Assessing the Value of Explainable Artificial Intelligence for Magnetic Resonance Imaging

Giada Frasson[1], Matteo Rizzo[1(✉)], Marco Salvatore Nobile[1], Amalia Lupi[2], and Emilio Quaia[2]

[1] Ca' Foscari University of Venice, Venice, Italy
869359@stud.unive.it, {matteo.rizzo,marco.nobile}@unive.it
[2] University of Padua, Padua, Italy
{amalia.lupi,emilio.quaia}@unipd.it

Abstract. Recent advancements in Artificial Intelligence (AI) have often improved the accuracy of medical diagnostics in several fields, such as cancer detection and diagnosing cardiovascular or neuromuscular diseases. However, the opaque nature of AI decision-making can limit its adoption in the clinical setting, as physicians require clear and interpretable explanations to trust these tools. To address this issue, the field of eXplainable Artificial Intelligence (XAI) aims to clarify the rationale behind AI predictions while ensuring compliance with ethical standards and advanced regulations such as the GDPR and the AI Act. This study applies multiple explainability methods to a diagnostic support model for Distal Myopathies (DMs), a rare neuromuscular disorder marked by subtle, early-stage tissue alterations. Beyond classification, our approach generates detailed explanations for the model's predictions. We propose novel techniques, including a hierarchical occlusion method and an ensemble framework that combines individual explanations to produce refined, interpretable visualizations. Feedback from expert radiologists is used to assess the effectiveness of these methods, highlighting their potential to enhance trust and usability in clinical practice. Our results show that pretrained convolutional networks achieve high classification accuracy, exceeding 88%, with perfect recall in identifying affected cases, while underscoring the need for adaptive and user-centric approaches to explainability in AI-driven diagnostic tools.

Keywords: Deep learning (DL) · eXplainable Artificial Intelligence (XAI) · Magnetic Resonance Imaging (MRI) · Neuromuscular Disorders (NMD) · Distal Myopathies

1 Introduction

Artificial Intelligence (AI) has revolutionized medicine in recent years, driving significant advancements across various domains, from drug design and discovery [14,26] to clinical decision support [6,23]. In particular, AI has demonstrated

G. Frasson and M. Rizzo—Equal contribution.

R. Guidotti et al. (Eds.): xAI 2025, CCIS 2576, pp. 423–447, 2026.
https://doi.org/10.1007/978-3-032-08317-3_20

remarkable potential as a decision-support tool in medical diagnostics [15,16]. A study by McKinney et al. [11] revealed that an AI system for breast cancer diagnosis used to interpret mammograms reduced false positives and false negatives by 5.7% and 9.4%, respectively. Multiple studies observed that AI systems can outperform human experts in some specific diagnostic tasks, enhancing physicians' capabilities through AI-assisted analysis. For instance, Kim et al. [7] showed an AI system with higher sensitivity in diagnosing breast cancer compared to radiologists, effectively identifying early-stage cases. Similarly, Haenssle et al. [5] showed that their AI model achieved superior diagnostic performance for melanoma cases, outperforming most, though not all, dermatologists involved in the study.

Despite these promising results, widespread adoption of AI in clinical practice is hindered by a critical challenge: physicians are unlikely to trust an algorithm's decision without a clear understanding of its reasoning process. To address this issue, the field of eXplainable Artificial Intelligence (XAI) has emerged, aiming to enhance the interpretability of AI models. XAI provides insights into AI decision-making, making models more transparent and comprehensible to human users. This topic is particularly sensitive in medicine, where ethical considerations and regulatory frameworks necessitate accountability and fairness. For example, the European Union's General Data Protection Regulation (GDPR, Article 15) and AI Act grant patients the right to understand how and why decisions affecting them are made. A comprehensive review by Van der Velden et al. [24] discusses various explainability methods applied to medical imaging across different anatomical regions, emphasizing the growing importance of XAI in healthcare. Our study focuses on deep-learning-based analysis of Magnetic Resonance Imaging (MRI) scans to diagnose Distal Myopathies (DMs), a rare Neuromuscular Disease (NMD). Radiological diagnosis of this condition requires significant expertise, as early-stage cases often exhibit subtle tissue alterations that can be challenging for less experienced observers to detect. AI systems can assist radiologists by identifying these patterns and providing supporting evidence for their predictions. Our objective is to move beyond classification and generate explanations that clarify the rationale behind the model's decisions and investigate the effectiveness of such explanations. However, the existing literature presents several gaps: (1) current explainability methods for MRI-based diagnosis have primarily focused on common diseases with large datasets, with limited attention given to rare conditions such as DMs; (2) existing saliency-based methods often generate noisy, low-resolution explanations that are difficult for clinicians to interpret; and (3) few studies have evaluated the practical clinical relevance of XAI outputs through direct user studies with radiologists. To fill these gaps, we introduce two novel explainability techniques tailored for MRI-based diagnosis of rare neuromuscular disorders: a hierarchical occlusion method and an ensemble explainability strategy. The hierarchical occlusion provides a multiscale view of regional importance by systematically masking image patches at various resolutions to improve the localization and clarity of the model's attention. The ensemble explainability strategy aggregates multiple explanation maps (e.g.,

GradCAM methods) to produce more robust and stable outputs that reduce artifacts and enhance interpretability. We benchmark these approaches against state-of-the-art methods and conduct a user study with expert radiologists to validate the clinical utility of the resulting explanations. Their feedback assesses the AI-generated explanations' trustworthiness, interpretability, and usability, providing critical insights into their potential adoption in real-world medical practice. The following research questions drive our study:

- **RQ1:** How accurately can Deep Learning (DL) models classify MRI scans for DMs, and what factors influence their misclassification?
- **RQ2:** How do expert radiologists perceive AI-generated explanations' interpretability and clinical relevance?
- **RQ3:** How does the radiologist's experience impact the interpretation and trust in explainability techniques?
- **RQ4:** What improvements are needed to enhance AI explainability for clinical adoption?

Our code is publicly available: https://github.com/matteo-rizzo/xai-for-mri.

2 Related Work

NMDs comprise a vast and heterogeneous group of pathologies affecting muscles and the nerves that control them [12]. These conditions manifest in childhood and adulthood and present significant diagnostic challenges due to their variable clinical features. Diagnosis involves an evaluation of patient history and symptoms, supplemented by instrumental examinations such as electromyography, muscle imaging, genetic analyses, and muscle biopsy.

Recent research has explored the potential of AI to improve diagnostic accuracy for NMDs. Pineros et al. [17] and related work on muscle MRI [19] underscore the utility of AI in this domain. Verdù-Dìaz et al. [25] analyzed patterns of muscle fatty replacement in T1-weighted MRIs of 976 pelvic and lower limb scans quantifying fatty infiltration with the Mercuri score and applying a Random Forest classifier to achieve an accuracy of 95.7% compared to experts. Yang et al. [27] developed a model for differentiating dystrophinopathies from other muscular diseases using 432 thigh-focused MRI cases, with the ResNet50 architecture achieving 91% accuracy, surpassing expert diagnoses ranging between 80% and 84%. Complementary studies include Felisaz et al. [4], who compared multiple Machine Learning (ML) models for predicting fat fraction and muscle water T2 from MRI texture analysis, and Fabry et al. [3], who employed a 1-Lipschitz neural network on whole-body MRI examinations to distinguish facioscapulohumeral dystrophy from myositis with accuracies between 69% and 77%. While these studies demonstrate the promise of AI in diagnosing NMDs, they also highlight a critical gap: the explainability of model predictions. Only a few works, notably Yang et al. [27], have integrated explainability techniques into their models. This gap motivates the need to explore XAI methods in rare conditions such as DM systematically. XAI comprises a range of techniques and

methodologies to interpret complex models' decision-making processes, particularly deep neural networks that are often regarded as opaque *black boxes*. Despite their impressive predictive performance, these models lack transparency, hindering clinical adoption. The field of XAI remains underdeveloped, lacking a universally accepted taxonomy a situation partly attributed to divergent definitions of *interpretability* and *explainability* [20,21,28]. In this work, we adopt the taxonomy proposed by Linardatos et al. [9], which categorizes methods based on dimensions such as model specificity (model-specific vs model-agnostic) and explanation scope (local vs global).

Among the well-established approaches for XAI, Class Activation Maps (CAMs) and their extensions have received significant attention. CAMs, introduced by Zhou et al. [30], are post-hoc, local, and model-specific techniques that visualize the discriminative regions used by a CNN for prediction. By performing global average pooling on the final convolutional feature maps and projecting the resulting weights back onto these maps, CAM highlights the regions most influential to the final decision. However, this method is limited to specific network architectures and only provides explanations from the last convolutional layer. To address these limitations, GradCAM [22] was developed. GradCAM extends CAM by incorporating the gradients of the class score concerning feature maps from any convolutional layer, thereby generating a class-discriminative localization map via global averaging of these gradients. Nonetheless, GradCAM may struggle to localize multiple instances of an object within an image accurately. GradCAM++ [1] refines this approach by computing a weighted average of the pixel gradients, while HiResCAM [2] further improves explanation fidelity by highlighting only the regions actively contributing to the class score. Comparative analyses indicate that GradCAM tends to produce broader explanations, whereas GradCAM++ and HiResCAM, mainly when applied to architectures such as ResNet50v, may occasionally highlight extraneous background regions though HiResCAM consistently yields more focused and detailed explanations.

In a contrasting paradigm, SHapley Additive Explanations (SHAP) [10] adopts a game-theoretic perspective to assign an importance value to each input feature based on its marginal contribution to a prediction. As a post-hoc, model-agnostic method, SHAP approximates complex models with an additive explanation model that satisfies properties such as local accuracy, missingness, and consistency. Despite the computational challenges inherent in exact Shapley value computation, practical approximations have rendered SHAP a powerful tool for both local and global interpretability.

Another noteworthy approach within XAI is occlusion, a sensitivity analysis methodology that evaluates the impact of masking specific input regions on model predictions. Initially introduced by Zeiler et al. [29], occlusion systematically masks parts of an input image using a sliding window, thereby identifying regions whose absence leads to a significant decrease in prediction confidence or a change in classification outcome. Although conceptually straightforward, occlusion is computationally intensive, requiring multiple forward passes through the network. Thus, careful selection of parameters such as window size, stride,

and occlusion value is crucial; larger windows reduce computational cost at the expense of granularity, while smaller windows offer finer resolution but require more computations. Occlusion is particularly relevant to our work as we propose a novel occlusion algorithm that works at multiple levels of granularity.

Collectively, these approaches from CAM-based visualizations to SHAP and occlusion methods provide a robust foundation for interpreting the decisions of complex DL models. These particular strategies are the base for our ensemble method.

3 The Use Case: Distal Myopathies

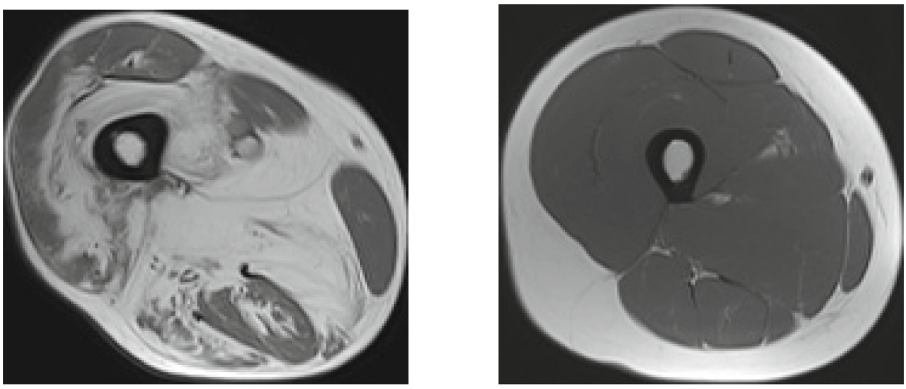
In the context of NMDs, the utility of muscle MRI has already been assessed in the diagnostic work-up and in monitoring the progression of muscle involvement. In fact, given the rarity of these disorders, a thorough clinical, histological, and imaging investigation should be carried out since the clinical heterogeneity and broad genetic spectrum of these conditions often make it difficult to reach a defined molecular diagnosis. In this scenario, muscle MRI proves helpful in identifying different patterns of muscle involvement. Nonetheless, such clinically and genetically heterogeneous conditions require knowledge about radiological characteristics based on distinct genetic mutations to improve diagnostic accuracy. Specific patterns are most recognizable in patients presenting mild phenotypes, with individual muscles selectively affected. In contrast, extensive and severe muscle involvement and very mild and initial involvement do not allow for clear pattern detection, even if an expert radiologist can identify them. For this purpose, different AI approaches could be implemented to improve diagnostic performance, and their explainability will be discussed in this work.

3.1 Dataset

Our proprietary dataset comprises 529 T1-weighted MR images of the lower limbs, capturing each patient's right and left sides. It contains seven patients affected by DM and six healthy controls. To augment the dataset, each image is divided into two separate images corresponding to the left and right lower limbs. Although this division could introduce bias, mainly when the disease affects only one side, consultation with an experienced radiologist confirmed that the benefits of a larger dataset outweigh this risk. An example of an affected lower limb is shown in Fig. 1a, while an example of a healthy lower limb is presented in Fig. 1b. A side-by-side comparison of both cases can be seen in Fig. 1.

3.2 Preprocessing

Our preprocessing pipeline is illustrated in Fig. 2. Central to the pipeline is a cropping algorithm designed to extract the minor crop that encloses the patient's body from each MRI. Initially, the algorithm enhances the image contrast using Contrast Limited Adaptive Histogram Equalization (CLAHE) [18] to address



(a) (b)

Fig. 1. Comparison of affected (a) and healthy (b) lower limb MRIs.

the uneven brightness and contrast inherent in the original MRIs. The enhanced image is then binarized by applying a threshold set at the mean intensity, isolating the image's significant regions. Despite producing a clear binary outline of the patient's body, this process can introduce internal holes and noise. To resolve these issues, the outer boundaries of the white regions are detected, with smaller contours filtered out in favor of retaining the two largest contours, which typically correspond to the pelvis and the legs. Subsequently, the rough contours are refined by computing their convex hulls, yielding smoother and more accurate boundaries. The background, which in MRIs often appears as shades of dark gray rather than pure black, is removed by multiplying the original image by its binary mask, thus preventing any confusion between background and anatomical structures. Finally, the smallest bounding box enclosing the refined contours is determined using OpenCV's `boundingRect`, and the image is cropped accordingly. Another significant challenge is the heterogeneity in image dimensions, as the scans range from pelvises to calves. Since the model requires fixed-size inputs of 224×224 pixels, directly resizing the images is not viable because it could introduce artifacts and distort anatomical proportions. To address this, expert guidance was followed in splitting pelvis images into left and right sections, given that the central pelvis area primarily contains organs rather than muscles. Images smaller than 224×224 pixels are padded to achieve the desired dimensions, while those larger than 224×224 pixels are segmented into 224×224 tiles. If an image exceeds the required size in one dimension only, it is divided into two tiles; if it exceeds in both dimensions, it is partitioned into four tiles. This approach minimizes the number of generated tiles, thereby reducing the risk of bias from mislabeling healthy regions in patients with the disease. The significant overlap between the tiles further ensures that critical border features are preserved. Table 1 displays the dataset structure after completing these preprocessing operations.

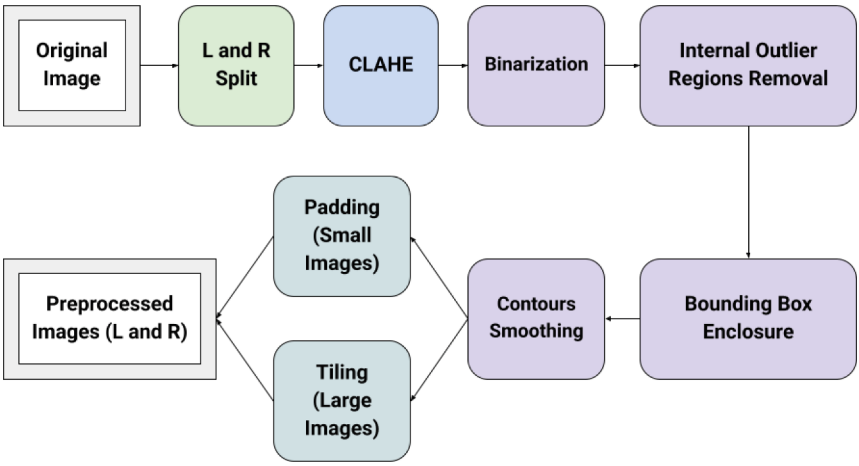


Fig. 2. Workflow of the preprocessing pipeline.

Table 1. Dataset structure after preprocessing.

	Before preprocessing	After preprocessing	Number of tiles
Healthy	202	404	438
Affected	327	654	969
Total	**529**	**1058**	**1407**

4 Deep Learning Models

Due to the limited size of the dataset, a transfer learning approach was adopted. ResNet50 was chosen based on its strong performance in a similar application reported in [27]. Although both studies address a binary classification task, the previous work focused on differentiating between two diseases, whereas our objective is to distinguish between healthy and affected individuals. ResNet18 the lightest Residual network available in PyTorch was also experimented with. Given the dataset's size, this choice maintained consistency within the model family while reducing complexity and mitigating the risk of overfitting.

Several modifications were made to adapt the pre-trained models to our specific task. First, the input layer was adjusted because the PyTorch pre-trained models are designed for 3-channel images, while our MRI images are in grayscale (1-channel). To incorporate the pre-trained weights appropriately, they were summed across the channels, following the intuition that for an RGB image with equal channel values, $R \cdot w_0 + G \cdot w_1 + B \cdot w_2$ simplifies to $R \cdot (w_0 + w_1 + w_2)$. Next, the global average pooling layer was removed from the network architecture. This change was necessary because techniques such as GradCAM and HiResCAM converge to CAM when applied to networks without global average pooling on the last convolutional feature maps. Accordingly, the architectures described in

[2] were modified by replacing the global average pooling with an additional convolutional layer. Finally, the output classifier was replaced. Since the original pre-trained models were configured to predict 1,000 classes (as trained on ImageNet), the final layer was substituted to enable binary classification. The resulting modified models are ResNet18v and ResNet50v, with "v" denoting the variant.

Before training, the dataset was partitioned into training and test sets using a patient-based split to prevent data leakage. This strategy ensured that the data from a single patient did not appear in both sets, thereby preserving the integrity of the evaluation process. However, given the limited number of patients, the test set contained only one individual per class. To address the potential sensitivity of the performance metrics to this selection, an experienced radiologist recommended including a healthy patient with a higher body fat percentage in the test set, thereby challenging the model and providing a conservative lower bound for evaluation.

Due to the dataset imbalance, ROSE oversampling was applied to the training set to equalize the class instances [13]. Furthermore, the feature extraction layers were frozen, and only the network's classifier layer was trained using a cross-validation framework with early stopping to reduce the risk of overfitting. Since the number of patients was limited, it was not feasible to reserve a separate validation set with one healthy and one affected individual; instead, each patient was treated as a separate fold in the cross-validation process. During training, data augmentation techniques provided by PyTorch such as random brightness and contrast modifications were applied to artificially increase the size and diversity of the training dataset. These augmentations, chosen in consultation with a domain expert, were designed to reflect the natural variations typically encountered in MRI images, thereby improving the model's generalization capability.

5 Novel XAI Methods

5.1 Hierarchical Occlusion

As previously described, Occlusion is a practical yet computationally intensive sensitivity analysis technique, particularly when a high level of detail is desired. Moreover, selecting appropriate parameters such as window size, stride, and occlusion value is nontrivial, as these parameters may require extensive tuning and may not be universally optimal for all inputs. To address these challenges, a hierarchical occlusion algorithm was developed. The central idea is to start with relatively large occlusion windows and progressively reduce their size, thereby achieving a balance between computational cost and the desired granularity

of the analysis. The initial concept involved leveraging an existing occlusion function, such as the `Occlusion` class provided by Captum [8]. However, this implementation exhibited two significant limitations. First, it does not permit the selection of an alternative metric, as it defaults to using the difference in the model's output. Second, it lacks the flexibility to perform occlusion on a designated subarea of the image, which is a requirement for the hierarchical approach. Consequently, a custom design and implementation were pursued.

The proposed hierarchical occlusion algorithm performs occlusion at multiple levels of granularity. Still, it restricts the refinement process to those windows that, at a coarser level, induce a change in the network's output. Several strategies were explored in the development process. Inspired by Captum, an initial approach employed the difference in the model's raw output before and after occlusion as the metric. However, combining results across different levels proved problematic, as each level inherently possesses a distinct value range; larger windows tend to produce more pronounced differences than smaller windows, thereby complicating direct comparisons. Using the difference in probability bound in the interval $[0, 1]$ appeared to be a more interpretable alternative, similar issues in merging results across varying granularities persisted. Smaller windows naturally yield more minor differences and may erroneously be interpreted as less significant. Furthermore, establishing a universal threshold for further refinement is challenging, given that each image exhibits its own range of output differences. To overcome these issues, an alternative metric focused on identifying windows that cause actual change in the network's classification. This approach is more stringent, as it disregards minor variations in output and concentrates solely on those occlusion windows that result in a class switch. By assigning a binary value (1 for windows that induce a change in the predicted class and 0 for those that do not), the results from different levels of granularity can be aggregated by summation. The outcome is a composite map that indicates, at various levels of detail, the regions whose occlusion induces a change in the network's prediction.

The algorithm initially computes occlusions using larger windows across the entire image to contain the computational cost. It then refines the analysis by applying occlusion with progressively smaller windows exclusively on those regions where the initial occlusions resulted in a change in the network's output. Algorithm 1 provides a high-level algorithm description. Several parameters are critical in our implementation. We provide the model used for classification, the input image, and the target class for occlusion. The initial window size and stride are chosen based on the input dimensions (in our case, 224×224 pixels), and they are halved at each successive level of granularity until a predefined minimum window size is reached. The occlusion window is filled with a specified value (zero in our implementation).

Selecting optimal parameters is challenging due to the heterogeneous nature of the dataset, which contains images of body parts with varying sizes, and the necessity of balancing computational efficiency with the quality of the resulting occlusion maps. For instance, given that the network input is 224×224 pixels,

the initial parameters were set to a window size of 56 and a stride of 28, with a minimum window size of 7. These settings are intended to capture relevant details without being excessively small, which might fail to cover more extensive regions of interest. In practice, the hierarchical algorithm frequently yielded void outputs, meaning that none of the occlusion windows produced a change in the network's prediction. This phenomenon was particularly notable for images of pelves and thighs and, to a lesser extent, calves and knees. In response, the window size was increased for images that initially produced void outputs, thereby reducing computational overhead by avoiding unnecessary recalculations across all images while preserving finer occlusion maps when available.

Algorithm 1. Hierarchical Occlusion

Input: Model, image, target class, initial window size, stride, minimum window size, occlusion value
hierarchical_map, *areas* ← compute occlusion at level n over the entire image
while *areas* $\neq \emptyset$ **do**
    single_map, *areas* ← compute occlusion at level $n-1$ restricted to the regions in `areas`
    hierarchical_map ← *hierarchical_map* + *single_map*
    $n \leftarrow n-1$
end while
return *hierarchical_map*

Table 2 summarizes the results obtained using the hierarchical occlusion methods. As expected, increasing the window size generally reduces the occurrence of void outputs. However, an increase in window size does not necessarily correlate with increased importance, as many new activations may result from occluding a large portion of the patient's body. An additional experiment was conducted with an exaggeratedly large window size of 200. Although a larger window is more likely to affect network predictions, Table 2 indicates that, particularly for the *affected* class, many images still do not exhibit a response to occlusion. This observation is counterintuitive since it would be expected that occluding regions in an affected image would more readily switch the prediction to *healthy*, whereas the converse should be more difficult.

A classical, non-hierarchical occlusion method was implemented to investigate this phenomenon further using the difference in probability as the metric. Figures 3a and 3b display the results of the initial occlusion windows for images classified as *affected* and *healthy*, respectively, using ResNet18. The numerical values annotated on the images represent the percentage difference between the original and occluded probabilities. A positive difference signifies a decrease in the network's confidence, whereas a negative value indicates an increase. Notably, the initial occlusions in Fig. 3a remained classified as *affected* with high confidence despite the occluded regions corresponding to healthy fat. A similar pattern was observed for the *healthy* images in Fig. 3b. This suggests that the

network may misinterpret subcutaneous fat as infiltrated fat when analyzed in isolation, potentially indicating an inherent bias toward predicting the *affected* class.

Table 2. Occlusion results across different occlusion window sizes.

Model	Prediction	Occlusion Window Size							
		56		86		112		200	
		Void	Not Void	Void	Not Void	Void	Not Void	Void	Not Void
Resnet18v	Affected	109	62	93	78	84	87	81	90
	Healthy	17	39	7	49	0	56	0	56
Resnet50v	Affected	102	67	88	81	88	81	55	114
	Healthy	22	36	12	46	6	52	0	58

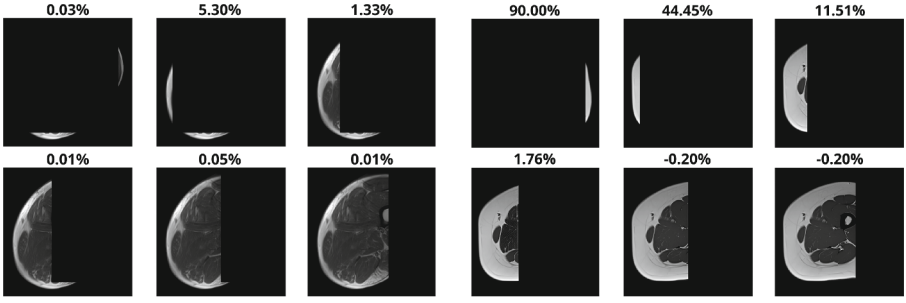


Fig. 3. Occlusion windows comparison for affected (a) and healthy (b) MRIs.

5.2 Ensemble of Explainability Methods

Given that no single XAI technique consistently outperforms the others, an ensemble approach is adopted to integrate multiple explainability methods. By aggregating the outputs of diverse models, the ensemble capitalizes on the strengths of each technique while compensating for their limitations, yielding more robust and reliable explanations. In this work, the ensemble is constructed by combining the non-zero heat maps produced by the various explainability methods with void occlusion maps excluded from the aggregation. Preliminary experiments combined outputs from all the explainability techniques; however, GradCAM was ultimately excluded from the final ensemble because its tendency to produce wider activation regions was found to dilute the more focused

insights provided by the other methods. The proposed ensemble strategy therefore involves GradCAM++, HiResCAM, SHAP, and Hierarchical Occlusion. Before integration, a preprocessing step is applied to the heatmaps to reduce noise and enhance interpretability. Rather than operating at the pixel level, the heatmaps are partitioned into 7×7 square blocks, with each block assigned a value equal to the average of its constituent pixels. Negative values are removed to focus exclusively on areas that contribute positively to the model's prediction. Finally, the data are normalized to the interval $[0, 1]$, ensuring that the outputs from different techniques are on a comparable scale. Three ensemble strategies were investigated, each with a different degree of restrictiveness. Figure 4 compares the base explainability methods and our three proposed ensemble strategies. The first strategy computes the average of the heatmaps and selects regions where the average exceeds 0.5, a procedure analogous to majority voting; this approach tends to yield broader areas of evidence. The second strategy is based on an intersection approach: heatmaps are first filtered to retain only activation values above 0.2, and the ensemble is then defined as the common regions across all methods. While emphasizing regions with unanimous support, this intersection approach may exclude significant areas according to all but one method. The third strategy focuses on the relevance of saliency by considering only those pixels with values above 0.7 and aggregating pixels selected by at least $n - 1$ of the n methods. This more selective approach highlights smaller, more precise regions of interest. This study adopted a threshold of 0.7 to achieve a stringent ensemble that emphasizes only the most salient areas.

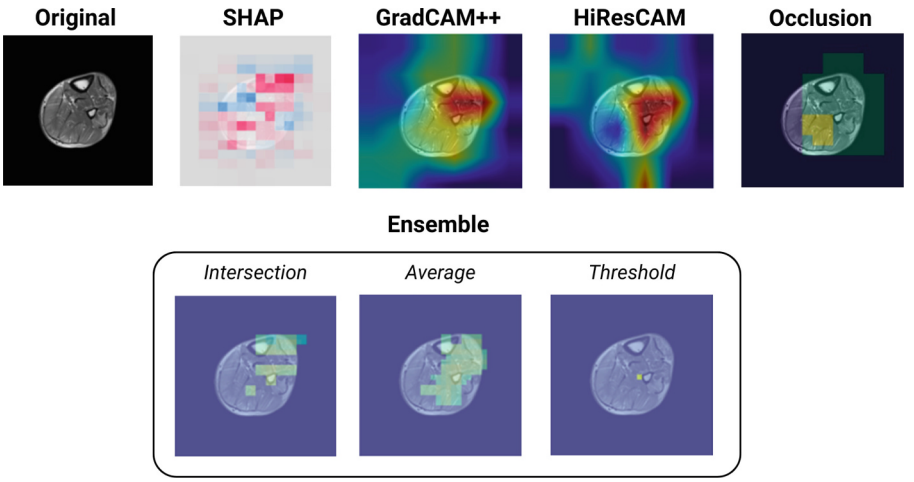


Fig. 4. Comparison of base explainability methods and proposed ensemble strategies.

6 Results

6.1 Model Accuracy

The performance of the AI models was quantified using standard classification metrics, including accuracy, precision, recall, and F1-score. As summarized in Table 3, both ResNet18v and ResNet50v demonstrated strong classification performance, achieving accuracies close to 90%. The high recall scores indicate that both models reliably identified all positive class instances while precision remained competitive, leading to robust F1 scores.

Table 3. Performance metrics for the final models evaluated on the test set.

Model	Accuracy	Precision	Recall	F1-score
ResNet18v	88.55%	84.80%	100%	91.77%
ResNet50v	89.43%	85.80%	100%	92.36%

Analysis of the confusion matrices (Fig. 5) reveals that both networks consistently identified all instances of the positive class, while misclassifications predominantly occurred within the *healthy* class. A closer inspection suggests that images misclassified as *affected* often exhibited prominent subcutaneous fat, which may have contributed to the incorrect classification. This indicates that the model may rely on fat distribution patterns as a proxy for pathology. This unintended bias could be addressed through refined preprocessing techniques such as explicit muscle segmentation or feature calibration.

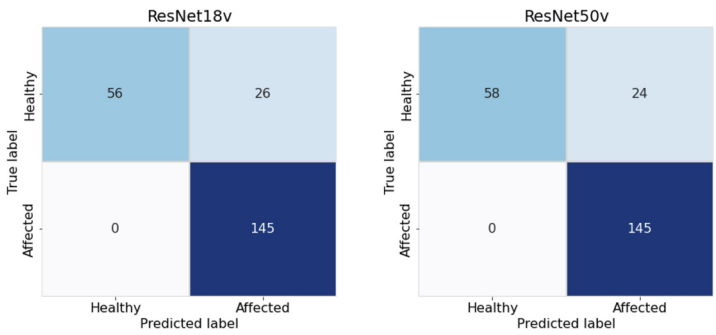


Fig. 5. Confusion matrices for ResNet18v and ResNet50v.

6.2 Explainability Methods

Beyond quantitative performance, this study investigated the interpretability and clinical relevance of AI-generated explanations through a structured evaluation involving seven radiologists, six residents, and one experienced specialist. The assessment required participants to classify the original MRIs and rate the explainability outputs on a five-point Likert scale. The key aspects assessed were diagnostic usefulness, appropriateness of highlighted regions, ease of interpretation, and overall reliability.

Diagnostic Accuracy and Observer Performance. Table 4 summarizes the accuracy of the radiologists compared to the most experienced observer (Observer G). The average diagnostic accuracy was 80%, with notable variability among residents. While some observers closely aligned with the network's predictions, others misclassified most cases. This variability suggests that some instances are inherently ambiguous, even for human experts, highlighting the potential for AI assistance in diagnostic workflows. Despite the high recall of the AI models, misclassified instances occurred among the network and human observers. Analysis of confusion matrices (Fig. 6) reveals that ambiguous cases lacked conclusive explainability outputs, which may explain neutral or negative ratings regarding their reliability. This finding underscores the subjectivity inherent in XAI techniques and the challenge of ensuring trust in AI-generated explanations.

Table 4. Diagnostic accuracy of radiologists on the test subset.

Physician	Experience (years)	Accuracy
Observer A	1	60%
Observer B	1	90%
Observer C	2	60%
Observer D	2	80%
Observer E	2	100%
Observer F	3	90%
Observer G	9	100%

Perception and Evaluation of Explainability Methods. Figure 7 illustrates the votes for different explainability techniques. No method emerged as the clear favorite, as all received at least three votes. GradCAM and SHAP were among the most frequently selected methods. However, a deeper examination of individual preferences shows that Observer F strongly favored SHAP, while GradCAM was chosen by only three out of seven observers. Notably, experienced

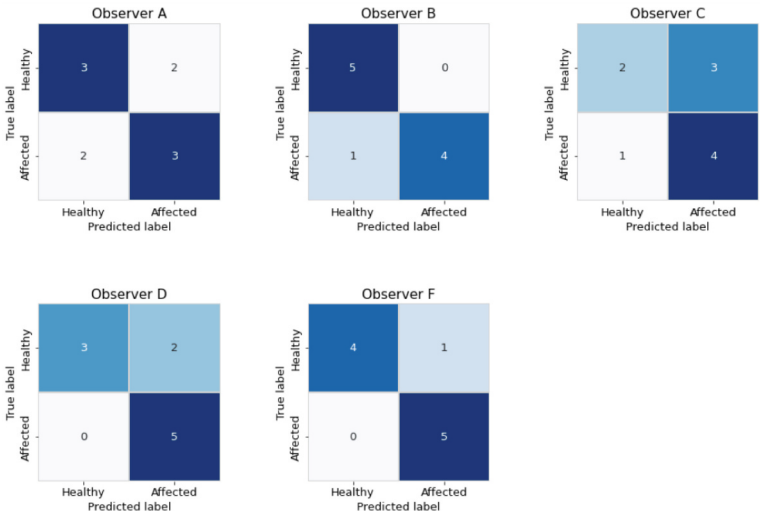


Fig. 6. Confusion matrices for the observers who committed classification errors.

radiologists did not favor these methods but slightly preferred GradCAM++. This discrepancy suggests that experience influences how explainability methods are interpreted, with expert users valuing refined and localized attributions over broader attention-spanning visualizations. The evaluations highlight significant variability in how different explainability techniques were perceived. Some observers consistently selected GradCAM, while those preferring ensemble methods tended to exclude it. Ensemble-based approaches, though receiving fewer overall votes, were regarded as producing more diagnostically relevant explanations by the experienced radiologist. This preference divergence underscores the importance of tailoring XAI methods to different user expertise levels. Further analysis of the explainability evaluations is presented in Fig. 8, illustrating the distribution of observer votes for each image. The chart highlights the high variability in observer preferences, ranging from a maximum agreement of 42% (three out of seven observers) for images 1, 9, and 10, to complete disagreement for image 3, where each radiologist selected a different explainability method. This lack of consensus underscores the subjective nature of explainability assessments and the need for more tailored, adaptable XAI techniques.

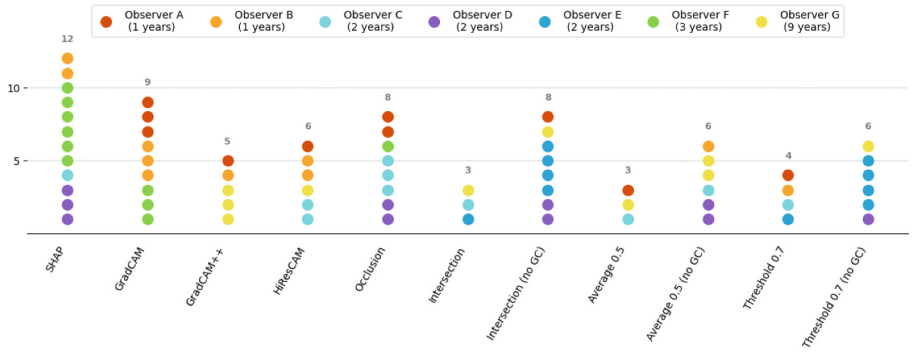


Fig. 7. Preferences for explainability methods across dataset. GC stands for GradCAM.

Explainability Ratings and Observer Preferences. To further analyze the perceptions of explainability techniques, we examined the distribution of scores across four evaluation criteria: (i) *Usefulness of the highlighted area for diagnosis*, (ii) *Appropriateness of the highlighted region size*, (iii) *Ease of interpretation*, and (iv) *Perceived diagnostic reliability*. Figure 9 presents the score distributions. The median ratings were predominantly neutral or low, indicating limited confidence in the AI-generated explanations. Notably, SHAP and the "Average 0.5" ensemble received the highest median ratings (4) for diagnostic usefulness. In contrast, occlusion-based methods received lower scores, likely due to their occasional failure to generate clear explanations. For perceived diagnostic reliability, most methods received scores concentrated in the lower range, except for the SHAP and "Average 0.5" ensemble, which exhibited symmetric distributions around neutral values. This suggests an overall hesitation in fully trusting the AI-generated explanations. During the evaluation, there were five cases where the AI correctly suggested a diagnosis to an observer who initially misclassified the input. However, in none of these cases did the observer change their decision after reviewing the AI explanations, reinforcing the challenge of aligning AI interpretability with clinical reasoning.

Comparison of Individual vs. Ensemble Methods. We grouped the evaluations into Individual and Ensemble methods to assess the effectiveness of ensemble approaches. Figure 10 shows the proportion of observer preferences for each category. Although Individual methods received slightly more votes overall, experienced radiologists slightly preferred ensemble techniques. Figure 11 displays score distributions for the two groups. Individual methods exhibited higher variability, with a skew towards lower values, while Ensemble methods had a more stable and neutral distribution. Interestingly, despite observers showing a preference for Individual methods, they rated Ensemble methods higher for perceived reliability, suggesting that while ensemble techniques are less familiar, they may provide more clinically meaningful insights.

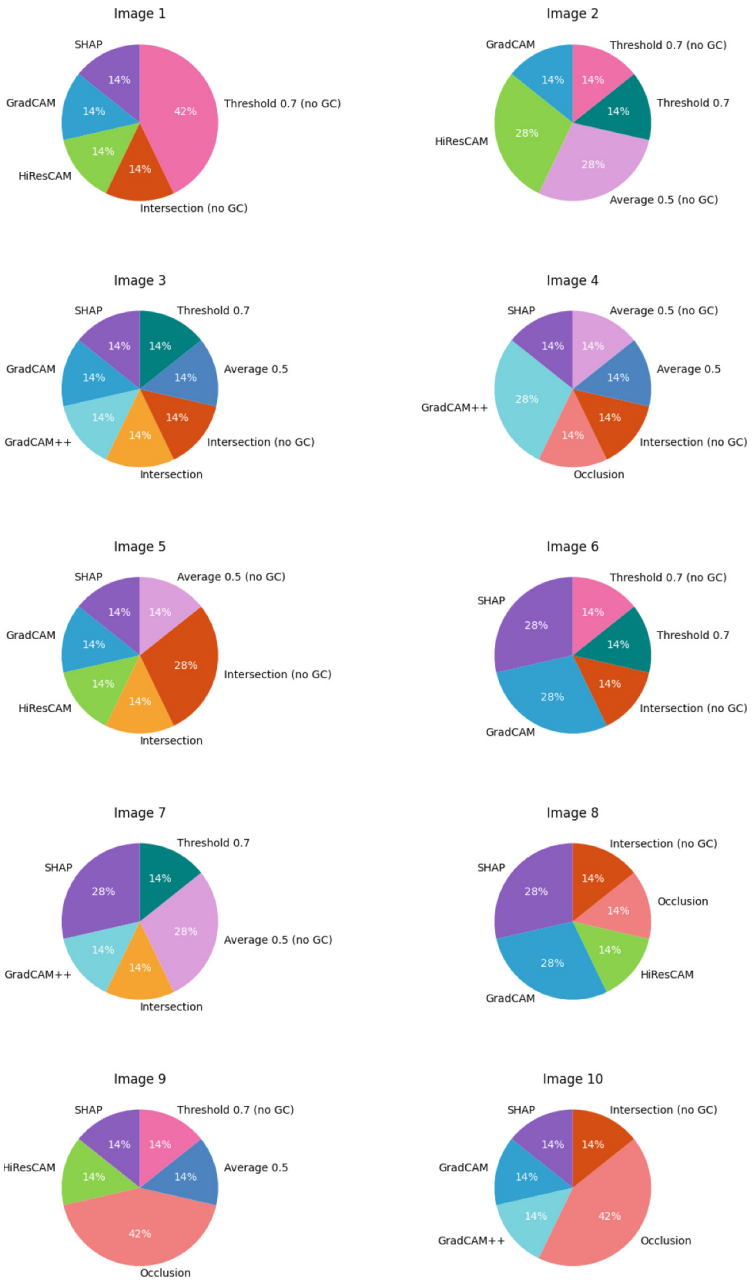


Fig. 8. Observer preferences for explainability methods across different images.

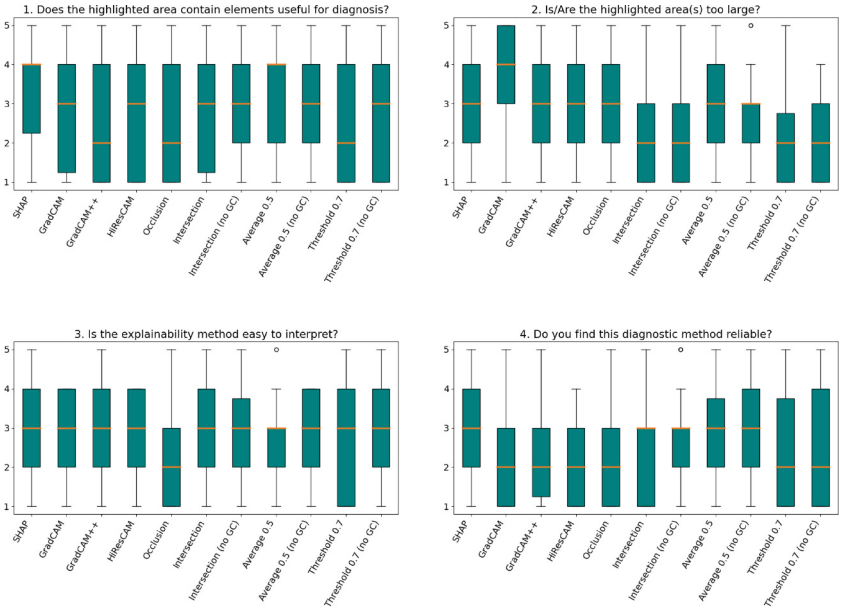


Fig. 9. Distributions of observer ratings (1 to 5) across different explainability methods. GC stands for GradCAM.

6.3 Discussion

Our evaluation shows that both ResNet18v and ResNet50v achieve strong classification performance, with accuracies exceeding 88% and perfect recall in identifying affected cases. However, a closer examination of misclassified instances reveals that these models occasionally base their predictions on secondary visual cues such as fat distribution that do not directly correspond to pathological markers. This reliance on non-specific features suggests the presence of unintended biases. It underscores the potential value of incorporating refined preprocessing techniques, such as explicit muscle segmentation or domain-aware feature extraction, to enhance model robustness and medical relevance.

We conducted a structured user study involving seven radiologists to assess the practical utility of AI-generated explanations. Participants evaluated different explainability methods regarding diagnostic usefulness, appropriateness, and reliability. Results indicate substantial variability in how different techniques were perceived. While methods like GradCAM and SHAP were found to be helpful in some instances, their effectiveness was highly dependent on both the observer's experience and the specific image context. Despite receiving fewer total votes, experienced radiologists consistently saw ensemble-based approaches producing more focused and diagnostically relevant explanations. A key insight from the study is the divergence in preferences between radiology residents and experienced practitioners. Less experienced users tended to favor broad and visually prominent explanation maps, which offered a general sense

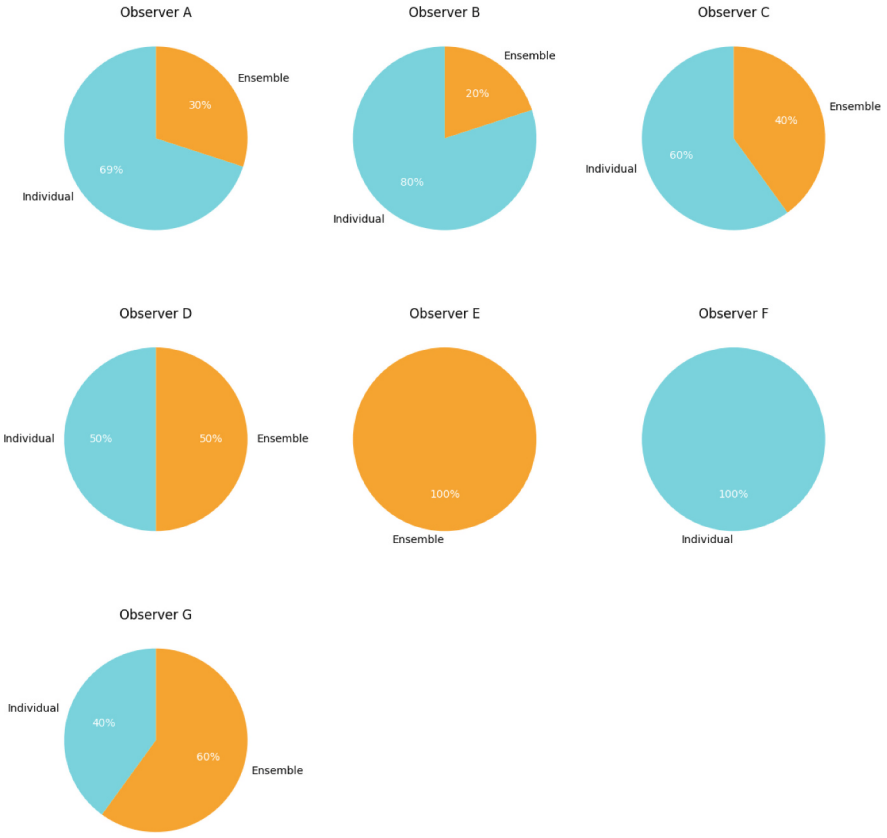


Fig. 10. Proportional distribution of observer preferences.

of model focus. In contrast, senior radiologists preferred precise and localized attributions that aligned more closely with clinically relevant structures. This distinction highlights the critical role of domain expertise in interpreting explainability outputs and suggests the need for adaptable visualization strategies that cater to different levels of clinical experience. Our findings suggest that no single explainability method can universally satisfy all users. Instead, there is a clear need for adaptive, user-centric approaches that balance clarity, precision, and flexibility. Potential improvements include: (i) integrating anatomical priors to reduce model reliance on irrelevant features, (ii) refining saliency techniques to reflect clinical reasoning patterns better, and (iii) developing interactive, customizable explainability tools that allow clinicians to tailor outputs to their diagnostic preferences.

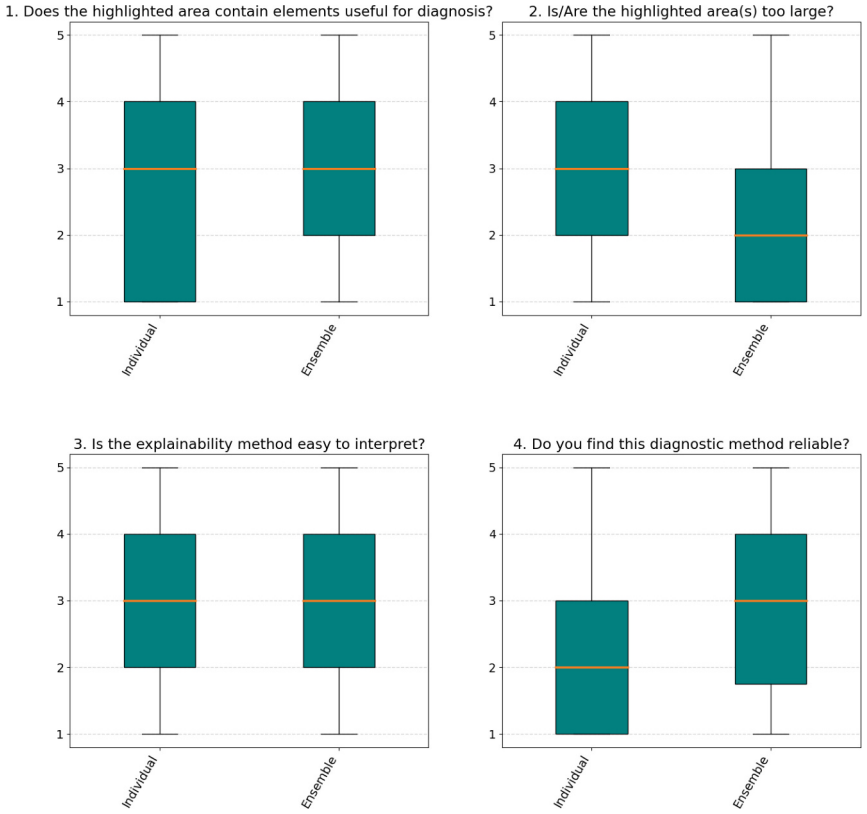


Fig. 11. Score distributions for Individual and Ensemble explainability methods.

7 Limitations

While this study provides valuable insights into AI models' diagnostic performance and explainability for DMs, several limitations must be acknowledged. First, though representative, the dataset used for training and evaluation may not fully capture the diversity of real-world MRI scans. Variability in imaging protocols, scanner models, and patient demographics could influence model generalizability. Future studies should validate these findings on larger, more heterogeneous datasets to ensure robustness across clinical settings. Second, the structured evaluation of explainability methods involved a limited number of radiologists, with only one experienced specialist. While this provided valuable perspectives on how expertise influences the interpretation of AI explanations, a broader sample of radiologists with varying experience levels would be necessary to derive more generalizable conclusions. Additionally, inter-observer variability suggests that user preferences for explainability methods may be highly individualized, emphasizing the need for adaptive and customizable XAI frameworks. Another limitation relates to the reliance on retrospective image assessments

rather than real-time clinical workflows. The physicians reviewed AI-generated explanations in a controlled experimental setting, which may not fully reflect how these methods would be used in actual diagnostic practice. Future studies should investigate the impact of AI explanations in prospective settings where radiologists integrate them into routine clinical decision-making. Furthermore, while the study examined various explainability techniques, it did not explore the full spectrum of available XAI methods or assess how combinations of techniques could enhance interpretability. Given the variability in observer preferences, future research should investigate hybrid approaches that dynamically combine multiple explanation strategies to better align with radiologists' diagnostic reasoning. Finally, the study primarily focused on interpretability and diagnostic performance without considering the proposed methods' computational efficiency and real-time feasibility. Some explainability techniques, particularly ensemble approaches, may be computationally expensive, potentially limiting their practical deployment in clinical environments. Future work should explore optimization strategies to balance explainability quality with computational constraints to ensure seamless integration into medical imaging workflows.

8 Conclusion and Future Work

In this work, we applied a DL approach to diagnose DM, a rare NMD. We investigated the feasibility of AI-based diagnostic support coupled to explainability methods to give physicians a rationale for predictions. Due to the limited size of the dataset, we exploited transfer learning. According to our results, two variants of the ResNet neural network achieved high classification performance, with nearly 90% accuracy on the test set, despite the inherent challenges posed by heterogeneous imaging data. Notably, misclassifications were often associated with increased subcutaneous fat, suggesting that the models may rely on secondary visual cues rather than strictly pathological features. The evaluation of explainability techniques including SHAP, GradCAM, occlusion-based methods, and ensemble approaches revealed no universally superior method. Instead, effectiveness varied significantly across different images and observers, underscoring the subjective nature of AI interpretability. Less experienced radiologists often favored broader visualizations, while more experienced specialists preferred refined, localized explanations, as observed in their inclination toward ensemble methods. This discrepancy highlights the importance of tailoring explainability frameworks to different expertise levels, reinforcing the need for adaptive, user-centric AI interfaces. Moreover, the study reveals an essential gap in trust between radiologists and AI-generated explanations. Even when AI models correctly identified cases misclassified by human observers, physicians rarely adjusted their diagnoses based on the provided explanations. This finding suggests that increasing the transparency and clinical alignment of explainability techniques is crucial for fostering trust in AI-assisted diagnostics.

While our findings demonstrate the potential of AI-based diagnostic support for DMs, several challenges warrant further investigation. A key priority is

refining explainability techniques to better align with radiologists' diagnostic reasoning. This includes improving saliency methods to highlight pathologically relevant regions more precisely and exploring multimodal feature attributions that combine spatial, textual, and temporal information for enhanced interpretability. Additionally, the reliance of AI models on secondary visual cues, such as fat distribution, suggests the need for more sophisticated preprocessing strategies, such as anatomical priors and domain-aware feature extraction, to ensure that model predictions are based on clinically meaningful features. Another critical direction is the development of adaptive explainability frameworks that account for varying levels of radiological expertise. Our study indicates that radiology residents and experienced specialists interpret AI explanations differently, suggesting that a one-size-fits-all approach may not be optimal. Future research should focus on designing interactive diagnostic tools that allow users to customize explainability outputs based on their preferences and expertise. Such tools could include user-adjustable explanation granularity, interactive overlays, and AI-guided annotation support to bridge the gap between automated analysis and human decision-making. Furthermore, improving trust in AI-generated explanations remains an essential challenge. Even when AI models correctly identified cases misclassified by human observers, radiologists did not modify their decisions based on the provided explanations. This underscores the need for human-centered AI design that fosters interpretability, transparency, and trust. Future studies should investigate how to optimize the presentation of AI explanations to encourage meaningful engagement and integration into clinical workflows. This may involve usability testing with radiologists in real-world diagnostic settings and iterative refinement of explanation delivery mechanisms. Finally, the clinical deployment of AI-driven diagnostic tools necessitates robust validation across diverse patient populations and imaging protocols. While our study provides insights into using DL models for DM diagnosis, further validation on larger and more heterogeneous datasets is needed to ensure generalizability. Future research should also explore prospective clinical trials where AI-assisted diagnostic recommendations are evaluated in real-time clinical decision-making scenarios. Establishing standardized evaluation metrics and regulatory frameworks for AI explainability in medical imaging will be crucial for facilitating safe and effective adoption in practice. While DL models show strong potential for supporting DM diagnosis, their successful clinical integration depends on performance improvements and refining interpretability methods to meet medical professionals' expectations. Bridging this gap will require a multidisciplinary effort involving advances in XAI, domain-specific feature extraction, and targeted user training to ensure that AI-driven diagnostic tools are accurate and clinically reliable.

References

1. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847. IEEE (2018)
2. Draelos, R.L., Carin, L.: Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. arXiv preprint arXiv:2011.08891 (2020)
3. Fabry, V., et al.: A deep learning tool without muscle-by-muscle grading to differentiate myositis from Facio-Scapulo-Humeral dystrophy using MRI. Diagn. Interv. Imag. **103**(7–8), 353–359 (2022)
4. Felisaz, P.F., et al.: Texture analysis and machine learning to predict water t2 and fat fraction from non-quantitative MRI of thigh muscles in facioscapulohumeral muscular dystrophy. Eur. J. Radiol. **134**, 109460 (2021)
5. Haenssle, H.A., et al.: Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann. Oncol. **29**(8), 1836–1842 (2018)
6. Jang, H., et al.: Classification of Alzheimer's disease leveraging multi-task machine learning analysis of speech and eye-movement data. Front. Hum. Neurosci. **15** (2021). https://doi.org/10.3389/fnhum.2021.716670
7. Kim, H.E., et al.: Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digital Health **2**(3), e138–e148 (2020)
8. Kokhlikyan, N., et al.: Captum: a unified and generic model interpretability library for PyTorch. arXiv preprint arXiv:2009.07896 (2020)
9. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: a review of machine learning interpretability methods. Entropy **23**(1), 18 (2020)
10. Lundberg, S.: A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874 (2017)
11. McKinney, S.M., et al.: International evaluation of an AI system for breast cancer screening. Nature **577**(7788), 89–94 (2020)
12. MedlinePlus: Neuromuscular disorders. https://medlineplus.gov/neuromusculardisorders.html
13. Menardi, G., Torelli, N.: Training and assessing classification rules with imbalanced data. Data Min. Knowl. Disc. **28**, 92–122 (2014)
14. Multari, S., Özçelik, R., Mazzolari, A., Nobile, M.S., Grisoni, F.: Predicting metabolic reactions with a molecular transformer for drug design optimization. In: 2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1–8. IEEE (2024)
15. Nobile, M.S., et al.: Unsupervised neural networks as a support tool for pathology diagnosis in MALDI-MSI experiments: a case study on thyroid biopsies. Expert Syst. Appl. **215**, 119296 (2023)
16. Papetti, D.M.: An accurate and time-efficient deep learning-based system for automated segmentation and reporting of cardiac magnetic resonance-detected ischemic scar. Comput. Methods Programs Biomed. **229**, 107321 (2023)
17. Piñeros-Fernández, M.C.: Artificial intelligence applications in the diagnosis of neuromuscular diseases: a narrative review. Cureus **15**(11) (2023)
18. Pizer, S.M., et al.: Adaptive histogram equalization and its variations. Comput. Vis. Graph. Image Process. **39**(3), 355–368 (1987). https://doi.org/10.1016/S0734-189X(87)80186-X

19. Radiopaedia: MRI sequences (overview). https://radiopaedia.org/articles/mri-sequences-overview
20. Rizzo, M., Marcuzzo, M., Zangari, A., Schiavinato, M., Albarelli, A., Gasparetto, A.: Stop overkilling simple tasks with black-box models, use more transparent models instead. In: Wallraven, C., Liu, C.L., Ross, A. (eds.) Pattern Recognition and Artificial Intelligence, pp. 279–293. Springer, Singapore (2025)
21. Rizzo, M., Veneri, A., Albarelli, A., Lucchese, C., Nobile, M., Conati, C.: A theoretical framework for AI models explainability with application in biomedicine. In: 2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1–9 (2023). https://doi.org/10.1109/CIBCB56990.2023.10264877
22. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
23. Soroski, T., et al.: Differentiating memory clinic patients and healthy volunteers using machine-learning analysis of speech and eye movements during a reading task. Alzheimer's Dementia **17**(S6), e055717 (2021). https://doi.org/10.1002/alz.055717
24. Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A.: Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med. Image Anal. **79**, 102470 (2022)
25. Verdú-Díaz, J., et al.: Accuracy of a machine learning muscle MRI-based tool for the diagnosis of muscular dystrophies. Neurology **94**(10), e1094–e1102 (2020)
26. Vora, L.K., Gholap, A.D., Jetha, K., Thakur, R.R.S., Solanki, H.K., Chavda, V.P.: Artificial intelligence in pharmaceutical technology and drug delivery design. Pharmaceutics **15**(7), 1916 (2023)
27. Yang, M., et al.: A deep learning model for diagnosing dystrophinopathies on thigh muscle MRI images. BMC Neurol. **21**, 1–9 (2021)
28. Zangari, A., Marcuzzo, M., Rizzo, M., Albarelli, A., Gasparetto, A.: Crossing the divide: designing layers of explainability. In: Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., Zurada, J.M. (eds.) Artificial Intelligence and Soft Computing, pp. 253–265. Springer, Cham (2025)
29. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13, pp. 818–833. Springer (2014)
30. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)