AI for Referable Knee Radiograph Detection in Primary Care

A Pathway-Specific Taxonomy for Dataset Generation from Reports

Imanol Pinto, Álvaro Olazarán, David Jurio, Miguel Sainz, Natalia Álvarez, Mikel Galar

Presenter: Imanol Pinto

Institute of Smart Cities, Public University of Navarre (UPNA)
General Directorate of Telecommunications and Digitalization, Government of Navarre

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Imanol Pinto López

AI Engineer & Researcher
PhD Student

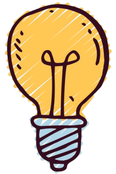
Contact: imanolpinto@proton.me

Imanol is a Software Engineer (UPNA, 2007) with over 15 years of experience in **software development**, specializing in **healthcare AI** since 2018. He has lead a team developing custom AI solutions for the Navarre Health Service (SNS-O) and is currently pursuing a PhD in AI applied to healthcare.

He has **developed two clinical AI systems** —NaIA-RD, NaIA-DMAE— which are currently used at the University Hospital of Navarre to improve clinical decision-making. His work has received multiple honors, including awards from Novartis (2022), Inforsalud (2023), and Roche Startup Creasphere (2024).

Motivation

1. High demand & general practitioner training gap [1][2].
2. Radiologist scarcity.
3. Musculoskeletal knee radiographs are an ideal AI entry point [3].



→ **AI for Referable Knee Radiograph Detection in Primary Care**

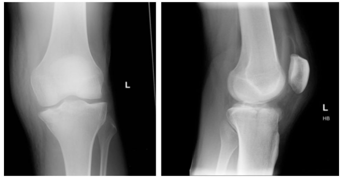
[1] R. Haas et al., “Prevalence and characteristics of musculoskeletal complaints in primary care: An analysis from the population level and analysis reporting (POLAR) database,” BMC Primary Care, vol. 24, no. 1, pp. 1–10, 2023.

[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.

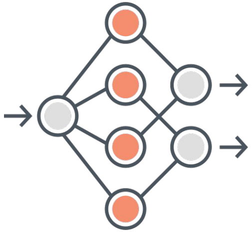
[3] F. C. Oettl et al., “Artificial intelligence-assisted analysis of musculoskeletal imaging—a narrative review of the current state of machine learning models,” Knee Surgery, Sports Traumatology, Arthroscopy, vol. 33, no. 1, pp. 24–38, 2025

A Decision-Support System for Primary Care

Plain Knee Radiograph

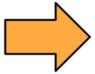


AI System



Referable

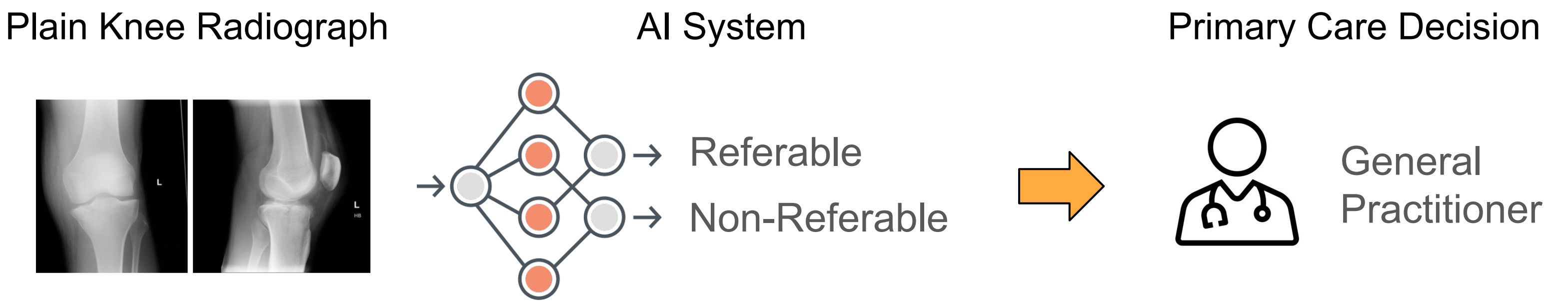
Non-Referable



Primary Care Decision

General Practitioner

A Decision-Support System for Primary Care



Yes, but:

Current solutions are **radiologist-oriented**, lacking the general MSK assessment required for primary care.

Public datasets focus on **specific pathologies** rather than holistic, preliminary evaluations (data scarcity).

Objectives

1. Generate a large-scale dataset from routine clinical data, following the precedence of *PadChest* [4] and *CheXPert* [5].

2. Train an AI model to distinguish referable vs. non-referable knee radiographs (a high order clinical task).

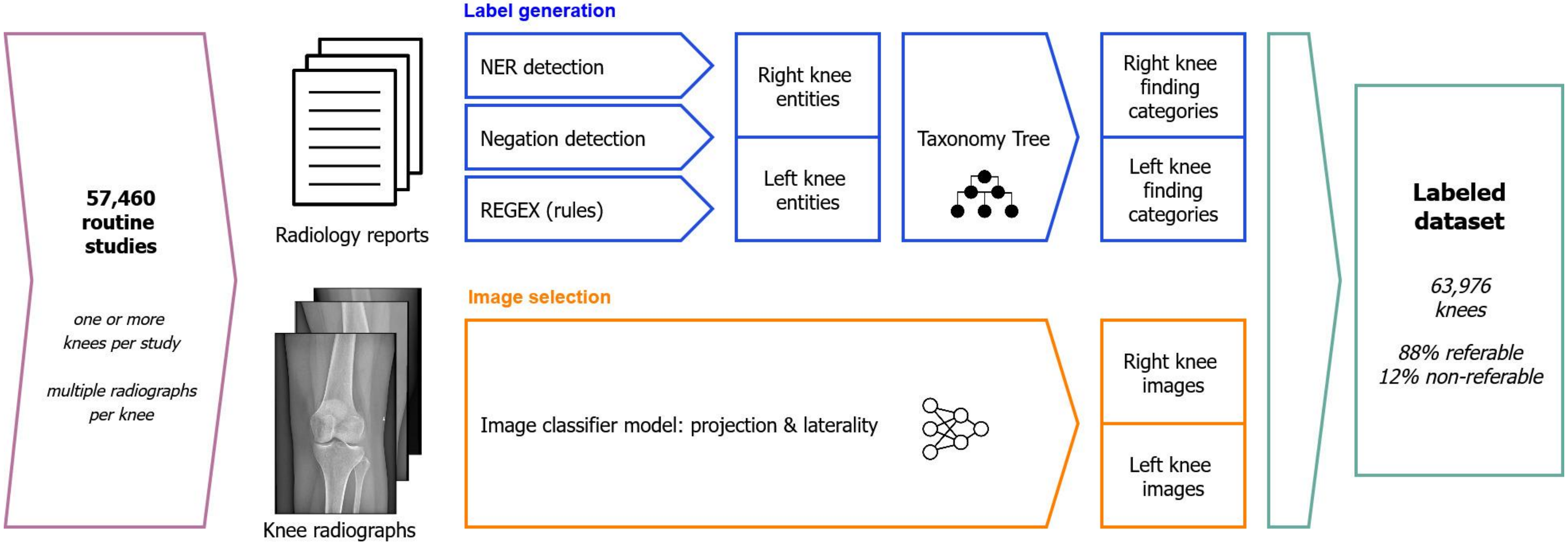
3. Validate both the labeling pipeline and the imaging model.

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

[5] 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.

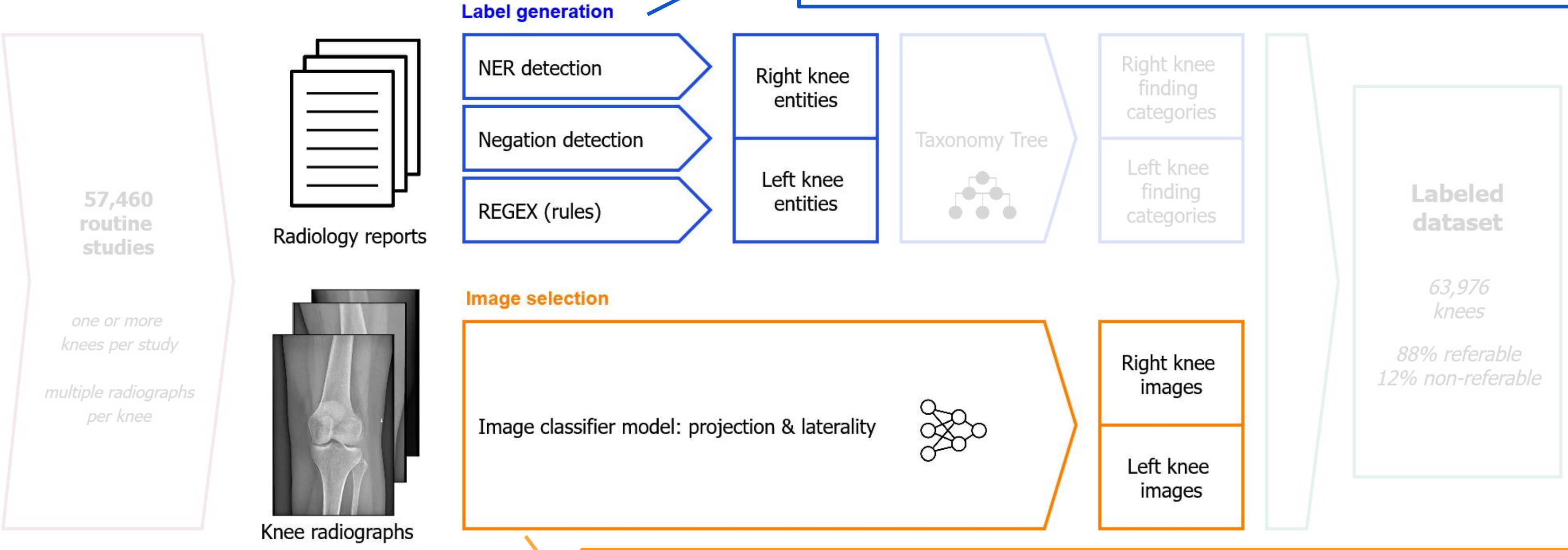
Dataset Creation

Dataset: 57,460 anonymized knee studies (2010–2024).
Source: Navarre Health Service; 63 primary care centers and 3 hospitals.
Content: Multi-projection X-rays (anteroposterior, lateral, axial) paired with free-text radiologist reports.



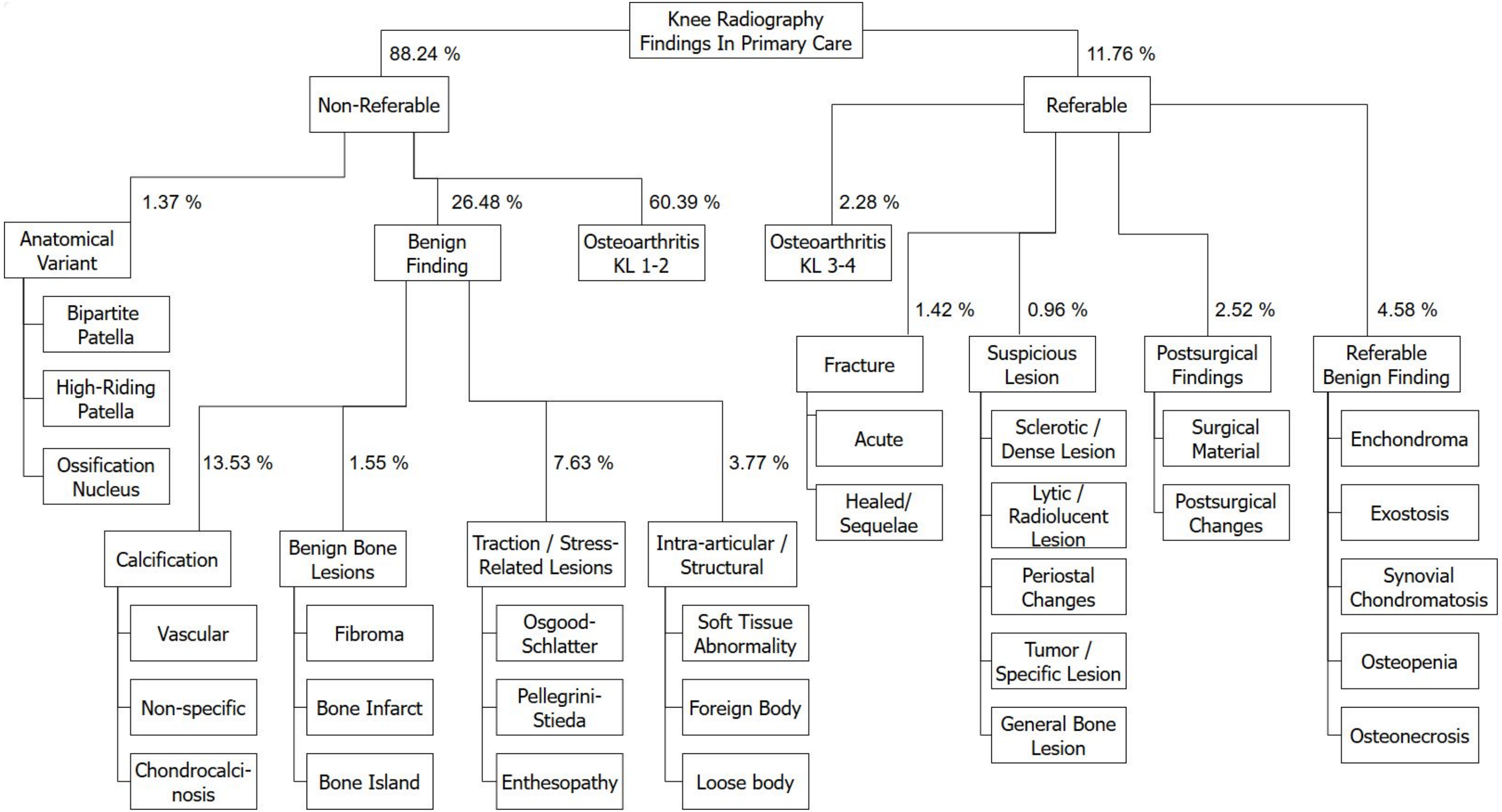
Dataset Creation

NER: Entity Recognition by RoBERTa model (Spanish) [6].
Negation: Scope detection by fine-tuned mBERT (Spanish) [7].
REGEX: Rule-based report splitting by laterality.



Motivation: DICOM metadata often lacked specific knee identification.
Dataset: We manually annotated 2,214 images for projection and laterality.
Model: ConvNeXt-Tiny, 97.4% accuracy on validation set.

Taxonomy Tree



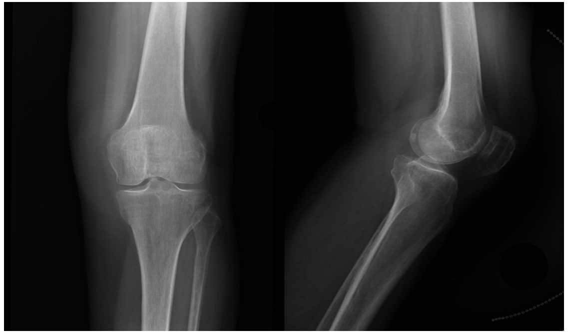
Automated Image Labeling Example

INPUT

Radiologist Report

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

Radiographs



Routine Knee Study

Label Generation → **NLP**

Image Selection → **CNN**

OUTPUT

Findings: osteopenia, chondrocalcinosis, fracture sequelae.
Label: Referable.

2 suitable images
Laterality: left.
Projections: anteroposterior, lateral.

2 Labeled Knee Images

Model for Referable Knee Detection

Architecture: ConvNeXt-Small

Dataset: 73,020 automatically labeled images (62,309 training / 10,711 validation)

- **Input:** 900x900px anteroposterior or lateral radiograph (padded & DICOM native windowing)
- **Output:** Referable/Non-referable

Training: 6 epochs | LR: 0.01 | cross-entropy loss | Fast.ai default augmentations

Performance: 0.841 AUC (Validation)

Hyperparameter and architecture variations yielded no significant gains.

Test set

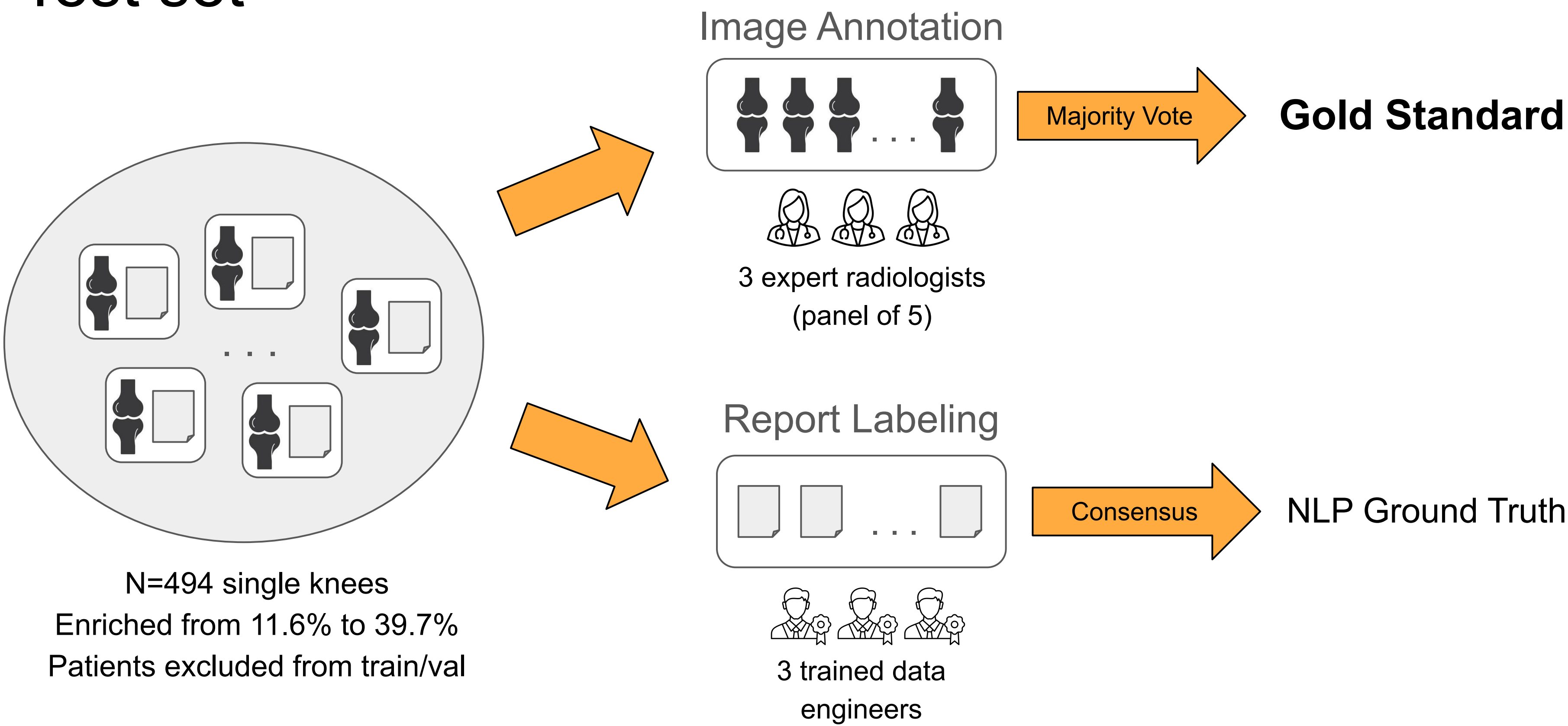


Image Annotation
Majority Vote
Gold Standard
3 expert radiologists
(panel of 5)
Report Labeling
Consensus
NLP Ground Truth
3 trained data
engineers
N=494 single knees
Enriched from 11.6% to 39.7%
Patients excluded from train/val

Performance: NLP Pipeline

High Overall Accuracy

0.92 Macro F1-score. Most categories >0.73 F1.

Labeling Challenge

Performance dropped for KL 3–4 (F1: 0.17).

Reason: clinical descriptions and implicit KL grades.

However

The **image model** successfully identified KL 3–4.

Finding Category	Precision	Recall	F1	Support
Osteo. KL 1-2	0.757	0.978	0.854	182
Normal study	0.946	1.000	0.972	88
Enthesopathy	0.971	0.971	0.971	68
Osteo. KL 3-4	0.857	0.095	0.171	63
Acute fracture	0.962	1.000	0.981	51
Soft tissue abnormality	1.000	0.956	0.977	45
Surgical material	0.884	0.905	0.894	42
Chondrocalcinosis	1.000	1.000	1.000	35
...				
Micro Average	**0.897**	**0.897**	**0.897**	**814**
Macro Average	**0.960**	**0.915**	**0.922**	**814**

Performance: Model for Referable Knee Detection

Model vs Manually Annotated Reports on the Gold Standard

95% CI

	AUC	Kappa	Sensitivity	Specificity
Model	**0.8800 (0.84–0.91)**	0.6098 (0.54–0.68)	81.94% (75.0–87.5)	**83.14% (79.1–86.9)**
Reports	0.7983 (0.76–0.84)	0.5393 (0.47–0.62)	81.94% (75.7–88.2)	77.71% (73.4–82.0)

➔ The vision model outperforms the reports (p = 0.00019).

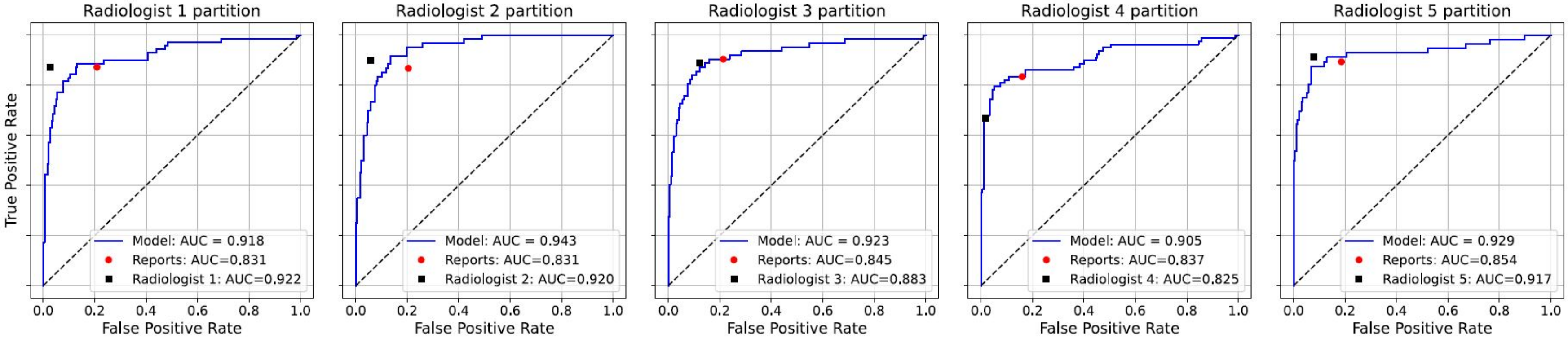
➔ All six suspicious lesions correctly referred.
Recall >70% for all referable categories*.

** Please refer to the original paper for more details.*

Performance: Model for Referable Knee Detection

Model vs Individual Radiologists on Gold Standard Partitions

** Please refer to the original paper for more details.*



Similar or slightly higher sensitivity than individual radiologists.
Slightly lower specificity, and lower Cohen's Kappa*.

Model errors were mostly straightforward misclassifications.
Report errors were frequently due to textual ambiguity.

Conclusions

1 **The model is accurate.**

Potential →
- Actionable guidance to general practitioners.
- Alleviate radiologist workload.
- Shorten diagnostic times.

2 **Scalable methodology.**

3 **Limitations.**

Conclusions

1. **The model is accurate.**
2. **Scalable methodology.**
   Contribution
   - Pathway-specific taxonomies enable clinically relevant datasets from reports.
   - Adaptable to other modalities and anatomical regions.
3. **Limitations.**

Conclusions

1. **The model is accurate.**
2. **Scalable methodology.**
3. **Limitations.** Such as → Small test set, lack of external validation, report noise.

Future Directions: LLMs & multi-label models.

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] A. Bustos, A. Pertusa, J. M. Salinas, and M. de la Iglesia-Vayá, “PadChest: A large chest x-ray image dataset with multilabel annotated reports,” Medical Image Analysis, vol. 66,p. 101 797, 2020.

[5] 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.

[6] 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.

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

THANK YOU!

Contact:
imanolpinto@proton.me