AI for Referable Knee Radiograph Detection in Primary Care: A Pathway-Specific Taxonomy for Dataset Generation from Reports

[1,2,*]Imanol Pinto, [2]Álvaro Olazarán, [2]David Jurio, [2]Miguel Sainz, [4]Natalia Álvarez, [1,3]Mikel Galar

[1]Institute of Smart Cities, Public University of Navarre, Campus Arrosadia, 31006 Pamplona, España

[2] General Directorate of Telecommunications and Digitalization, Government of Navarre, C/ Cabarceno 6, Sarriguren, 31621, Spain

[3]IdiSNA, Navarre Institute of Health Research, C/ Irunlarrea 3, Pamplona, 31008, Spain

[4]University Hospital of Navarre, Musculoskeletal Radiology Service, C/ Irunlarrea 3, Pamplona 31008, Spain

[*]Corresponding author. E-mail: imanolpinto@proton.me

Abstract—**Radiologist shortages often delay musculoskeletal radiograph interpretation in primary care, with the knee among the most frequently imaged regions. To our knowledge, we developed the first Artificial Intelligence (AI) system for knee radiograph assessment in this setting, using a pathway-specific taxonomy to automatically label reports, thereby advancing the field of medical disease extraction. A total of 57,460 knee radiology reports (2010–2024) from the Public Healthcare Service of Navarre, Spain, were retrospectively processed with a Natural Language Processing (NLP) pipeline guided by this taxonomy. The pipeline extracted thousands of frequent and relevant findings, reliably linked them to images through a projection and laterality classifier, and organized them into 43 hierarchical categories for dataset generation. To assess feasibility, we trained a ConvNeXt-Small model to classify radiographs as referable (requiring specialist review) or non-referable, and validated it on an independent test set of 494 studies (39.7% referable). Ground truth was defined by consensus of three radiologists from a panel of five experts. On this test set, the model achieved an Area Under the ROC Curve (AUC) of 0.880 (95% Confidence Interval [CI]: 0.843–0.915), with 81.9% sensitivity and 83.1% specificity, significantly outperforming routine reports (AUC 0.798; p=0.0002). Compared with the individual radiologists, the model achieved comparable sensitivity but lower specificity (3.6–11% below radiologists). These results support the potential of our deep learning algorithm as a primary care decision-support tool, helping reduce unnecessary referrals and radiologist workload, while showing how pathway-specific taxonomies enable scalable, efficient AI in data-limited settings.**

Keywords-Deep Learning for radiology; Medical disease extraction; AI-based health systems and applications; Knee Radiography; Primary Care

I. INTRODUCTION

Musculoskeletal (MSK) conditions are one of the most frequent reasons for patient visits in primary care, where plain radiography constitutes the primary modality for initial imaging assessment [1][2]. Nevertheless, delays in expert radiological interpretation represent an obstacle in many healthcare systems, often postponing diagnosis and subsequent treatment. Therefore, addressing these delays is essential for improving administrative efficiency in primary care. In this context, general practitioners frequently assume responsibility for preliminary image review despite lacking specialized training to identify the full spectrum of MSK conditions.

Recent advances in Artificial Intelligence (AI), particularly in Deep Learning (DL), have demonstrated strong performance in the analysis of MSK radiographs [3]. Building on this progress, we present an AI-based framework designed specifically for the evaluation of knee radiographs in primary care —one of the most frequently imaged anatomical regions in this clinical context, second only to the spine [1][2]. Knee radiographs offer an ideal starting point for AI-based decision support due to their relatively simple anatomical structure and standardized diagnostic criteria, in contrast to the spine, which presents greater complexity both anatomically and diagnostically [4][5].

Despite this opportunity, there are no studies focusing on AI tools for preliminary interpretation of knee radiographs in primary care settings. Most existing research concentrates on specific MSK disorders, such as osteoarthritis [6], fractures [7], bone age [8] and tumors [9], using manually annotated datasets and specialist-oriented labels [10].

To address this gap, we implemented a data pipeline that takes advantage of routinely collected clinical data from the Public Healthcare Service of Navarre (Servicio Navarro de Salud–Osasunbidea, SNS-O) to create a large-scale training set. The pipeline uses an image classifier to reliably identify target knee radiographs, and applies Natural Language Processing (NLP) to radiology reports to extract clinical findings, which are then organized into categories using a custom taxonomy that reflects primary care decision-making needs.

We assessed the effectiveness of this approach by training a DL model on the resulting dataset and evaluating it on a radiologist-annotated test set, achieving promising results.

To our knowledge, this is the first AI system specifically designed for preliminary knee radiograph assessment in primary care. In this context, emulating the diagnostic approach of a radiologist, the proposed AI system analyzes a broad range of radiological findings. Our methodological framework advances the field of medical disease extraction and illustrates how NLP-based labeling, guided by a taxonomy aligned with the clinical pathway, can enable the development of AI-based health systems and applications in settings with limited annotated data.

This paper is organized as follows: Section II reviews related work, Section III describes the methodology and dataset creation, Section IV presents validation results, and Section V discusses conclusions, limitations, and future work.

Copyright (c) IARIA, 2026.   ISBN: 978-1-68558-362-0

II. RELATED WORK

The application of AI to MSK radiology has advanced rapidly, particularly for trauma detection [7], bone-age assessment [8], osteoarthritis grading [6] and implant evaluation [10]. Meta-analyses and large studies report high diagnostic accuracy for fracture detection on plain radiographs (pooled sensitivities/specificities often >91%) and demonstrate that AI assistance can improve clinician performance in routine practice [7][11][12]. For the knee, DL models have achieved expert-level performance in Kellgren–Lawrence (KL) osteoarthritis grading, with reported AUCs up to $\approx$ 0.96 and Cohen's kappa around 0.86 [6][13]. While large MSK repositories exist, such as MURA for upper limbs [14] and the Osteoarthritis Initiative for knee osteoarthritis [15], to our knowledge, no comparable large-scale, multi-label (multiple findings per image) knee datasets are currently available.

To address the scarcity of annotated data, prior thoracic imaging studies have applied NLP to radiology reports to generate large-scale, weakly supervised labels (e.g., PadChest [16], CheXpert [17]). Since chest imaging taxonomies cannot be directly applied to knee radiographs, we developed a novel taxonomy for primary care, grounded in real-world finding prevalence and clinical relevance. This taxonomy enables the construction of large-scale, multi-label knee datasets and the development of AI models tailored to general practitioners' needs.

III. METHODS

This study focuses on developing and evaluating an AI framework for knee radiograph assessment in primary care. Rather than aiming for maximal model performance, our primary goal was to demonstrate the feasibility of creating a clinically meaningful decision-support system from automatically labeled retrospective data.

A. *Dataset Creation*

The SNS-O comprises 63 primary care centers and 3 hospitals, serving a population of over 600,000 patients. From its radiology information system, we retrieved 263,763 knee radiograph studies requested by general practitioners (in primary care) between 2010 and 2024, of which 57,460 (21%) had an associated formal radiology report and were available for analysis.

Each study typically contained multiple radiograph images of one or both knees, often captured in multiple projections (anteroposterior, lateral, axial and others) and of varying laterality (right, left, or bilateral), and interpreted by a radiologist in a single report. To transform this unstructured material into an AI-ready dataset, three tasks were required: (1) mapping radiological findings from the free-text reports into structured categories, (2) reliably identifying projection and laterality for each image, and (3) linking each image with its corresponding findings.

Accordingly, as illustrated in Figure 1, we developed a data pipeline with two main branches: label generation and image selection. Label generation involved the use of NLP techniques to extract radiological findings from reports, which were selected and organized using our primary care–oriented taxonomy, detailed in Section III-B. As image metadata was usually missing, we developed a neural network to perform image selection, specifically to determine projection and laterality.

The resulting dataset comprised 63,976 single-knee radiograph studies from 28,719 patients (54% female; mean age 58.2 years). For each knee, findings were mapped to a category within our taxonomy and assigned a global grade of *referable* or *non-referable*, indicating whether expert radiologist evaluation was recommended. The pipeline yielded a dataset of 56,152 non-referable studies (87.77%) and 7,824 referable studies (12.23%), each consisting of a single knee report linked to its corresponding radiographs.

Figure 2 illustrates an example study alongside its radiology report (translated to English), extracted finding categories, and identified projections and laterality. In this case, since the study involved only a single knee, no laterality-based splitting was required. The two core components of the pipeline (label generation and image selection) are described below.

Label generation. Our label generation pipeline processed each report in four main steps: (1) laterality analysis, splitting reports that described both knees into separate reports, one per knee; (2) biomedical named entity recognition, extracting all mentioned medical entities; (3) negation detection, discarding entities appearing in a negated context; and (4) taxonomy mapping, filtering entities, assigning them to relevant finding categories, and generating a final binary label (referable/non-referable).

Reports were split by laterality using regular expressions, since most included distinct sections for each knee. Single-laterality reports were left intact. Entity recognition used a Spanish RoBERTa-base biomedical model [18], while negation detection relied on a fine-tuned BERT-base multilingual model trained on the NUBES Spanish clinical dataset [19][20].

When applied to our retrospective dataset, the NLP pipeline extracted 102,127 entity mentions (6,676 unique entities). Of these, 87,189 (85%) were successfully mapped using the taxonomy detailed in Section III-B. A report was labeled as referable if at least one referable finding category was present.

Image selection. Most radiographs lacked reliable DICOM metadata to accurately determine laterality and projection. To address this, we manually annotated 2,214 knee radiographs, assigning one of the following nine labels: anteroposterior (right/left), lateral (right/left), axial (right/left), bilateral, biaxial, or other. We then trained a ConvNeXt-Tiny model on this task using Fast.ai and Timm libraries [21][22], with ImageNet pretraining and input images resized to 350×350px. The network was trained for 22 epochs with a learning rate of 0.006, using standard augmentations, such as resizing, brightness/contrast adjustments, random erasing, and geometric transformations. Horizontal and vertical flips were deliberately excluded to preserve anatomical laterality. The model achieved an AUC of 0.9984 and an accuracy of 97.4%,

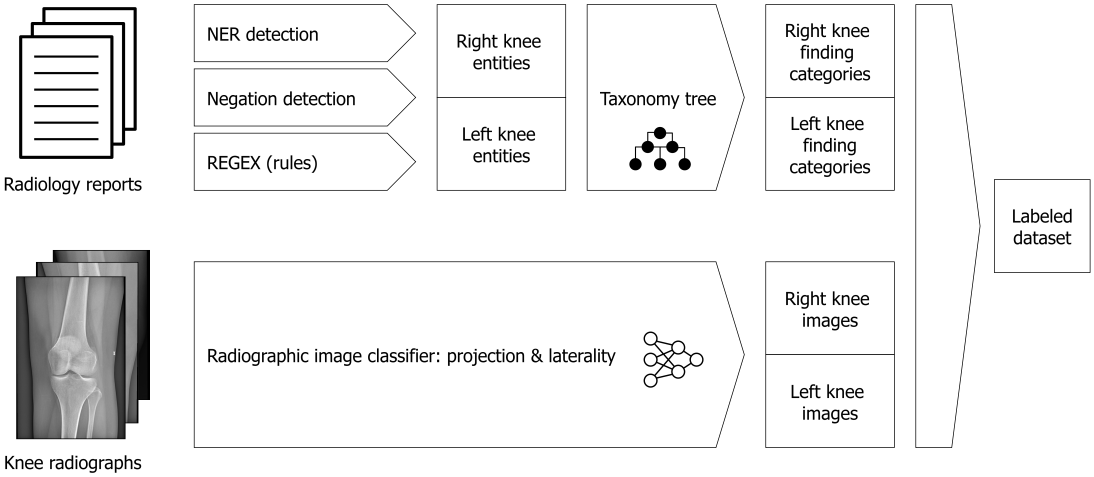
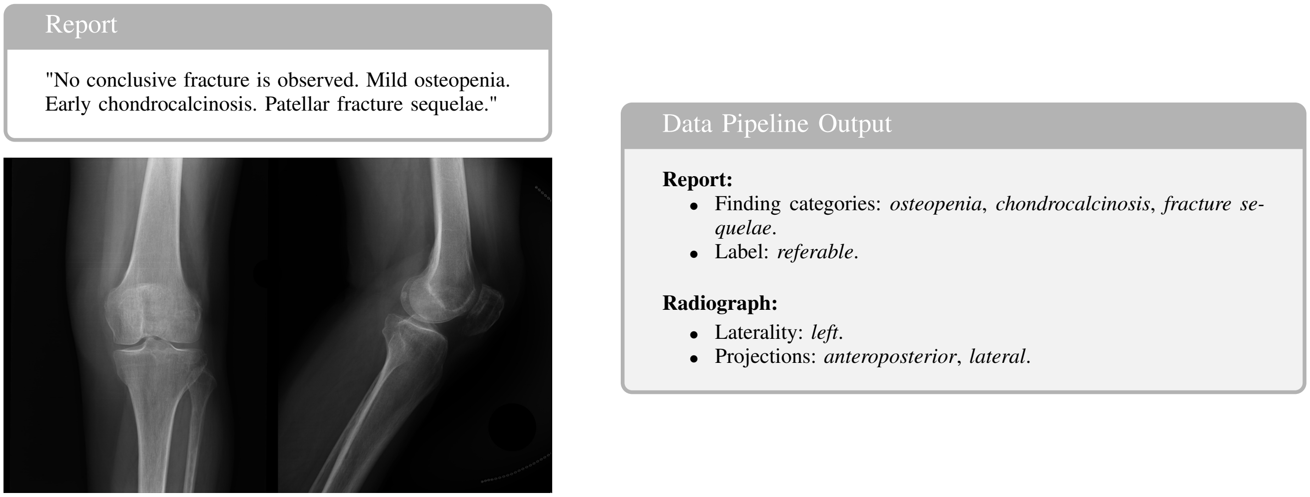


Figure 1. Data pipeline for dataset creation. The NLP pipeline (top) and the image selection (bottom) branches are merged to produce our dataset.



Report

"No conclusive fracture is observed. Mild osteopenia. Early chondrocalcinosis. Patellar fracture sequelae."

Data Pipeline Output

Report:
- Finding categories: *osteopenia*, *chondrocalcinosis*, *fracture sequelae*.
- Label: *referable*.

Radiograph:
- Laterality: *left*.
- Projections: *anteroposterior*, *lateral*.

Figure 2. Example knee study with corresponding radiology report and two projections (left). The outputs of the data pipeline are also displayed (right).

enabling reliable assignment of radiographic laterality and projection across the dataset. For the purposes of this study, only anteroposterior and lateral projections were retained, as they are the standard views for clinical knee assessment. When a radiograph was classified as bilateral, the image was divided in two standard projections.

B. A Taxonomy for Primary Care Knee Radiography

A radiographic finding taxonomy was developed to categorize the reports. The complete taxonomy is presented in Figure 3, which shows the resulting finding categories and their relative frequencies, selected based on entity frequency and clinical relevance in primary care.

Finding categories were initially defined by a senior MSK radiologist based on SNS-O clinical pathways and subsequently reviewed and refined within a multidisciplinary workgroup comprising five MSK radiologists and one senior general practitioner. Categories were retained if they appeared in at least 0.9% of the dataset or posed significant clinical risk. Findings were classified as referable when they typically required radiologist review, advanced imaging, or management changes beyond primary care. In the absence of formal guidelines, decisions were guided by routine SNS-O practice to safely reduce radiologist workload.

Using this taxonomy, extracted clinical entities were consolidated into broader categories and classified as referable or non-referable, providing practical guidance for general practitioners (e.g., “chondrocalcinosis” and “calcification” were mapped to non-specific calcification, classified as non-referable). Detailed entity mapping is provided in the Supplementary Material.

C. Model Training

To evaluate the utility of the generated dataset, we trained a baseline ConvNeXt-Small model to classify knee radiographs as referable or non-referable. The dataset included 62,309 anteroposterior and lateral radiographs, each labeled according to the overall referability grade of its corresponding knee. A

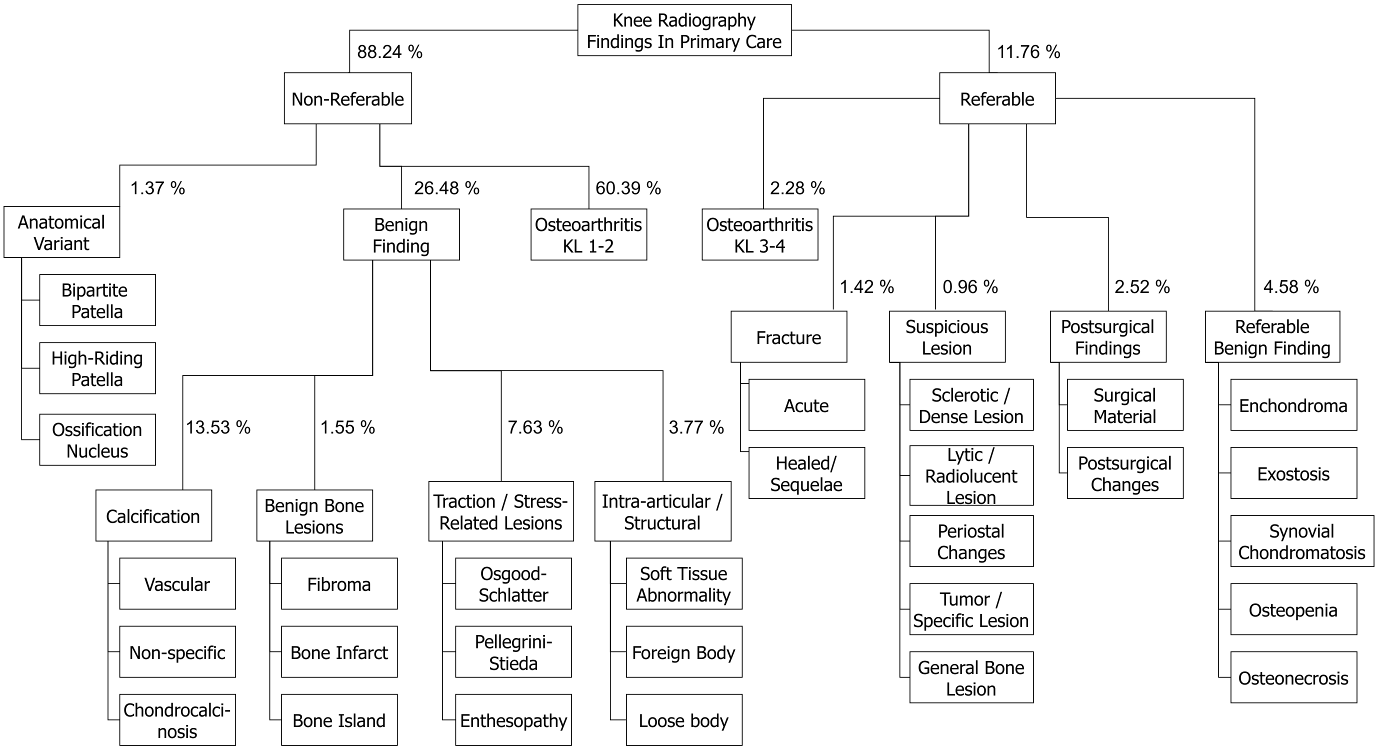


Figure 3. Taxonomy tree of knee radiological findings in primary care, with relative frequencies calculated from the 87,189 mapped entities.

validation set of 10,711 images was reserved, ensuring patient-level separation. During DICOM-to-JPG conversion, we applied the windowing configuration stored in each DICOM study, to preserve clinically optimal brightness and contrast.

Leveraging the Fast.ai library [21], the model was initialized with ImageNet weights and trained on 900×900 padded images using cross-entropy loss, default augmentations, and a learning rate of 0.01. The best performance was achieved at epoch 6, with an AUC of 0.8413 and an accuracy of 89.85% on the validation set.

We tried several variations in architecture, image size, learning rate, and training duration, but none yielded meaningful improvements. We chose a ConvNeXt-Small architecture for its strong ImageNet results, fast inference, and modern design, serving as a robust baseline. Further gains are expected with targeted model refinement, improved preprocessing and enhanced training strategies tailored to our dataset.

D. Reference Standard Test Set

We created a random test set of 494 radiographic single-knee studies (anteroposterior and lateral images), enriched to increase the proportion of referable knees from 11.60% to 39.70%, ensuring one knee per patient and strict patient-level separation to prevent data leakage. Each study was independently reviewed by three expert radiologists drawn from a pool of five, with each radiologist reviewing a different but overlapping random test set partition (around 250 studies). Reviewers identified all relevant findings according to the predefined categories of the taxonomy (Section III-B), and a study was deemed referable if any finding mapped to a referable category.

The image model was evaluated against the three-radiologist majority vote, which served as the clinical reference standard, and was also compared with each individual radiologist, using the majority vote of the other two as ground truth (excluding cases where the two remaining radiologists disagreed). In parallel, all test set reports were manually annotated by engineers based solely on their clinical text, providing the reference standard for evaluating the NLP pipeline.

IV. RESULTS

A. NLP pipeline validation

The NLP pipeline, when evaluated against the manually annotated reports from the test set, achieved micro- and macro-averaged F1-Scores of 0.8968 and 0.9219, respectively, across all finding categories. The lowest performance was observed for osteoarthritis KL 3–4 category (a referable category), with an F1-Score of 0.1714, while all other findings achieved F1-Scores above 0.7368. Detailed per-category results are provided in the Supplementary Material.

Error analysis revealed that misclassifications mainly arose from complex degenerative cases where the KL grade was not explicitly indicated in the report. For instance, reports often document "gonarthrosis" without explicit grading, despite providing a descriptive context that implies advanced osteoarthritis. In such cases, the pipeline tended to assign a KL 1–2 label, although these reports would likely be interpreted by a radiologist as KL 3–4. This systematic bias was more frequent in the enriched test set, where KL grades were less often documented explicitly, than in the training set. Nevertheless, many KL 3–4 cases were correctly captured when the grade was explicitly stated, allowing the trained image model to successfully classify most KL 3–4 radiographs as referable in the test set (see Section IV-B).

B. Model validation

Image classification performance was evaluated against the majority vote of the radiologists (Table I). The model achieved a higher AUC than the manually annotated reports (0.8800 vs. 0.7983), with significance confirmed by 100,000 paired bootstrap iterations (p = 0.00019). At a matched sensitivity of 81.94%, the model achieved higher specificity than the reports, reaching 83.14% compared to 77.71%.

TABLE I. PERFORMANCE ON THE TEST SET FOR CLASSIFYING KNEE RADIOGRAPHS AS REFERABLE OR NON-REFERABLE.

Source	AUC	Kappa	Sensitivity[2]	Specificity
Model	0.8800 0.8434–0.9147	0.6098 0.5428–0.6778	81.94% 75.00–87.52	83.14% 79.14–86.86
Reports[1]	0.7983 0.7560–0.8361	0.5393 0.4669–0.6185	81.94% 75.69–88.19	77.71% 73.43–82.00

The ground truth was obtained with the majority vote of three expert radiologists. All confidence intervals are 95%, estimated via 1,000-iteration bootstrap resampling with replacement.

[1] Radiology reports, produced during routine clinical practice, were manually labeled here without image review.

[2] The model's operating point was adjusted to match the sensitivity of the radiology reports.

At a more granular level, Table II breaks down detection performance on the test set by grouped finding categories, where recalls reflect the proportion of correctly identified studies. Both the model and reports show minor differences identifying referable findings, and notably, they both detected all six suspicious lesions (osteosarcoma and periosteal reactions). However, the model was more effective at identifying normal cases, achieving higher recall for these than the reports (95% vs. 85%).

TABLE II. COMPARISON OF MODEL AND MANUALLY ANNOTATED REPORTS IN THE TEST SET, STRATIFIED BY FINDING CATEGORY.

Finding Category[1]	**Total**	**Model Recall**	**Report Recall**
Non-referable			
Benign finding	272	55.15%	52.57%
Osteoarthritis (KL 1–2)	208	66.83%	65.39%
Anatomical variant	145	66.21%	64.83%
No finding / Normal	80	95.00%	85.00%
Referable			
Osteoarthritis (KL 3–4)	59	83.05%	76.27%
Acute fracture	20	80.00%	90.00%
Suspicious lesion	6	100.00%	100.00%
Referable benign finding	24	70.83%	70.83%
Postsurgical findings	47	100.00%	97.87%
Healed / Sequelae fracture	9	100.00%	88.89%

[1] Please refer to the taxonomy tree (Figure 3) for a detailed decomposition.

Then, the model's performance was compared with individual radiologists and the manually annotated reports (Table III). Across all partitions (i.e., the subsets of studies reviewed by each radiologist), it consistently outperformed the reports in AUC, sensitivity, and specificity. Compared with individual radiologists, the model reached similar or slightly higher sensitivity (up to +14%), but generally lower specificity (3.6–11% below) and kappa (0.66–0.72 vs. 0.70–0.86). As shown in Figure 4, its ROC curves indicated strong performance, though still below 3 of 5 radiologists.

These results prompted a detailed error analysis, in which we reviewed each discordant case by reading the original reports and comparing them with the model outputs. We found that model errors were usually straightforward misclassifications—such as false negatives of fractures or positives for benign calcifications. In contrast, report errors often stemmed from textual ambiguity, incomplete descriptions or lack of sufficient clinical context, which either prevented the reader from extracting findings or rendered them too ambiguous to be reliably mapped into the taxonomy. Therefore, despite the measured metrics, the nature of the errors between the model and the reports differed.

V. CONCLUSION AND FUTURE WORK

We introduced a novel system for AI-assisted general interpretation of knee radiographs in primary care, leveraging automatically generated labels from real-world radiology reports through NLP. To support this, we developed a pathway-specific taxonomy that organizes findings into 43 categories and evaluated the resulting model against a reference test set annotated by expert radiologists. Importantly, the taxonomy was grounded in retrospective finding prevalence, ensuring the meaningfulness of both the categories and the resulting labels.

The model, trained to classify a knee radiograph as referable or non-referable, outperformed manually annotated routine radiology reports (AUC 0.880 vs. 0.798) and successfully referred all suspicious lesions—the most critical referable cases—while maintaining high specificity. Although its overall performance was generally below that of individual expert radiologists, it achieved comparable agreement in certain cases, suggesting its potential as a decision-support tool for general practitioners in primary care.

Error analysis showed that most report-related errors stemmed from the inherent challenges of clinical text interpretation (ambiguity, incomplete descriptions, or lack of context), while model errors reflected more straightforward misclassifications. In fact, our NLP pipeline, based on entity recognition and regular expressions, particularly struggled to assign osteoarthritis grades when they were not explicitly stated—introducing a performance ceiling for the model. These limitations highlight opportunities for improvement through advanced NLP methods, such as recent large language models [23].

While the test set was enriched to 39.70% referable cases to ensure statistical power for rare pathologies, we acknowledge that in a natural primary care prevalence (12.23%), the 3.6–11% specificity gap compared to radiologists might lead to increased false-positive referrals. Future deployment would require calibration and a configurable threshold to prioritize

TABLE III. PERFORMANCE COMPARISON PER INDIVIDUAL RADIOLOGIST ON TEST SET PARTITIONS.

	Radiologist 1 partition 252 knees, 27.8% referable			**Radiologist 2 partition** 246 knees, 24.4% referable			**Radiologist 3 partition** 259 knees, 23.9% referable			**Radiologist 4 partition** 253 knees, 30.8% referable			**Radiologist 5 partition** 250 knees, 22.4% referable		
	AUC	Kappa	Sens / Spec	AUC	Kappa	Sens / Spec	AUC	Kappa	Sens / Spec	AUC	Kappa	Sens / Spec	AUC	Kappa	Sens / Spec
Radiologist	0.9220	0.8591	87.14 / 97.25	0.9204	0.8178	90.00 / 94.09	0.8826	0.6995	88.71 / 87.82	0.8248	0.7073	66.67 / 98.29	0.9167	0.7836	91.07 / 92.27
Model	0.9175	0.6963	88.57 / 86.26	0.9429	0.6883	88.33 / 87.10	0.9231	0.6553	90.32 / 84.26	0.9049	0.7220	80.77 / 91.43	0.9292	0.6902	89.29 / 87.63
Reports	0.8313	0.5877	87.14 / 79.12	0.8312	0.5664	86.67 / 79.57	0.8450	0.5755	90.32 / 78.68	0.8367	0.6393	83.33 / 84.00	0.8536	0.5941	89.29 / 81.44

Each radiologist annotated a different partition of the test set, resulting in six overlapping partitions whose union constitutes the entire test set.
For each partition, the ground truth was defined by the majority vote of the other two radiologists; samples without consensus were excluded.
Sens / Spec denote sensitivity and specificity, respectively; values are reported as percentages.

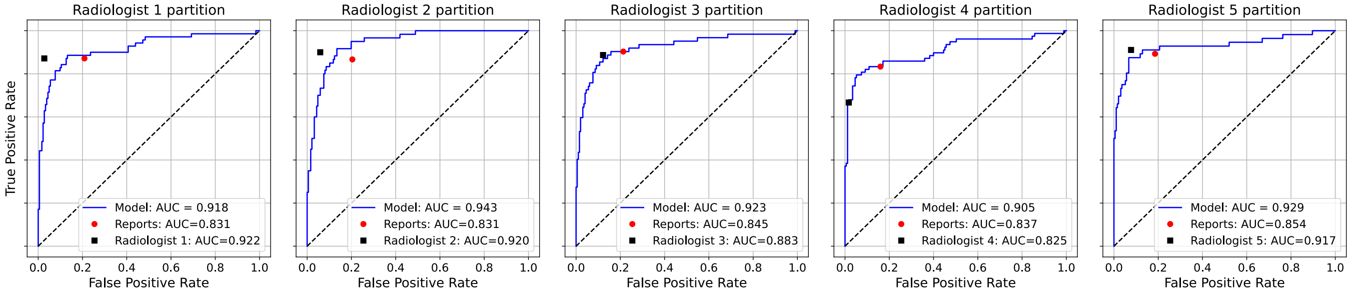


Figure 4. ROC curve of the model on the test set partitions. The operating points of radiologists and manually labeled radiology reports are overlaid.

either high-sensitivity screening or high-specificity workload reduction.

As a first step toward clinical use, we aimed to provide general practitioners with a tool that could help avoid unnecessary referrals, reduce radiologist workload, and shorten diagnostic times. We acknowledge that this approach simplifies decision-making in primary care by grouping important findings together, regardless of urgency or pathway—for example, advanced osteoarthritis and suspicious lesions are both classified as referable. Nevertheless, our results illustrate the feasibility of developing clinically meaningful AI systems based on NLP pipelines guided by pathway-specific taxonomies, offering a scalable strategy to extend AI to other body regions or imaging modalities, especially in contexts with limited labeled data but abundant radiology reports.

This study has several limitations: the small test set size and rarity of many findings limited per-finding evaluation; no external dataset was available to assess generalizability; weak labels derived from radiology reports may miss visible but unreported findings; plain radiographs have inherent diagnostic limitations; and potential selection bias arises from including only studies requested by general practitioners.

Future work will focus on developing multilabel models to stratify findings and enhance explainability, further improving the NLP pipeline, expanding the test set, and using pseudo-labeling to incorporate images without radiology reports. Prospective, multicenter, and multimodal studies will be essential to evaluate real-world impact and ensure safe integration into clinical workflows.

ETHICS STATEMENT

This work was approved by the SNS-O. No patient consent was required, as all clinical data were fully anonymized.

ACKNOWLEDGMENTS

This work was made possible thanks to the contributions of Borja de la Osa, Aritz Oscoz, Jimena Angarita, Marta Tirapu, Ana Andueza and Ainhoa Ovelar.

REFERENCES

[1] R. Haas *et al.*, “Prevalence and characteristics of musculoskeletal complaints in primary care: An analysis from the population level and analysis reporting (POLAR) database,” *BMC Primary Care*, vol. 24, no. 1, pp. 1–10, 2023.

[2] K. P. Jordan *et al.*, “Annual consultation prevalence of regional musculoskeletal problems in primary care: An observational study,” *BMC Musculoskeletal Disorders*, vol. 11, no. 1, pp. 144–152, 2010.

[3] F. C. Oettl *et al.*, “Artificial intelligence-assisted analysis of musculoskeletal imaging—a narrative review of the current state of machine learning models,” *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 33, no. 1, pp. 24–38, 2025.

[4] G. U. Kim, M. C. Chang, D. H. Sung, J. B. Song, and H. J. Park, “Diagnostic modality in spine disease: A review,” *Asian Spine Journal*, vol. 14, no. 6, pp. 910–920, 2020.

[5] S. Newman, H. Ahmed, and N. Rehmatullah, “Radiographic vs. MRI vs. arthroscopic assessment and grading of knee osteoarthritis—are we using appropriate imaging?” *Journal of Experimental Orthopaedics*, vol. 9, no. 1, pp. 2–11, 2022.

[6] K. A. Thomas *et al.*, “Automated classification of radiographic knee osteoarthritis severity using deep neural networks,” *Radiology: Artificial Intelligence*, vol. 2, no. 2, pp. 2638–6100, 2020.

[7] A. Nowroozi *et al.*, “Artificial intelligence diagnostic accuracy in fracture detection from plain radiographs and comparing it with clinicians: A systematic review and meta-analysis,” *Clinical Radiology*, vol. 79, no. 8, pp. 579–588, 2024.

[8] A. A. Bajjad, F. S. Al-Shehri, and S. M. Al-Malki, “Artificial intelligence in bone age assessment of healthy individuals: A scoping review,” *Journal of the World Federation of Orthodontists*, vol. 13, no. 2, pp. 95–102, 2024.

[9] C. E. von Schacky *et al.*, “Multitask deep learning for segmentation and classification of primary bone tumors on radiographs,” *Radiology*, vol. 301, no. 2, pp. 398–406, 2021.

ISBN: 978-1-68558-362-0

[10] S. Gitto *et al.*, "AI applications in musculoskeletal imaging: A narrative review," *European Radiology Experimental*, vol. 8, no. 1, pp. 22–35, 2024.

[11] R. Lindsey *et al.*, "Deep neural network improves fracture detection by clinicians," *Proceedings of the National Academy of Sciences*, vol. 115, no. 45, pp. 11 591–11 596, 2018.

[12] P. G. Anderson *et al.*, "Deep learning assistance closes the accuracy gap in fracture detection across clinician types," *Clinical Orthopaedics and Related Research*, vol. 481, no. 3, pp. 580–588, 2023.

[13] A. Tiulpin and S. Saarakkala, "Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks," *Diagnostics*, vol. 10, no. 11, pp. 932–945, 2020.

[14] P. Rajpurkar *et al.*, "MURA: Large dataset for abnormality detection in musculoskeletal radiographs," *arXiv preprint arXiv:1712.06957*, pp. 1–11, 2018.

[15] F. Eckstein, W. Wirth, and M. C. Nevitt, "Recent advances in osteoarthritis imaging: The osteoarthritis initiative," *Nature Reviews Rheumatology*, vol. 8, no. 10, pp. 622–630, 2012.

[16] A. Bustos, A. Pertusa, J. M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest x-ray image dataset with multi-label annotated reports," *Medical Image Analysis*, vol. 66, p. 101 797, 2020.

[17] J. Irvin *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 590–597, 2019.

[18] C. P. Carrino *et al.*, "Pretrained biomedical language models for clinical NLP in Spanish," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, Association for Computational Linguistics, 2022, pp. 193–199.

[19] A. J. Tamayo, *Negation scope detection in spanish clinical texts using mBERT fine-tuned on the NUBEs dataset*, https://github.com/ajt/NegScope, 2025.

[20] S. Lima López, N. Perez, M. Cuadros, and G. Rigau, "NUBes: A corpus of negation and uncertainty in spanish clinical texts," in *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, 2020, pp. 5772–5781.

[21] J. Howard and S. Gugger, "Fastai: A layered API for deep learning," *Information*, vol. 11, no. 2, pp. 108–122, 2020.

[22] R. Wightman, *Pytorch image models*, https://github.com/huggingface/pytorch-image-models, version 1.2.2, 2019.

[23] S. H. Kim *et al.*, "Benchmarking the diagnostic performance of open-source LLMs in 1,933 Eurorad case reports," *npj Digital Medicine*, vol. 8, no. 1, pp. 1–9, 2025.