American Journal of Smart Technology and Solutions (AJSTS)

ISSN: 2837-0295 (ONLINE)

VOLUME 5 ISSUE 2 (2026)



PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Smart Technology and Solutions (AJSTS)

Volume 5 Issue 2, Year 2026
ISSN: 2837-0295 (Online)
DOI: https://doi.org/10.54536/ajsts.v5i2.8870
https://journals.e-palli.com/home/index.php/ajsts

Weakly Supervised Multi-Label Knee MRI Abnormality Detection with ConvNeXt and Window-Attention MIL

Bajpayee Singha[1], Md. Sulayman[2*]

Article Information

Received: August 12, 2026

Accepted: October 24, 2026

Published: November 27, 2026

Keywords

ConvNeXt, Knee MRI, Multiple-Instance Learning, Multi-Label Classification, Weak Supervision

ABSTRACT

Automated analysis of knee magnetic resonance imaging (MRI) is challenging because relevant abnormalities may be distributed across many slices, while comprehensive expert annotations are expensive to obtain. This study examines a weakly supervised multi-label framework combining a ConvNeXt-Small encoder, finding-specific spatial attention and window-attention multiple-instance learning (MIL). The supplied training archive contains 4,407 unique study identifiers and twelve report-derived target scores. Of these, 4,349 non-gold studies have complete soft-target values; 4,263 were included in the final all-data training pool, while 58 designated gold studies were excluded from training and 86 non-gold identifiers were absent from the frozen local image pool. MRI windows comprise three adjacent slices sampled at a nominal physical interval of 3.5 mm, with full-field and anatomy-centred local views. The recorded model was trained for 16 epochs using soft-target binary cross-entropy and AdamW optimisation. Training loss decreased from 0.55166 to 0.39110, a reduction of 29.1%. Descriptive analysis showed substantial variation in the frequency of target scores at or above 0.5, from 6.8% for fracture to 58.8% for effusion, and a strong correlation between effusion and synovitis scores (Spearman's $\rho = 0.903$). At study level, 72.3% of studies had at least two scores ≥ 0.5; the largest fold-wise range in a target's high-score proportion was 4.24 percentage points. These results describe the training labels and optimisation behaviour, not clinical diagnostic performance. The supplied checkpoint was trained on all eligible studies and has no held-out evaluation. Independent testing against appropriate reference labels is therefore required before diagnostic accuracy or clinical utility can be claimed.

INTRODUCTION

Knee magnetic resonance imaging (MRI) supports the assessment of ligament and meniscal injuries, degenerative joint disease, joint fluid, synovial abnormalities and osseous lesions. Its value lies in its ability to depict soft tissues and bone marrow without ionizing radiation. However, interpretation of a complete examination requires integration of information across multiple image sequences, anatomical planes and slices. Some abnormalities are spatially limited, whereas others, including degenerative changes, may involve several compartments. Multiple findings can coexist in a single study, so a useful automated system must account for both the local appearance of an abnormality and the broader examination context. These demands make multi-label knee MRI analysis a more complex task than classification of a single image or one diagnosis.

Deep learning has shown potential for supporting knee MRI interpretation, but reported performance depends on the target definition, reference standard, cohort composition and evaluation design. The MRNet study established a prominent early benchmark for automated detection of general abnormalities, ACL tears and meniscal tears, and included an assessment of whether model assistance could influence clinician interpretation (Bien *et al.*, 2018). Subsequent work has expanded towards multi-abnormality prediction. Qiu *et al.* (2024) developed a co-plane attention model for twelve knee abnormalities, while Xie *et al.* (2025) reported a multi-task system for nine common abnormalities using a large, multicentre, stepwise validation design. These studies demonstrate the importance of evaluating broader target sets and clinically relevant validation, but their performance estimates cannot be transferred directly to a different model or label-generation process.

A persistent constraint is the cost of producing comprehensive, reliable annotations. Slice-level or region-level annotation requires expert time, and the precise image location supporting a report-level diagnosis is not always readily available. Radiology reports provide a scalable source of supervision because they summarise findings at examination level. Nevertheless, reports may omit negative findings, express uncertainty, use variable terminology or differ in detail between institutions. Labels derived from reports should therefore be treated as imperfect proxies rather than equivalent to a reference standard established through independent expert adjudication. Recent work on weakly supervised radiology learning has emphasised this distinction and the need to understand how label quality affects downstream models (Misera *et al.*, 2024).

Multiple-instance learning (MIL) provides a natural

[1] Master's of Computer Science and Technology, School of Artificial Intelligence, Yanshan University, Qinhuangdao, China
[2] B.Sc. in Electronic and Electrical Engineering, Department of Electrical and Electronic Engineering, Manarat International University, Dhaka, Bangladesh
[*] Corresponding author's e-mail: engrmdsulayman@gmail.com

formulation when the label belongs to an entire examination but the relevant image region is unknown. An MRI study can be represented as a bag of image windows, each encoding a limited part of the examination. An aggregation mechanism then converts window-level representations into study-level scores. Attention-based MIL learns weights over instances, allowing a model to prioritise informative windows without requiring explicit region annotations (Ilse *et al.*, 2018). For knee MRI, this design is attractive because evidence for one finding may be concentrated in a small subset of slices, while another finding may be supported by broader anatomical patterns. This study examines a framework combining a ConvNeXt-Small image encoder, finding-specific spatial attention, and window-attention MIL. Spatial attention weights feature locations within each window, while window attention combines information across windows for each target. The model predicts twelve findings: ACL, MCL, medial meniscus, lateral meniscus, medial osteoarthritis, lateral osteoarthritis, patellofemoral osteoarthritis, effusion, synovitis, Baker's cyst, contusion and fracture. Training uses continuous report-derived soft labels, rather than treating all targets as independently verified binary outcomes.

The objectives are to document the supplied data and label construction; describe the model and preprocessing pipeline sufficiently for reproducibility; quantify target-score distributions and their associations; and assess what can and cannot be inferred from the available training record. The primary research questions are whether the report-derived targets provide a sufficiently informative basis for multi-label model development, how their distribution varies between findings, and what additional experiments are required to establish diagnostic performance. The work is intentionally framed as a transparent analysis of the available package rather than as proof that the model outperforms existing systems.

The distinction between optimisation and generalisation is central. The available checkpoint was trained on all 4,263 eligible studies and its completion record reports zero held-out studies. A decreasing training loss confirms that the training objective was optimised, but it does not quantify performance on unseen examinations. The manuscript therefore reports descriptive label analysis and training behaviour while specifying a rigorous evaluation protocol for future held-out testing. This approach is consistent with current reporting guidance that emphasises transparent dataset provenance, model evaluation, uncertainty and limitations in medical-imaging AI research (Tejani *et al.*, 2024; Collins *et al.*, 2024).

LITERATURE REVIEW

Knee MRI Classification and Multi-Abnormality Assessment

Early deep-learning studies in knee MRI commonly addressed a small number of targets, such as general abnormality, ACL tear or meniscal tear. MRNet demonstrated that convolutional features from MRI series could be combined to generate examination-level predictions and that model assistance could be evaluated with clinical readers (Bien *et al.*, 2018). This was an important proof of concept, but a limited target set does not capture the full range of findings in routine musculoskeletal practice. Knee abnormalities differ in prevalence, spatial extent, visual conspicuity and clinical relevance, and one examination may contain several simultaneously.

The recent systematic review by Mead *et al.* (2025) included 54 studies of deep-learning models for knee MRI and found that the literature had focused mainly on ACL injuries, meniscal pathology, osteoarthritis and general abnormality detection. The review noted variation in data sources, study designs and validation methods, and called for larger and more diverse datasets. Zhao *et al.* (2024) similarly found that AI models were more accurate at detecting the presence of meniscal tears than identifying their precise location, illustrating that classification performance does not automatically imply reliable localisation. A further meta-analysis of meniscal-injury models reported substantial variation across studies and concluded that performance should be interpreted alongside methodological quality and clinical comparators (Mohammadi *et al.*, 2025). A recent systematic review of AI for ACL tear detection likewise reported promising pooled estimates but substantial between-study heterogeneity and continuing limitations in interpretability, reinforcing the need for study-specific validation (Gill *et al.*, 2026).

Qiu *et al.* (2024) advanced the field by using co-plane attention across MRI sequences for twelve abnormalities, showing that multi-label assessment is a realistic research objective rather than merely a theoretical extension of single-target classification. Xie *et al.* (2025) subsequently investigated a nine-target system using multicentre data and staged validation. These studies are close comparators for the present work because they consider several findings at examination level. However, they use different model designs and label/reference procedures. In particular, models trained with expert-confirmed labels should not be compared uncritically with models trained using report-derived soft scores. The present analysis therefore focuses on its own label provenance and on defining the evidence needed for a valid comparison.

Reviews of musculoskeletal AI also emphasise that clinical adoption depends on more than headline discrimination metrics. Guermazi *et al.* (2024) discuss opportunities for AI in musculoskeletal imaging while stressing the need for trusted evidence, accepted reference standards and collaboration between radiologists and technical researchers. Gitto *et al.* (2024) similarly describe applications across musculoskeletal imaging while identifying barriers related to data, validation and translation into clinical workflows. These considerations support a cautious approach: multi-label model development is worthwhile, but claims of clinical utility require evidence that is appropriate to the intended use.

Weak Supervision and Report-Derived Labels

Weak supervision encompasses several distinct situations: labels may be available for only part of a dataset, may apply to a complete examination rather than a precise image region, or may contain errors. These categories have different implications for model design and evaluation. Zeng *et al.* (2024) review self-supervised learning as a complementary response to annotation scarcity; unlike self-supervision, the present model is trained against report-derived target scores, so its predictions remain dependent on the label-generation process. Misera *et al.* (2024) describe report-derived labels as one route to scaling radiology AI, because existing clinical text can be converted into structured targets without requiring a new annotation exercise for every image. The same scalability introduces uncertainty, however, because reports are written for clinical communication rather than to satisfy a machine-learning label ontology.

Radiology report labelling has been investigated in other imaging domains. CheXpert used a large chest-radiograph collection with uncertainty labels and expert comparison, demonstrating the importance of retaining uncertainty rather than forcing every report statement into a binary category (Irvin *et al.*, 2019). CheXbert used a BERT-based model to combine automated labellers and expert annotations for report classification, showing how natural-language processing can support structured label extraction (Smit *et al.*, 2020). MIMIC-CXR-JPG provides another example of large-scale image-report resources that enable research on weak labels (Johnson *et al.*, 2019). These systems concern chest radiographs, so their label semantics cannot be transferred directly to knee MRI, but they establish relevant principles for report-derived supervision.

The supplied knee dataset uses two report-derived score sources blended at study-target level. When both source scores are available, their arithmetic mean is used; when the JEV value is missing, the value from the other source is retained. This procedure preserves continuous values and avoids discarding every row with a missing component. It does not, however, guarantee that the blended score is calibrated or that disagreement between sources is clinically meaningful. The model learns to approximate these scores under the chosen loss function, so evaluation against the same label construction assesses consistency with the labelling pipeline rather than diagnostic truth.

Label uncertainty is especially important in multi-label settings. The absence of a report mention may mean that a finding is absent, clinically unimportant, outside the report's focus or simply not documented. Moreover, related findings may be discussed together, producing correlated scores even when their underlying clinical status is not identical. Consequently, score thresholds should be described as analytical choices, not as validated diagnostic cut-offs. The present analysis uses 0.5 only to summarise label-score distributions; it does not treat that threshold as a confirmed definition of disease presence.

ConvNeXt, Attention and Multiple-Instance Learning

ConvNeXt adapts several design choices from contemporary vision architectures to a convolutional network, providing a modern feature encoder while retaining the computational and inductive-bias characteristics of convolution (Liu *et al.*, 2022). The supplied model initialises ConvNeXt-Small from an ImageNet-pretrained variant and fine-tunes it on knee MRI. Transfer learning can be useful where the target dataset is not large enough to support training a high-capacity network from random initialisation. Nevertheless, pretraining on natural images does not guarantee that all learned representations are appropriate for MRI, and any claimed benefit should be measured against relevant baselines.

Attention mechanisms allow feature aggregation to depend on the content of the input. The Transformer architecture popularised attention-based sequence processing (Vaswani *et al.*, 2017), while attention-based MIL adapted learned weighting to bags of instances with only bag-level labels (Ilse *et al.*, 2018). In the present framework, the two attention stages have distinct roles. Spatial attention weights positions in a feature map for each finding; window attention weights window-level representations across a study. This separation is conceptually appropriate for an examination containing many slices, but the usefulness of the design depends on empirical comparisons with global pooling and simpler instance aggregation.

Recent work in other medical-image domains continues to examine attention-based MIL. Gadermayr and Tschuchnig (2024) review the state of MIL in digital pathology, highlighting its usefulness when slide-level labels are available but region-level annotations are not. Cai *et al.* (2025) revisit attention-based MIL through instance attributes and discuss limitations in differentiating relevant and irrelevant instances. Although histopathology differs substantially from MRI, the methodological problem of aggregating many local instances under weak supervision is shared. Such work provides context for the architecture but does not establish that the same mechanism will improve knee MRI performance.

Training configuration also affects model behaviour. The supplied pipeline uses AdamW, which decouples weight decay from the gradient-based parameter update (Loshchilov & Hutter, 2019), mixed precision, gradient checkpointing and a warm-up/cosine learning-rate schedule. These are implementation choices rather than independent scientific contributions. A robust study should report them clearly and compare architecture variants under matched training conditions. Without matched ablations, it is not possible to determine whether performance differences arise from the encoder, spatial attention, window aggregation, preprocessing or

optimisation settings.

Evaluation, Generalisation and Reporting Standards

Discrimination on an internal dataset is not sufficient evidence of clinical readiness. Medical images can differ across scanners, institutions, acquisition protocols and patient populations. Yang *et al.* (2024) showed that medical-imaging models may exploit shortcuts and that behaviour can change under distribution shift, underscoring the need for external evaluation and attention to subgroup performance. Guermazi *et al.* (2024) likewise argue that musculoskeletal AI must be assessed using an appropriate reference standard and within a realistic clinical context. The implications are direct for knee MRI: a model trained on one frozen dataset may learn both anatomy-related features and acquisition-specific patterns.

Reporting guidelines have become more explicit about these issues. CLAIM 2024 provides recommendations for transparent reporting of medical-imaging AI studies, including data provenance, partitioning, model development and evaluation (Tejani *et al.*, 2024). TRIPOD+AI addresses the reporting of clinical prediction models that use regression or machine learning (Collins *et al.*, 2024), while STARD-AI extends diagnostic-accuracy reporting to AI-centred diagnostic studies (Sounderajah *et al.*, 2025). Earlier guidance, including STARD 2015, remains relevant where the study is designed as a diagnostic-accuracy evaluation (Bossuyt *et al.*, 2015). These frameworks do not replace good experimental design, but they help make omissions visible.

Methodological analyses have also warned that medical AI studies may overestimate performance when data splitting, sample selection or reference labels are inadequate. Nagendran *et al.* (2020) compared AI and clinician studies and identified limitations in design and reporting. Kelly *et al.* (2019) described key challenges in translating AI into clinical impact, including validation, workflow and implementation. Varoquaux and Cheplygina (2022) discussed methodological pitfalls in medical imaging machine learning, particularly the risk of drawing broad conclusions from limited data. Roberts *et al.* (2021), in a review of COVID-19 imaging models, similarly identified bias, weak reproducibility and inadequate validation; although their clinical task differs from knee MRI, the methodological warnings are relevant. Together, this literature supports separating training behaviour, internal test performance, external generalisation and clinical impact as distinct evidence levels.

Metrics should be selected to reflect the evaluation question. ROC-AUC measures ranking discrimination across thresholds, whereas precision–recall analysis can be more informative when positives are uncommon (Fawcett, 2006; Saito & Rehmsmeier, 2015). If two correlated ROC curves are compared, methods such as DeLong's test may be appropriate under their assumptions (DeLong *et al.*, 1988). Bootstrap methods can quantify sampling uncertainty (Efron & Tibshirani, 1993), and calibration should be assessed separately from discrimination because high ranking performance does not guarantee accurate probability estimates (Guo *et al.*, 2017). Under dataset shift, predictive uncertainty can also change, as shown by Ovadia *et al.* (2019). These principles inform the proposed evaluation protocol, although such metrics cannot be calculated as clinical accuracy without independent labels and held-out predictions.

MATERIALS AND METHODS

Study Design and Dataset Provenance

This work is a retrospective computational analysis of the supplied knee MRI model-development package. The archive contains training scripts, configuration files, label tables, study manifests, a completion record and a trained checkpoint. It does not include the original DICOM images, radiology report text, all external anatomy-model assets or model prediction arrays. The analysis therefore distinguishes direct inspection of the package and its metadata from experiments that would require access to the original image environment.

The target table contains 4,407 unique study identifiers. It includes 4,349 non-gold rows with report-derived soft-target values and 58 designated gold rows with blank target values. The frozen local pool contains 4,321 studies: 4,263 labelled studies used in the final all-data training run and 58 gold studies excluded from training. A further 86 non-gold identifiers are absent from the local pool. The package documentation attributes these absences to local DICOM deletion to release storage rather than to a clinical or image-quality exclusion criterion. These counts are summarised in Figure 1.

The data package is associated with the RSNA Knee Abnormality Detection task, which defines twelve study-level targets and uses MRI studies containing multiple series (Radiological Society of North America [RSNA], n.d.). The official dataset version, access date, permitted use, acquisition characteristics and source-platform citation must be confirmed by the authors before submission. The supplied table does not contain sufficient demographic or patient-level information to describe age, sex, scanner distribution or patient-level clustering reliably. These characteristics should not be inferred from the study identifiers.

Study selection was determined by the frozen pool and preprocessing-success criteria recorded by the supplied pipeline. The manifest reports no precomputation failures among the 4,321 studies in the final pool. However, a successful preprocessing manifest does not establish population representativeness or absence of bias. The 86 missing local studies should be reported transparently because their exclusion changes the set available for training and may affect generalisability if missingness is associated with site, acquisition protocol or another relevant factor.

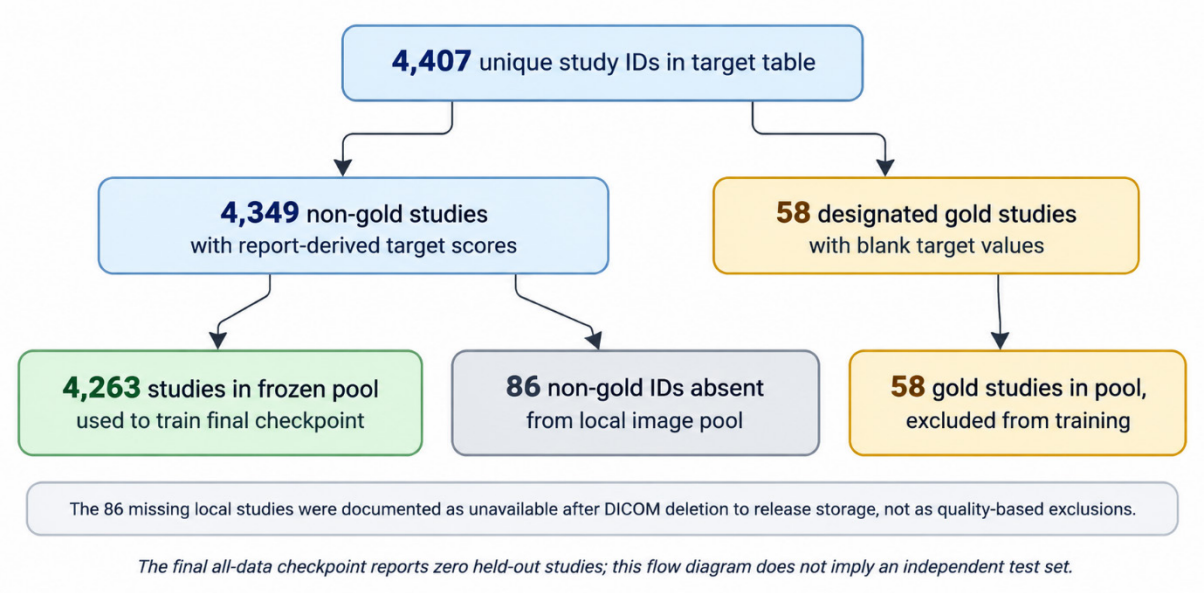


Figure 1: Study selection and the frozen local pool. The diagram describes file and label availability, not a clinical recruitment flow or an independent test cohort.

Target Definitions and Soft-Label Construction

The twelve target categories are ACL, MCL, medial meniscus, lateral meniscus, medial osteoarthritis, lateral osteoarthritis, patellofemoral osteoarthritis, effusion, synovitis, Baker's cyst, contusion and fracture. They represent different anatomical and pathological concepts, with different expected spatial distributions and label frequencies. The target table contains both `prior` and `strict` columns; the strict columns are blank in the supplied file, whereas the prior columns contain the scores used by the final `E_prior` configuration.

The prior labels were created by combining two report-derived score sources, identified in the package as `labels_v1_blend` and JEV hybrid labels. When both values were available for a study-target pair, the stored value was their arithmetic mean. When the JEV value was missing, the value from `labels_v1_blend` was retained. This procedure yielded complete prior scores for the 4,349 non-gold rows. It is a cell-wise combination, not a requirement that both sources be present for every target. The supplied archive does not provide sufficient clinical adjudication information to treat the resulting scores as verified diagnoses.

Soft targets retain degrees of label confidence or support rather than converting all training targets to zero or one. Binary cross-entropy with logits can be defined for target values between zero and one, and the model is optimised to match the supplied scores. This preserves information in the original labelling process but makes the interpretation of model outputs dependent on the quality and calibration of those scores. In particular, the arithmetic mean of two label sources is not necessarily a calibrated probability, and two correlated label sources can share the same errors.

For descriptive analysis, a score threshold of 0.5 was used to summarise the number and percentage of training studies with a score at or above that value. This threshold was chosen for transparent summary and is not claimed to be clinically validated. The continuous score distribution was also examined using means, medians, standard deviations and pairwise Spearman correlations. These analyses describe the training labels and should not be reported as disease prevalence or confirmed clinical co-occurrence.

MRI Preprocessing and Window Sampling

The training configuration specifies the `v4` view setup, which combines full-field-of-view images with anatomy-centred local views. The local views are guided by anatomical anchors produced through an external SKM-TEA anatomy V-Net pipeline. This design aims to preserve context around the knee while providing higher-resolution views of relevant anatomical regions. The exact external model version, weights, preprocessing library versions and licensing details should be recorded in the final reproducibility statement because these assets are not included in the supplied archive.

The pipeline renders images at 320 × 320 pixels and constructs three-slice windows. Slices are ordered using physical slice-position information. For thin stacks, adjacent slices are averaged into slabs of approximately 3.5 mm, and the window sampler uses slice-spacing metadata to approximate consistent physical sampling. The physical interval is important because a fixed number of slices can represent different anatomical distances across MRI acquisitions. Nevertheless, uniform physical sampling cannot fully remove variation in sequence contrast, field strength or acquisition protocol.

Each model input contains three slices as the three image channels. During training, candidate windows are drawn across available series and views. When more than 64 candidate windows are available for a study, 64 are selected by uniform random sampling without replacement. A random 288 × 288 crop is then taken from each 320 ×

320 window. The supplied configuration does not enable additional rotation, scale, gamma or contrast augmentation, although alternative augmentation options exist in the code. These settings define the recorded experiment and should not be changed retrospectively when describing it.

Window metadata include the series slot, view identifier, relative position within a series, log-transformed slice spacing, normalised field-of-view extent and a fluid-sensitive-sequence indicator. These features provide contextual information alongside image-derived representations. The sequence indicator refers to acquisition metadata and is not the effusion target. As the DICOM images and full preprocessing environment are not present in the archive, the present work verifies the documented settings rather than re-running the image rendering and window-generation stages.

ConvNeXt-Small and Finding-specific Spatial Attention

The image encoder is the `convnext_small` variant initialised from `convnext_small.fb_in22k_ft_in1k` in the `timm` model family. The encoder is fine-tuned for the knee MRI task. Pixel values are scaled to the range zero to one and normalised using the configured ImageNet channel means and standard deviations. The encoder generates a spatial feature map for each three-slice window. The architecture retains spatial information for attention pooling rather than immediately collapsing each window into a single global vector.

Let F(i,w,p) denote the feature vector at spatial position p for window w in study i. For each target c, the spatial-attention module applies feature normalisation, a learned linear projection and a projected two-dimensional positional encoding. A target-specific query scores each position, and a softmax over spatial positions produces weights alpha(i,w,p,c). The target-specific representation for a window is the weighted sum of feature vectors across spatial positions. In this way, one target can assign weight to one part of a window while another target can emphasise a different region.

The architecture is compatible with the premise that findings differ in spatial extent. A ligament abnormality may require attention to a narrow structure, while joint effusion or degenerative changes may be represented by broader patterns. However, the learned weights are optimisation parameters and are not automatically valid lesion-localisation maps. Their anatomical plausibility would need to be evaluated independently, for example through expert review or localisation annotations. The current archive does not include such validation, so the manuscript describes attention as a feature-aggregation mechanism rather than evidence of explainability (figure 2).

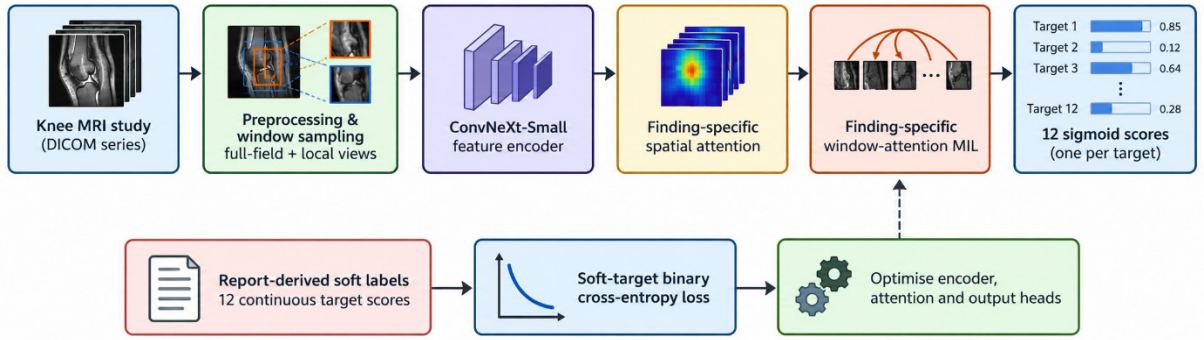


Figure 2: Conceptual model framework derived from the supplied configuration. The schematic illustrates preprocessing, feature encoding, spatial attention, window-level aggregation and twelve outputs; it is not a visualisation of learned attention weights.

Window-Attention Multiple-Instance Learning

After spatial pooling, the model aggregates window representations to obtain a study-level feature for each target. Series-slot and view embeddings are combined with a learned projection of the geometric and sequence metadata. Dropout is applied before the window-attention operation. For target c, a learned query scores each transformed window representation. A softmax across valid windows yields target-specific window weights, while padded windows are excluded from the normalisation.

The study-level representation is formed as the weighted sum of metadata-conditioned window features, followed by layer normalisation. A target-specific linear read-out produces a logit, and the sigmoid function maps the logit to a score between zero and one. The final output contains twelve scores per study. Spatial attention and window attention operate at different levels: the former combines spatial positions within an image window, whereas the latter combines windows across the study. The design is therefore hierarchical and allows different targets to use different spatial and window-level evidence.

Conceptually, this formulation follows attention-based MIL, in which a study is a bag of instances and a target is attached to the bag rather than to every individual instance (Ilse *et al.*, 2018). The formulation is appropriate for weak labels but does not guarantee that the model identifies a causal or clinically relevant feature. A model can assign attention to correlated acquisition artefacts, anatomical context or other shortcuts. Ablation experiments and independent localisation review are required to test whether the attention components contribute useful evidence rather than merely additional capacity.

Training Configuration and Optimisation

The final checkpoint uses the `E_prior` configuration and the twelve prior soft targets. The loss is binary cross-entropy with logits. For each target, the loss is averaged over labelled studies and the target-level losses are then averaged across the twelve targets. Sample weights are set to one in the supplied configuration. This target-wise normalisation gives each target a defined contribution to the overall objective, although it does not eliminate label uncertainty or guarantee equal clinical importance.

The optimiser is AdamW with a learning rate of 0.0001 for the trainable encoder parameters and 0.001 for the prediction head. Weight decay is 0.05. The learning-rate schedule uses a 5% linear warm-up followed by cosine decay. Training uses a batch size of two studies, up to 64 windows per study and 16 epochs (table 1). Mixed-precision bfloat16 autocasting and gradient checkpointing are enabled, gradient norms are clipped at five, and the random seed is 42. Exponential moving average of model weights is disabled.

The completion record reports 34,096 optimiser steps. Training loss declined from 0.55166 in epoch 1 to 0.39110 in epoch 16. The checkpoint contains approximately 50.02 million tensor elements across 362 tensors. These figures document the supplied run and its optimisation history. They do not indicate how well the model performs on data that did not contribute to parameter updates. The checkpoint was trained on all 4,263 eligible studies, and its completion record reports zero held-out studies.

Table 1: Principal training settings recorded in the supplied configuration

Setting	Recorded value	Role in the experiment
Encoder	ConvNeXt-Small, ImageNet initialisation	Feature extraction and fine-tuning
Input representation	Three-slice windows; 320 × 320 rendered images	Image input
Crop size	288 × 288 pixels	Random training crop
Physical sampling interval	3.5 mm nominal	Window/slice sampling
Maximum sampled windows	64 per study	Training-time bag construction
Epochs	16	Training duration
Batch size	2 studies	Optimisation batch
Encoder / head learning rates	0.0001 / 0.001	AdamW parameter groups
Weight decay	0.05	Regularisation
Random seed	42	Run reproducibility
Held-out studies in final checkpoint	0	No independent evaluation in this run

Descriptive Analysis and Proposed Diagnostic Evaluation

The archive analysis was performed on the supplied target CSV, frozen study pool, training completion record and checkpoint manifest. Study identifiers were checked for uniqueness and joined against the frozen pool to distinguish all labelled rows from those used by the final checkpoint. For each target, the number and percentage of training studies with a soft-label score of at least 0.5 were calculated. Continuous score distributions were summarised, and pairwise Spearman correlations were calculated across the twelve target columns. The number of target scores ≥0.5 was also counted per study; target-specific proportions were calculated within each supplied fold; and scores were grouped into four descriptive bins (<0.20, 0.20–<0.50, 0.50–<0.80 and ≥0.80). Training loss was extracted from the recorded per-epoch loss history. These additional summaries describe the label table and fold composition, not diagnostic performance.

Because no held-out predictions are included, diagnostic ROC-AUC, precision–recall AUC, sensitivity, specificity, positive predictive value, negative predictive value and calibration against independent diagnoses cannot be calculated from the supplied archive. If independent binary reference labels become available, the preferred evaluation should report per-target ROC-AUC and average precision with confidence intervals, as well as the number of positive and negative studies for each target. Precision–recall analysis is especially important for less frequent targets because a high ROC-AUC may coexist with low precision at practical operating points (Fawcett, 2006; Saito & Rehmsmeier, 2015).

A valid internal evaluation requires predictions from models that did not train on the corresponding studies. The package contains fold information, but the all-data checkpoint cannot be used to create unbiased out-of-fold predictions after training. Each fold model should be trained without its held-out fold, with all model selection and threshold tuning restricted to the training and validation portions. If multiple studies belong to the same patient, splitting must be performed at patient level; the supplied target table does not provide sufficient patient identifiers to verify this condition. A full five-fold out-of-fold run or a prespecified held-out split is therefore needed.

Confidence intervals should be calculated at the study or patient level using bootstrap resampling, with patient-level resampling preferred where multiple examinations

per patient are present. For paired comparisons of ROC-AUC, a method such as DeLong's test may be considered when its assumptions are satisfied (DeLong *et al.*, 1988). Calibration should be assessed independently of discrimination, using reliability plots and a proper scoring rule such as the Brier score (Guo *et al.*, 2017). Threshold-dependent metrics should use thresholds selected on validation data rather than tuned on the test set. These analyses are proposed for a future evaluation; they were not performed in the supplied all-data run.

Reproducibility also requires versioned code, fixed split files, a record of external assets, and a clear account of excluded studies. The training package includes a SHA-256 manifest, and the supplied checksums were reported to match the recorded code files and checkpoint. Nevertheless, the original DICOMs and external anatomy-model assets are absent. The full preprocessing and inference pipeline cannot therefore be reproduced from the ZIP archive alone. This limitation should be resolved or clearly documented before submission, in line with CLAIM, TRIPOD+AI and STARD-AI guidance (Tejani *et al.*, 2024; Collins *et al.*, 2024; Sounderajah *et al.*, 2025).

RESULTS AND DISCUSSION

Study Composition and Data Availability

The supplied target table contained 4,407 unique study identifiers. Of these, 4,349 non-gold rows had complete values for all twelve prior soft targets, while 58 designated gold rows had blank target values. The frozen local pool contained 4,321 studies, comprising 4,263 non-gold studies used to train the final checkpoint and the 58 gold studies excluded from training. The remaining 86 non-gold identifiers were not represented in the frozen local pool. According to the package documentation, those images had been deleted locally to release storage rather than excluded on the basis of a clinical quality criterion.

These counts establish the data available to the supplied training run, but they do not describe the demographic or clinical representativeness of the cohort. No reliable summary of patient age, sex, scanner vendor, field strength, or site distribution can be derived from the included label table alone. It is therefore not possible to determine whether the missing 86 studies are exchangeable with the included studies or whether their absence disproportionately affects particular findings. A submission-ready dataset table should add acquisition and patient-level characteristics once the source dataset and permitted metadata are available.

The 58 designated gold studies should not be described as a clinical test set on the basis of their name alone. Their target columns are blank in the supplied file, and the archive does not include independent adjudication records or model predictions for them. A valid test cohort requires defined reference labels and predictions from a model that has not been trained on those cases. The study flow in Figure 2 should therefore be interpreted as a data-availability diagram rather than evidence of completed diagnostic validation.

Distribution of Report-Derived Target Scores

The proportion of training studies with soft-label scores of at least 0.5 varied substantially between targets. Effusion had the largest proportion (2,506 studies; 58.8%), followed by patellofemoral osteoarthritis (1,911; 44.8%), medial meniscus (1,730; 40.6%) and medial osteoarthritis (1,550; 36.4%). Fracture had the smallest proportion (292; 6.8%), followed by synovitis (527; 12.4%), MCL (616; 14.4%) and lateral meniscus (655; 15.4%). Figure 3 summarise these distributions. The results show that the label scores are not evenly distributed across target

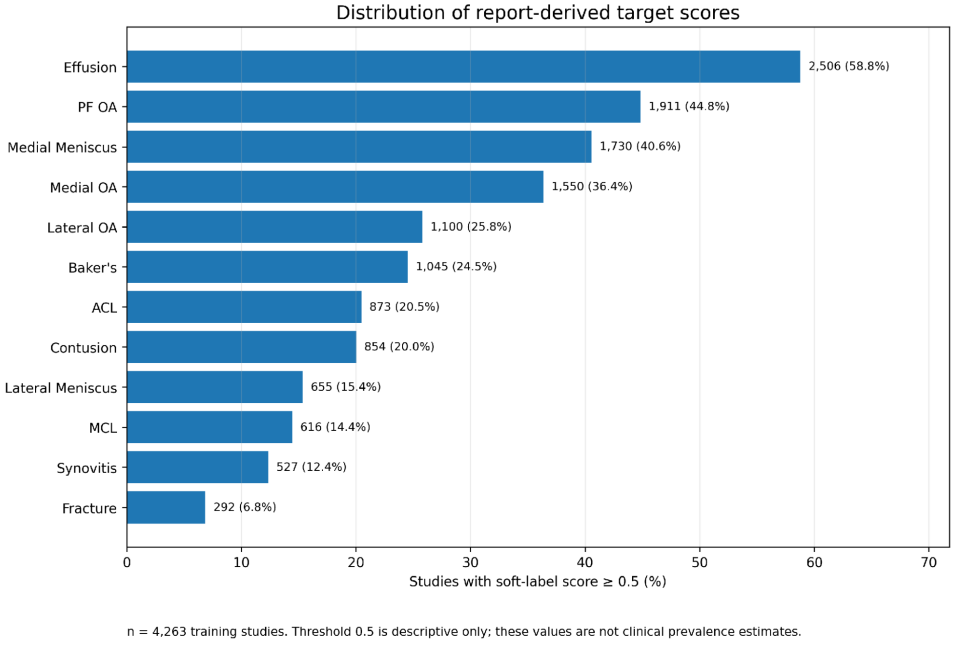


Figure 3: Number and percentage of training studies with a soft-label score ≥ 0.5 for each target. The threshold is descriptive and does not estimate clinical disease prevalence.

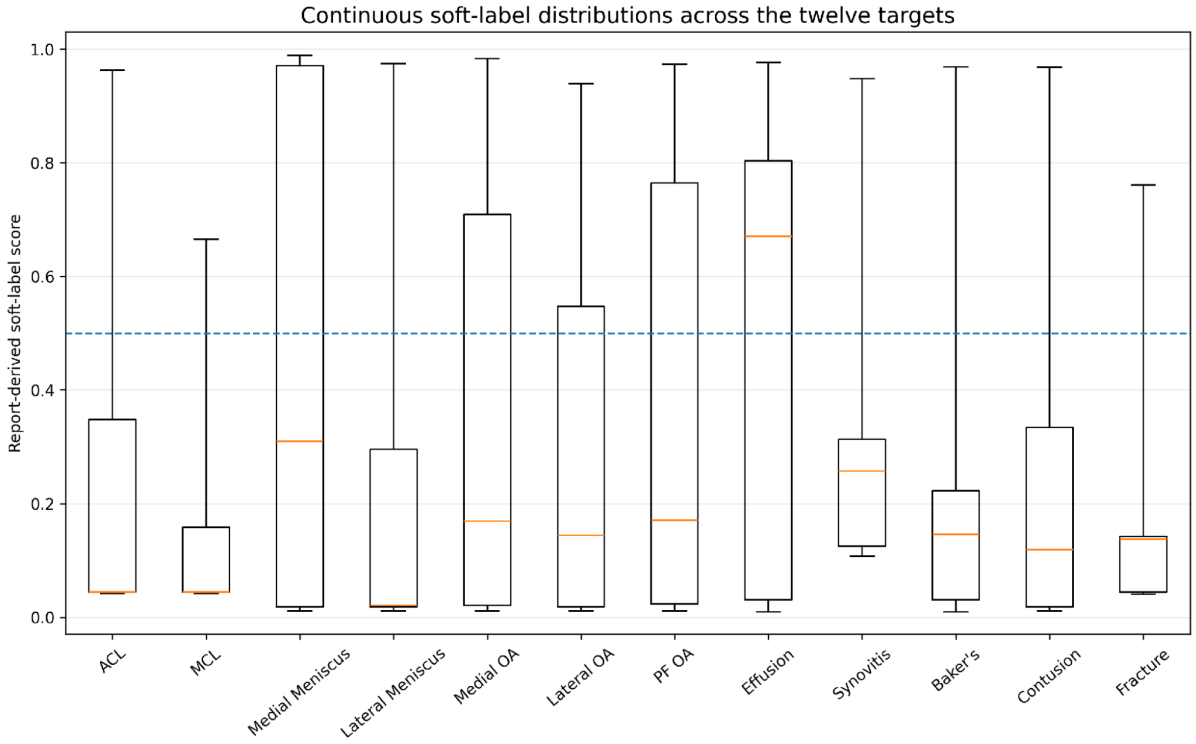


Figure 4: Continuous report-derived soft-label distributions for the twelve targets. The dashed line indicates the exploratory threshold of 0.5.

categories.

The threshold-based summary is useful for illustrating the relative frequency of high soft-label scores in the training data, but it must not be interpreted as the prevalence of confirmed abnormalities. The labels were generated from report-derived sources, and the threshold of 0.5 was selected for descriptive purposes rather than validated against an independent clinical reference. Differences between targets may reflect actual variation in findings, report-writing conventions, the behaviour of the label extraction systems or some combination of these factors. A clinically meaningful prevalence estimate would require a representative sample with independently established disease status.

The continuous distributions in Figure 4 provide complementary information. A threshold compresses a continuous score into two groups and can conceal uncertainty near the cut-off. Conversely, means and medians alone can conceal skewness, multimodality or source disagreement. Future analyses should examine source-specific scores, the distribution of differences between label sources, and performance stratified by score confidence. If a subset of cases can be independently adjudicated, that subset could quantify the relationship between soft scores and clinical truth, including whether calibration differs by target.

Associations Between Target Scores

Spearman analysis identified a strong positive association between effusion and synovitis scores ($\varrho = 0.903$). Other prominent associations included medial and lateral osteoarthritis ($\varrho = 0.653$), medial osteoarthritis and patellofemoral osteoarthritis ($\varrho = 0.612$), ACL and MCL ($\varrho = 0.575$), and lateral osteoarthritis and patellofemoral osteoarthritis ($\varrho = 0.564$). The leading associations are summarised in Table 4, and the full matrix is shown in Figure 5. These relationships indicate that the report-derived scores contain substantial dependencies rather than behaving as twelve statistically independent targets. The effusion–synovitis relationship is plausible in clinical reporting because these findings may be described together, but the observed correlation cannot be attributed solely to biological co-occurrence. Shared wording, shared report context or correlated errors from the label-generation pipeline may also contribute. Likewise, correlations between osteoarthritis compartments may reflect the clinical distribution of degenerative change, the structure of report descriptions, or both. Without independently adjudicated labels, the analysis cannot separate these explanations.

Target dependence matters for model development. A shared encoder can benefit from common anatomical representations, while finding-specific attention allows each target to select different evidence. However, strong correlations can also make a model appear successful on one target by exploiting features associated with another. Ablation studies, conditional error analysis and independent reference labels would help establish whether each target is detected from relevant evidence. In particular, performance should be examined separately for correlated findings and for cases where the report-derived labels disagree.

Table 2: Largest observed pairwise Spearman correlations between soft-label scores

Target pair	Spearman's ϱ	Interpretive caution
Effusion – Synovitis	0.903	May reflect co-reporting and shared label-source behaviour
Medial OA – Lateral OA	0.653	May reflect multi-compartment degeneration or report structure
Medial OA – Patellofemoral OA	0.612	Not proof of clinical co-occurrence
ACL – MCL	0.575	May reflect associated injury patterns and report co-mention
Lateral OA – Patellofemoral OA	0.564	Requires validation against independent labels

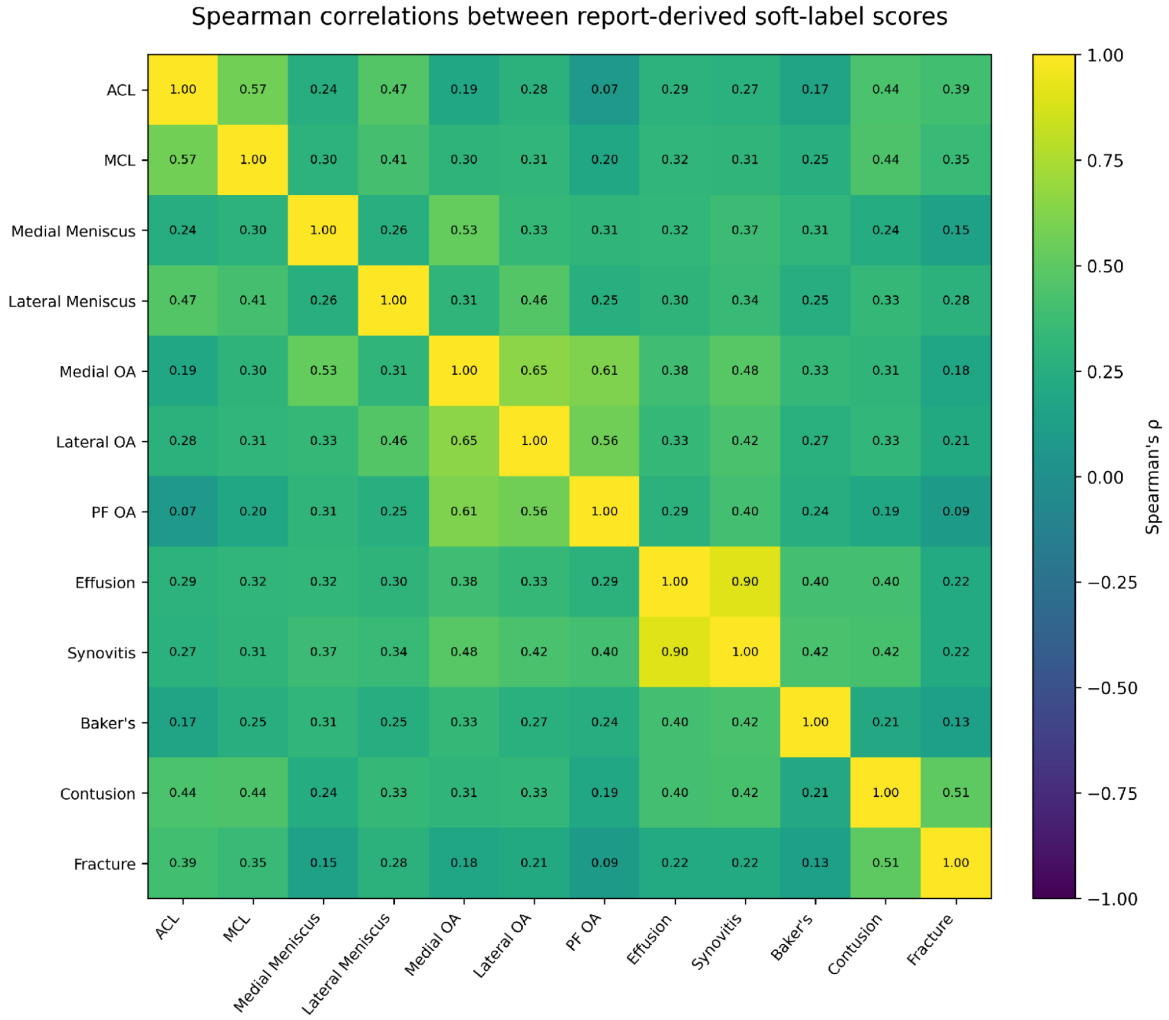


Figure 5: Spearman correlation matrix for the twelve report-derived soft-label scores. Correlations describe score relationships and do not establish clinical co-occurrence

Training Convergence and Model Integrity

The recorded training loss declined from 0.55166 in epoch 1 to 0.39110 in epoch 16, corresponding to a 29.1% reduction. The completion record reports 34,096 optimiser steps. The loss decreased across the recorded epochs, with smaller incremental reductions in the later epochs. Figure 6 shows the recorded curve. This pattern is consistent with optimisation of the chosen soft-target binary cross-entropy objective, but it does not establish that the model has learned clinically generalisable features or that it performs well on unseen studies.

The archive integrity checks support the identity of the analysed artefacts. The ZIP archive passed its integrity check, the checkpoint hash matched the manifest, and the 25 training files listed in the checksum file matched their recorded SHA-256 hashes. The checkpoint contained 362 tensors and approximately 50.02 million tensor elements, consistent with a high-capacity ConvNeXt-Small-based model and attention head. These checks reduce uncertainty about file corruption but do not validate the scientific claims associated with the checkpoint.

The package documentation reports a public leaderboard score of 0.950. That value was not independently reproduced in the present analysis, and the archive does not provide the prediction file, scoring script output or held-out reference labels needed to verify it. It should therefore



Figure 6: Recorded training loss over 16 epochs. The 29.1% reduction indicates optimisation progress only; the all-data checkpoint has no held-out evaluation

be described, if retained at all, as a score reported by the package documentation rather than as an independently measured result of this study. This distinction is important because benchmark scores depend on the test set, target definition and evaluation implementation.

Additional Analyses of Target-Label Structure

To characterise the multi-label structure beyond per-target summaries, the number of report-derived target scores ≥0.5 was counted for each of the 4,263 training studies (Figure 7). The mean was 3.20 high-scoring targets per

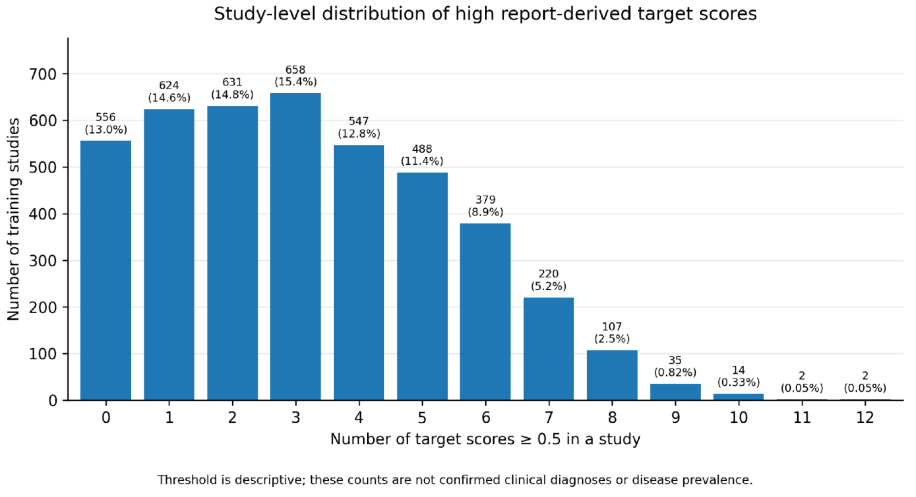


Figure 7: Study-level distribution of the number of report-derived target scores ≥0.5 across 4,263 training studies. The threshold is descriptive and does not represent confirmed clinical diagnoses

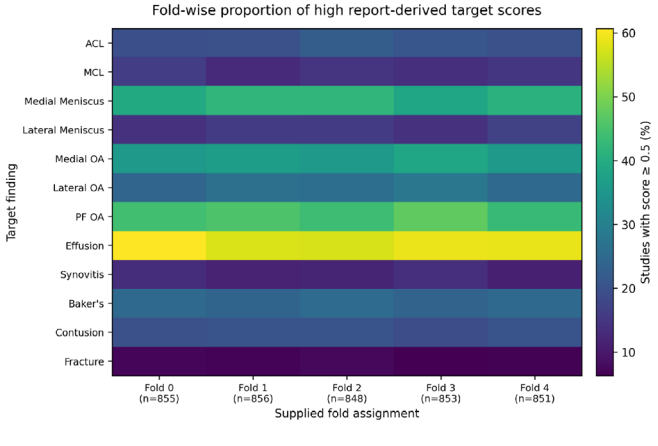


Figure 8: Fold-wise percentage of studies with a report-derived target score ≥0.5 for each finding. The fold assignments describe label composition and do not establish patient-level independence or diagnostic performance

study. In total, 556 studies (13.0%) had no scores ≥0.5, 624 (14.6%) had one, and 3,083 (72.3%) had at least two; 2,452 (57.5%) had at least three. Eighteen studies (0.42%) had ten or more. These counts describe the supplied scores, not confirmed co-occurring abnormalities.

The supplied fold assignments contained 848–856 eligible studies per fold (Figure 8). For each target, the proportion with a score ≥0.5 was calculated within each fold. The largest maximum-to-minimum range was 4.24 percentage points for PF OA. This describes variation in label-score composition across the provided folds; it does not establish patient-level independence, held-out diagnostic performance or clinical generalisation.

Figure 9 further separates each target's continuous scores into four descriptive intervals. The score composition was heterogeneous: 92.4% of fracture scores were below 0.20, whereas 47.7% of synovitis scores fell between 0.20 and <0.50. For effusion, 33.8% of scores were between 0.50 and <0.80 and 25.0% were ≥0.80. These patterns show why a single threshold can conceal differences in label-score composition. The intervals are analytical bins, not clinically calibrated probability bands.

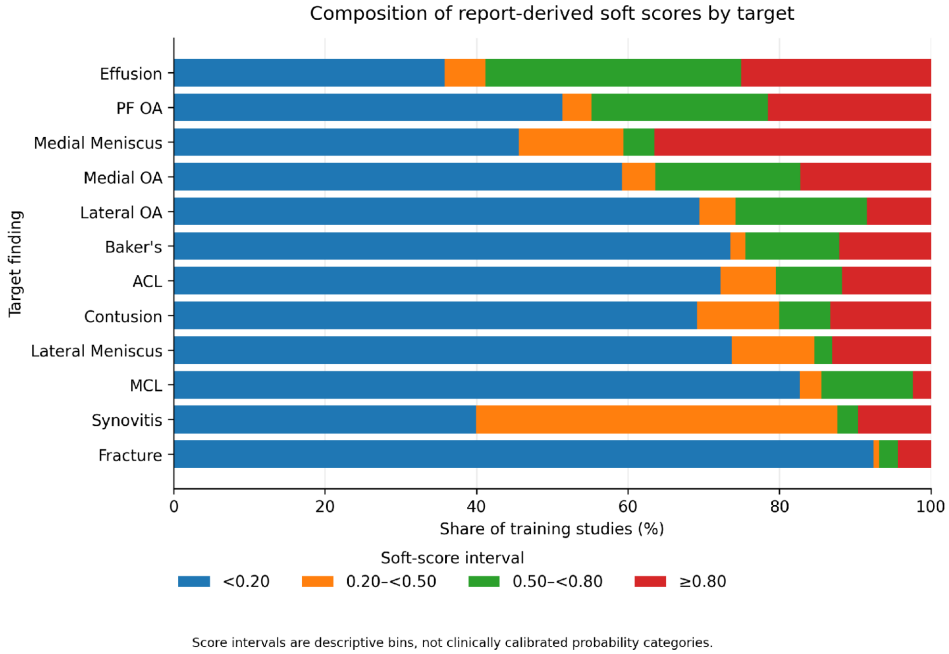


Figure 9: Composition of report-derived soft scores across four descriptive intervals for each target. The intervals are analytical bins, not clinically calibrated probability categories

Implications for Model Development and Clinical Interpretation

The descriptive results support several practical conclusions for further model evaluation. First, target-specific analyses are necessary because the label-score distribution differs substantially between findings. Rare high scores, such as those for fracture, may create a different operating regime from frequent high scores, such as effusion. Evaluation should therefore report per-target metrics and the number of positive and negative reference cases rather than only a macro-average. Precision–recall curves may be particularly informative for less frequent findings, but only if the labels are sufficiently reliable and the evaluated predictions are genuinely held out (Saito & Rehmsmeier, 2015).

Secondly, the correlations indicate that multi-label learning should be examined in terms of both shared and target-specific representations. A common encoder may exploit recurring anatomical patterns, while target-specific spatial and window attention may help accommodate differences in location and spatial extent. Nonetheless, attention should not be assumed to guarantee interpretability. Qualitative examples should be reviewed by musculoskeletal radiologists, and any localisation claim should be tested against appropriate annotations. If such annotations are unavailable, the paper should describe attention weights as internal model coefficients rather than explanatory evidence.

Thirdly, independent evaluation must be designed before finalising the paper's diagnostic claims. A fold-based evaluation should produce predictions only for studies excluded from the corresponding model's training. The available all-data checkpoint cannot serve as its own test model, and in-sample scores would be optimistically biased. A complete out-of-fold run would allow estimation of agreement with report-derived targets; it would still not establish clinical accuracy unless the evaluation labels were independent of the training label pipeline. An adjudicated reference subset or external cohort would provide stronger evidence.

Fourthly, generalisation needs to be examined across data sources and acquisition conditions. Knee MRI studies can differ in scanner hardware, sequences, spatial resolution and reporting practice. The wider medical-imaging literature shows that distribution shifts can affect model reliability and subgroup performance, even when internal

metrics are strong (Yang *et al.*, 2024). If metadata become available, analyses should consider performance by site, scanner, sequence and relevant patient subgroups. Such analyses should be planned in advance and should not be performed on demographic variables that are absent or unreliable in the source package.

Finally, the intended clinical role should remain modest until evidence improves. A study-level model may eventually support triage, second reading or structured reporting, but its use could introduce automation bias if clinicians over-rely on a prediction that is not well calibrated or robust to unfamiliar data. The literature on clinical translation stresses workflow integration, prospective evaluation and clear accountability, rather than technical accuracy alone (Kelly *et al.*, 2019; Guermazi *et al.*, 2024). The current evidence supports continued evaluation of the model, not clinical deployment.

CONCLUSION

This study provides a transparent descriptive analysis of a weakly supervised, multi-label knee MRI model package combining ConvNeXt-Small, finding-specific spatial attention and window-attention multiple-instance learning. The supplied archive contains 4,263 studies used for final all-data training and twelve report-derived soft targets whose distributions vary substantially across findings. Effusion and synovitis scores show the strongest observed association, while the recorded training loss decreases by 29.1% over sixteen epochs. These findings document the training data and optimisation process, but they do not establish clinical diagnostic performance. The available checkpoint has no held-out evaluation, and the soft labels are not independent clinical reference diagnoses. Further work should generate leakage-controlled out-of-fold predictions, evaluate the model against independently established labels, compare the attention components with appropriate baselines and report uncertainty and calibration. The framework should be considered a research model pending such validation, and claims of clinical utility or superiority should be withheld until supported by independent evidence.

REFERENCES

Bien, N., Rajpurkar, P., Ball, R. L., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B. N., Yeom, K. W., Shpanskaya, K., Halabi, S., Zucker, E., Amanatullah, D. F., Beaulieu, C. F., Riley, G. M., Stewart, R. J., Blankenberg, F. G., Larson, D. B., Jones, R. H., … Lungren, M. P. (2018). Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLOS Medicine, 15*(11), e1002699. https://doi.org/10.1371/journal.pmed.1002699

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C. W., & Kressel, H. Y. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ, 351,* h5527. https://doi.org/10.1136/bmj.h5527

Cai, L., Huang, S., Zhang, Y., Lu, J., & Zhang, Y. (2025). AttriMIL: Revisiting attention-based multiple instance learning for whole-slide pathological image classification from a perspective of instance attributes. *Medical Image Analysis, 103*, 103631. https://doi.org/10.1016/j.media.2025.103631

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, e078378. https://doi.org/10.1136/bmj-2023-078378

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837–845. https://doi.org/10.2307/2531595

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Gadermayr, M., & Tschuchnig, M. (2024). Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations and future potential. *Computerized Medical Imaging and Graphics, 112*, 102337. https://doi.org/10.1016/j.compmedimag.2024.102337

Gill, S. S., Haq, T., Zhao, Y., Ristic, M., Amiras, D., & Gupte, C. M. (2026). AI demonstrates comparable diagnostic performance to radiologists in MRI detection of anterior cruciate ligament tears: A systematic review and meta-analysis. *European Radiology, 36*, 2500–2517. https://doi.org/10.1007/s00330-025-12020-2

Gitto, S., Serpi, F., Albano, D., Risoleo, G., Fusco, S., Messina, C., & Sconfienza, L. M. (2024). AI applications in musculoskeletal imaging: A narrative review. *European Radiology Experimental, 8*, Article 22. https://doi.org/10.1186/s41747-024-00422-8

Guermazi, A., Omoumi, P., Tordjman, M., Fritz, J., Kijowski, R., Regnard, N.-E., Carrino, J., Kahn, C. E., Knoll, F., Rueckert, D., Roemer, F. W., & Hayashi, D. (2024). *How AI may transform musculoskeletal imaging. Radiology, 310*(1), e230764. https://doi.org/10.1148/radiol.230764

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330). PMLR.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Ilse, M., Tomczak, J. M., & Welling, M. (2018). Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2127–2136). PMLR.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 590–597. https://doi.org/10.1609/aaai.v33i01.3301590

Johnson, A. E. W., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-Y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., & Horng, S. (2019). MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *Scientific Data, 6*, Article 317. https://doi.org/10.1038/s41597-019-0322-0

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine, 17*, Article 195. https://doi.org/10.1186/s12916-019-1426-2

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11976–11986). https://doi.org/10.1109/CVPR52688.2022.01167

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Mead, K., Cross, T., Roger, G., Sabharwal, R., Singh, S., & Giannotti, N. (2025). MRI deep learning models for assisted diagnosis of knee pathologies: A systematic review. *European Radiology, 35*, 2457–2469. https://doi.org/10.1007/s00330-024-11105-8

Misera, L., Müller-Franzes, G., Truhn, D., & Kather, J. N. (2024). Weakly supervised deep learning in radiology. *Radiology, 312*(1), e232085. https://doi.org/10.1148/radiol.232085

Mohammadi, S., Jahanshahi, A., Shahrabi Farahani, M., Salehi, M. A., Frounchi, N., & Guermazi, A. (2025). Diagnosis of knee meniscal injuries using artificial intelligence: A systematic review and meta-analysis of diagnostic performance. *PLOS ONE, 20*(6), e0326339. https://doi.org/10.1371/journal.pone.0326339

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E. J., Ioannidis, J. P. A., Collins, G. S., & Maruthappu, M. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ, 368*, m689. https://doi.org/10.1136/bmj.m689

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems* (Vol. 32).

Qiu, Z., Xie, Z., Lin, H., Li, Y., Ye, Q., Wang, M., Li, S., Zhao, Y., & Chen, H. (2024). Learning co-plane attention across MRI sequences for diagnosing twelve types of knee abnormalities. *Nature Communications, 15*, Article 7637. https://doi.org/10.1038/s41467-024-51888-4

Radiological Society of North America. (n.d.). RSNA Knee Abnormality Detection [Data set]. Kaggle. *Retrieved October 21*, 2026, from https://www.kaggle.com/competitions/rsna-knee-abnormality-detection/data

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, J., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., AIX-COVNET, Rudd, J. H. F., Sala, E., Escudero Sánchez, L., Delgado Ortet, M., Martin Gonzalez, P., … Schönlieb, C.-B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence, 3*, 199–217. https://doi.org/10.1038/s42256-021-00307-0

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A. Y., & Lungren, M. P. (2020). CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 1500–1519). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.117

Sounderajah, V., Guni, A., Liu, X., Collins, G. S., Karthikesalingam, A., Markar, S. R., Golub, R. M., Denniston, A. K., Shetty, S., Moher, D., Bossuyt, P. M., Darzi, A., Ashrafian, H., & STARD-AI Steering Committee. (2025). The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature Medicine, 31*, 3283–3289. https://doi.org/10.1038/s41591-025-03953-8

Tejani, A. S., Klontzas, M. E., Gatti, A. A., Mongan, J. T., Moy, L., Park, S. H., & Kahn, C. E., Jr. (2024). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiology: Artificial Intelligence, 6*(4), e240300. https://doi.org/10.1148/ryai.240300

Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: Methodological failures and recommendations for the future. *npj Digital Medicine, 5*, Article 48. https://doi.org/10.1038/s41746-022-00592-y

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, A.

(2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30).

Xie, Z., Qiu, Z., Li, Y., Chen, X., Ye, Q., Wang, M., Chen, X., Shao, Y., Liu, C., Song, L., Lin, H., Li, S., Zhang, R., Li, T., Chen, L., Duan, Y., Li, C., Chen, H., & Zhao, Y. (2025). Development of a multi-task deep learning system for classification of nine common knee abnormalities on MRI: A large-scale, multicentre, stepwise validation study. *eClinicalMedicine, 89*, 103534. https://doi.org/10.1016/j.eclinm.2025.103534

Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D., & Ghassemi, M. (2024). The limits of fair medical imaging AI in real-world generalization. *Nature Medicine, 30*, 2838–2848. https://doi.org/10.1038/s41591-024-03113-4

Zeng, X., Abdullah, N., & Sumari, P. (2024). Self-supervised learning framework application for medical image analysis: A review and summary. *BioMedical Engineering OnLine, 23*, Article 107. https://doi.org/10.1186/s12938-024-01299-9

Zhao, Y., Coppola, A., Karamchandani, U., Amiras, D., & Gupte, C. M. (2024). Artificial intelligence applied to magnetic resonance imaging reliably detects the presence, but not the location, of meniscus tears: A systematic review and meta-analysis. *European Radiology, 34*, 5954–5964. https://doi.org/10.1007/s00330-024-10625-7