IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

KneeSight: Supervised Multi-Sequence Knee MRI Abnormality Detection Using Multilingual Report Labels

Lenny Dany Derek D[1], Sidhanth Bibi[2], Ashish Ranjan[3], S. Poornima[4]

Dept. of CSE (Emerging Technologies), SRM Institute of Science and Technology, Vadapalani Campus Chennai[1-4]

Abstract: Knee magnetic resonance imaging (MRI) is the reference examination for soft-tissue knee injury, but expert image annotation is scarce and expensive. In the RSNA Knee Abnormality Detection challenge, only 58 of 4,407 training studies carry expert labels, every training study has a free-text radiology report written in one of seven languages, and the hidden test set contains images only. This paper presents KneeSight, a weakly supervised pipeline that treats the reports as a training-time teacher. A deterministic multilingual rules labeller converts each report into soft labels, an explicit not-addressed mask and a confidence weight. Raw DICOM series are ordered by patient-space geometry, routed into six canonical sequence slots and encoded as 2.5D adjacent-slice triplets by a shared pretrained backbone, followed by slice- and series-level aggregation into twelve abnormality scores. Training uses a confidence-weighted, masked binary cross-entropy in which expert labels receive four times the weight of report labels, evaluated with duplicate-aware grouped five-fold cross-validation. Against the 58 expert-labelled studies, the final labeller reaches a macro AUROC of 0.706, compared with 0.679 for the baseline rules, 0.663 for a literal-rubric variant and 0.500 for a random labeller, with the largest gains on ACL (+0.062), MCL (+0.124) and effusion (+0.098). The implementation passes 1,556 automated tests and is packaged as a hash-verified offline Kaggle bundle; full-data image-model results are identified as the next step.

Keywords: knee MRI, weak supervision, radiology report labelling, multi-label classification, 2.5D convolutional networks, multiple-instance learning, macro AUROC

I. INTRODUCTION

The knee is the most commonly injured and most frequently imaged joint in the human body. Osteoarthritis alone affects an estimated 654 million people worldwide, and acute knee injuries account for 15–40% of sports-related trauma. MRI shows ligaments, cartilage, menisci, synovium and bone marrow in fine detail without ionising radiation, which is why it is the standard examination whenever a soft-tissue knee injury is suspected. A knee MRI study is not a single image: it typically contains five to eight series of 20–40 slices acquired in sagittal, coronal and axial planes with different contrasts, which a radiologist must mentally fuse. Reading these studies is slow, needs musculoskeletal expertise that is scarce outside large centres, and is not always consistent between readers. Models that pre-read a study and highlight likely abnormalities can therefore act as decision-support tools while the final decision remains with the clinician.

In 2026 the Radiological Society of North America (RSNA) released the Knee Abnormality Detection challenge [1], the first RSNA AI challenge in which every imaging study is paired with its original radiology report. The task is to predict, for each study, the probability of twelve abnormalities: anterior and medial collateral ligament (ACL, MCL) tears, medial and lateral meniscal tears, medial, lateral and patellofemoral osteoarthritis (OA), effusion, synovitis, Baker's cyst, bone contusion and fracture. Submissions are scored by the macro-averaged area under the ROC curve (AUROC), so rare targets count as much as common ones. Four properties make the problem harder than standard image classification: only 58 of the 4,407 training studies have expert image labels; reports are written in English, Dutch, German, French, Spanish, Turkish and Greek and often omit normal findings; studies contain variable, sometimes missing or corrupt series in an archive of roughly 570 GB; and the hidden test set has no reports, with inference required to run inside an Internet-disabled notebook.

Published knee MRI systems [2], [3] assume plentiful expert labels, whereas report-mining work [4] has mostly targeted English chest radiographs. KneeSight sits at the intersection of these threads: an image model whose supervision comes mostly from multilingual text. The main contributions of this paper are:

- a deterministic, auditable multilingual rules labeller that outputs soft labels, confidences and not-addressed masks, measured against the 58-study expert panel;
- a geometry-first preprocessing pipeline that routes heterogeneous series into six canonical sequence slots with

 This work is licensed under a Creative Commons Attribution 4.0 International License

explicit missing-data masks;
- a hierarchical 2.5D slice- and series-level model trained with a confidence-weighted masked loss in which expert labels dominate report labels;
- a leakage-safe evaluation protocol with duplicate-aware grouped folds and preregistered falsifiers, delivered as a tested, offline-reproducible implementation.

Section II reviews related work, Section III describes the data, Section IV presents the method, Section V the implementation, Section VI the results, and Section VII concludes.

TABLE I. SUMMARY OF THE COMPETITION DATASET

Component	Description
Training studies	4,407 knee MRI studies (≈570 GB of DICOM)
Series per study	Sagittal, coronal and axial series with varied contrasts
Radiology reports	One free-text report per training study, seven languages
Expert labels	Image-derived labels for 58 training studies
Targets	12 binary abnormalities
Hidden test set	Image-only studies; no report column
Metric	Macro-averaged AUROC over the 12 targets

II. RELATED WORK

A. *Knee MRI Classification*

MRNet [2] was the landmark study for deep learning on knee MRI. Using 1,370 examinations, it trained a separate AlexNet-based slice encoder for each of three series, max-pooled the slice features and fused the series-task models with logistic regression, reporting AUROC 0.965 for ACL tear and 0.847 for meniscal tear. KneeMRI [5] released sagittal proton-density volumes graded for ACL injury, and ELNet [6] matched MRNet-level accuracy with far fewer parameters using multi-slice normalisation. CoPAS [3] is the published work closest to ours: on 1,748 subjects from five centres it predicted twelve knee abnormalities with a 3D ResNet-18 per plane, co-plane cross-attention and sequence-gated fusion, reporting an internal average AUROC of 0.812 and 0.72–0.73 on external validation. These results show that a twelve-target knee task is genuinely hard and that performance drops on unseen hospitals.

B. *Slice Aggregation and Pretrained Backbones*

A study is a bag of slices and series with one label set, which makes multiple-instance learning (MIL) a natural framing. Attention-based MIL pooling [7] learns how much each instance should contribute, although simple mean pooling often performs comparably on 3D medical data, so attention should be tested as an ablation rather than assumed to win. Transformer attention [8] underlies our optional label-query aggregation. For encoders, ResNet [9] remains a strong, efficient baseline, DINOv2 [10] provides self-supervised vision-transformer features that transfer well, and RadImageNet [11] showed the value of radiology-specific pretraining. Top solutions in recent RSNA Kaggle competitions converged on a 2.5D recipe: three adjacent slices fed as an RGB image into a 2D ImageNet backbone, followed by aggregation across slices.

C. *Learning From Reports and Noisy Labels*

CheXpert [4] mined 224,316 chest radiographs with a rule-based labeller that outputs positive, negative or uncertain for each finding. Snorkel [12] formalised weak supervision by combining noisy labelling functions into probabilistic labels, RadGraph [13] extracted clinical entities and relations from reports, and recent work [14] evaluated large language models for labelling musculoskeletal reports. A recurring finding is that report labels are useful but imperfect: reports omit normal structures, use institution-specific vocabulary and may disagree with what an image reader would grade. KneeSight addresses this by modelling uncertainty and “not addressed” explicitly instead of forcing every finding into a hard label.

III. DATASET AND PROBLEM FORMULATION

KneeSight uses the official RSNA Knee Abnormality Detection dataset [1], summarised in Table I. The data are de-identified DICOM studies, a training table with one radiology report per study, expert labels for a small subset and a sample submission that fixes the output format.

Formally, given a study X consisting of a variable number of series, the goal is to output a vector $p \in [0, 1]^{12}$ of abnormality probabilities. This is multi-label binary classification: one study may be positive for several targets at once, such as an ACL tear with a bone contusion and an effusion. All eligible training studies are divided into five folds at the study level, so that every study receives exactly one out-of-fold (OOF) prediction; the 58 expert studies are distributed across folds and serve both as high-weight training targets and as the most trustworthy validation labels. DICOM pixels, reports, identifiers and row-level predictions are never committed to the code repository, and no dataset images are reproduced in this paper.

IV. PROPOSED METHOD

A. Preprocessing and Sequence Routing

Raw DICOM data arrive unordered, with scanner-specific intensity scales, inverted (MONOCHROME1) series and occasional truncated files. Every header is first indexed without pixels into a Parquet manifest with quality counters. Slices are then sorted by their position projected onto the slice normal, computed from ImagePositionPatient and ImageOrientationPatient, rather than by file name or instance number. Rescale slope/intercept are applied, MONOCHROME1 images inverted, and intensities clipped to the 1st–99th percentile of each series and scaled to [0, 1].

Each series is routed by its plane, fluid sensitivity and fat-suppression metadata into one of six canonical slots: sagittal, coronal and axial fluid-sensitive fat-suppressed, sagittal fluid-sensitive without fat suppression, coronal T1 and sagittal T1; a deterministic quality ranking resolves competing series. Routed series are resized to 288 × 288 pixels (cropped to 280× 280), resampled to 24 cached slices and sampled as 16 adjacent-slice triplets, where slices $k-1$, k and $k+1$ form the three colour channels of one image. Missing slices or series are zero-padded and flagged in a validity mask, so the model can distinguish dark tissue from absent data. Horizontal flipping is deliberately not used for augmentation, because it would swap the medial and lateral compartments while medial and lateral OA and menisci are separate targets.

B. Multilingual Report Labeller

A deterministic rules labeller, shown in Fig. 1, reads each report and produces, for each of the twelve findings, a soft label value, a reason code and a confidence. Rules were preferred over a black-box model for the baseline because every decision can be inspected, tested and corrected. The labeller normalises Unicode and case and matches a lexicon of clinical terms in seven languages; limits negation and uncertainty to the local clause, so "no effusion, ACL torn" negates only the effusion; applies the competition rubric, mapping borderline size or grade wording to a soft negative of 0.20, uncertain wording to 0.55 and a clear positive to 0.95;
treats silence as a weak negative (0.10) only for ACL, MCL, both menisci and Baker's cyst, where reports reliably mention abnormalities, and marks other unmentioned findings as not addressed; and treats words such as "intact" and "unremarkable" in eight languages as negations of the structure they describe. Where an expert label exists it overrides the report label.

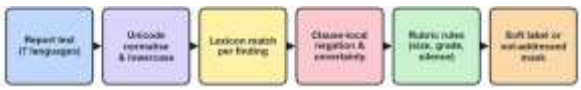


Fig. 1. Multilingual rules-labeller pipeline.

C. Hierarchical 2.5D Model

Fig. 2 shows the model. Each 3 × 280 × 280 triplet is encoded by a pretrained backbone shared across all series and slices: ResNet34 [9] (512-dimensional pooled features) as the fast baseline, or DINOv2-S/14 [10] (384-dimensional tokens). A learned encoding of the normalised physical slice position is added, and slices are aggregated into one token per series. For series s with slice features f_k and validity mask m_k, masked-mean aggregation averages valid slices only:

$$h_s = \Sigma_k \, m_k \, f_k \, / \, \max(1, \Sigma_k \, m_k) \quad (1)$$

Slot, plane, sequence and presence embeddings tell the network which series each token came from, so a study with an entirely missing slot still receives a finite prediction. Series tokens are then aggregated into twelve logits, either by masked mean or by optional label-query attention, in which each target j owns a learned query q_j that attends over the series tokens:gradients when a batch happens to contain few supervised examples.

$$z_j = \text{softmax}(q_j K_j^T / \sqrt{d}) V_j, \quad \text{logit}_j = w_j^T z_j + b_j \qquad (2)$$

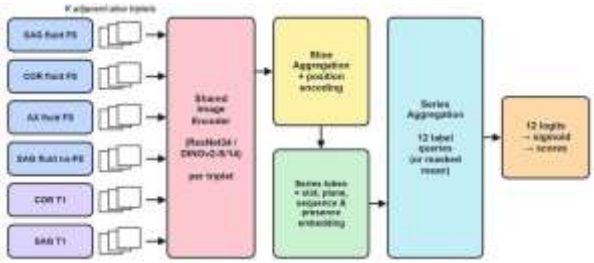


Fig. 2. Hierarchical 2.5D slice- and series-level model.

D. Confidence-Weighted Masked Loss

Let y be the soft label, $\hat{y} = \sigma(\text{logit})$ the prediction, m the supervision mask (0 when not addressed) and w the source weight, which is 4.0 for expert labels and $\max(0.25, 2|y - 0.5|)$ for report labels. A confident report label near 0 or 1 therefore receives nearly full weight, while an uncertain label near 0.5 contributes a quarter. The per-target loss is normalised by the training-set coverage c_j of that target, and all twelve targets are averaged with equal weight:

$$L = (1/12) \Sigma_j [(1/N) \Sigma_i m_{ij} w_{ij} \text{BCE}(y_{ij}, \hat{y}_{ij})] / c_j \qquad (3)$$

Normalising by fixed training coverage rather than minibatch counts prevents rare targets from producing erratic

E. Validation Protocol and Metric

Validation uses five study-level folds. Duplicate and near-duplicate studies, and studies with identical normalised reports, are grouped so that they always fall in the same fold; otherwise a model could be validated on a near-copy of a training study. An additional scanner-held-out stress fold estimates robustness to unseen hardware. Before any run, four competing approaches were written down with a hypothesis and a falsifier: (A) a plain CNN with mean pooling, (B) hierarchical DINOv2 with label queries, (C) denser slice geometry with a fixed physical crop, and (D) better report labels. A configuration is promoted only if it improves one-fold macro AUROC by at least 0.003 and full cross-validation by at least 0.002 without collapsing any major target. For each target, the AUROC equals the probability that a randomly chosen positive study is ranked above a randomly chosen negative one [15]; the final score is their mean:

$$\text{Score} = (1/12) \Sigma_{i=0}^{11} \text{AUC}_i \qquad (4)$$

F. Training Configuration

Table II lists the base configuration. Training follows a staged funnel (smoke, partial, full, five-fold, three-seed) so that cheap runs catch software errors before GPU time is spent, with early stopping on validation macro AUROC.

TABLE II. TRAINING CONFIGURATION

Hyper-parameter	Value
Optimiser	AdamW [16]
Learning rate (backbone / head)	3×10^{-5} / 3×10^{-4}
Weight decay	0.05
Effective batch size	8 (gradient accumulation)
Epochs / early stopping	15 / patience 3
Mixed precision	Enabled (AMP)
Expert label weight	4.0
Sequence / slice dropout	0.15 / 0.10
Input	6 slots × 16 triplets × 3 × 280 × 280

V. SYSTEM IMPLEMENTATION

Fig. 3 shows the complete system. The image stream indexes DICOM studies, routes series into canonical slots and builds the 2.5D cache; the text stream, used only for training, converts reports into soft labels and masks. The hierarchical model consumes both, its OOF predictions produce the metrics, and its exported checkpoint drives

offline inference that writes a twelve-score submission file. If a test study has no routable series at all, a governed prior computed from cross-fitted OOF means is substituted; any other missing or non-finite prediction is treated as a hard failure.

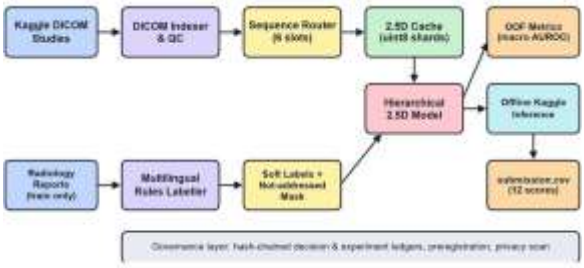


Fig. 3. Architecture of the KneeSight system.

KneeSight is implemented as an installable Python package built on PyTorch [17], timm, pydicom, pandas and PyArrow, with one command-line entry point covering every stage from indexing to submission. The codebase comprises roughly 31,000 lines of source and 23,500 lines of tests across 76 test modules, checked by continuous integration on Python 3.11 and 3.12. Experiments are preregistered with their hypothesis, single change and acceptance criterion in append-only, hash-chained ledgers, and a tracked-file privacy scanner keeps clinical data out of the repository. For deployment, a content-hashed offline bundle of about 104 MB containing the package, dependency wheels, weights and configuration is verified inside an Internet-disabled Kaggle notebook before use.

VI. RESULTS AND DISCUSSION

Following the project's governance rules, measured label-agreement scores, synthetic software-test results and figures published by others are kept clearly separate. The full five-fold image-model training had not been completed at the time of writing, so no image-model AUROC is claimed; the results below concern label quality and system verification.

A. Label-Extraction Results

The agreement between the labeller's soft labels and the 58 expert-labelled studies was measured as AUROC per target for three versions (Table III, Fig. 4). Version v3 was the baseline. Version v4 applied the competition rubric literally, treating "mild" as negative for every target. Version v5 applied the rubric only where it is the actual criterion (size for effusion and Baker's cyst, grade for ACL and MCL), added the silence rule, fixed the "intact" negation and completed the multilingual lexicon.

TABLE III. LABELLER AGREEMENT WITH THE 58 EXPERT STUDIES (AUROC)

Target	**v3**	**v4**	**v5**	**v5 − v3**
ACL	0.739	0.743	0.801	+0.062
MCL	0.724	0.738	0.848	+0.124
Medial meniscus	0.619	0.621	0.614	−0.005
Lateral meniscus	0.714	0.707	0.730	+0.017
Medial OA	0.713	0.597	0.705	−0.008
Lateral OA	0.665	0.598	0.669	+0.004
PF OA	0.588	0.548	0.591	+0.003
Effusion	0.650	0.748	0.748	+0.098
Synovitis	0.625	0.569	0.625	0.000
Baker's cyst	0.774	0.774	0.800	+0.026
Contusion	0.657	0.628	0.657	0.000
Fracture	0.681	0.681	0.681	0.000
Macro AUROC	**0.679**	**0.663**	**0.706**	**+0.027**
Sensitivity	0.980	0.627	0.721	–
Specificity	0.433	0.778	0.788	–
Not-addressed rate	0.709	0.465	0.465	–

 This work is licensed under a Creative Commons Attribution 4.0 International License

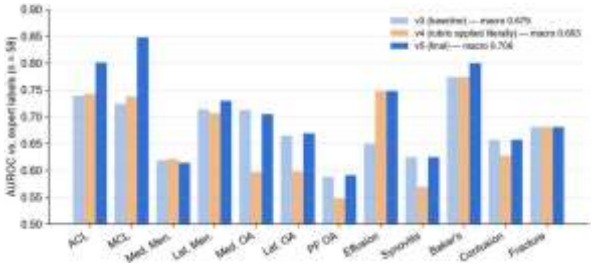


Fig. 4. Per-target label agreement for labeller versions v3, v4 and v5.

v5 improves macro AUROC from 0.679 to 0.706, with the largest gains on MCL (+0.124), effusion (+0.098) and ACL (+0.062). It also balances the operating point: v3 was highly sensitive (0.98) but unspecific (0.43) because it over-called or masked many findings, whereas v5 reaches 0.72 sensitivity and 0.79 specificity and lowers the not-addressed rate from 70.9% to 46.5%, giving the image model more usable supervision.

B. Comparison With Baselines and Published Label Sets

Table IV and Fig. 5 place the labeller against internal baselines, a random labeller and two label sets published by other competitors. v5 is the best of the measured labellers,
0.043 above the literal-rubric variant and 0.206 above chance. The published regex (0.814) and LLM (0.878) label sets agree more closely with the expert panel; these figures are competitor-reported and were not reproduced on our pipeline, and the same competitors report that better label agreement did not always yield a better image model.

TABLE IV. LABEL-QUALITY COMPARISON

Label source	Macro AUROC	Evidence
KneeSight v5 (final)	**0.706**	**Measured**
KneeSight v3 (baseline)	0.679	Measured
KneeSight v4 (literal rubric)	0.663	Measured
Random labeller	0.500	Chance level
Published regex label set	0.814	Competitor-reported
Published LLM label set	0.878	Competitor-reported

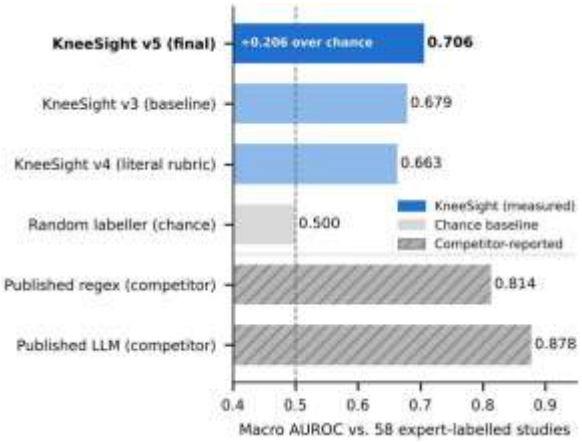


Fig. 5. Label agreement of KneeSight labeller versions versus a chance baseline and published label sets.

C. Software Verification

Because no image-model result may be claimed until it is recorded in the experiment ledger, every component was verified on controlled synthetic data. The suite reports 1,556 passed and 64 skipped tests (49 requiring CUDA); linting and static typing pass on 83 source files; both ledger hash chains verify; the duplicate audit scores 9,708,621 candidate pairs for a synthetic 4,407-study set in under 2 s; and a 100-study ResNet34 diagnostic bundle loads pretrained weights offline and passes BF16 forward/backward and checkpoint-restore checks on Kaggle GPUs.

 This work is licensed under a Creative Commons Attribution 4.0 International License

D. Discussion

Three lessons emerge. First, domain-aware rules beat literal rules: applying "mild means negative" to every target lowered agreement to 0.663, because a radiologist's "mild osteoarthritis" often already meets the image threshold, while restricting the rule to size- and grade-based targets raised it to 0.706. Second, a single bug can dominate: before v5 the word "intact" was not treated as a negation, so "ACL intact" was labelled as an ACL finding; about 1,200 of 4,407 reports use such a word, and fixing it raised ACL agreement from 0.739 to 0.801 and MCL from 0.724 to 0.848. Third, the v5 changes were designed after inspecting v4 on the same 58 studies, so the +0.027 gain is optimistically biased, and on a panel of this size differences below about 0.02 are hard to separate from noise. The decisive test is therefore an image model trained on v3 versus v5 labels under identical conditions, which is exactly what Approach D's falsifier specifies.

The main limitations are that full-data OOF results are pending, rule-based labels remain below published label sets, the small expert panel has been used for labeller development and is no longer an independent test, free GPU quotas restrict the number of full experiments, and the system has not been validated on external hospitals and must not be used for patient care.

VII. CONCLUSION AND FUTURE WORK

KneeSight is a complete, tested pipeline for detecting twelve knee abnormalities from multi-series MRI under sparse expert supervision. It converts heterogeneous DICOM studies into a canonical six-slot 2.5D representation, turns multilingual reports into soft, uncertainty-aware labels with an explicit not-addressed state, and trains a hierarchical network with a loss that trusts expert labels four times more than report labels. Careful, domain-aware label engineering raised label agreement with the expert panel from 0.679 to 0.706 macro AUROC, 0.206 above chance, with large gains on ACL, MCL and effusion. Future work will complete the five-fold ResNet34 baseline with bootstrap confidence intervals, run the preregistered comparison of Approaches A–D, test denser slice geometry and local language-model labels against the rules baseline, explore report–image contrastive pretraining, and add slice- and series-level attention maps to explain predictions to radiologists.

ACKNOWLEDGMENT

The authors thank Dr. P. Chitra, Professor and Head, Department of Computer Science and Engineering (Emerging Technologies), SRM Institute of Science and Technology, Vadapalani Campus, for her guidance and support, and the RSNA and Kaggle for providing the challenge data.

REFERENCES

[1] Radiological Society of North America, "RSNA Knee Abnormality Detection," Kaggle competition, 2026. [Online]. Available: https://www.kaggle.com/competitions/rsna-knee-abnormality-detection

[2] N. Bien et al., "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet," *PLOS Med.*, vol. 15, no. 11, Art. no. e1002699, 2018.

[3] X. Qiu et al., "Learning co-plane attention across MRI sequences for diagnosing twelve types of knee abnormalities," *Nat. Commun.*, vol. 15, Art. no. 7637, 2024.

[4] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 590–597.

[5] I. Štajduhar, M. Mamula, D. Miletić, and G. Ünal, "Semi-automated detection of anterior cruciate ligament injury from MRI," *Comput. Methods Programs Biomed.*, vol. 140, pp. 151–164, 2017.

[6] C.-H. Tsai, N. Kiryati, E. Konen, I. Eshed, and A. Mayer, "Knee injury detection using MRI with efficiently-layered network (ELNet)," in *Proc. Med. Imaging Deep Learn. (MIDL)*, PMLR 121, 2020, pp. 784–794.

[7] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, PMLR 80, 2018, pp. 2127–2136.

[8] A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[10] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, 2024, arXiv:2304.07193.

[11] X. Mei et al., "RadImageNet: An open radiologic deep learning research dataset for effective transfer learning," *Radiol. Artif. Intell.*, vol. 4, no. 5, Art. no. e210315, 2022.

[12] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with

 This work is licensed under a Creative Commons Attribution 4.0 International License

weak supervision," *Proc. VLDB Endow.*, vol. 11, no. 3, pp. 269–282, 2017.

[13] S. Jain et al., "RadGraph: Extracting clinical entities and relations from radiology reports," in *Proc. NeurIPS Datasets Benchmarks Track*, 2021, arXiv:2106.14463.

[14] Kreutzer et al., "Large-language-model labelling of musculoskeletal radiograph reports for downstream model training," *Eur. Radiol.*, 2025, doi: 10.1007/s00330-025-12102-1.

[15] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[16] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[17] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.

 This work is licensed under a Creative Commons Attribution 4.0 International License