EfficientNet-Based Myocarditis Detection in T2-weighted CMR Images: A Patient-Level Evaluation Strategy

Robert G. V. de Souza[1], **Joventino O. Campos**[1], **Luiz M. S. Maciel**[1]

[1]Department of Computer Science
Universidade Federal de Juiz de Fora (UFJF) – Juiz de Fora, MG – Brazil

robert.souza@estudante.ufjf.br, {joventino.campos, luiz.maciel}@ufjf.br

Abstract. *Myocarditis diagnosis via Cardiac Magnetic Resonance is challenged by subtle tissue alterations and observer subjectivity. This study evaluates four EfficientNet architectures for automated classification using exclusively non-contrast T2-weighted images. Ensuring methodological rigor, we implemented a 5-fold cross-validation with strict patient-level separation on 45 participants, applying data augmentation to balance classes. EfficientNetB1 achieved the best performance across 10 independent repetitions, with 96.73% average image-level. Aggregating slice predictions via majority voting yielded 99.56% patient-level accuracy. Results demonstrate a reliable non-invasive diagnostic support tool that simplifies clinical protocols by eliminating contrast agents.*

1. Introduction

Myocarditis is an inflammatory process of the myocardium associated with cellular injury and cardiac dysfunction, potentially evolving into heart failure, severe arrhythmias, dilated cardiomyopathy, and even sudden death [Hatfaludi et al. 2024]. The incidence of the disease has increased significantly, mainly associated with viral infections such as influenza, parvovirus, and, more recently, the COVID-19 pandemic [Łajczak and Jóźwik 2024]. Despite its clinical severity and epidemiological relevance, early diagnosis remains a challenge in medical practice, since electrocardiographic changes are often disparate among patients. The method considered most accurate is endomyocardial biopsy, capable of detecting inflammatory infiltrate and myocyte necrosis [Friedrich et al. 2009]. However, it is an invasive, technically complex procedure associated with risks such as cardiac perforation, tamponade, and thrombus formation. In this context, non-invasive techniques, primarily Cardiac Magnetic Resonance (CMR), have been widely used for diagnosis [Ferreira et al. 2018]. Nevertheless, the interpretation of CMR is still observer-dependent and can be influenced by the radiologist's experience and the subtlety of myocardial alterations, which increases the risk of diagnostic errors.

Given this scenario, deep learning-based models have been proposed as diagnostic aid tools, aiming to reduce the subjectivity of visual analysis [Łajczak and Jóźwik 2024]. Among these models, Convolutional Neural Networks (CNNs) stand out for achieving good results in consolidated medical computer vision applications, such as the detection of schizophrenia from brain magnetic resonance imaging [Oh et al. 2019]. However, their application in myocarditis is still limited by the scarcity of treated, labeled, and publicly available datasets, as well as the difficulty of standardization across image sequences.

Another critical point lies in the data treatment and splitting method. Preliminary studies employing all image series from a dataset suggested excessively optimistic results

[Łajczak and Jóźwik 2024]. However, such metrics may have been heavily inflated by a critical methodological flaw: the absence of a strict patient-level split. While patient isolation between training and testing sets is a fundamental and necessary rigor in medical machine learning, it has been frequently overlooked in previous works utilizing this dataset [Sharifrazi et al. 2022]. This methodological flaw allows the inclusion of images from the same individual in both the training and test sets, leading the network to memorize idiosyncratic anatomical features or acquisition artifacts rather than learning generalizable pathological patterns [Zech et al. 2018, Yagis et al. 2021].

Although this effect may lead to apparently satisfactory metrics, when testing the model on images from patients never seen before, performance tends to drop significantly, revealing low generalization power. In a real clinical scenario, where the goal is to correctly identify new individuals, this methodological bias renders the practical application of the model unfeasible.

In the current landscape, several deep learning approaches have been proposed for myocarditis diagnosis, although methodological heterogeneity among studies hinders direct comparisons. Comprehensive systematic reviews and meta-analyses have demonstrated that AI-augmented cardiovascular magnetic resonance can provide high diagnostic accuracy [Łajczak et al. 2025]. Recent works explore various paradigms, ranging from advanced clustering frameworks [Sharifrazi et al. 2022] to ensemble techniques [Kasmaee et al. 2024]. Other strategies utilize high-throughput radiomics to quantitatively distinguish inflammatory conditions [Baeßler et al. 2019].

More recently, Hatfaludi et al. (2024) proposed a Deep Learning framework with patient separation, achieving high performance using the Few-Shot Learning technique [Hatfaludi et al. 2024]. However, this study presents distinct characteristics that must be considered: the dataset used is severely imbalanced (231 myocarditis cases versus only 38 normal controls), which may influence model generalization to balanced populations. Furthermore, the high accuracy reported by the authors was concentrated on contrast-enhanced sequences (LGE), while performance on T2-weighted images was considerably lower, falling below 70% in some configurations.

Unlike these methodologies that rely on the use of contrast or deal with imbalanced databases, this study proposes an approach focused on the T2-weighted sequence. The objective is to investigate whether, through adequate class balancing, it is possible to extract inflammatory edema biomarkers with high diagnostic precision, dispensing with the need for contrast agents and simplifying the clinical protocol. The choice of this specific sequence is justified by its high sensitivity to myocardial edema, which is the primary pathophysiological marker of the acute phase of inflammation. Unlike LGE, which identifies necrosis and fibrosis, or CINE, which focuses on motor function, T2-weighted allows for the early identification of reversible injury, simplifying the model by focusing on the most-determinant feature of active inflammation.

In this context, the present work proposes a myocarditis classification based on CNNs, using exclusively T2-weighted images. We evaluated model efficacy through a cross-validation strategy with strict patient separation. The proposed pipeline covers everything from data curation to the consolidation of the final patient-level prediction.

The main contributions of this work can be summarized as follows:

- Development of a model based on the EfficientNetB1 architecture optimized for myocarditis classification using exclusively T2-weighted images.
- Implementation of a cross-validation protocol with strict patient-level split and experimental repetition, mitigating data leakage bias and reinforcing the statistical reliability of the results.
- The realization of a patient-level prediction strategy, which significantly reduces classification errors observed in isolated slices, approximating the model's behavior to real clinical analysis protocols.

2. Materials and Methods

The methodology adopted in this study was structured into a sequential pipeline composed of four main stages, as illustrated in Figure 1. The process begins with the curation and preprocessing of raw data from the Z-Alizadeh dataset [Sharifrazi et al. 2022]. In this stage, the dataset was reviewed by specialists, and T2-weighted images were selected.

In the subsequent stage, to mitigate class imbalance, a data augmentation technique restricted to the minority class (normal individuals) was applied, yielding a balanced dataset for training. The training phase consists of a stratified 5-fold cross-validation protocol with strict patient-level separation, ensuring that the model is evaluated on patients never seen during the learning phase.

Finally, we conducted performance evaluation at two levels: image-level and patient-level (combining predictions for the final diagnosis). The following sections detail each stage of the pipeline.

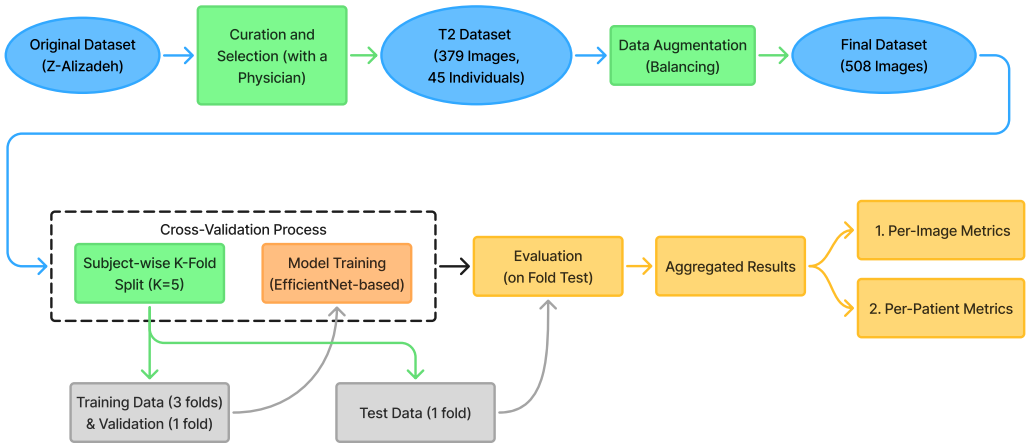


Figura 1. Comprehensive diagram of the experimental pipeline. The stages encompass: (1) Data Curation: Selection of T2-weighted images from the original Z-Alizadeh dataset, which results in 379 images from 45 individuals. (2) Preprocessing: Application of data augmentation techniques to the class of Normals to correct imbalance, totaling 508 images. (3) Cross-Validation: Implementation of a 5-fold protocol with strict patient separation; in each iteration, the EfficientNetB1 model is trained on specific folds and evaluated on a distinct test fold. (4) Evaluation: Performance analysis is conducted at the image-level and patient-level (using the majority voting on the class predicted).

2.1. Dataset

The Z-Alizadeh Sani Myocarditis dataset[1] was utilized, originally comprising 98,898 CMR images from 47 individuals. Initially, the dataset contained 37,564 images from normal controls and 61,334 from myocarditis patients. The acquisition protocol encompassed the following sequences: CINE Segmented (TrueFISP), T2-weighted (TIRM), and LGE high-resolution PSIR [Sharifrazi et al. 2022]. To ensure a comprehensive volumetric assessment, images were acquired covering the standardized Short Axis (SAX) and Long Axis (LAX) anatomical planes, including 2-, 3-, and 4-chamber views. It is noteworthy that, due to acquisition-specificities and field-of-view (FOV) adjustments for each patient, the original images present varying spatial resolutions.

Preliminary studies utilizing the entire image dataset indicated high accuracy values [Sharifrazi et al. 2022]. However, as previously discussed in the introduction, the absence of patient-level separation may have inflated these results. Therefore, for all experiments in this study, we employed a cross-validation strategy with strict patient-level separation. This strategy ensures complete patient isolation between training, validation, and testing sets.

To refine the experimental process, we classified all images manually to correctly associate each series with its corresponding sequence (CINE, LGE, or T2-weighted). In the original distribution, CINE sequences accounted for the majority of data (36,355 normal and 36,090 myocarditis images), followed by LGE (1,080 normal and 24,994 myocarditis images). T2-weighted sequences, our primary focus, comprised only 129 images for the normal class and 250 for the myocarditis class. This analysis was necessary because the available dataset does not provide a clear distinction between the series. During this curation process, two patients from the myocarditis group were excluded due to incomplete T2-weighted series or severe acquisition artifacts, reducing the initial cohort from 47 to 45 individuals. A cardiac imaging cardiologist from the University Hospital of Juiz de Fora reviewed the material and contributed to selecting the most informative sequences, aligning the study with clinical practice priorities.

Although the literature suggests the use of CINE and LGE sequences in myocarditis detection models [Łajczak and Jóźwik 2024], T2-weighted images are relevant for identifying myocardial edema, the primary characteristic of the acute phase of the disease, in individual images [Ferreira et al. 2018]. The decision to use this sequence exclusively aims to reduce model complexity and prevent learning from being biased by features from other sequences (e.g., contractility in CINE or fibrosis in LGE), which are less specific for diagnosing acute inflammation at the image level [Hatfaludi et al. 2024, Friedrich et al. 2009].

Following this analysis and refinement process, the final working set comprised 379 T2-weighted images, sourced from 45 individuals (15 healthy and 30 with myocarditis). A total of 129 images belong to the healthy group, while 250 correspond to the myocarditis patient group.

2.2. Balancing and Data Augmentation

Due to class imbalance, with a significantly lower number of images from healthy individuals, we adopted a data augmentation process applied exclusively to the T2-weighted

[1] https://www.kaggle.com/datasets/danialsharifrazi/myocarditis-dataset

images of this class. The objective was to generate new variations of existing images to balance the dataset and prevent the network from developing a bias toward the majority class.

The transformations were designed based on an analysis of variations in the dataset, ensuring that the artificial augmentations remained consistent with clinical acquisition patterns. The process included:

- **Random Horizontal Flip:** Horizontal mirroring.
- **Random Rotation:** Rotations of up to 15 degrees to simulate slight variations in patient positioning.
- **Random Contrast:** Contrast adjustments (factor of 0.5) to mimic intensity differences between exams.
- **Random Zoom:** Zoom of up to 20% to simulate field-of-view variations.

Each original image from the normal class was expanded into a new version using these transformations, balancing the number of images in the healthy and disease classes. The augmented set was then saved and used in all subsequent experiments.

At the conclusion of this process, the final T2-weighted image dataset consisted of 508 images from 45 distinct individuals, comprising 15 healthy subjects and 30 with confirmed myocarditis. A total of 258 images belong to the healthy group, while 250 belong to the myocarditis patient group.

2.3. Preprocessing and Cross-Validation

All images were resized to 224×224 pixels, and Z-score normalized before being fed into the model. We computed the mean and standard deviation on the training set only and applied them to the training, validation, and test sets.

To ensure a robust and clinically relevant evaluation, we implemented a stratified 5-fold cross-validation protocol with strict patient-level separation. The 45 individuals were divided into 5 mutually exclusive folds, each comprising 9 patients (3 healthy and 6 with myocarditis), as detailed in Table 1. In each iteration of the cross-validation, one fold was designated for testing, one for validation, and the remaining three for training, as illustrated in Table 2. Furthermore, the data augmentation process was performed strictly at the patient level. Since all augmented variations remain exclusively associated with their original patient, this setup inherently guarantees that no augmented images from the test set are exposed during the training phase, completely preventing data leakage. This approach ensures that the model is trained, tuned, and evaluated on completely distinct patient sets, applying a necessary methodological rigor that was lacking in prior literature, thereby providing a more accurate estimate of its generalization capability.

2.4. Architecture and Training

For the classification task, an initial exploratory phase evaluated several consolidated Convolutional Neural Network architectures, including ResNet, VGG, and DenseNet. However, recent analytical studies highlight that while ResNet remains a robust baseline, the EfficientNet family consistently achieves superior accuracy-to-computation ratios in constrained medical tasks [Kailani 2025]. Consequently, experiments focused on variants B0 to B3. This architecture is distinguished by its compound scaling mechanism [Tan and Le 2019], which optimally and simultaneously adjusts the network's depth,

Tabela 1. Stratification of individuals and number of CMR images across the five folds of the cross-validation protocol, detailing the distribution between normal and myocarditis patients.

Fold	Individuals (IDs)		Number of images		
	Normal	Myocarditis	Normal	Myocarditis	Total
Fold 1	1, 2, 3	17, 18, 19, 20, 22, 23	52	47	99
Fold 2	4, 5, 6	24, 25, 26, 27, 28, 29	48	53	101
Fold 3	7, 8, 9	30, 31, 32, 33, 34, 35	58	53	111
Fold 4	10, 11, 12	36, 37, 38, 39, 40, 41	52	50	102
Fold 5	13, 14, 15	42, 43, 44, 45, 46, 47	48	47	95

Tabela 2. Experimental design of the 5-fold cross-validation protocol with strict patient separation, detailing the assignment of data splits and the resulting distribution of samples across training, validation, and testing sets.

Split	Set	Folds	Number of individuals		Number of images (Samples)	
			Normal	Myocarditis	Normal	Myocarditis
Split 1	Test	Fold 1	3	6	52	47
	Validation	Fold 2	3	6	48	53
	Training	Folds 3, 4, 5	9	18	158	150
Split 2	Test	Fold 2	3	6	48	53
	Validation	Fold 3	3	6	58	53
	Training	Folds 4, 5, 1	9	18	152	144
Split 3	Test	Fold 3	3	6	58	53
	Validation	Fold 4	3	6	52	50
	Training	Folds 5, 1, 2	9	18	148	147
Split 4	Test	Fold 4	3	6	52	50
	Validation	Fold 5	3	6	48	47
	Training	Folds 1, 2, 3	9	18	158	153
Split 5	Test	Fold 5	3	6	48	47
	Validation	Fold 1	3	6	52	47
	Training	Folds 2, 3, 4	9	18	158	150

width, and resolution. Among the evaluated architectures, EfficientNetB1 presented the most satisfactory balance between performance and computational cost in preliminary experiments and was, therefore, selected for conducting the detailed final analysis in this work.

The models were trained using transfer learning, leveraging the weights pre-trained on the ImageNet dataset. The final architecture was customized and trained as follows:

- Convolutional Base: EfficientNet was utilized as the base feature extractor. A full fine-tuning approach was adopted, in which all layers of the pre-trained base model were kept unfrozen. Consequently, all weights were continuously updated

throughout the training process, allowing the network's prior knowledge to be directly adapted to the specific features of the T2-weighted CMR images.

- Classification Layers: Following the convolutional base, a custom classification head was added, with all its weights randomly initialized prior to training. This module comprises: a Flatten layer to convert the feature maps into a one-dimensional vector; a Dense layer with 100 neurons and ReLU activation for learning complex non-linear patterns; a Dropout layer with a rate of 0.5 for regularization, preventing overfitting; and a Dense output layer with Softmax activation for the final binary prediction.

Figure 2 shows the schematic representation of the adapted EfficientNetB1 used in this work. We modified the other EfficientNet versions similarly.

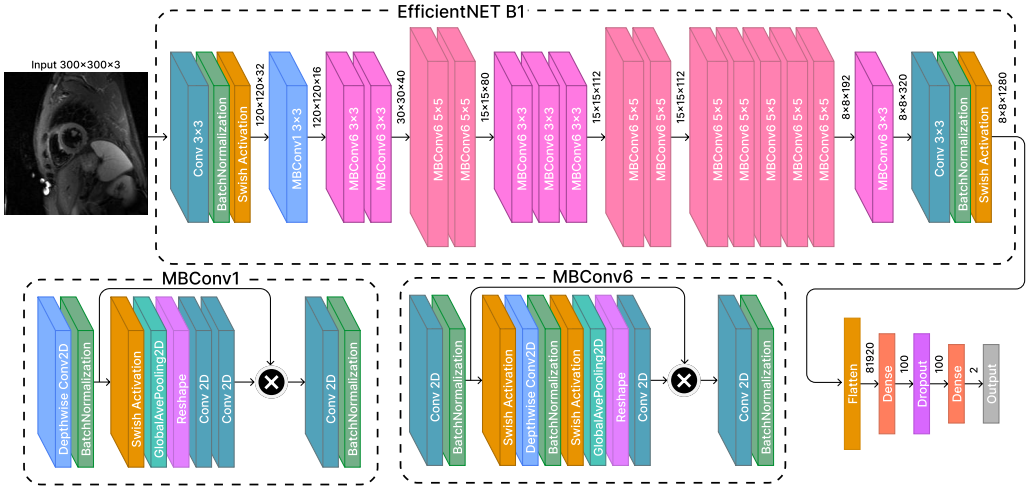


Figura 2. Schematic representation of the proposed architecture. The pipeline integrates a pre-trained EfficientNetB1 as the feature extractor (Convolutional Base) with a custom classification head. The classification block comprises a flattening layer, a dense layer (100 neurons, ReLU activation), a dropout layer (rate 0.5) for regularization, and a final Softmax layer for binary prediction (normal or myocarditis).

The AdamW optimizer [Loshchilov and Hutter 2019] was employed for model training; this variant of Adam enhances regularization by decoupling weight decay from gradient updates. A weight decay factor of 0.004 was adopted. Categorical cross-entropy was selected as the loss function, and performance was monitored using the categorical accuracy metric.

Training was executed for each of the 5 splits over 30 epochs, with a batch size of 32. At each epoch, we evaluated model performance on the validation set and saved the model version with the highest validation accuracy, ensuring that the final model for each split reached optimal performance on the validation set.

2.5. Evaluation Metrics and Prediction Strategy

Model performance was evaluated at two distinct and complementary levels: *image-level* and *patient-level*. To quantify performance, standard metrics including accuracy, precision, recall (sensitivity), and F1-score were utilized, all derived from the confusion matrix

of the complete dataset. The matrix terms are defined as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Accuracy reflects the overall effectiveness of the classifier, indicating the probability of the model making a correct prediction, regardless of the class:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{1}$$

Precision evaluates the reliability of positive predictions. This metric is relevant for measuring the model's ability to distinguish pathology without generating excessive false positives:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2}$$

Recall quantifies detection completeness. In diagnostic systems, its relevance lies in the ability to identify the majority of existing positive cases, minimizing the occurrence of undiagnosed cases (false negatives):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3}$$

F1-Score is the harmonic mean of Precision and Recall. This metric is used to assess model stability and is fundamental for comparing performance when there is an imbalance between precision and sensitivity:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

In this work, besides the image-level prediction, we propose to analyze the patient-level prediction strategy. This method simulates the real diagnostic process, in which a specialist analyzes multiple images to form a consolidated opinion. The final prediction for each patient was obtained by majority voting on the class predicted for each T2-weighted slice of the individual; the class assigned to the most images was designated as the final diagnosis. While the image-level analysis validates the network's technical capability, the patient-level evaluation best reflects its potential for clinical application.

2.6. Computational Environment

All computational experiments, including data preprocessing and neural network training, were conducted using high-performance cloud computing resources provided by the Kaggle platform. Leveraging this environment accelerated the training process by using a dedicated GPU.

The hardware configuration consisted of an NVIDIA Tesla T4 GPU with 16GB of VRAM, paired with an Intel(R) Xeon(R) CPU @ 2.00GHz and 13GB of RAM. The underlying operating system was a Linux distribution (containerized Docker environment).

The algorithms were implemented in the Python programming language (version 3.10), using TensorFlow (version 2.15.0) as the primary framework in conjunction with the Keras API for constructing, compiling, and training the EfficientNet architectures.

3. Results

The experiments were designed to evaluate and compare the performance of the EfficientNet architectures (variants B0 to B3) in classifying myocarditis images. To ensure result consistency and mitigate the stochastic variability inherent to neural network training, we repeated the entire 5-fold cross-validation evaluation 10 times for each architecture.

A summary of the image-level performance is presented in Table 3, consolidating the results of the 10 execution runs for each model. The EfficientNetB1 architecture demonstrated superior, more stable performance, achieving the highest average accuracy (96.73%) with the lowest standard deviation (0.01) across runs and was therefore selected as the final model for this study.

Tabela 3. Comparative performance analysis of the EfficientNet architectures (B0–B3). The table presents the classification accuracy (mean ± standard deviation across 10 runs) by evaluation level (image-wise and individual-wise) and diagnostic class (overall, normal, and myocarditis).

Model	Average Accuracy per Image			Average Accuracy per Individual		
	Overall	Normal	Myocarditis	Overall	Normal	Myocarditis
EfficientNet B0	94.76% ± 0.01	91.31% ± 0.02	99.09% ± 0.01	95.57% ± 0.02	100% ± 0.00	88.53% ± 0.05
EfficientNet B1	96.73% ± 0.01	97.17% ± 0.02	96.32% ± 0.01	99.56% ± 0.01	100% ± 0.00	98.75% ± 0.03
EfficientNet B2	96.04% ± 0.01	94.46% ± 0.02	97.84% ± 0.01	97.56% ± 0.01	100% ± 0.00	93.35% ± 0.04
EfficientNet B3	96.04% ± 0.01	93.75% ± 0.02	98.74% ± 0.01	98.44% ± 0.01	100% ± 0.00	95.70% ± 0.04

Table 3 shows that although EfficientNetB0 achieved the highest accuracy for the myocarditis class (99.09%), its performance on the normal class was lower (91.31%), suggesting a bias that may increase the false-positive rate. In contrast, EfficientNetB1 exhibited the most balanced behavior, leading in both overall accuracy and the normal class (97.17%), while maintaining a high accuracy rate for myocarditis (96.32%).

The most significant result was observed in the patient-level evaluation. By applying the prediction aggregation strategy, the EfficientNetB1 model achieved high performance, with an average accuracy of 99.56% in classifying the 45 individuals. It is worth highlighting a specific behavior observed at this level: the model obtained 100% accuracy in identifying normal individuals across all execution runs. Consequently, all residual errors were concentrated on the myocarditis class (False Negatives).

However, it is imperative to contextualize the 98.75% accuracy metric observed for the myocarditis class. Given the dataset contains 45 individuals, the misclassification of a single patient in a specific run – averaged across 10 repetitions – disproportionately penalizes the final numeric score. Therefore, this slight percentage reduction reflects the high statistical sensitivity of the accuracy metric in small cohorts regarding isolated errors, rather than a systemic inability of the model to detect the pathology.

Beyond accuracy, an in-depth analysis of the complementary metrics, presented in Table 4, confirms the efficacy of the selected architecture. While models such as EfficientNetB0, B2, and B3 prioritized Precision at the expense of Recall in image-level analysis – which could lead to a higher false-negative rate – EfficientNetB1 exhibited the best balance between these metrics. Notably, it achieved a Recall of 97.17% per image and 98.75% per individual, indicating superior sensitivity in disease detection. Consequently, EfficientNetB1 achieved the highest F1-Score at both evaluation levels (96.77% per

image and 99.35% per patient), validating its superior consistency in balancing correctly identifying positive cases and minimizing false alarms.

Tabela 4. Comparative performance metrics across the EfficientNet architectures (B0–B3). The table presents the classification Recall, Precision, and F1-Score, presented as mean percentages ± standard deviation over 10 independent runs by evaluation level (image-wise vs. individual-wise).

Model	Metrics per Image			Metrics per Individual		
	Recall	Precision	F1-Score	Recall	Precision	F1-Score
EfficientNet B0	91.31% ± 0.02	99.19% ± 0.01	95.07% ± 0.01	88.53% ± 0.05	100.00% ± 0.00	93.84% ± 0.03
EfficientNet B1	97.17% ± 0.02	96.40% ± 0.01	96.77% ± 0.01	98.75% ± 0.03	100.00% ± 0.00	99.35% ± 0.01
EfficientNet B2	94.46% ± 0.02	97.98% ± 0.01	96.18% ± 0.01	93.35% ± 0.04	100.00% ± 0.00	96.51% ± 0.02
EfficientNet B3	93.75% ± 0.02	98.84% ± 0.01	96.22% ± 0.01	95.70% ± 0.04	100.00% ± 0.00	97.76% ± 0.02

4. Discussion

This study demonstrated the capability of a deep learning model, based on the EfficientNetB1 architecture, for classifying myocarditis in T2-weighted CMR images. The average accuracy of 96.73% in image-level classification and 99.56% at the patient level represent a significant outcome, suggesting a high diagnostic potential for the proposed approach. Beyond the raw metric, the selection of the B1 variant was justified by its inter-run stability and class balance. Unlike the B0 variant, which demonstrated a bias towards the positive class, thereby impairing the identification of healthy individuals (91.31% accuracy), EfficientNetB1 maintained high performance in both leves: image-level and patient-level, validating its consistency for clinical application.

The analysis of the results in comparison with recent literature highlights the potential of the proposed approach. While systematic reviews indicate that deep learning models for myocarditis frequently reach accuracies between 91% and 95% [Łajczak et al. 2025], our model achieved 99.56% in the patient-level evaluation. However, it is important to note that methodological differences and database compositions prevent a direct comparison.

When contextualizing with the study by Hatfaludi et al. (2024), who also adopted patient-level separation, we observe that the authors reported a weighted accuracy of 96.9%. However, they obtained this result primarily LGE images. When evaluated specifically on the T2-weighted sequence, the methodology of Hatfaludi et al. yielded more modest metrics, with accuracy ranging from 67% to 72% across the anatomical views [Hatfaludi et al. 2024]. Although the cited study used a larger cohort, the marked imbalance between classes (approximately 6:1 in favor of the pathology) imposes challenges distinct from those faced in this work, which used balanced classes. The superiority of our performance in the T2-weighted modality suggests that the balancing strategy, combined with the EfficientNetB1 architecture, was effective in identifying subtle edema patterns without contrast agents.

An error analysis reveals an asymmetric behavior of the model at the patient level: the identification of healthy individuals was absolute (100% specificity). Consequently, the rare residual errors were concentrated on the non-detection of myocarditis cases (false negatives). It is important to acknowledge that the 100% specificity may be partially attributed to the data augmentation process. Because artificial transformations were applied

prior to fold splitting to correct class imbalance, augmented variations of the normal class were evaluated within the test folds, potentially artificially increasing the model's confidence on synthetic features. Although maximum sensitivity is ideal in screening tools, it is fundamental to interpret these data in light of the sample size: in a group of only 45 patients, the misclassification of a single diseased individual in a specific execution, among the 10 repetitions performed, abruptly impacts the overall accuracy. Therefore, the slight fluctuation observed in the metric reflects more the statistical sensitivity of small cohorts to unitary errors than a systemic inability of the model to detect the pathology.

This study presents some limitations that should be considered when interpreting the results. First, the dataset is derived from a single center and comprises a small number of individuals (45 patients), which reflects the current scarcity of publicly available, well-annotated CMR datasets specifically focused on myocarditis. To mitigate this constraint, we adopted a strict patient-level cross-validation protocol combined with multiple experimental repetitions, aiming to provide a more reliable estimate of model performance.

While the patient-level accuracy of 99.56% is promising, further validation on independent external multi-center datasets is necessary to confirm the generalization capability of the proposed approach. Moreover, the experimental comparison could be expanded by incorporating more recent architectures, such as Vision Transformers or ConvNeXt-based models, to better position the approach against current state-of-the-art methods. Finally, future investigations should integrate explainability analyses (e.g., Grad-CAM) to provide additional evidence that the network bases its predictions on clinically meaningful myocardial regions, thereby increasing confidence in its clinical applicability.

5. Conclusion

This work proposes and validates a methodology to diagnose myocarditis using a deep learning model based on the EfficientNetB1 architecture, relying exclusively on T2-weighted cardiac magnetic resonance images. By employing patient-level cross-validation and multiple execution runs to mitigate stochastic variability, we demonstrated that the approach is highly effective and promising.

The model achieved remarkable performance, culminating in an accuracy of 96.73% for image-level classification and an accuracy of 99.56% for patient-level classification. These results highlight the method's potential as a diagnostic aid that reduces the subjectivity of human visual interpretation and accelerates clinical decision-making, while employing a simpler, non-invasive imaging protocol that does not require contrast agents.

Despite the achieved success, validation on larger, independent external datasets remains necessary to confirm the model's generalization in real-world clinical settings. Future work should focus on expanding the method's capabilities toward multi-class classification, distinguishing myocarditis from other cardiac pathologies. Additionally, exploring contemporary architectures like Vision Transformers and incorporating explainability techniques, such as Grad-CAM, will be essential steps to enhance both the comparative robustness and clinical transparency of the proposed tool.

Referências

Baeßler, B. et al. (2019). Radiomics for distinguishing myocardial infarction from myocarditis at late gadolinium enhancement at mri: Comparison with subjective visual analysis. *Radiology: Cardiothoracic Imaging*, 1(3):e180026.

Ferreira, V. M. et al. (2018). Cardiovascular magnetic resonance in nonischemic myocardial inflammation: expert recommendations. *Journal of the American College of Cardiology*, 72(24):3158–3176.

Friedrich, M. G. et al. (2009). Cardiovascular magnetic resonance in myocarditis: A jacc white paper. *Journal of the American College of Cardiology*, 53(17):1475–1487.

Hatfaludi, C.-A. et al. (2024). Deep learning automatically distinguishes myocarditis patients from normal subjects based on mri. *The International Journal of Cardiovascular Imaging*, 40(12):2617–2629.

Kailani, N. M. S. (2025). Comparison between resnet and efficientnet for image classification: An analytical study of performance and efficiency. *North African Journal of Scientific Publishing (NAJSP)*, 3(4):18–25.

Kasmaee, A. M. M., Ataei, A., Moravvej, S. V., Alizadehsani, R., Gorriz, J. M., Zhang, Y.-D., Tan, R.-S., and Acharya, U. R. (2024). Elrl-md: a deep learning approach for myocarditis diagnosis using cardiac magnetic resonance images with ensemble and reinforcement learning integration. *Physiological Measurement*, 45(5):055002.

Łajczak, P. M. and Jóźwik, K. (2024). Artificial intelligence and myocarditis: a systematic review of current applications. *Heart Failure Reviews*, 29(6):1217–1234.

Łajczak, P. M., Sahin, O. K., Matyja, J., et al. (2025). Machine learning for myocarditis diagnosis using cardiovascular magnetic resonance: a systematic review, diagnostic test accuracy meta-analysis, and comparison with human physicians. *The International Journal of Cardiovascular Imaging*.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations*.

Oh, K., Chung, Y.-C., Kim, K., Kim, W., and Oh, I. (2019). Deep learning for schizophrenia classification. *IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 113–116.

Sharifrazi, D., Alizadehsani, R., Joloudari, J. H., Band, S. S., Hussain, S., Sani, Z. A., Hasanzadeh, F., Shoeibi, A., Dehzangi, A., Sookhak, M., et al. (2022). Cnn-kcl: Automatic myocarditis diagnosis using convolutional neural network combined with k-means clustering. *Mathematical Biosciences and Engineering*, 19(3):2381–2402.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114.

Yagis, E., Atnafu, S. W., and Garcia Seco de Herrera, A. (2021). Effect of data leakage in brain mri classification using 2d convolutional neural networks. *Scientific Reports*, 11(1):22544.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683.