric	Noise	Zipper	Positioning	Banding	Motion	Contrast	Distortion
Weighted							
Precision	0.863 $\uparrow_{0.238}$	0.871 $\uparrow_{0.021}$	0.902 $\uparrow_{0.037}$	0.936 $\uparrow_{0.001}$	0.755 $\uparrow_{0.025}$	0.811 $\downarrow_{0.012}$	0.870 $\uparrow_{0.100}$
Recall	0.876 $\uparrow_{0.086}$	0.867 $\uparrow_{0.010}$	0.838 $\downarrow_{0.048}$	0.905 $\downarrow_{0.057}$	0.781 $\uparrow_{0.029}$	0.810 $\downarrow_{0.019}$	0.867 $\uparrow_{0.057}$
F1-score	0.865 $\uparrow_{0.167}$	0.863 $\uparrow_{0.010}$	0.860 $\downarrow_{0.011}$	0.919 $\downarrow_{0.028}$	0.747 $\uparrow_{0.012}$	0.799 $\downarrow_{0.023}$	0.861 $\uparrow_{0.085}$
F2-score	0.871 $\uparrow_{0.120}$	0.864 $\uparrow_{0.009}$	0.844 $\downarrow_{0.036}$	0.910 $\downarrow_{0.045}$	0.765 $\uparrow_{0.020}$	0.803 $\downarrow_{0.022}$	0.863 $\uparrow_{0.069}$
Accuracy	0.876 $\uparrow_{0.086}$	0.867 $\uparrow_{0.010}$	0.838 $\downarrow_{0.048}$	0.905 $\downarrow_{0.057}$	0.781 $\uparrow_{0.029}$	0.810 $\downarrow_{0.019}$	0.867 $\uparrow_{0.057}$
Macro							
Precision	0.747 $\uparrow_{0.483}$	0.865 $\uparrow_{0.214}$	0.713 $\uparrow_{0.037}$	0.593 $\downarrow_{0.061}$	0.716 $\uparrow_{0.080}$	0.705 $\downarrow_{0.036}$	0.725 $\uparrow_{0.177}$
Recall	0.728 $\uparrow_{0.396}$	0.684 $\uparrow_{0.055}$	0.794 $\uparrow_{0.177}$	0.643 $\uparrow_{0.088}$	0.598 $\uparrow_{0.013}$	0.738 $\uparrow_{0.030}$	0.636 $\uparrow_{0.215}$
F1-score	0.725 $\uparrow_{0.430}$	0.731 $\uparrow_{0.098}$	0.732 $\uparrow_{0.097}$	0.605 $\uparrow_{0.012}$	0.625 $\uparrow_{0.022}$	0.698 $\downarrow_{0.022}$	0.657 $\uparrow_{0.217}$
F2-score	0.725 $\uparrow_{0.408}$	0.698 $\uparrow_{0.068}$	0.761 $\uparrow_{0.138}$	0.622 $\uparrow_{0.053}$	0.605 $\uparrow_{0.014}$	0.716 $\uparrow_{0.004}$	0.641 $\uparrow_{0.216}$
Accuracy	0.728 $\uparrow_{0.395}$	0.684 $\uparrow_{0.055}$	0.793 $\uparrow_{0.177}$	0.643 $\uparrow_{0.088}$	0.598 $\uparrow_{0.013}$	0.738 $\uparrow_{0.030}$	0.636 $\uparrow_{0.215}$
Micro							
Precision	0.876 $\uparrow_{0.086}$	0.867 $\uparrow_{0.010}$	0.838 $\downarrow_{0.048}$	0.905 $\downarrow_{0.057}$	0.781 $\uparrow_{0.029}$	0.810 $\downarrow_{0.019}$	0.867 $\uparrow_{0.057}$
Recall	0.876 $\uparrow_{0.086}$	0.867 $\uparrow_{0.010}$	0.838 $\downarrow_{0.048}$	0.905 $\downarrow_{0.057}$	0.781 $\uparrow_{0.029}$	0.810 $\downarrow_{0.019}$	0.867 $\uparrow_{0.057}$
F1-score	0.876 $\uparrow_{0.086}$	0.867 $\uparrow_{0.010}$	0.838 $\downarrow_{0.048}$	0.905 $\downarrow_{0.057}$	0.781 $\uparrow_{0.029}$	0.810 $\downarrow_{0.019}$	0.867 $\uparrow_{0.057}$
F2-score	0.876 $\uparrow_{0.086}$	0.867 $\uparrow_{0.010}$	0.838 $\downarrow_{0.048}$	0.905 $\downarrow_{0.057}$	0.781 $\uparrow_{0.029}$	0.810 $\downarrow_{0.019}$	0.867 $\uparrow_{0.057}$
Accuracy	0.876 $\uparrow_{0.086}$	0.867 $\uparrow_{0.010}$	0.838 $\downarrow_{0.048}$	0.905 $\downarrow_{0.057}$	0.781 $\uparrow_{0.029}$	0.810 $\downarrow_{0.019}$	0.867 $\uparrow_{0.057}$
Mean	0.826 $\uparrow_{0.216}$	0.822 $\uparrow_{0.040}$	0.818 $\uparrow_{0.019}$	0.814 $\downarrow_{0.020}$	0.725 $\uparrow_{0.027}$	0.778 $\downarrow_{0.012}$	0.797 $\uparrow_{0.113}$

than that of other artifacts, probably due to its prevalence in the dataset. Zipper achieved the highest macro F1 score, improving by nearly 10% over baseline.

4 Discussion

Is One Backbone Architecture Enough? Performance in Table 2 suggests that a one-size-fits-all architecture may not be ideal for MRI artifact detection. Instead, leveraging the complementary strengths of diverse backbones could offer improved robustness across artifact types. The consistent variability in per-artifact performance across architectures, especially for more challenging categories like Distortion and Motion, indicates that certain models are more sensitive to specific artifact patterns. Rather than seeking a universally strong backbone, it may be more effective to utilize this diversity and design adaptive frameworks that combine multiple models to capitalize on their respective strengths. We hypothesize that the performance difference comes from how each network propagates features. ResNet adds features through residual connections, capturing global structure and performing better on artifacts affecting the whole image, like motion, contrast, and banding. DenseNet concatenates features, preserving fine detail, which helps with localized or textural artifacts such as zipper lines and positioning shifts.

Is Cross-Entropy Enough? The baseline experiments show that standard cross-entropy loss is suboptimal when compared to both weighted cross-entropy and ordinal loss. These alternative loss functions are more effective in addressing label imbalance and capturing ordinal relationships between classes, resulting in

higher macro- and weighted scores. However, when paired with BRIQA, the standard cross-entropy loss achieves the highest overall performance compared to other losses. This improvement comes from the rotating batch configuration’s ability to expose the model to diverse artefact combinations across training iterations, helping the model generalise better across underrepresented classes. In addition to weighting the loss based on gradient contributions, which eliminates the need for explicit reweighting mechanisms in focal loss. In this context, standard cross-entropy benefits from a more uniform and representative training distribution, making it competitive, even surpassing more specialized loss functions.

Does Scan Rotation Always Help? If rotation-based augmentation is applied, the benefits appear inconsistent in different settings. While it yields marginal gains in some configurations at baseline, such as standard cross-entropy, it can degrade performance in more optimized settings like regularized training, where the non-rotated variant achieved a substantially better mean (0.780 vs. 0.750) and the highest micro-average overall. In contrast to baseline, in batch configuration settings, applying rotation with standard cross entropy degraded the performance. This suggests that rotation may inject noise or disrupt spatial integrity in some representations, especially when models already have strong regularisation or batching restrictions.

Can Custom Batching Improve Learning from Imbalanced Data? Our findings highlight the significant impact of batching design on model performance under class imbalance. Unlike standard batching, which randomly samples data and may repeatedly draw from overrepresented classes, rotating batching enforces a fixed class ratio within each batch while systematically cycling through majority class (class 0) samples across epochs. This strategy ensures that minority classes are consistently represented in every batch, while the majority class is varied to maintain diversity and prevent oversaturation. By balancing the gradient signal across classes, rotating batching helps stabilize training and reduces the tendency to overfit to dominant class patterns. This is particularly important in the context of artifact detection, where severe artifact cases (class 2) are underrepresented. Without careful batching, the model could learn to ignore rare artifacts in favor of more frequent, clean scans. Rotating batching ensures that learning remains attentive to all severity levels, improving generalization and classification robustness.

Does Frequency Domain Help? The observed performance of the DFT fusion setup suggests that incorporating frequency-domain features offers complementary benefits to spatial representations. Although the mean score of the DFT fusion model falls slightly below that of the best regularized configuration, its strong performance in macro-averaged metrics indicates improved generalization to underrepresented severity levels. This is particularly relevant for artifacts, where frequency patterns may be more informative than spatial textures alone. However, the relatively modest gains compared to those of other setups suggest that a simple fusion strategy may not fully exploit the potential of spectral infor-

mation. More advanced integration mechanisms, such as attention-based fusion, may be necessary to effectively combine spatial and frequency-domain features. **Computational Requirements.** BRIQA takes 0.008–0.016 s per sample with 740 MB peak GPU memory for DenseNet-based variants (Noise, Zipper, Positioning), 0.021 s and 4.3 GB for ResNet18 models (Banding, Contrast, Motion), and 0.041 s with 7.2 GB for the MedNextS Distortion model.

5 Conclusion

In this work, we addressed the challenge of automatic classification of MRI artifact severity under class imbalance by proposing BRIQA, which integrates axis prediction, gradient-based loss reweighting, and a rotation-based batch construction strategy. Our findings show that architectural diversity can be leveraged for better performance across different artifact categories, while rotating batching significantly enhances generalization by ensuring consistent exposure to minority classes. Future work may explore dynamic ensemble methods based on artifact type or severity distribution, as well as artifact-specific expert models trained to handle visually distinct patterns. While BRIQA demonstrates improved performance, several limitations warrant consideration. First, the multi-architecture approach requires training and maintaining multiple models, increasing computational overhead compared to single-model solutions. Second, the relatively small dataset size and reliance on simulated artifacts may limit generalization to other low-field MRI systems or diverse patient populations. Clinical validation on larger, multi-center datasets is necessary to confirm BRIQA’s utility in real diagnostic workflows.

Disclosure of Interests. The authors have no competing interests.

References

1. Thomas Campbell Arnold, Colbey W Freeman, Brian Litt, and Joel M Stein. Low-field mri: clinical promise and challenges. *Journal of Magnetic Resonance Imaging*, 57(1):25–44, 2023.
2. Hyunwook Kim, Jinew Seo, Seiyoun Ryu, Joon Hyung Park, Sungchul On, and Jinwha Choi. Axis-guided quality assessment and multi-label hippocampal and ventricular segmentation in low-resolution pediatric brain mri. In Natasha Lepore and Marius George Linguraru, editors, *Proceedings of the Low Field Pediatric Brain Magnetic Resonance Image Segmentation and Quality Assurance (LISA 2024)*, volume 15515 of *Lecture Notes in Computer Science*, pages 53–62. Springer, Cham, 2025.
3. Rhoshel K Lenroot and Jay N Giedd. Brain development in children and adolescents: insights from anatomical magnetic resonance imaging. *Neuroscience & biobehavioral reviews*, 30(6):718–729, 2006.
4. Natasha Lepore and Marius George Linguraru. Low field pediatric brain magnetic resonance image segmentation and quality assurance: First miccai challenge, lisa 2024, held in conjunction with miccai 2024, marrakesh, morocco, october 10, 2024, proceedings, 2025.

5. Yiwei Lou, Jiayu Zhang, Dexuan Xu, Yongzhi Cao, Hanpin Wang, and Yu Huang. No-reference mri quality assessment via contrastive representation: Spatial and frequency domain perspectives. In *2024 IEEE International Conference on Multi-media and Expo (ICME)*, pages 1–6. IEEE, 2024.
6. I Putu Gede Yoga Pramana Putra, Ni Wayan Jeri Kusuma Dewi, Putu Surya Wedra Lesmana, I Gede Totok Suryawan, and Putu Satria Udyana Putra. Comparison of resnet-50 and densenet-121 architectures in classifying diabetic retinopathy. *Indonesian Journal of Data and Science*, 6(1):64–72, 2025.
7. Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F. Jäger, and Klaus H. Maier-Hein. Mednext: Transformer-driven scaling of convnets for medical image segmentation. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 405–415, Cham, 2023. Springer Nature Switzerland.
8. Uroosa Saman, Anwarul Haque, Namaya Hussain, and Bushra Shamim. Utility of magnetic resonance imaging of brain in neurocritically ill children in pediatric intensive care unit: A single-center retrospective observational study. *Journal of Pediatric Critical Care*, 11(1):6–9, 2024.
9. Thomas Sanchez, Oscar Esteban, Yvan Gomez, Elisenda Eixarch, and Meritxell Bach Cuadra. Fetmrqc: automated quality control for fetal brain mri. pages 3–16, 2023.
10. Xintong Shi, Wenzhi Cao, and Sebastian Raschka. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications*, 26(3):941–955, 2023.
11. Vaanathi Sundaresan and Nicola K Dinsdale. Automated quality assessment using appearance-based simulations and hippocampus segmentation on low-field paediatric brain mr images. pages 41–52, 2024.
12. Tomoki Uemura, Janne J Näppi, Toru Hironaka, Hyungseop Kim, and Hiroyuki Yoshida. Comparative performance of 3d-densenet, 3d-resnet, and 3d-vgg models in polyp detection for ct colonography. In *Medical Imaging 2020: computer-aided diagnosis*, volume 11314, pages 736–741. SPIE, 2020.
13. Firehiwot Workneh, Theresa Inez Chin, Kalkidan Yibeltal, Krysten North, Nebiyou Fasil, Workagegnhu Tarekegn, Betelhem Haimanot Abate, Sarem Mulugeta, Gellila Asmamaw, Atsede Teklehaimanot, et al. Feasibility and acceptability of magnetic resonance imaging and electroencephalography for child neurodevelopmental research in rural ethiopia. *Frontiers in Public Health*, 13:1551982, 2025.
14. Yuan Yang, Lin Zhang, Mingyu Du, Jingyu Bo, Haolei Liu, Lei Ren, Xiaohe Li, and M Jamal Deen. A comparative analysis of eleven neural networks architectures for small datasets of lung images of covid-19 patients toward improved clinical decisions. *Computers in Biology and Medicine*, 139:104887, 2021.
15. Wenhao Zhang, Xin Zhang, Lingyi Li, Lufan Liao, Fenqiang Zhao, Tao Zhong, Yuchen Pei, Xiangmin Xu, Chaoxiang Yang, He Zhang, et al. A joint brain extraction and image quality assessment framework for fetal brain mri slices. *NeuroImage*, 290:120560, 2024.
16. Yueyue Zhu, Haotian Jiang, Rongqing Cai, and Geng Chen. Multi-label mambaout for quality assessment of low-field pediatric brain mr images. In *MICCAI Challenge on Low Field Pediatric Brain Magnetic Resonance Image Segmentation and Quality Assurance*, pages 3–11. Springer Nature Switzerland Cham, 2024.