Explainable Classification System for Hip Fractures: A Hybrid CBR+LLM Surrogate Approach

Enrique Queipo-de-Llano[1,*], Marius Ciurcau[1], Alejandro Paz-Olalla[1], Belén Díaz-Agudo[1,2,*] and Juan A. Recio-Garcia[1,2]

[1]Department of Software Engineering and Artificial Intelligence, Universidad Complutense de Madrid, Spain
[2]Instituto de Tecnologías del Conocimiento, Universidad Complutense de Madrid, Spain

Abstract

Hip fractures, mainly displaced femoral neck fractures, pose a significant diagnostic challenge due to their elusive nature on conventional radiographs. These fractures often evade detection in initial assessments, underscoring the critical necessity for refined classification systems. This paper presents a neural network model that can classify X-ray hip fractures accurately and focuses on the need for explanations to increase trust and interpretability. As a first contribution, we evaluate the suitability of several model-specific explanation methods for this concrete classification task. To enhance the limited explainability of these model-specific methods, we also present a surrogate model applying Case-Based Reasoning (CBR) to perform explanations by examples enriched with textual descriptions. This CBR surrogate uses a clustered case base with prototypical cases that include textual medical reports of the fractures. The last contribution of the paper is the integration of LLMs into the reuse stage of the CBR cycle to adapt and personalize these textual reports to the query image and target user.

Keywords

Hip fracture classification, eXplainable AI, Case-Based Reasoning,

1. Introduction

Interpretability and trust have become a requirement for black box models applied to real-world tasks like diagnosis or decision-making processes. In the medical domain, interpretability ensures that medical practitioners can validate and explain the reasoning behind AI-driven diagnoses. This transparency not only enhances the credibility and acceptance of AI technologies within the medical community but also facilitates collaboration between human experts and machine learning systems.

Our research aims to enrich AI-driven diagnoses with different types of explanations. Firstly, we have explored the applicability of several model-specific explanation methods, such as Grad-CAM, that highlight crucial regions of an image that influence the network's prediction based on the gradients of the last convolutional layer.

However, this approach has limited explainability, and therefore, we propose a surrogate system based on CBR [1, 2] to provide explanations by example. The explanation cases also contain a textual component that associates medical reports with the most significant (or prototypical) examples. Then, as a novel contribution to the XCBR field, we integrate large language models (LLMs) into the reuse stage of the XCBR system to adapt these textual reports to the target user and the X-ray image being explained. To the best of our knowledge, this approach is a novelty, and it presents both a challenge and an opportunity for the impact of synergies of CBR and LLMs in the eXplainable AI (XAI) landscape. The proposed approach could prove highly beneficial for providing textual explanations in domains with small sets of training data (where generative approaches are not possible). Besides, a CBR methodology is useful for capturing new textual explanations from experts and even for the training process of future experts in similar domains.

ICCBR XCBR'24: Workshop on Case-Based Reasoning for the Explanation of Intelligent Systems at ICCBR2024, July 1, 2024, Mérida, Mexico

*Corresponding author.
equeip01@ucm.es (E. Queipo-de-Llano); mciurcau@ucm.es (M. Ciurcau); alpaz@ucm.es (A. Paz-Olalla); belend@ucm.es (B. Díaz-Agudo); jareciog@fdi.ucm.es (J. A. Recio-Garcia)
0000-0003-2818-027X (B. Díaz-Agudo); 0000-0001-8731-6195 (J. A. Recio-Garcia)

Section 2 describes related work on the hip fracture detection problem and its interpretability. Section 3 describes our fracture classification system, which uses a deep learning approach trained from a heterogeneous dataset of X-ray images. As it is a black-box system, we describe gradient-based explanations in section 4. In Section 5, we define a CBR system to generate natural language explanations without an extensive corpus of textual data for training. The system first retrieves a prototypical explanation that is extended using a generative LLM. Finally, Section 6 concludes the paper and discusses open lines of future work.

2. Related Work

The classification of hip fractures is a critical aspect of orthopedic diagnosis, guiding treatment decisions and patient care. With the advent of artificial intelligence (AI), there has been a surge in research aimed at automating and improving the accuracy of hip fracture classification. In this section, we review relevant literature, including the pioneering work of Krogue et al. [3] and Tanzi et al. [4], along with other significant contributions in the field.

Justin Krogue et al. [3] conducted groundbreaking research aimed at automating the identification and classification of hip fractures using deep learning techniques. Their study involved the analysis of hip and pelvic radiographs from a substantial dataset of 3026 hips gathered throughout 20 years and a total of 1118 different studies. Through meticulous labeling and classification, they trained a deep learning–based (DenseNet169) object detection model to automatically identify and classify hip fractures into various categories, including displaced and nondisplaced femoral neck fractures, intertrochanteric fractures, as well as previous surgical interventions such as open reduction and internal fixation or arthroplasty.

The results of Krogue et al.'s research were highly promising, with the developed model achieving remarkable accuracy rates. The binary accuracy for detecting a fracture reached 93.7%, with a sensitivity of 93.2% and a specificity of 94.2%. Moreover, the multiclass classification accuracy was reported at 90.8%. Remarkably, the model's performance was comparable to or even surpassed that of human experts, including fellowship-trained radiologists and orthopedists, under various conditions. As for explainability, point-based gradient explanations were applied to the model. It is noted that the model appears to pay attention to cortical outlines to make its classification, while the lucent fracture line appears to receive very little attention.

Furthermore, Krogue and his team demonstrated that the use of the deep learning model significantly improved human performance in hip fracture classification. When aided by the model's predictions, resident physicians achieved classification accuracies approaching those of unaided fellowship-trained attending physicians.

The work of Tanzi et al. [4] represents a significant advancement in femur fracture classification, particularly in sub-fracture classification. The study utilized a ViT-based architecture to classify femur fractures with remarkable accuracy, substantially improving baseline methods such as InceptionV3 and a hierarchical network. Through extensive experimentation and evaluation, Tanzi et al. demonstrated that ViT outperformed existing approaches and provided valuable insights into fracture localization and feature extraction. The attention maps generated by ViT revealed the network's focus on specific anatomical regions, enhancing interpretability and clinical relevance. Moreover, the study highlighted the synergistic effect of combining ViT predictions with specialist expertise as a Computer-Aided Diagnosis (CAD) system, resulting in significantly improved diagnostic accuracy compared to either approach alone. Importantly, Tanzi et al.'s work emphasizes the clinical significance of accurate sub-fracture classification, addressing a critical need in medical diagnostics and attaining satisfactory results on the matter. This pioneering use of ViT in femur fracture classification sets a new standard for future research in this domain, with potential applications extending to more complex levels of sub-fractures in the AO/OTA classification system.

The work conducted by Twinprai et al. [5] shows the performance of a YOLO-v4-Tiny model trained with a carefully elaborated dataset of 900 training images. A bounding box was manually drawn onto the

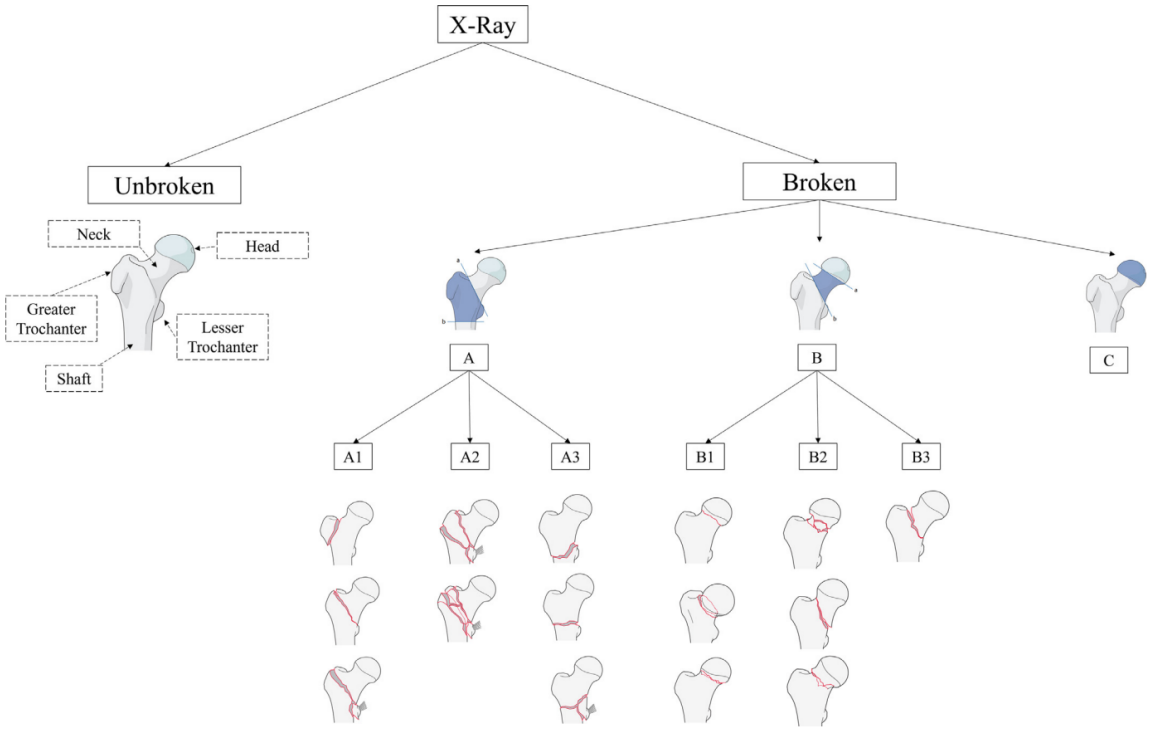

Figure 1: AO/OTA classification. Type A, type B, and type C are subsequently divided into different subgroups, by Tanzi et al. [4].

images in a rectangular shape with an annotation tool. The article shows the performance comparison between the model and several human doctors of different expertise.

Previous successful hip fracture detection studies include the models of Cheng [6], which used DenseNet-121 for a sensitivity of 98% and an accuracy of 91% and also implemented Grad-CAM explanations. Lee [7] successfully used the meta-learning deep neural network GoogLeNet (Inception v3) to classify femoral fractures in pelvic radiographs. Adams [8] compared AlexNet and GoogLeNet for femoral neck fracture detection with an accuracy of 88.1% and 89.4%, respectively.

Regardless of the quality of diagnosis, trust and acceptance are vital to the adoption of medical AI systems, and these depend on the ability to explain the system's decisions. Our focus lies in pushing the boundaries of application by integrating a CBR model to generate diagnostic insights from radiographic images.

The work conducted by Kim et al. [9] resonates with our work. Their approach is to use prototype representations and clustering in a dataset to perform joint inference on cluster labels, prototypes and important features. Similarly, the studies by Schank and Leake [10] and Li et al. [11] offer valuable insights into the use of case-based reasoning (CBR) to generate explanations by retrieving and adapting prior explanations in memory. We believe their work highlights the importance of leveraging expert knowledge and experience to enhance the interpretability of AI systems in the field of medical diagnostics.

Our approach proposes ANN-CBR Twins [12] where a deep-learning, black box classifier is twinned with CBR as an interpretable model. The successful demonstrations of ANN-CBR Twins to provide *post-hoc explanations* by example that is enriched with the generation of textual explanations personalized according to the user context and expertise, make this a valuable approach for explainability. As outlined in Ford and Keane [13] very few eXplainable AI studies consider how users understanding of explanations might change depending on whether they differ in their context or expertise. Expertise is a critical facet of human decision-making in the health domain, and understanding differs from a trainee doctor, an experienced consultant, or a patient.

3. Hip fracture classification system

Our hip fracture classification system uses a deep learning approach trained from a heterogeneous dataset of X-ray images. Next, we describe this dataset, its preprocessing, and the resulting classification model.

3.1. Dataset

Our starting point is the AO/OTA[1] classification for bone fractures [14]. Following their scheme, hip fractures are labeled as: "A" for fractures in the trochanteric region, "B" for femoral neck fractures, and "C" for femoral head fractures. Figure 1 illustrates such classification.

Our primary goal was to train a classification model for hip fractures. To do this, we needed to acquire correctly labeled data in the form of hip X-rays. Thanks to the Virgen de la Victoria Hospital and Müller Foundation Spain, we have collected this data from three different sources:

- The YOLO Roboflow project available online at [15]. This dataset consisted of both complete hip X-rays (Bilateral) and some single hip (Unilateral) X-rays. We were able to find a total of 640 X-ray images.
- A private research collection of 246 hip fractures provided by the Virgen de la Victoria Hospital (Málaga, Spain).
- X-ray fractures provided by the Maurice E. Müller Foundation. We were given a total of 132 images labeled according to the aforementioned classification.

No personal data was collected, all the data was completely anonymous when received.

3.2. Preprocessing

Since our collection of X-rays came from three different sources, we had to deal with the problem of obtaining a uniform and homogeneous dataset with which to train our model. To achieve this, we performed various methods of preprocessing on the images. The data from the Müller foundation was the most challenging to preprocess as a portion of the data (n=62) had an inverted color palette, meaning the image displayed a black bone on a white background. In contrast, the rest of our images had a black background with white bones. Furthermore, most of these images came with two X-rays: one showing the bone fracture and another showing the prosthesis or implant used to treat the patient. This is shown in Figure 2.

An additional problem with the collected images is the cropping of the femur area. As Figure 2 (right) illustrates, most of them were general X-ray images where the femur area was only a small portion of the whole image. Therefore, we required a method to crop the hip area.

For this task, we obtained the femur crops using a custom YOLOv5 model trained on our complete dataset. Ultimately, we implemented Python scripts to flip right femurs into left femurs, unify color scale into gray, and resize images to 224x224 pixels, which are the dimensions required for the input channels of our residual neural network.

A diagram representing the obtention of the whole dataset can be found in Figure 3. As shown in the diagram, we finally obtained 891 images corresponding to unbroken hips, 212 to femoral neck fractures, and 294 to hips presenting a fracture in the trochanteric region.

The collection of such a challenging dataset led us to take the following additional preprocessing steps:

- Given that the data is generally only labeled as type A, type B, or type C, it was not possible for us to train a classifier for the given subtypes of A, B, or C. Even if we had the complete labels, we

[1] AO/Orthopaedic Trauma Association. AO stands for Arbeitsgemeinschaft für Osteosynthesefragen or the Association of the Study of Internal Fixation (ASIF), established in 1958 by a group of Swiss general and orthopedic surgeons to strive to transform the contemporary treatment of fractures in Switzerland.

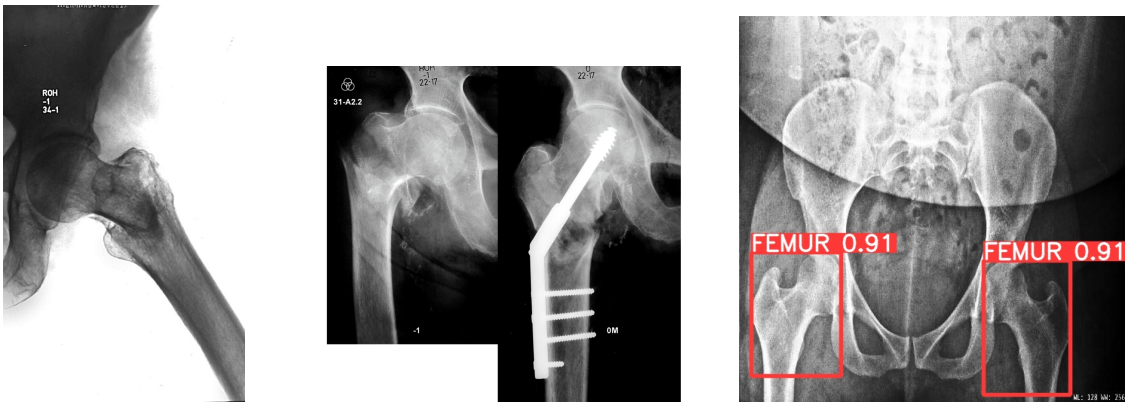

Figure 2: Black on white X-ray (left), X-ray showing hardware (center) and Bilateral femur crops with YOLOv5 (right)

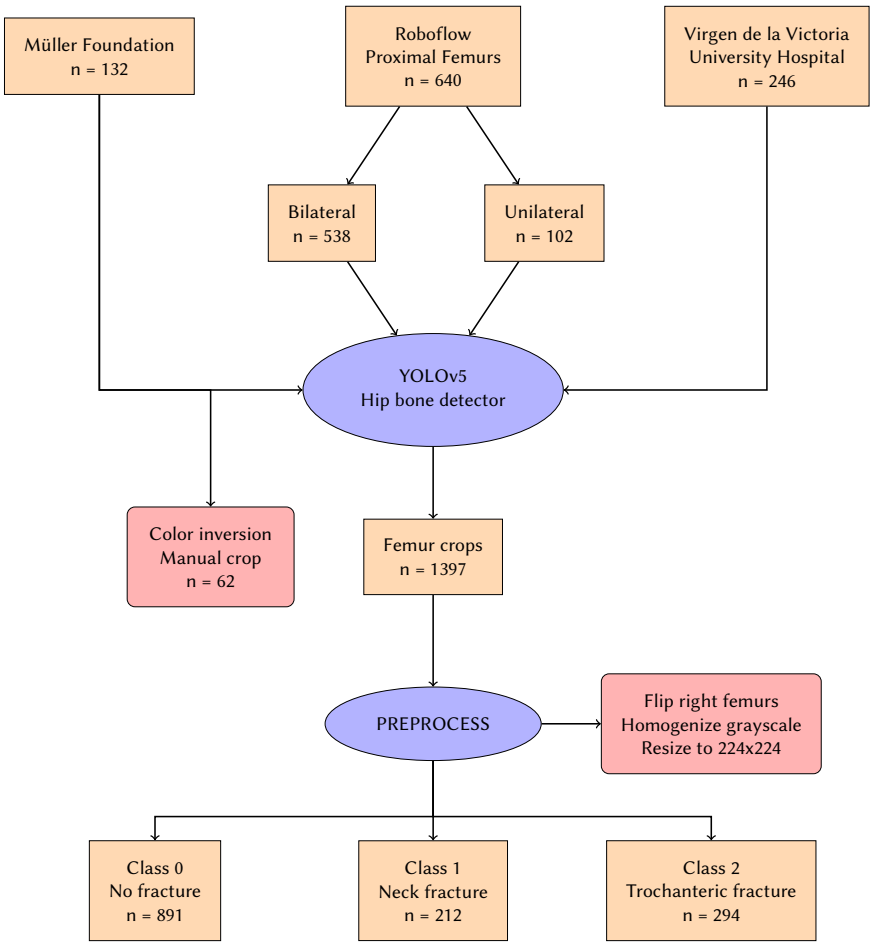

Figure 3: Diagram representing the collection and preprocessing of the dataset

did not have enough data to implement a model able to predict the subtypes, so we decided to focus on just the general type, forgetting the subtypes. Moreover, type C fractures are unusual, which results in a three-class model, predicting no fracture, type A fractures, or type B fractures. Working beyond this problem has no easy solution, as it involves experts manually labeling each X-ray into more complicated types and gathering more data.

- Looking at the data, it is obvious we have an imbalanced class problem: for types 1 and 2, we have a similar amount of data, but for type 0 (unbroken), we have about three times as much data. To solve this, we used the *Albumentations* library, which has been designed for image augmentations,

primarily focused on deep learning tasks. It offers a wide range of transformations from where we implemented uniform angle rotations between 0 and 15 degrees and random gamma, adjusting the global brightness of the image. These transformations were applied with a probability of 50% and 10%, respectively. We perform three augmentations for classes 1 and 2, and this way, we finally obtain a balanced dataset with about 2600 images ready to train our model with 891 images for class 0 and about 800-900 for each of classes 1 and 2.

3.3. Classification model

We performed tests on different architectures of neural networks, finally opting for a Residual Network with 18 convolutional layers (Resnet18). This model has an 18-layer structure that incorporates residual blocks with skip connections, addressing the vanishing gradient problem and enabling effective training of deeper networks. Pre-trained weights available for ResNet18 further enhance its utility through transfer learning, allowing fine-tuning on medical datasets with reduced labeled data requirements. By automatically extracting relevant features from medical images, ResNet18 proves adept at capturing bone structures, edges, and textures crucial for fracture identification. We have trained this model for 10 epochs, implementing cross-validation to avoid over-fitting. We split the dataset into 80% to train and 20% to test, and we obtained satisfactory results as reported in Table 1 and the confusion matrix shown in Figure 4. Although the score of the model is very high, we have ruled out the posibility of over-fitting for two main reasons: firstly, the results on the test set are impressively satisfactory, and secondly, helped by experts of the University Hospital Virgen de la Victoria, we have noticed that the explanations generated by Grad-CAM, and detailed in the following section, highlight regions close to the location of the fracture.

	Precision	Recall	F1-score	Support
No Fracture	0.97	0.94	0.95	178
Femoral Neck	0.94	0.98	0.96	178
Trochanteric	0.97	0.95	0.96	178
Accuracy			0.96	534
Macro avg	0.96	0.96	0.96	534
Weighted avg	0.96	0.96	0.96	534

Table 1
Accuracy report for the Resnet classification model

4. Explainability Model

As a classifier, the model offers no means of interpretability, a fundamental cornerstone in medical systems. One common approach for explainability when working with image data is applying Gradient-weighted Class Activation Mapping (Grad-CAM) or Integrated Gradients, and it was our first focus.

Grad-CAM analyzes gradients in the last convolutional layer to decode the importance of each feature map for a specific class. It generates a heat map, highlighting crucial regions of an image that influence the network's prediction, providing interpretability without compromising accuracy. By visualizing how models arrive at predictions, we gain insights, debug models, and enhance performance. Several heat-map explanation methods from the libraries *Xplique* [2] and *pytorch_grad_cam* [3] have been implemented and compared. The ones from *Xplique* were discarded because, under the criteria of experts at Virgen de la Victoria Hospital, the results were not satisfactory, as those explanations did not

[2]https://github.com/deel-ai/xplique
[3]https://github.com/jacobgil/pytorch-grad-cam

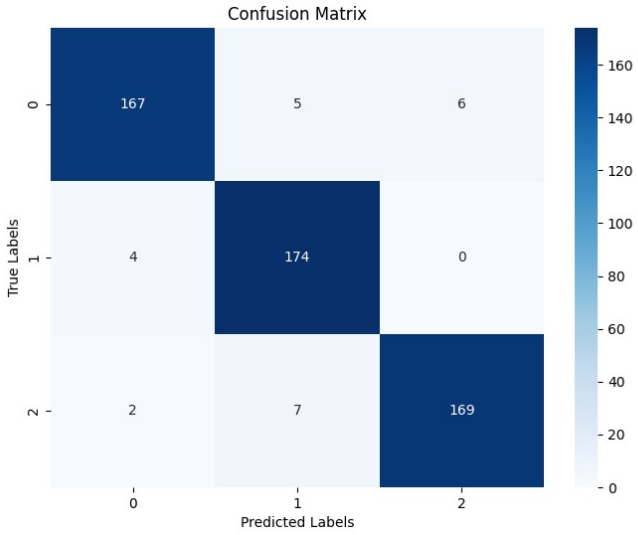

Figure 4: Confusion Matrix

seem to highlight the regions where fractures were visible. To choose between GradCam, Gradcam++, EigenCam, and Ablation, the ROAD (RemOve And Debias) score was implemented. Fundamentally, this metric gives a score to a heatmap by obscuring the more important regions and performing a new prediction to understand how hiding such regions affects the output. This way, if hiding the red zone of a heatmap means a change in the model's prediction, we can affirm this region is relevant to the model so that this explanation will have a high score. More information about this metric can be found in Rong et al. [16]. In Figure 5, a batch of different explanations and scores is shown for some X-rays. A thourough revision was conducted by an expert in dedicated meetings, where he would be shown fractures and discussed which of the explanations helped most. According to the expert, Grad-CAM seems to be the most helpful explanation method, so we finally decided to use this one. We also tested Random-CAM, which consists of generating a random heat map, to prove the ROAD metric works. The information conveyed by gradient-based explanations is what part of the image our model focuses on to generate the prediction. However, this information is not helpful when providing a diagnosis by giving an X-ray. Other authors have also outlined this issue [17]. Textual explanations can be more beneficial than integrated gradients in X-ray images due to their ability to provide contextual understanding, and educational value, and address subjectivity and uncertainty. While integrated gradients may highlight important regions in an image, textual explanations offer a comprehensive understanding by describing anatomical structures, abnormalities, differential diagnoses, and treatment options, and acknowledging limitations. Furthermore, textual explanations offer a level of customization that can be tailored to the expertise level of the explainee. They can provide detailed insights for experts in the field, while also offering simplified explanations for those with less experience in radiology interpretation. This personalized approach ensures that the information provided is accessible and relevant to the individual's level of knowledge, further enhancing comprehension and aiding in effective decision-making.

Here, LLMs appear as a direct and effective solution. However, current state-of-the-art generative models, such as GPTs, are of no use in this case since they have not been trained with medical images, and fine-tuning requires a large corpus of textual explanations associated with X-ray images. As the manual collection of this corpus is unviable, we proposed a hybrid solution using CBR, as presented next.

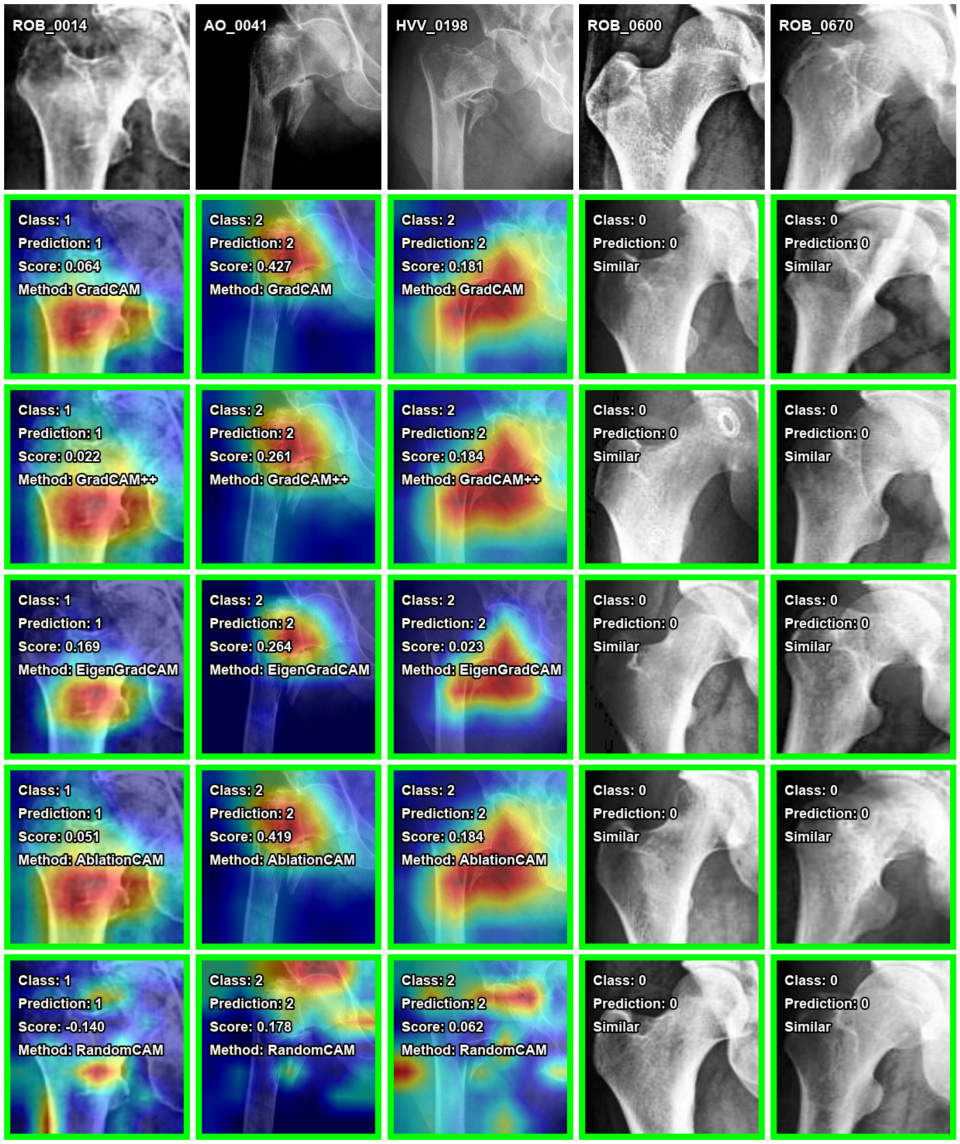

Figure 5: ROAD Metric to compare explanation methods.

5. Hybrid CBR-LLM explainer

Our research faces a pivotal question here: besides gradient-based explanations, we have so far, *can we effectively generate natural language explanations for this AI-driven fracture classification without an extensive corpus of textual data for training*?

As we only have a limited number of textual explanations associated with prototypical fractures, CBR raises as a suitable approach to, given a new X-ray image, retrieve the most similar prototypical case and reuse its textual explanation. The adoption of the CBR paradigm has the additional benefit of enabling the use of the most similar images as explanation examples. This way, our explanation methodology is three-fold: attribution-based (Grad-CAM), textual (LLMs), and explanation-by-example. All these explanation modalities are presented to the user through the graphical interface presented in Figure 7.

Therefore, we propose a twin surrogate architecture to explain the classification model. Figure 6 depicts the two parallel pipelines. The first (top) pipeline performs the classification process using the *Resnet model* after preprocessing the images with YOLOv5 (see Section 3), and provides a visual explanation using Grad-CAM (see Section 4). In the second pipeline, we use a CBR system to re-

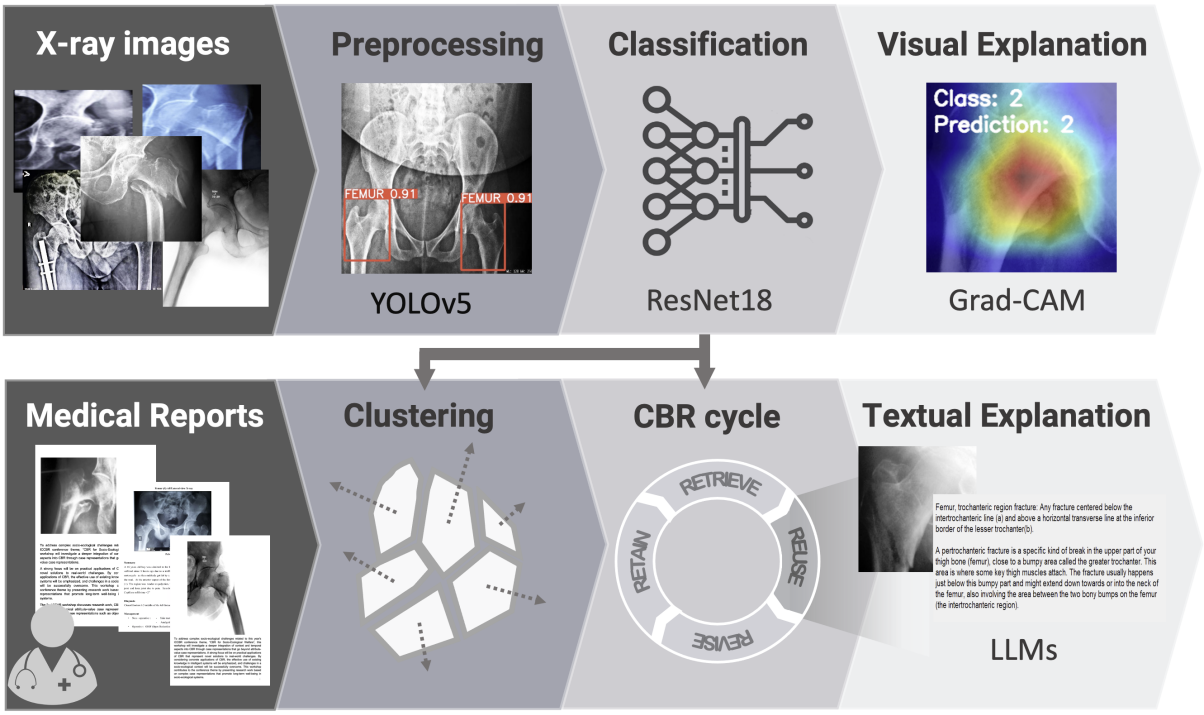

Figure 6: Twin surrogate architecture.

trieve prototypical explanations followed by an LLM that adapts the resulting text to generate textual explanations. These processes are described in the next sections.

5.1. Case Base

The case base has been elicited by an expert who provided highly representative images and the corresponding medical reports, which we will also call prototypical cases. These will be the seed cases of our case base, defined as $Case = \langle I, C, R \rangle$, where I is the X-ray image, C the classification given by the ResNet18 model, and R is the associated textual report.

Initially, we have about ten seed cases that have been included with the help of experts: one image of an unfractured femur, three images of subtypes of femoral neck fractures, and two images of subtypes of trochanteric fractures with their corresponding texts. Each of these texts is a definition for the given subtype of fracture. We also included three general explanations for femoral neck fractures and two for trochanteric fractures. Figure 8 shows an example of a prototypical case. In this text, the part corresponding to a general description of a femoral neck fracture is highlighted in blue, while below, in gray, appears the text explaining the subtype of fracture that, according to the expert, is recognized in the prototypical X-ray shown on the right.

5.2. Retrieval

Due to the limited number of seed cases, we propose a retrieval strategy using a clustered organization of the case base. Having limited textual explanations, we aim to spread such texts to a wider range of images, expanding our base cases. We implement an algorithm to cluster images within the same type of fracture, keeping at least one prototypical case in each resulting cluster.

The idea is to group around the prototypical cases, which would serve as "centroids", providing their explanation to the associated cases. Initially, our idea was to implement a KMeans algorithm, but unfortunately, this would move the centroids, meaning that prototypical cases could end up in a different cluster and some clusters may not get a prototypical case at all.

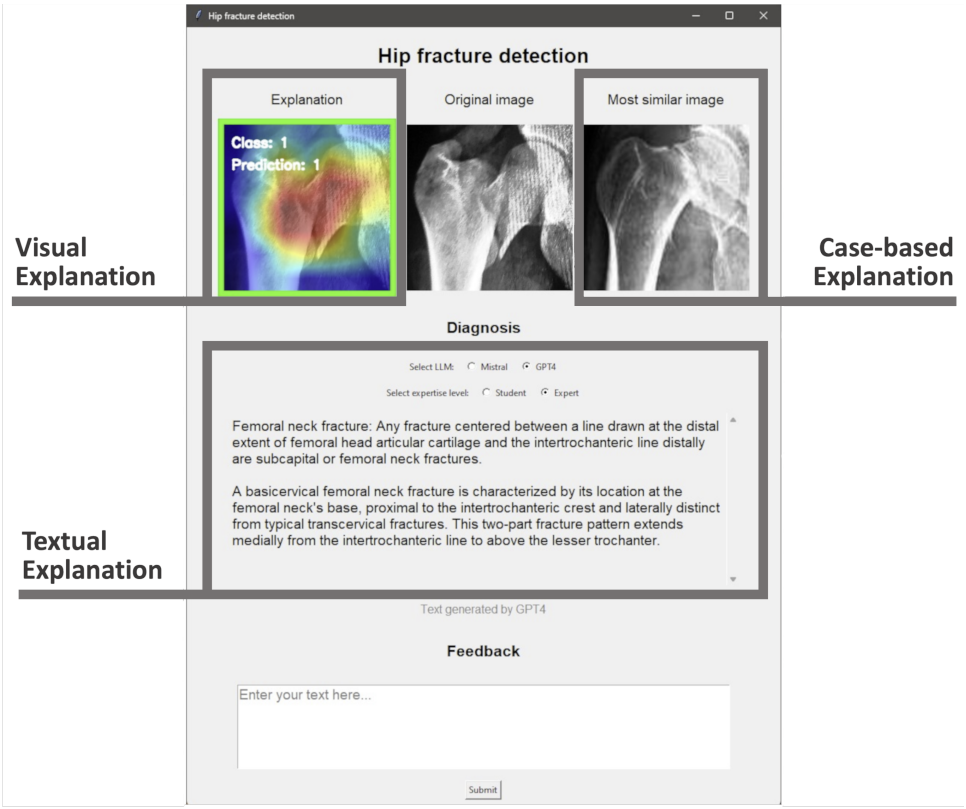

Figure 7: GUI including the two explanation modalities

Femoral neck. This area is located at the top of the femur, just below the spherical part (head of the femur) of the ball-and-socket joint. Fractures of the femoral neck are divided into basicervical or transcervical depending on the distance to the femoral head, basicervical fractures occur at the junction of the head and the neck, and transcervical fractures occur in the neck itself. Since the blood supply to the femoral head passes through the neck, this type of fracture can cause a complication because the break often blocks the passage of blood to the femoral head, leading to avascular necrosis, which is equivalent to a heart attack but in bone. These fractures can occur in young people in high-energy accidents but are common among older individuals who fall from their height due to osteoporosis.

A transcervical femoral neck fracture is another specific type of fracture that occurs in the femur bone, in the region of the femoral neck. In this case, however, the fracture traverses the femoral neck from side to side, completely dividing it. It can have serious consequences due to the disruption of blood flow to the femoral head.

Figure 8: X-ray prototypical case and textual explanation

The clustering process is as follows. Wanting to respect the label restriction, every image of a given label is compared only with the prototypical cases of such label, associating it with the one to which it is most similar, based on the Structural Similarity Index Measure (SSIM) metric [18]. This allows for a separation of the images into as many subgroups as cases we originally had. Although each cluster does represent a certain type of subfracture, this should be taken as an attempt to approximate grouping

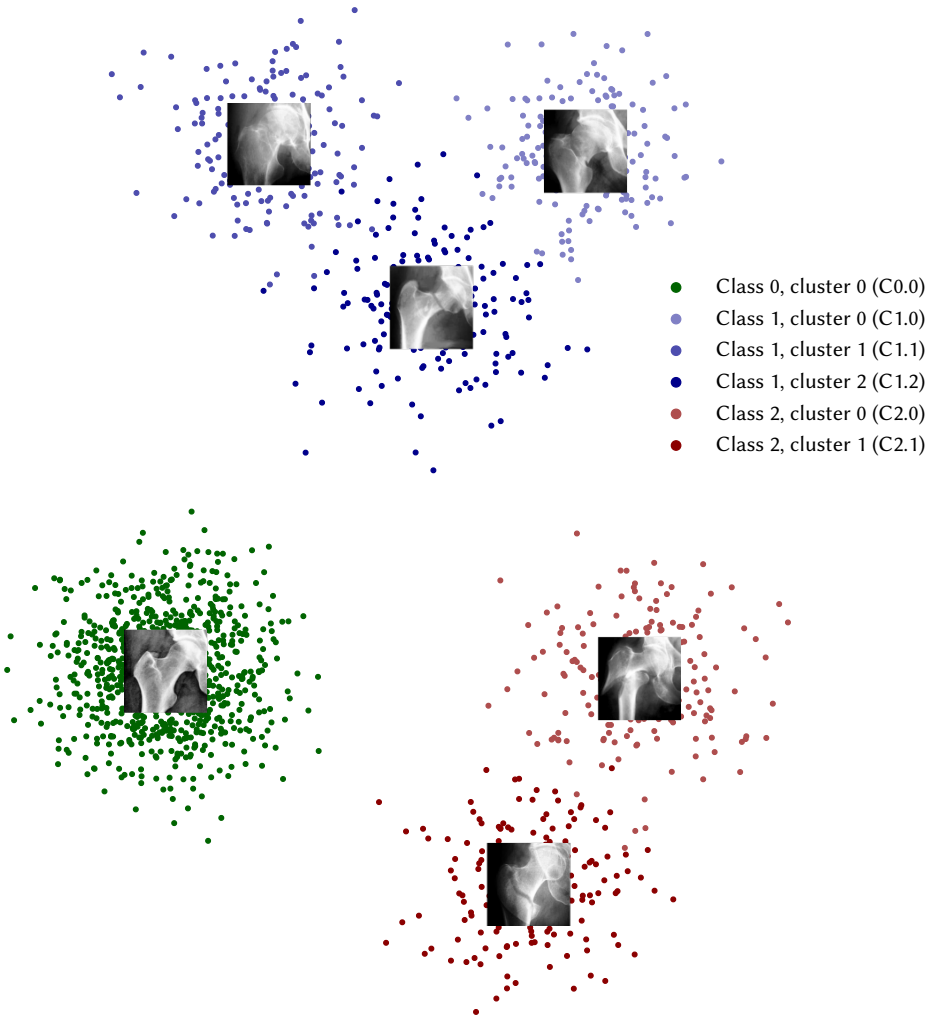

Figure 9: Clusters representation

with restrictions and it should not be taken as an attempt to predict types of subfractures, as searching by SSIM is not realiable enough for this task. In this sense, even though we talk about clusters, we are not implementing an unsupervised algorithm and the word cluster should be taken as group.

In this stage, we can retrieve the prototype case for the cluster where the query image is classified. Firstly, the query goes through our Neural Network classifier, providing a predicted label for it. Secondly, once the label is obtained, the image is compared to all the cases in that same label. As a result, we retrieve the most similar image, and, since it already belongs to a cluster, we can provide its explanation for our query.

In Figure 9, the different clusters generated for each of the classes are shown, along with the image representing the centroid of each one. Each cluster is built around a centroid that represents a certain subtype of a fracture, blue being subtypes of femoral neck fractures and red ones for fractures in the trochanteric region. As mentioned in this section, the idea is that images sharing the subtype within the same class will exhibit some similarity among them and, therefore, should belong to the same cluster or be close to each other. This way, we manage to propagate the explanations we have to new images.

As we have mentioned, for comparing images, we use the metric known as the Structural Similarity Index Measure (SSIM), which is measured between -1 (completely different images) and 1 (identical images). This measure is particularly interesting for our domain because, in addition to considering parameters such as luminosity, contrast, or brightness of the images to determine their similarity, it also takes into account the structure of these images, which is perhaps the most crucial aspect when comparing X-ray images.

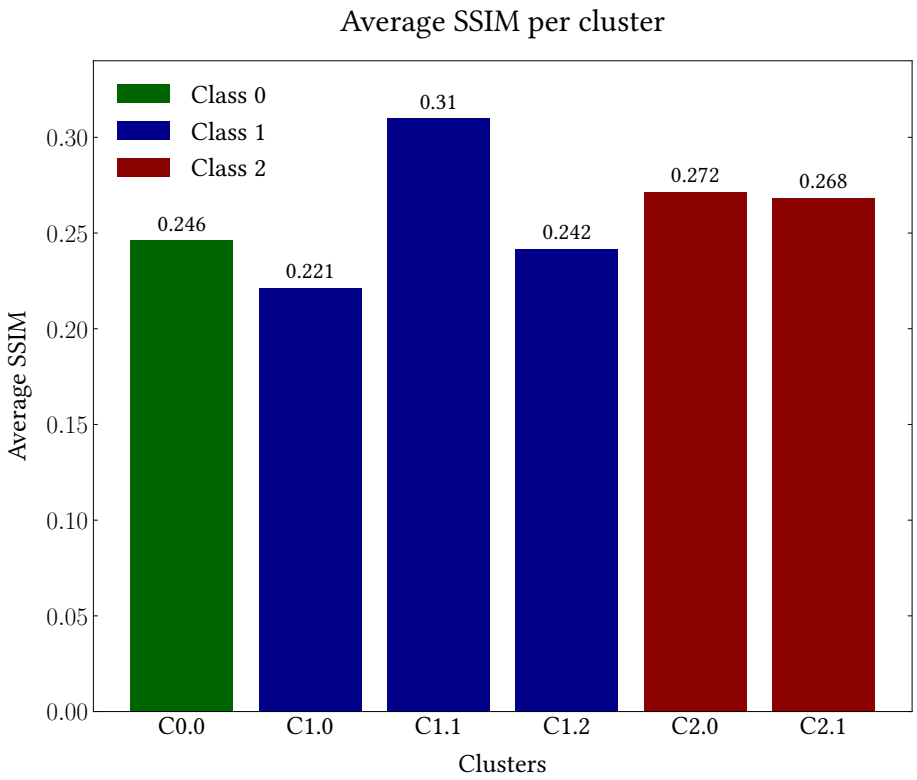

Figure 10: Average SSIM of the resulting clusters

Figure 10 reports an evaluation of the quality of the resulting clusters, measured through the average SSIM similarities within each cluster. As we can observe, all the clusters have an equivalent similarity, denoting a balanced distribution of the cases.

Using the clustering approach has the additional advantage that the image corresponding to the prototypical case of the cluster can be presented to the user as an explanation example. This way, our explanation approach not only uses the heat maps provided by Grad-CAM and the textual explanations generated by the LLMs but also provides an explanation case to ease the understanding of the fracture.

Recall that a simple search by SSIM will not provide an accurate classification of a given subtype of fracture. SSIM retrieves images that are structurally similar, but it may not recognize the underlying differences between fractures. Since the subtype explanations are generated based on an SSIM search, they may not be reliable or precise while the general ones will always be, as the images are taken from the same label.

Finally, Figure 11, shows an example of the retrieval stage: the most similar image to a given one is displayed, along with the cluster where this latter image is located. The classifier's prediction is used to determine among which images to search for the most similar one since, as mentioned earlier, it is only compared with images within the same class.

5.3. Reuse

The reuse stage is coupled to an LLM to adapt the textual report R of the retrieved case. This is useful to prevent the system from being too repetitive, especially in the beginning when there are few cases. Note that the use of generative models to provide textual explanations has the risk that it may generate some invalid explanations, which is why the review phase of the CBR system is crucial in our proposal.

To implement this stage, we have tested Mistral and GPT-4 through the HuggingFace and OpenAI APIs, respectively. The prompts include the textual report R obtained by the retrieval stage and various "knowledge levels" of the user, with which we intend to generate outputs adapted to different types of users (medical experts or patients/students).

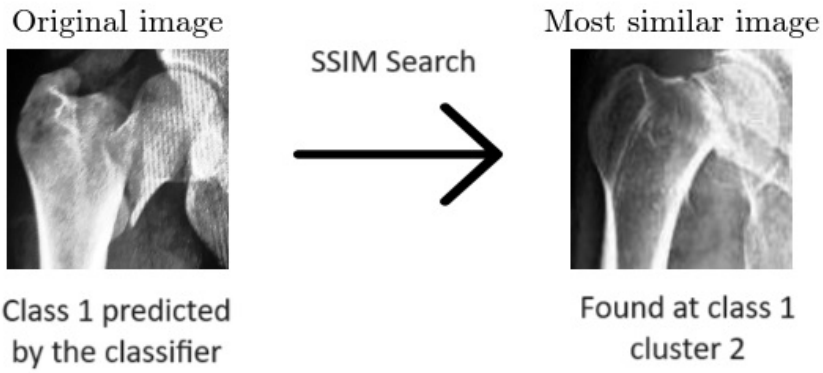

Figure 11: Most similar image for a given X-ray

Table 2 shows two different versions generated by GPT-4 of a case base text created by an expert to explain a basicervical femoral neck fracture. In the second column, the original text is presented, and in the fourth column, text versions for students and experts, respectively, generated by GPT-4 are shown. This approach helps us avoid repetitive explanations, as GPT-4 generates different texts each time. Furthermore, we can adapt the original text depending on the user, whether they are a student or an expert, thereby enhancing the usability of the system.

Expertise	Specific diagnosis	Prompt	Generated diagnosis
Student	A basicervical femoral neck fracture is a specific type of fracture that affects the base of the neck of the femur, at the junction with the trochanteric mass. It is a two-part fracture at the base of the femoral neck that was medial to the intertrochanteric line, exiting above the lesser trochanter but was more lateral than a classic transcervical fracture.	Given the following medical explanation: «**Specific diagnosis**», generate a text of EQUAL or SHORTER length that is suitable for a person with **Student** knowledge.	A basicervical femoral neck fracture is a specific type of fracture that affects the base of the neck of the femur, at the junction with the trochanteric mass. It is a two-part fracture at the base of the femoral neck that was medial to the intertrochanteric line, exiting above the lesser trochanter but was more lateral than a classic transcervical fracture.
Expert		Given the following medical explanation: «**Specific diagnosis**», generate a text of EQUAL or SHORTER length that is suitable for a person with **Expert** knowledge.	A basicervical femoral neck fracture is a break in the bone that occurs where the thigh bone's neck and its wider part at the top meet. Unlike a common neck fracture that happens further inside, this type takes place closer to the side, above a bony prominence but not as far to the side as some other types.

Table 2
Example prompts and generated diagnosis for different expertise levels.

5.4. Revise and Retain

In our approach, the revision phase is not merely a formality but a critical mechanism for refining and enhancing the performance of our CBR system. With an initial case base comprising only a handful of seed cases and a reuse stage reliant on LLMs, which may occasionally produce inaccurate textual descriptions, the revision stage emerges as a pivotal component of our methodology.

Our system has been designed with a user-centric approach, empowering experts and practitioners to engage in the learning process actively. One of the key features of our system is its capability to solicit alternative textual descriptions directly from users during the revision process. This functionality enables users to provide nuanced insights, corrections, or additional context that our system may have overlooked or inaccurately represented.

By incorporating this interactive feedback loop, we not only facilitate user involvement but also foster a process of continuous improvement. Every correction, refinement, or enhancement suggested by users serves as a valuable input for the system, contributing to its iterative learning process. Over time, as the system accumulates diverse user-contributed insights, it evolves into a more robust and adaptive tool for diagnosing hip fractures from X-ray images.

This way, the review stage serves as a cornerstone of our CBR approach, facilitating ongoing refinement and enhancement through active user engagement. By embracing user feedback and leveraging the collective intelligence of the medical community, we strive to create a system that not only meets the needs of its users but also evolves and adapts to deliver increasingly accurate and reliable diagnostic capabilities.

Due to the low number of seed cases in the current state of the case base, we allow the direct addition of new cases to the case base and we *retain* each revised case as a new case for future iterations. We plan to extend the research to evaluate how these new cases can affect the system behavior and other approaches for the retain phase. It is important to note that the inclusion of new cases potentially implies recalculating the clusters that organize the case base.

6. Conclusions and Future Work

Hip fractures on conventional radiographs pose a significant diagnostic challenge. This paper first presents a neural network model that can classify X-ray hip fractures accurately and focuses on the need for explanations to increase the trust and interpretability of such a black-box classifier. We propose a twin surrogate architecture to elucidate the classification model. This architecture comprises two parallel pipelines. The first pipeline conducts classification and provides a visual explanation using Grad-CAM. The second pipeline consists of a CBR system to retrieve prototypical explanations, followed by an LLM that adapts the resulting text to generate textual explanations.

We use a seed case base of prototypical X-ray images for hip fractures manually annotated with textual explanations by an expert. The CBR system retrieves a prototypical case and uses an LLM to generate extended textual explanations. This way, our explanation approach is three-fold: attribution-based (Grad-CAM), textual (LLMs), and explanation-by-example. While integrated gradients serve as a valuable tool for highlighting important regions in X-ray images, textual explanations offer a more comprehensive and personalized approach to understanding. By providing contextual insights, the interpretation process is more effective. Integrating both approaches can enhance the overall interpretation process, ensuring accurate diagnosis and informed decision-making in clinical practice.

As future work, as the case base grows, scaling up this CBR system poses several challenges and considerations. First of all, storing and managing a vast number of diagnostic texts requires robust infrastructure and efficient data-handling techniques. Therefore, it is essential to develop a method for selecting which textual explanations provided by experts are crucial for refining the CBR system or for combining similar cases. This selection process could involve prioritizing explanations based on their relevance to common diagnostic scenarios, their impact on improving model performance, or their alignment with evolving clinical practices. The use of different types of prompts and a deeper study on how to personalize the textual explanation to the query are both a challenge and an opportunity for the impact of synergies of CBR and LLMs in the eXplainable AI (XAI) landscape.

Furthermore, with an expanding dataset of diagnostic texts, another aspect to address is the difficulty of reorganizing clusters maintaining an efficient and meaningful clustering structure as new cases are added. For this reason, it would be necessary to decide when the image groups are recalculated to generate better results as the system is used.

Another challenge related to the dataset growth lies in the process of searching for similar images. Given an image, this CBR system has to compare it with all the others with the same label. Too much time could be spent on this task, so a way of speeding up this process becomes necessary.

On the other hand, collaboration between domain experts and programmers will be crucial in addressing errors generated by LLMs, which present a significant challenge. Despite their advanced capabilities, LLMs are not immune to inaccuracies or misinterpretations, so correcting these errors is essential for ensuring the reliability and accuracy of generated textual explanations. To make the most of experts' feedback and refine stored diagnostics, it's crucial to establish a method promptly once this CBR system gains widespread usage. In line with this, it's essential to share explanations with experts via meetings and surveys. This not only confirms the accuracy of the explanations but also ensures the effectiveness of our approach for patients. This step of the process is designed specifically for medical expert users capable of precisely drafting medical reports.

Acknowledgments

This work is supported by the PERXAI project PID2020-114596RB-C21 funded by MCIN/AEI/10.13039/501100011033 and the Complutense University of Madrid (Group 921330). Special thanks to the Orthopedic Surgery and Traumatology department at Hospital Universitario Virgen de la Victoria and to the Maurice E. Müller Foundation.

References

[1] L. Malburg, D. Verma (Eds.), Proceedings of the Workshops at the 31st International Conference on Case-Based Reasoning (ICCBR-WS 2023) co-located with the 31st International Conference on Case-Based Reasoning (ICCBR 2023), Aberdeen, Scotland, UK, July 17, 2023, volume 3438 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3438.

[2] D. B. Leake, D. McSherry, Introduction to the special issue on explanation in case-based reasoning, Artif. Intell. Rev. 24 (2005) 103–108. doi:10.1007/s10462-005-4606-8.

[3] J. D. Krogue, K. V. Cheng, K. M. Hwang, P. Toogood, E. G. Meinberg, E. J. Geiger, M. Zaid, K. C. McGill, R. Patel, J. H. Sohn, A. Wright, B. F. Darger, K. A. Padrez, E. Ozhinsky, S. Majumdar, V. Pedoia, Automatic hip fracture identification and functional subclassification with deep learning, Radiology: Artificial Intelligence 2 (2020) e190023. doi:10.1148/ryai.2020190023, pMID: 33937815.

[4] L. Tanzi, A. Audisio, G. Cirrincione, A. Aprato, E. Vezzetti, Vision transformer for femur fracture classification, Injury 53 (2022) 2625–2634. doi:https://doi.org/10.1016/j.injury.2022.04.013.

[5] N. Twinprai, A. Boonrod, A. Boonrod, J. Chindaprasirt, W. Sirithanaphol, P. Chindaprasirt, P. Twinprai, Artificial intelligence (ai) vs. human in hip fracture detection, Heliyon 8 (2022) e11266. doi:10.1016/j.heliyon.2022.e11266.

[6] Y. Gao, N. Y. T. Soh, N. Liu, G. Lim, D. Ting, L. T.-E. Cheng, K. M. Wong, C. Liew, H. C. Oh, J. R. Tan, N. Venkataraman, S. H. Goh, Y. Y. Yan, Application of a deep learning algorithm in the detection of hip fractures, iScience 26 (2023) 107350. doi:https://doi.org/10.1016/j.isci.2023.107350.

[7] C. Lee, J. Jang, S. Lee, Y. S. Kim, H. J. Jo, Y. Kim, Classification of femur fracture in pelvic x-ray images using meta-learned deep neural network, Scientific reports 10 (2020) 13694.

[8] M. Adams, W. Chen, D. Holcdorf, M. W. McCusker, P. D. Howe, F. Gaillard, Computer vs human: Deep learning versus perceptual training for the detection of neck of femur fractures, Journal of Medical Imaging and Radiation Oncology 63 (2019) 27–32. doi:https://doi.org/10.1111/1754-9485.12828.

[9] B. Kim, C. Rudin, J. Shah, The bayesian case model: A generative approach for case-based reasoning and prototype classification, 2015. arXiv:1503.01161.

[10] R. C. Schank, D. B. Leake, Creativity and learning in a case-based explainer, Artificial Intelligence 40 (1989) 353–385. URL: https://www.sciencedirect.com/science/article/pii/0004370289900532. doi:https://doi.org/10.1016/0004-3702(89)90053-2.

[11] O. Li, H. Liu, C. Chen, C. Rudin, Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions, 2017. arXiv:1710.04806.

[12] E. M. Kenny, M. T. Keane, Explaining deep learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI, Knowl. Based Syst. 233 (2021) 107530. doi:10.1016/J.KNOSYS.2021.107530.

[13] C. Ford, M. T. Keane, Explaining classifications to non-experts: An XAI user study of post-hoc explanations for a classifier when people lack expertise, in: J. Rousseau, B. Kapralos (Eds.), Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges - Montreal, QC, Canada, August 21-25, 2022, Proceedings, Part III, volume 13645 of *LNCS*, Springer, 2022, pp. 246–260. URL: https://doi.org/10.1007/978-3-031-37731-0_15.

[14] E. G. Meinberg, J. Agel, C. S. Roberts, M. D. Karam, J. F. Kellam, Fracture and dislocation classification compendium—2018, Journal of orthopaedic trauma 32 (2018) S1–S10.

[15] T. v8, Proximal femur fracture detection and classification dataset, https://universe.roboflow.com/thesisyolo-v8/proximal-femur-fracture-detection-and-classification, 2023. URL: https://universe.roboflow.com/thesisyolo-v8/proximal-femur-fracture-detection-and-classification, visited on 2024-03-23.

[16] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, E. Kasneci, A consistent and efficient evaluation strategy for attribution methods 162 (2022) 18770–18795. URL: https://proceedings.mlr.press/v162/rong22a.html.

[17] S. Suara, A. Jha, P. Sinha, A. A. Sekh, Is grad-cam explainable in medical images?, CoRR abs/2307.10506 (2023). doi:10.48550/ARXIV.2307.10506. arXiv:2307.10506.

[18] Zhou Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612. doi:10.1109/TIP.2003.819861.