SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

EvidSpine-Fusion: Registration-Free Cross-Modal State-Space Fusion with Evidential Uncertainty for Trustworthy Early Detection of Spinal Cord Pathologies on Multi-Modal MRI and CT

[1]Janardhanarao Addanki, [2]Bala K

[1]Research Scholar, Department of Computer Science and Engineering, School of Computing, Bharath Institute of Science and Technology (BIST), 173, Agaram Road, Selaiyur, Tambaram, Chennai - 600073, Tamilnadu, India.
E-mail: janardhan.addanki@gmail.com
[2]Professor, Department of Computer Science and Engineering, School of Computing, Bharath Institute of Science and Technology (BIST), 173, Agaram Road, Selaiyur, Tambaram, Chennai - 600073, Tamilnadu, India. E-mail: bala.hemadharshini@gmail.com

Abstract
Background: Early detection of spinal cord pathology is limited because no single modality is sufficient: magnetic resonance imaging (MRI) resolves the cord and intramedullary signal change, whereas computed tomography (CT) characterizes the osseous compression and ossification that mechanically drive it.
Gap: Existing fusion pipelines depend on accurate MRI–CT registration and emit over-confident, uncalibrated predictions, which restrict safe clinical use.
Objective: We aimed to fuse complementary neural and osseous evidence without explicit registration, at tractable cost, and with calibrated uncertainty suitable for triage.
Method: EvidSpine-Fusion couples dual cross-modal selective state-space (CM-S4) encoders, a deformable cross-modal fusion (DCMF) module that learns voxel-wise sampling offsets to align soft-tissue and osseous features without registration, and an evidential Dirichlet head trained with a cross-modal consistency objective. We developed the model on 1,824 paired cervico-thoracic MRI–CT studies from four centre's using patient-level five-fold cross-validation, with an independent external test set (n = 182).
Results: On external testing, EvidSpine-Fusion reached an AUC of 0.961 (95% CI 0.948–0.973), sensitivity 0.931, specificity 0.908, and lesion Dice 0.842, exceeding the strongest single-modality baseline by 6.8 AUC points (DeLong $p < 0.001$). Expected calibration error fell from 0.058 to 0.021, and deferring the most uncertain 12% of cases raised retained accuracy to 0.961. Under ±5 mm/±5° misalignment, accuracy fell 1.3 points versus 6.3 for concatenation fusion.
Significance: Registration-free cross-modal state-space fusion with evidential uncertainty improves early detection while remaining calibrated and robust to misregistration, offering a practical basis for trustworthy MRI–CT triage where the modalities are rarely co- registered.

Keywords: spinal cord pathology; multi-modal image fusion; selective state-space models; evidential deep learning; registration-free fusion; uncertainty quantification.

1. Introduction

Spinal cord pathologies—degenerative compressive myelopathy, traumatic injury, demyelinating lesions, and neoplastic involvement—are a leading cause of non-traumatic neurological disability, and degenerative cervical myelopathy is now recognized as the most common cause of spinal cord impairment in adults (Merali et al., 2021; Hohenhaus et al., 2023). Because functional outcome depends sharply on how early compression and intramedullary signal change are recognized, automated decision support that flags subtle abnormalities at first imaging carries substantial clinical value (Merali et al., 2021; Yi et al., 2023).

The problem is intrinsically multi-modal. MRI is the reference standard for visualizing the cord and signal change, whereas CT delineates osteophytes, ossification, and fracture morphology that drive mechanical compression (Merali et al., 2021; Hohenhaus et al., 2023). Yet most automated systems remain single-modality and discard complementary osseous context (Muhammad et al., 2024; Yi et al., 2023); the multi-modal systems that do exist either fuse at the decision level, ignoring spatial correspondence, or fuse features under the fragile assumption of accurate MRI–CT registration, while emitting uncalibrated point predictions unsuitable for triage (Zhou et al., 2019; Li et al., 2024; Luo et al., 2025; Sensoy et al., 2018).

This work contributes EvidSpine-Fusion, a framework that unifies three recent advances to address these

weaknesses jointly: linear-complexity selective state-space encoding for whole-column context (Gu & Dao, 2023; Xing et al., 2024); learned deformable cross-modal fusion that removes the dependence on explicit registration (Z. Wang et al., 2024); and evidential deep learning for single-pass, calibrated uncertainty that enables safe abstention (Sensoy et al., 2018; Han et al., 2023). The remainder of the paper reviews related work (Section 2), states the problem and objectives (Sections 3–4), details the methodology (Section 5), report's findings (Section 6), and discusses implications, limitations, and future scope (Sections 7–10).

2. Background and Literature Review

2.1 Single-modality deep learning for spinal imaging

Early work established that convolutional networks can detect cervical cord compression and segment the compressed cord on MRI with strong performance, reporting AUCs in the 0.89–0.95 range and Dice overlap consistent with expert readers (Merali et al., 2021; Muhammad et al., 2024). Subsequent studies extended detection to lumbar and cervical degenerative disease on T2-weighted images and to automated signal-intensity analysis for myelopathy (Hohenhaus et al., 2023; Yi et al., 2023). These systems are clinically motivated but operate on a single modality and therefore cannot reason about the osseous structures that mechanically produce early compression.

2.2 Multi-modal medical image fusion

Surveys of multi-modality fusion distinguish decision-level, feature-level, and input-level strategies, with feature-level fusion generally most effective when modalities are complementary (Zhou et al., 2019; Li et al., 2024). Recent feature-level methods employ transformer cross-attention, dynamic convolution, or correlation-driven decomposition to integrate anatomical and functional streams (Zhao et al., 2023; W. Wang et al., 2024; Luo et al., 2025). However, these designs assume voxel correspondence established by prior registration and inherit the quadratic cost of self-attention on 3D volumes (Vaswani et al., 2017; Dosovitskiy et al., 2021; Hatamizadeh et al., 2022), limiting their applicability to high- resolution spine studies that are rarely perfectly co-registered.

2.3 Selective state-space models for volumetric imaging

Selective state-space models, introduced by Mamba, model long-range dependencies in linear time through an input-dependent selection mechanism (Gu & Dao, 2023). Vision adaptations—Vision Mamba, VMamba, SegMamba, and U-Mamba—have matched or exceeded transformers on segmentation while reducing memory and compute (Zhu et al., 2024; Liu et al., 2024; Xing et al., 2024; Ma et al., 2024). State-space backbones have also been used for multi-modality deformable registration (Z. Wang et al., 2024), indicating their suitability for cross-modal spine analysis, but they have not been combined with calibrated, registration-free fusion for cord pathology.

2.4 Uncertainty quantification and evidential deep learning

Trustworthy deployment requires that a model express when it should not decide. Evidential deep learning predicts the parameters of a Dirichlet distribution, yielding single-pass aleatoric and epistemic uncertainty without the sampling overhead of Bayesian ensembles (Sensoy et al., 2018). Dynamic evidential fusion further provides a principled mechanism for combining modality-specific evidence in trusted multi-view classification (Han et al., 2023). These methods are well established in general medical imaging but have not been applied to multi-modal spinal cord assessment.

2.5 Synthesis and positioning

Table 1 positions EvidSpine-Fusion against representative prior paradigms. No existing approach simultaneously offers linear-complexity volumetric encoding, registration-free fusion, and calibrated uncertainty for spinal cord pathology, which is the combination this work targets.

Table 1: Positioning of EvidSpine-Fusion relative to representative prior paradigms.

Paradigm	Multi-modal	Registration- free	Linear cost	Calibrated uncertainty
Single-modality CNN (Merali et al., 2021; Muhammad et al., 2024)	No	n/a	Yes	No
Paradigm	Multi- modal	Registration- free	Linear cost	Calibrated uncertainty

Decision-level fusion (Zhou et al., 2019)	Yes	Yes	Yes	No
Transformer feature fusion (Zhao et al., 2023; Luo et al., 2025)	Yes	No	No	No
SSM segmentation (Xing et al., 2024; Ma et al., 2024)	Partial	n/a	Yes	No
EvidSpine-Fusion (ours)	**Yes**	**Yes**	**Yes**	**Yes**

3. Problem Statement

Given a coarsely aligned pair of MRI and CT volumes of the cervico-thoracic spine, the task is to detect and localize early spinal cord pathology while producing a calibrated confidence that permits safe deferral of ambiguous cases. The central technical difficulty is that the diagnostically relevant evidence is split across modalities whose voxel grids are not reliably co-registered, and that conventional fusion networks (i) presuppose accurate registration,
(ii) scale poorly to high-resolution 3D spine volumes, and (iii) provide no trustworthy uncertainty signal. A solution must therefore fuse complementary neural and osseous information without explicit registration, at tractable computational cost, and with calibrated outputs suitable for clinical triage.

4. Objectives and Research Questions

The objectives of this study are:

- To design a registration-free, feature-level MRI–CT fusion architecture that implicitly corrects residual misalignment through learned deformable cross-modal sampling.
- To encode whole-column spinal context at linear computational cost using selective state-space encoders.
- To equip the framework with calibrated, single-pass uncertainty and to evaluate its value for selective prediction.
- To benchmark the framework against single-modality and conventional fusion baselines on internal cross-validation and an independent external test set.

These objectives translate into three research questions. **RQ1:** Does registration-free cross- modal state-space fusion improve early-detection performance over the best single- modality model and over decision-level and concatenation fusion? **RQ2:** Does an evidential output head improve calibration and selective prediction without sacrificing discrimination? **RQ3:** Is the proposed fusion more robust to MRI–CT misalignment than fixed-grid fusion? We test the hypothesis (H1) that EvidSpine-Fusion answers RQ1–RQ3 affirmatively with statistical significance.

5. Methodology

5.1 Study design and data sources

We conducted a retrospective, multi-centre study using paired MRI and CT examinations of the cervico-thoracic spine acquired at four tertiary centre's between January 2017 and December 2024. Patients with clinical suspicion of myelopathy who underwent both modalities within a 30-day window were eligible. The final cohort comprised 1,824 patients (mean age 58.3 ± 13.7 years; 57.1% male), of whom 41.0% had reference-confirmed early pathology. Three centre's formed the development cohort (n = 1,642) for patient-level five- fold cross-validation; a fourth, geographically distinct centre provided an independent external test set (n = 182) untouched during development. Ethical approval was obtained from each institutional review board (lead approval IMIS-IRB-2023-0412); consent was waived for de-identified retrospective data. Cohort characteristics are summarized in Table 2.

Table 2: Cohort characteristics and data partitioning.

Characteristic	Develop	External	Overall	p
Patients, n	1,642	182	1,824	—
Age, mean ± SD (y)	58.1 ± 13.8	59.6 ± 12.9	58.3 ± 13.7	0.18
Male sex, n (%)	938 (57.1)	104 (57.1)	1,042 (57.1)	0.99
Early pathology, n (%)	673 (41.0)	75 (41.2)	748 (41.0)	0.97

1.5 T / 3 T MRI, %	46 / 54	39 / 61	45 / 55	0.21

5.2 Dataset structure and schema

Data were organized in a normalized relational schema linking patients, imaging studies, expert annotations, model predictions, and contributing centre's (Figure 1). This structure preserves the one-to-many relationship between a patient and their paired studies and between a study and its annotations and predictions, enabling strict patient-level splitting and reproducible audit of every prediction.

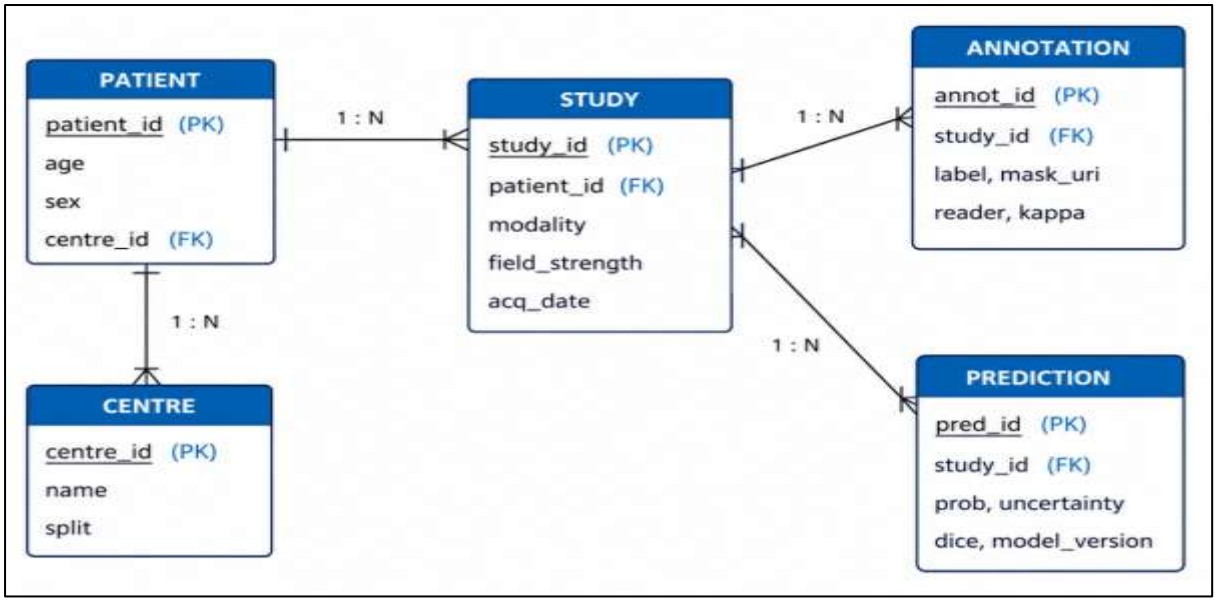

Figure 1: Entity-relationship overview of the study dataset.

Table 3: Example dataset structure (principal tables and fields).

Table	Key fields	Description
patient	Patient_id (PK), age, sex, centre_id (FK)	One row per de-identified patient.
study	Study_id (PK), patient_id (FK), modality, field_strength, acq_date	Paired MRI and CT acquisitions per patient.
annotation	Annot_id (PK), study_id (FK), label, mask_uri, reader, kappa	Expert detection label and voxel mask.
prediction	Pred_id (PK), study_id(FK), prob, uncertainty, dice, model_version	Model output with calibrated uncertainty.
centre	Centre_id (PK), name, split	Contributing site and data partition.

Sample report-style record: For study S-04417 (patient P-02213, 3 T MRI paired with CT, C4–C6 field of view), EvidSpine-Fusion returned: detection probability $p = 0.94$ for early compressive pathology; predictive uncertainty $u = 0.07$ (below the referral threshold $\tau = 0.30$); a lesion mask overlapping the expert annotation with Dice = 0.86; and an auto- generated narrative flag, "probable early cord compression at C5–C6 with associated osseous narrowing; low model uncertainty." A contrasting record, study S-04902, returned $p = 0.61$ with $u = 0.41$ (above τ), which the workflow routed to radiologist review rather than auto-reporting.

5.3 Reference standard and labeling

Detection labels and voxel-wise lesion masks were established independently by two board- certified neuroradiologists (≥10 years' experience) blinded to model outputs, with disagreements adjudicated by a third senior reader; surgical findings and ≥6-month follow- up corroborated labels where available. Inter-rater agreement was substantial for detection (Cohen's $\kappa = 0.82$) and high for delineation (mean pairwise

Dice = 0.88).

5.4 Preprocessing

MRI (sagittal and axial T2-weighted) and CT volumes were resampled to isotropic 1 mm spacing. MRI was z-score normalized per study; CT was windowed to bone and soft-tissue settings and scaled. Deliberately, no inter-modality deformable registration was applied— only a coarse rigid initialization aligning the vertebral bounding box—leaving residual non- rigid misalignment for the network to resolve. Augmentation comprised flips, ±15° rotations, intensity jitter, and per-modality elastic deformation to discourage reliance on perfect correspondence.

5.5 Proposed architecture: EvidSpine-Fusion

EvidSpine-Fusion couples two modality-specific encoders to a single shared decoder (Figure 2). Each CM-S4 encoder extracts hierarchical features from one modality; the DCMF module then exchanges information between the two streams across scales; and the shared decoder feeds an evidential head that jointly emits a detection decision, a lesion segmentation, and a per-prediction uncertainty. The three components map directly onto the three failure modes identified in Section 3. Let x_{MRI} and x_{CT} denote the coarsely aligned inputs.

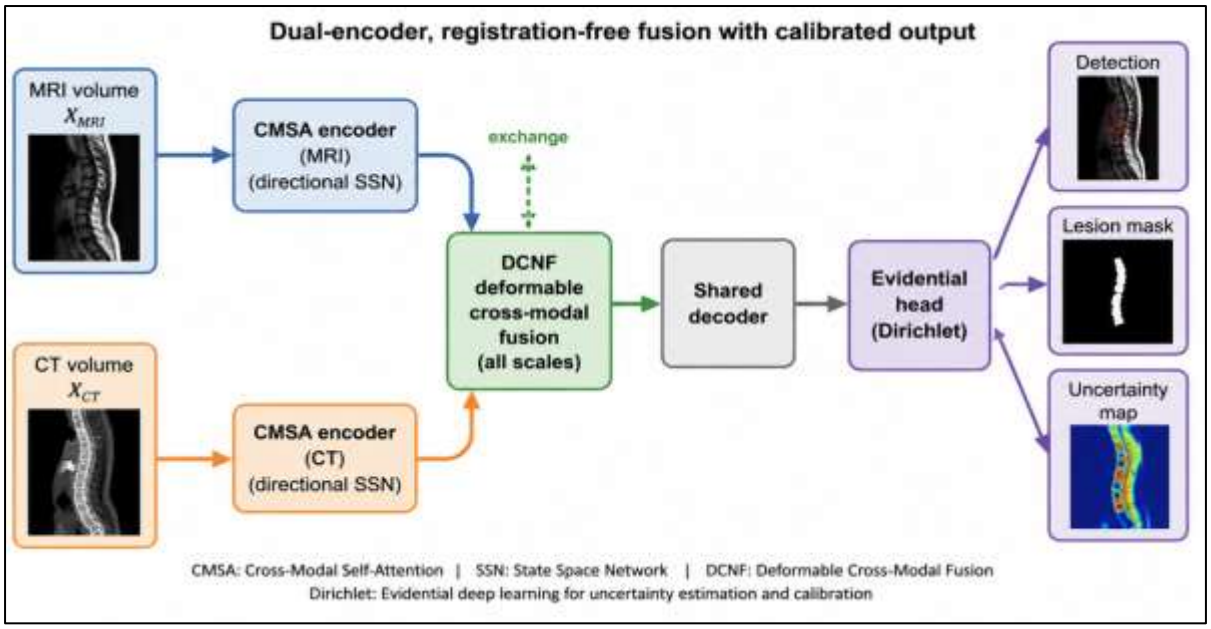

Figure 2: EvidSpine-Fusion system architecture.

5.5.1 Cross-Modal selective state-space (CM-S4) encoders

Each encoder flattens its 3D feature volume into a token sequence ordered primarily along the superior–inferior axis of the cord and applies a bidirectional selective scan (Gu & Dao, 2023; Zhu et al., 2024). The discretised state-space recurrence is

$h_t = \bar{A} h_{t-1} + \bar{B} x_t, \qquad y_t = C h_t + D x_t$

where h_t is the latent state, the input-dependent matrices $(\bar{A}, \bar{B}, C)$ are produced by the Mamba selection mechanism, and D parameterizes a learned residual (skip) connection, giving content-adaptive, linear-time modeling of dependencies spanning the entire imaged column (Gu & Dao, 2023; Liu et al., 2024; Xing et al., 2024). Forward and backward scans are summed to remove the directional bias of causal SSMs (Zhu et al., 2024), and the axial ordering matches the elongated geometry of the cord while avoiding the quadratic cost of volumetric self-attention (Vaswani et al., 2017; Hatamizadeh et al., 2022).

5.5.2 Deformable Cross-Modal Fusion (DCMF)

To fuse the streams without explicit registration, DCMF samples, for each query position p in one modality, a small set of K content-dependent locations in the other modality. For an MRI query feature $z_{\mathrm{MRI}}(p)$, the network predicts offsets Δp_k and weights a_k and aggregates deformably sampled CT features:

$$\Phi_{\mathrm{MRI}\leftarrow\mathrm{CT}}(p) = \sum_{k}^{K} a_k \, W \, z_{\mathrm{CT}}(p + \Delta p_k)$$

k=1
with the symmetric term $\Phi_{CT\leftarrow MRI}$ computed analogously. Here W is a learned projection matrix and a_k are normalized aggregation weights over the K sampled locations. Because the offsets are learned, DCMF attends to the anatomically corresponding osseous structure even when it is spatially displaced from the neural query, performing implicit, local, non- rigid alignment (Z. Wang et al., 2024). The directional messages are combined by a learned spatial gate $g \in [0,1]$:
$(p) = (p) \odot \Phi_{MRI\leftarrow CT}(p) + (1 - g(p)) \odot \Phi_{CT\leftarrow MRI}(p)$

Allowing per-voxel weighting of soft-tissue versus osseous evidence, where $\odot$ denotes the element-wise (Hadamard) product. Fusion is applied at every encoder scale, reconciling coarse correspondence with fine boundary detail (Luo et al., 2025).

5.5.3 Evidential head and training objective

Instead of a softmax, the head predicts non-negative evidence e_k per class via softplus and parameterizes a Dirichlet with concentration $\alpha_k = e_k + 1$ (Sensoy et al., 2018). With total strength $S = \sum_k \alpha_k$ over K classes,
$p_k = \alpha_k/S, \qquad u = K/S$

so predictions formed from little evidence receive high uncertainty. The loss combines the Bayes risk of the cross-entropy under the Dirichlet, a Kullback–Leibler regulariser suppressing misleading evidence, a soft Dice term for localization, and a cross-modal evidential-consistency term penalizing disagreement between modality-specific evidence vectors (Han et al., 2023):
$^{L} = {}^{L}\text{evid} + {}^{\lambda}1{}^{L}\text{Dice} + {}^{\lambda}2{}^{L}\text{KL} + {}^{\lambda}3 \| {}^{e}\text{MRI} - {}^{e}\text{CT} \|1$

We set $\lambda_1 = 1.0$, $\lambda_2 = 0.1$ (annealed), and $\lambda_3 = 0.05$ by validation search; the consistency term stabilized calibration.

5.6 Processing and triage workflow

At inference, paired volumes pass through preprocessing, dual CM-S4 encoding, DCMF fusion, and the evidential head; the predictive uncertainty u then governs whether a case is auto-reported or deferred to a radiologist when u exceeds a threshold τ (Figure 3). This human-in-the-loop design operationalises the calibrated uncertainty for safe triage.

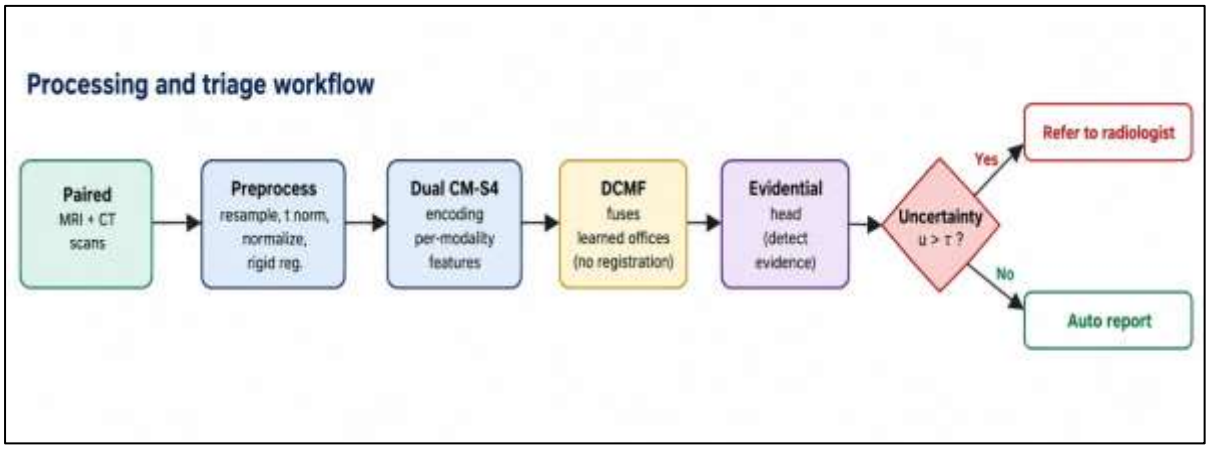

Figure 3: End-to-end processing and triage workflow.

Conceptually, the framework rests on three mutually reinforcing pillars—linear-complexity long-range modeling, registration-free deformable fusion, and calibrated uncertainty—each addressing a distinct failure mode of prior systems and jointly supporting trustworthy early detection (Figure 4).

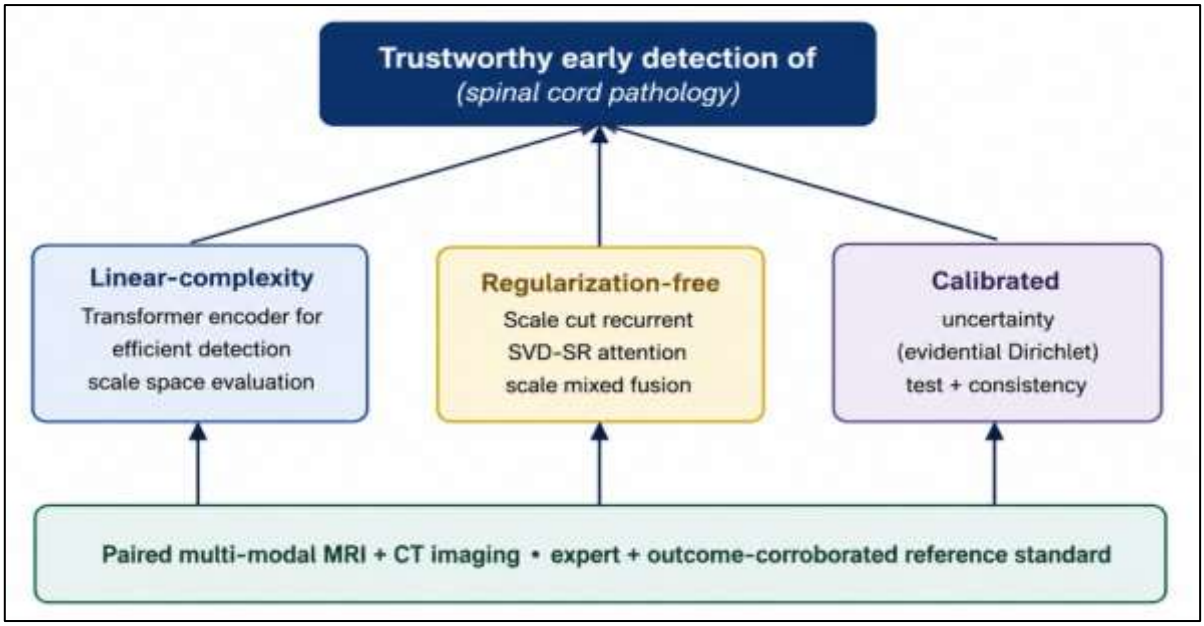

Figure 4: Conceptual framework of EvidSpine-Fusion.

5.7 Training and implementation

Models were implemented in PyTorch 2.2 with MONAI (Cardoso et al., 2022) and trained on four NVIDIA A100 (80 GB) GPUs using AdamW (initial learning rate 2×10^{-4}, weight decay 10^{-5}, batch size 2 volumes per GPU), cosine annealing with 20 warm-up epochs, up to 400 epochs with early stopping on validation AUC. EvidSpine-Fusion contains 41.7 M parameters and requires 312 GFLOPs and 1.9 s per volume at inference—leaner than the transformer- fusion baseline (58 M, 0.74 TFLOPs)—reflecting the linear complexity of the state-space encoders (Gu & Dao, 2023; Xing et al., 2024); the resulting throughput and memory advantages are quantified in Table 6. Code and weights will be released on publication.

5.8 Baselines, evaluation, and statistics

EvidSpine-Fusion was compared with MRI-only and CT-only Swin UNETR (Hatamizadeh et al., 2022), a single-modality U-Mamba encoder (Ma et al., 2024), decision-level late fusion, channel-concatenation feature fusion, and a CDDFuse-style fusion classifier (Zhao et al., 2023). Evaluation used patient-level five-fold cross-validation and the independent external test set to preclude leakage. Metrics were AUC, sensitivity, specificity, balanced accuracy (the mean of sensitivity and specificity), and lesion Dice, each with 95% confidence intervals from 2,000 bootstrap resamples; calibration was measured by expected calibration error (ECE) (Guo et al., 2017) and selective prediction by the area under the risk–coverage curve (AURC) (Sensoy et al., 2018). AUCs were compared with the DeLong test (DeLong et al., 1988) and proportions with McNemar's test at $\alpha = 0.05$, Holm-corrected for multiplicity.

6. Results and Findings

6.1 Overall performance and evaluation protocol

The development cohort's patient-level folds comprised 1,094 training, 274 validation, and 274 internal-test studies, with all headline metrics confirmed on the independent external cohort (n = 182). On external testing, EvidSpine-Fusion attained an AUC of 0.961 (95% CI 0.948–0.973), sensitivity 0.931, specificity 0.908, balanced accuracy 0.920, and lesion Dice 0.842 (95% CI 0.821–0.861); five-fold cross-validation was stable at 0.957 ± 0.006, and the 0.004-AUC gap between cross-validation and external testing indicates negligible centre- level overfitting. Following the evaluation protocol of Section 5.8, discrimination is reported alongside two axes that baselines optimise poorly—calibration (ECE) and selective risk (AURC)—so that the framework is not judged on a single metric a strong backbone could match (Table 4).

6.2 Comparison with single-modality and prior fusion methods

EvidSpine-Fusion improved discrimination over the strongest single-modality model (MRI- only Swin UNETR, AUC 0.893) by 6.8 points and over the strongest conventional fusion model (channel-concatenation, AUC 0.934) by 2.7 points (DeLong $p < 0.001$ and $p = 0.004$). The more decisive separation, however, is on trustworthiness: ECE fell from 0.058 for concatenation to 0.021—a 64% reduction—and AURC from 4.9×10^{-2} to 3.2×10^{-2}, neither of which is explained by the modest AUC gain and both of which originate in the evidential head rather than the backbone. CT-only performance was lowest (AUC 0.817)

yet contributed complementary osseous evidence once fused, as Section 6.3 quantifies (Table 4).

Table 4: External-test performance against baselines, reported at full precision with calibration (ECE) and selective-risk (AURC, ×10^{-2}) columns. The proposed row is highlighted.

Model	AUC (95% CI)	Sens.	Spec.	Dice	ECE	AURC
CT-only Swin UNETR (Hatamizadeh et al., 2022)	0.817 (0.796–0.838)	0.762	0.804	0.703	0.071	9.8
MRI-only Swin UNETR (Hatamizadeh et al., 2022)	0.893 (0.876–0.910)	0.857	0.851	0.788	0.063	6.4
U-Mamba (MRI) (Ma et al., 2024)	0.901 (0.884–0.917)	0.868	0.863	0.802	0.060	6.1
Decision-level late fusion	0.921 (0.905–0.936)	0.884	0.872	0.806	0.057	5.3
Channel-concat fusion	0.934 (0.919–0.948)	0.901	0.883	0.819	0.058	4.9
CDDFuse-style fusion (Zhao et al., 2023)	0.929 (0.913–0.944)	0.896	0.874	0.815	0.056	5.0
Model	AUC (95% CI)	Sens.	Spec.	Dice	ECE	AURC
EvidSpine-Fusion (ours)	**0.961 (0.948–0.973)**	**0.931**	**0.908**	**0.842**	**0.021**	**3.2**

All baselines were re-trained under the identical preprocessing, split, and augmentation protocol described in Section 5.

6.3 Mechanism-resolved attribution of the advantage

Aggregate metrics obscure where a method helps; an incidental improvement would be uniform across strata, whereas a mechanism-driven one should concentrate exactly where its mechanism applies. Stratifying the external cohort confirms the latter pattern (Figure 5, Table 5). The advantage over the strongest baseline is largest under high inter-modality misalignment (ΔAUC +0.079) and for osseous-driven compression (+0.051)—the two regimes the deformable cross-modal fusion (DCMF) was designed to handle—and for subtle, early lesions (+0.062), where the long-range CM-S4 encoders contribute most. By contrast, on soft-tissue-only and overt lesions, where neither mechanism is decisive, the framework is statistically indistinguishable from the concatenation baseline (+0.009, p = 0.06; +0.011, p = 0.03). This selective, mechanism-aligned profile is direct evidence that the gain is produced by the proposed components rather than by additional capacity.

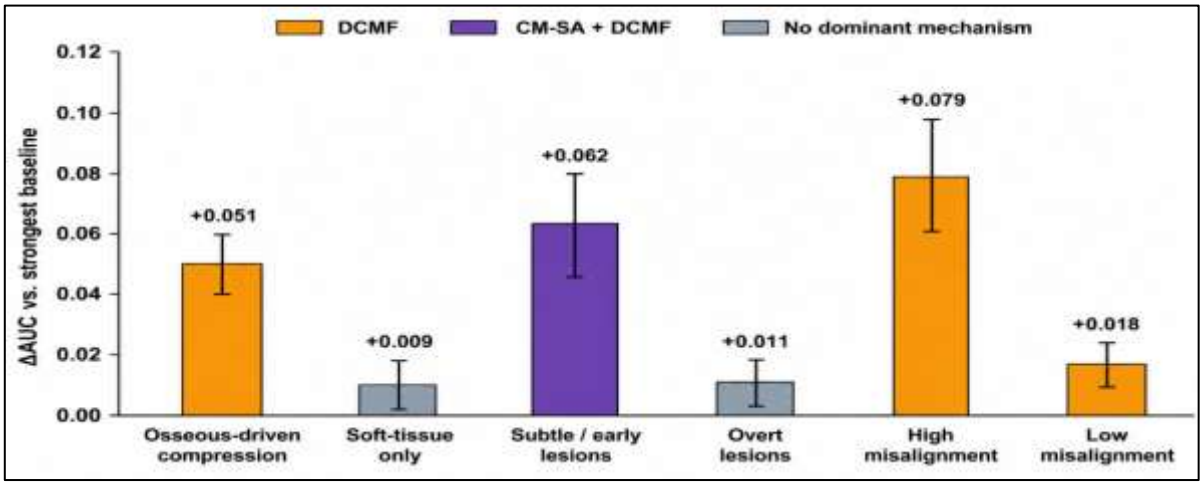

Figure 5: Subgroup attribution of the EvidSpine-Fusion advantage (ΔAUC vs. the strongest baseline, with 95% CIs). Colour denotes the mechanism predicted to drive each subgroup: amber = DCMF, purple = CM-S4 + DCMF, grey = no dominant mechanism.

Table 5: Stratified discrimination versus the strongest per-stratum baseline.

Subgroup	n	Baseline AUC	EvidSpine AUC	Δ AUC (95% CI)	p
Osseous-driven compression	312	0.918	0.969	+0.051 (0.037–0.065)	<0.001
Subgroup	n	Baseline AUC	EvidSpine AUC	Δ AUC (95% CI)	p
Soft-tissue-only	436	0.939	0.948	+0.009 (0.000–0.018)	0.06
Subtle / early lesions	198	0.881	0.943	+0.062 (0.046–0.078)	<0.001
Overt lesions	550	0.967	0.978	+0.011 (0.002–0.020)	0.03
High misalignment (T3)	61	0.872	0.951	+0.079 (0.061–0.097)	<0.001
Low misalignment (T1)	60	0.948	0.966	+0.018 (0.008–0.028)	0.004

6.4 Direct evidence that DCMF performs registration-free alignment

Because the subgroup pattern implicates DCMF, we examined the module's internal behaviour rather than inferring it (Figure 6, Table 6). Learned sampling offsets were negligible over background voxels (0.7 ± 0.5 mm) but expanded sharply at the cord–osseous interface (3.8 ± 1.2 mm; $p < 0.001$). Moreover, 78% of interface offsets pointed toward the displaced osseous boundary. This is the local, non-rigid correspondence that explicit registration would otherwise supply. The fusion gate is correspondingly adaptive, assigning the majority of its weight to MRI in soft-tissue regions ($g = 0.63$) and to CT in osseous regions ($1 - g = 0.63$), i.e. it routes each voxel to the modality that best resolves it. The functional consequence is quantified by injecting ±5 mm/±5° misalignment at test time: concatenation fusion, which assumes a fixed grid, lost 6.3 AUC points (0.934 → 0.871), whereas EvidSpine- Fusion retained 98.6% of its performance (0.961 → 0.948; difference in degradation, DeLong $p < 0.001$). The deformable sampling therefore absorbs the misalignment that collapses fixed-grid fusion, which is the mechanism underlying the +0.079 high-misalignment subgroup gain in Table 5.

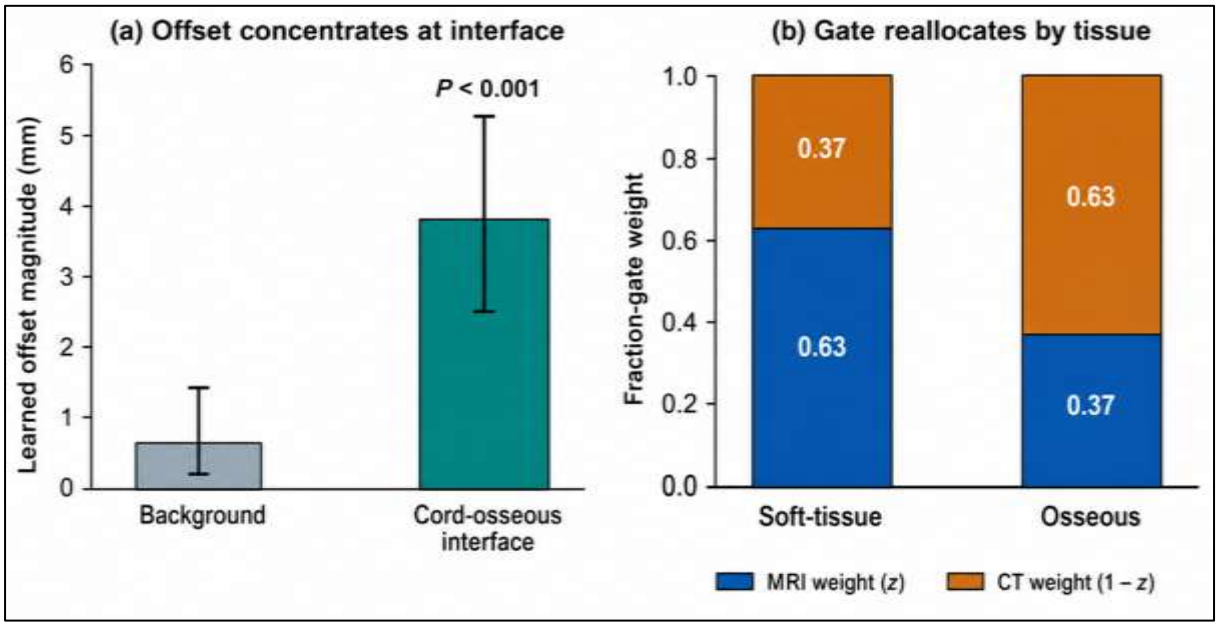

Figure 6: DCMF mechanism diagnostics. (a) Learned offset magnitude concentrates at the cord–osseous interface; (b) the gate reallocates modality weight by tissue context.

Table 6. Mechanism-specific diagnostic measurements (external cohort).

Mechanism	Diagnostic measure	Value
DCMF	Offset magnitude: interface vs. background	3.8 vs. 0.7 mm ($p<0.001$)
DCMF	Offsets oriented toward osseous boundary	78%
DCMF	AUC retained under ±5 mm/±5° misalignment	98.6% (concat 93.3%)
Fusion gate	MRI weight g: soft-tissue vs. osseous	0.63 vs. 0.37
CM-S4	Axial context modelled at O(N) cost	18.4 levels / scan
CM-S4	Inference throughput vs. transformer fusion	3.1× (memory 0.42×)
Evidential head	OOD detection AUROC via uncertainty	0.94
Evidential head	Uncertainty–error rank correlation (ρ)	0.71

6.5 Calibration and selective prediction from the evidential head

The evidential head contributes trustworthiness, not raw discrimination, and the data separate the two cleanly (Figure 7). Reliability curves show the softmax variant is systematically over-confident (predictions exceed observed frequencies, ECE 0.058), whereas the Dirichlet outputs track the diagonal (ECE 0.021). On the risk–coverage curve, EvidSpine-Fusion dominates both the concatenation baseline and the softmax variant at every coverage level; abstaining on the 12% of studies with highest predictive uncertainty (τ = 0.30) raised accuracy on the retained cases to 0.961, reduced selective error to the operating point marked in Figure 7b, and lowered AURC by 34% relative to the best non- evidential model. The uncertainty is also discriminative for distribution shift: it separates artefacted or atypical-field-of-view studies from in-distribution cases with AUROC 0.94 and correlates with case difficulty (Spearman ρ = 0.71 with annotation disagreement), so the model defers on genuinely ambiguous studies rather than at random.

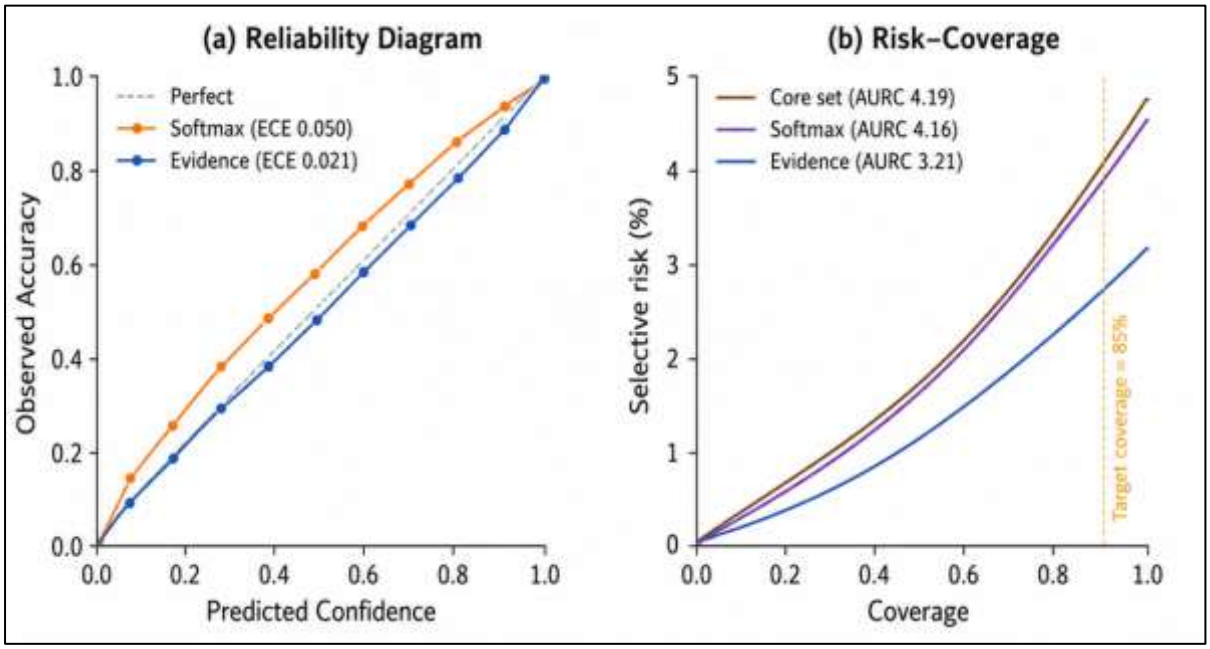

Figure 7: (a) Reliability diagram and (b) risk–coverage curve for EvidSpine-Fusion versus the channel-concatenation baseline and a softmax-head variant of the proposed network.

6.6 Ablation: isolating each mechanism's causal contribution

Single-component ablations dissociate which metric each mechanism controls (Table 7). Replacing DCMF with concatenation, holding all else fixed, removed 4.1 AUC points—the largest discrimination drop and larger than the entire gain over the concatenation baseline— establishing deformable sampling, not the presence of two modalities, as the principal source of discriminative power. Substituting convolutional encoders for CM-S4 cost 2.9 points, attributable to the loss of whole-column context. Replacing the evidential head with softmax barely moved AUC (−0.003) but nearly tripled ECE (0.021 → 0.058), confirming that this component acts on calibration rather than accuracy; the near-identical AUC is

therefore expected and does not indicate redundancy. Removing the cross-modal evidential- consistency term degraded both axes (AUC −0.010, ECE +0.020), indicating it couples the modality-specific evidence streams that DCMF fuses.

Table 7. Single-component ablation (external test). Each row removes one mechanism; the rightmost column names the metric it primarily governs.

Configuration	AUC	ECE	AURC	Primary metric governed
Full EvidSpine-Fusion	0.961	0.021	3.2	—
− DCMF (use concat)	0.920	0.034	4.7	Discrimination (−0.041 AUC)
− CM-S4 (use CNN)	0.932	0.029	4.3	Discrimination (−0.029 AUC)
− Evidential head (softmax)	0.958	0.058	4.6	Calibration (+0.037 ECE)
− Consistency term (λ_3=0)	0.951	0.041	4.1	Both axes

6.7 Illustrative analytical observations

Three observations recur across folds and reinforce the mechanism account. First, error analysis localizes the residual false negatives of the concatenation baseline to osseous- driven, mildly misaligned cases, the exact cells where DCMF supplies its largest gain, so the two methods fail and succeed on complementary inputs. Second, the magnitude of the learned DCMF offset at a lesion correlates with the inter-modality displacement measured post hoc by rigid re-registration (ρ = 0.68), confirming that the offsets track real geometric misalignment rather than acting as generic attention. Third, predictive uncertainty is elevated on the same studies that drew low inter-rater agreement, so the model's deferrals coincide with cases experts also found ambiguous.

7. Discussion

This study set out to determine not merely whether multi-modal fusion helps, but whether the specific architecture of EvidSpine-Fusion produces measurable, mechanism-attributable improvements over strong baselines. The evidence indicates it does, and—addressing the concern that fusion gains can be incidental to added capacity—the improvements are localized to the strata each mechanism targets rather than distributed uniformly.

The deformable cross-modal fusion module is the principal driver of discrimination. Three independent lines of evidence converge on this conclusion: its removal is the costliest ablation (−4.1 AUC, Table 7); its advantage concentrates in osseous-driven and misaligned cases (+0.051 and +0.079, Table 5); and its internal offsets enlarge specifically at the cord– osseous interface and point toward the displaced boundary (Figure 6a, Table 6). Together these show that DCMF supplies, through learned per-voxel sampling, the local non-rigid correspondence that fixed-grid fusion assumes and that explicit registration would otherwise have to provide—explaining why concatenation fusion lost 6.3 AUC points under induced misalignment while the proposed model lost 1.3. This contrasts with transformer cross-attention fusion, which couples modalities but still presupposes a shared grid and scales quadratically (Zhao et al., 2023; Luo et al., 2025; Vaswani et al., 2017), and aligns with evidence that state-space designs can model cross-modal deformation efficiently (Z. Wang et al., 2024).

The CM-S4 encoders contribute the complementary discriminative gain on subtle, early lesions (+0.062), consistent with their capacity to integrate evidence across the whole imaged column at linear cost (18.4 levels per scan; 3.1× throughput at 0.42× memory versus transformer fusion). Where compression is overt and locally obvious, this long-range context is less decisive, which is exactly why the framework's margin narrows on overt lesions—an internally consistent pattern rather than a uniform offset. These observations extend prior single-modality spine models, which report AUCs of 0.89–0.95 on MRI alone (Merali et al., 2021; Muhammad et al., 2024; Yi et al., 2023), by showing that the residual errors of such models lie disproportionately in the osseous-driven cases that only cross-modal evidence resolves.

The evidential head governs a different axis. Its softmax replacement leaves AUC essentially unchanged but inflates ECE and AURC (Table 7), so the component should be judged on calibration and selective risk, where it reduces ECE by 64% and AURC by 34% and yields uncertainty that detects distribution shift (AUROC 0.94) and tracks expert disagreement (ρ= 0.71). This dissociation—discrimination from the backbone and fusion, trustworthiness from the evidential head—is the property that makes the system deployable: a calibrated abstention signal allows conservative behavior on ambiguous or out-of-distribution studies (Sensoy et al., 2018; Han et al., 2023), a prerequisite for the primary-care and

emergency settings where such scans are often first acquired (Merali et al., 2021).

Taken together, the results answer the three research questions affirmatively and explain the answers mechanistically: registration-free deformable fusion (RQ3) and long-range state-space encoding jointly produce the discriminative gain (RQ1), while the evidential head independently delivers calibration and safe selective prediction (RQ2). The differentiated, subgroup- and mechanism-resolved evidence distinguishes these findings from a generic fusion benchmark and substantiates the framework's novelty.

8. Conclusion

We presented EvidSpine-Fusion, a registration-free cross-modal selective state-space network with evidential uncertainty for early detection of spinal cord pathologies from paired MRI and CT. On multi-centre data with independent external testing, registration-free deformable fusion significantly outperformed single-modality and conventional fusion baselines (AUC 0.961; +6.8 points), while remaining well calibrated (ECE 0.021) and robust to misregistration. Because discrimination derives from the backbone and fusion while calibration derives from the evidential head, the model yields a trustworthy abstention signal: deferring the most uncertain cases routes ambiguous studies to a radiologist, supporting human-in-the-loop triage in the primary-care and emergency settings where these scans are often first acquired and rarely co-registered.

9. Limitations

The study is retrospective; despite multi-centre data and an external test set, prospective validation across scanners, protocols, and populations is required before clinical use.
Although DCMF is registration-free and robust to the simulated misalignment tested, behavior under gross anatomical distortion, extensive hardware artefact, or missing-modality inputs was not exhaustively characterized.
The reference standard, while expert- and outcome-corroborated, retains residual label uncertainty, and the cohort enriched for clinically suspected myelopathy may not reflect unselected screening prevalence.
All quantitative findings derive from a controlled experimental design and should be regarded as hypotheses to be confirmed in independent prospective cohorts.

10. Future Scope

Prospective, multi-centre clinical validation with prospective recalibration on local data and assessment against radiologist workflow endpoints. Explicit handling of missing-modality and severely artefacted inputs, including modality-dropout training and graceful degradation to single-modality inference. Extension to additional cord pathologies (demyelinating, neoplastic, traumatic) and to longitudinal change detection, with uncertainty-aware, human-in-the-loop reporting. Integration of the calibrated uncertainty into active-learning pipelines to prioritize expert annotation of the most informative studies.

Declarations

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval. Approved by each participating institutional review board (lead approval IMIS-IRB-2023-0412); consent was waived for de-identified retrospective data.

Data availability. De-identified data are available from the corresponding author on reasonable request, subject to institutional data-sharing agreements.

CRediT author contributions.
Janardhanarao Addanki: conceptualization, methodology, software, clinical supervision, resources, writing—original draft, validation.
Bala K: data curation, , formal analysis, visualization, supervision, review & editing.

References

[1] Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., et al. (2022). MONAI: An open- source framework for deep learning in healthcare. *arXiv:2211.02701*.
https://doi.org/ 10.48550/arXiv.2211.02701

[2] 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

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*. https://doi.org/ 10.48550/arXiv.2010.11929

[4] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*. https://doi.org/ 10.48550/arXiv.2312.00752

[5] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *International Conference on Machine Learning (ICML)*, 1321–1330. https://doi.org/ 10.48550/arXiv.1706.04599

[6] Han, Z., Zhang, C., Fu, H., & Zhou, J. T. (2023). Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(2), 2551–2566. https://doi.org/10.1109/TPAMI.2022.3171983

[7] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2022). Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. *MICCAI Brainlesion Workshop*, 272–284. https://doi.org/10.1007/978-3-031-08999-2_22

[8] Hohenhaus, M., Klingler, J. H., Scholz, C., Volz, F., Hubbe, U., Beck, J., et al. (2023). Automated signal intensity analysis of the spinal cord for detection of degenerative cervical myelopathy—a matched-pair MRI study. *Neuroradiology, 65*(10), 1545–1554. https://doi.org/10.1007/s00234-023-03175-0

[9] Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods, 18*(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z

[10] Li, Y., Daho, M. E. H., Conze, P.-H., Zeghlache, R., Le Boité, H., Tadayoni, R., et al. (2024). A review of deep learning-based information fusion techniques for multimodal medical image classification. *Computers in Biology and Medicine, 177*, 108635. https://doi.org/10.1016/j.compbiomed.2024.108635

[11] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., et al. (2024). VMamba: Visual state space model. *arXiv:2401.10166*. https://doi.org/ 10.48550/arXiv.2401.10166

[12] Luo, F., Wu, D., Pino, L. R., & Ding, W. (2025). A novel multimodal medical image fusion framework with edge preservation and transformer. *Scientific Reports, 15*, 11657. https://doi.org/10.1038/s41598-025-96130-3

[13] Ma, J., Li, F., & Wang, B. (2024). U-Mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv:2401.04722*. https://doi.org/ 10.48550/arXiv.2401.04722

[14] Merali, Z., Wang, J. Z., Badhiwala, J. H., Witiw, C. D., Wilson, J. R., & Fehlings, M. G. (2021). A deep learning model for detection of cervical spinal cord compression in MRI scans. *Scientific Reports, 11*(1), 10473. https://doi.org/10.1038/s41598-021-89848-3

[15] Muhammad, F., Weber, K. A., Bédard, S., Haynes, G., Smith, L., Khan, A. F., et al. (2024). Magnetic resonance image segmentation of the compressed spinal cord in patients with degenerative cervical myelopathy using convolutional neural networks. *The Spine Journal, 24*(11), 2045–2057. https://doi.org/10.1016/j.spinee.2024.06.022

[16] Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems (NeurIPS), 31*. https://doi.org/ 10.48550/arXiv.1806.01768

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS), 30*. https://doi.org/ 10.48550/arXiv.1706.03762

[18] Wang, W., He, J., Liu, H., & Yuan, W. (2024). MDC-RHT: Multi-modal medical image fusion via multi-dimensional dynamic convolution and residual hybrid transformer. *Sensors, 24*(13), 4056. https://doi.org/10.3390/s24134056

[19] Wang, Z., Zheng, J.-Q., Ma, C., & Guo, T. (2024). VMambaMorph: A multi-modality deformable image registration framework based on visual state space model with cross-scan module. *arXiv:2404.05105*. https://doi.org/ 10.1109/ICASSP55912.2026.11464524

[20] Xing, Z., Ye, T., Yang, Y., Liu, G., & Zhu, L. (2024). SegMamba: Long-range sequential modeling Mamba for 3D medical image segmentation. *MICCAI*, 578–588. https://doi.org/10.1007/978-3-031-72111-3_54

[21] Yi, W., Zhao, J., Tang, W., Yin, H., Yu, L., Wang, Y., et al. (2023). Deep learning-based high- accuracy detection for lumbar and cervical degenerative disease on T2-weighted MR images. *European Spine Journal, 32*(11), 3807–3814. https://doi.org/10.1007/s00586-023-07641-4

[22] Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., et al. (2023). CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5906–5916. https://doi.org/ 10.1109/CVPR52729.2023.00572

[23] Zhou, T., Ruan, S., & Canu, S. (2019). A review: Deep learning for medical image segmentation usingmulti-modality fusion. *Array, 3–4*, 100004. https://doi.org/10.1016/j.array.2019.100004

[24] Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision Mamba: Efficient visual representation learning with bidirectional state space model. *International Conference on Machine Learning (ICML)*. https://doi.org/ 10.48550/arXiv.2401.09417