DOI: 10.37943/SJEN9984

Beibit Abdikenov
PhD, Director, Artificial Intelligence Science and Innovation Center
beibit.abdikenov@astanait.edu.kz, orcid.org/0000-0002-0284-0949
Astana IT University, Kazakhstan
Zhaidar Kairat
Master's Student, Junior Researcher, Artificial Intelligence Science and Innovation Center
255238@astanait.edu.kz, orcid.org/0009-0001-5151-6846
Astana IT University, Kazakhstan
Temirlan Karibekov
PhD, Director, MedTech Science and Innovation Center
t.karibekov@astanait.edu.kz, orcid.org/0009-0008-9801-1774
Astana IT University, Kazakhstan
Tomiris Zhaksylyk
PhD Student, Researcher, Artificial Intelligence Science and Innovation Center
zhaksylyk.tomiris@astanait.edu.kz, orcid.org/0009-0002-8749-1967
Astana IT University,Kazakhstan

DEEP LEARNING FOR LYMPH NODE DETECTION IN CT IMAGING: A BENCHMARKING STUDY OF OBJECT DETECTION FRAMEWORKS

Abstract: Accurate detection of lymph nodes on computed tomography (CT) scans is essential for cancer staging and treatment planning, yet manual identification on CT is time-consuming and subject to considerable inter-observer variability, motivating the development of reliable automated methods. Automating this task remains challenging because many nodes are very small, contrast with surrounding tissue is frequently low, and target instances constitute only a minuscule fraction of all voxels, an extreme class-imbalance problem. In this study, we evaluate five state-of-the-art object detection architectures (Faster R-CNN, YOLOv8-m, MULAN, YOLOv11-m, and RT-DETR-L) under standardized training and evaluation protocols on two datasets with different anatomical compositions: a heterogeneous multi-region dataset (Ver1) and a mediastinal-focused dataset (Ver2). Both datasets were processed using a memory-efficient 2.5D pipeline that feeds three consecutive CT slices as input, capturing volumetric context while avoiding the computational cost of full 3D models. On Ver2, Faster R-CNN and MULAN achieved the highest mAP50 of 0.641, while RT-DETR-L led in recall (0.646) and F1-score (0.630). Every model improved its mAP50 on the anatomically focused dataset. Because the same shift appeared across five structurally different detectors, we attribute it to the detection task itself rather than to any particular architectural design. Paired testing on Ver2 showed that MULAN significantly outperformed both YOLO variants in mAP50, with Faster R-CNN following the same tendency. On Ver1, no pair of models differed significantly, so performance converges once anatomical diversity increases. Taken together, the comparison gives clinicians practical grounds for choosing between speed, precision, and sensitivity, and supplies reference baselines for later work on automated lymph node detection in CT.

Keywords: CT imaging, deep learning, lymph node detection, object detection, 2.5D processing

Introduction

Lymph node metastasis remains one of the strongest prognostic indicators in oncology and has a direct impact on staging, treatment planning, and overall survival across almost every solid malignancy. In the widely used TNM staging system, nodal (N) status is given equal weight to the primary tumor (T) and distant metastases (M); this alone illustrates how crucial accurate nodal

Copyright © 2026, Authors. This is an open access article under the Creative Commons CC BY-NC-ND license
Received: 12.06.2026 Accepted: 17.07.2026 Published: 30.06.2026

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

assessment is in daily practice. The presence, number, and location of metastatic nodes can completely change the therapeutic approach and provide critical clues about a patient's likely outcome [1].

In breast cancer, for example, the degree of axillary involvement often decides whether a patient needs adjuvant chemotherapy, and how aggressive that treatment should be. In lung cancer, finding metastatic mediastinal nodes typically rules out surgery and moves patients toward chemoradiation instead [2]. In gastroesophageal cancers, nodal burden and distribution guide the use of neoadjuvant therapy and the extent of resection [3]. Because these decisions have major effects on morbidity, quality of life, and healthcare costs, accurate lymph node assessment remains essential in modern oncology.

Limitations of current clinical practice

Computed tomography (CT) remains the primary imaging modality of choice for lymph node evaluation in routine oncologic practice because of its widespread availability, rapid acquisition, whole-body coverage, and superior spatial resolution. Current radiological assessment relies primarily on size criteria. Nodes larger than 10 mm in short-axis diameter are considered suspicious according to RECIST guidelines [4]. This size-based approach has well-recognized limitations, with reported sensitivities of 57–76% and specificities of 55–82% across cancer types. The problem works in two directions: tiny metastases can appear in lymph nodes of normal size, leading to false negatives, while benign nodes may become enlarged due to inflammation, causing false positives. Additionally, manually identifying and measuring nodes across hundreds of image slices is time-consuming and shows significant inter-observer variability (15–35%), even among experienced radiologists [5].

Technical challenges in automated detection

Automated lymph node detection faces several unique technical challenges. Lymph nodes are small (typically 3–30 mm), placing them near the resolution limit of clinical CT scanners. They show low contrast against surrounding tissues, with Hounsfield unit values similar to muscle, vessels, and other structures. They occupy less than 1% of voxels in whole-body scans, creating severe class imbalance that biases standard algorithms toward the background. There is also substantial inter-patient variability in node size, shape, and anatomical distribution. Finally, lymph nodes often resemble nearby blood vessels and muscle bundles, which frequently leads to false positive detections even for expert readers.

Deep learning for object detection

Over the past decade, deep learning has driven major progress in medical image analysis. On the detection side, single-stage detectors such as YOLO [6] are valued for their speed and have been widely adopted across medical imaging tasks. Modern iterations including YOLOv8 [7] and YOLOv11 [8] further improve small-object detection through anchor-free heads and optimized feature pyramid networks. Two-stage detectors such as Faster R-CNN [9] remain competitive baselines, combining a region proposal network with a detection head for high-precision localization. On the transformer side, RT-DETR [10] introduces a real-time transformer-based detector that decouples multi-scale feature encoding and employs uncertainty-minimal query selection, achieving competitive accuracy with efficient inference. MULAN [11] is a universal lesion detection framework based on an improved Mask R-CNN with multi-task branches for detection, tagging, and segmentation across diverse body parts. When handling volumetric CT data, the 2.5D approach, which processes several consecutive slices at once, provides a reasonable balance between volumetric context and computational feasibility [12]. Despite these advances, automated lymph node detection remains challenging due to extreme class imbalance, limited annotated data, high anatomical variability, and strict clinical requirements.

Contributions and paper organization

Although deep learning has been applied to lymph node detection, few studies have systematically investigated how the anatomical composition of training datasets affects detection

performance across multiple modern frameworks. This work addresses that gap through the following main contributions:

1. **Systematic detection framework comparison:** We evaluate five state-of-the-art detection models (Faster R-CNN, YOLOv8-m, MULAN, YOLOv11-m, and RT-DETR-L) using standardized protocols across two datasets with contrasting anatomical coverage.
2. **Dual-dataset evaluation:** All models were evaluated on a heterogeneous multi-region dataset (Ver1) and a mediastinal-focused subset (Ver2), enabling quantitative analysis of how anatomical homogeneity influences detection performance.
3. **Memory-efficient 2.5D pipeline:** All experiments employ a 2.5D strategy that takes three consecutive CT slices as input, providing inter-slice context while remaining computationally tractable on consumer-grade GPUs.

The rest of the paper is organized as follows. Related Work surveys existing deep learning approaches to lymph node detection and relevant object detection frameworks. Methods and Materials details the datasets, preprocessing pipeline, model architectures, and training protocols. Results present quantitative detection performance across all models and datasets. Discussion interprets the findings, addresses limitations, and considers implications for clinical translation. Conclusion summarizes the key contributions and outlines directions for future research.

Related work

Automated lymph node detection in CT imaging has a long history, progressing from classical handcrafted feature methods toward fully learned deep learning pipelines. Early approaches relied on random forests and boosted classifiers operating on intensity, shape, and texture descriptors extracted from candidate regions, but these methods struggled with the high variability of lymph node appearance and the extreme foreground-background imbalance inherent to CT volumes.

The introduction of convolutional neural networks transformed the field. Roth et al. [12] demonstrated that training deep CNNs on randomly oriented 2.5D patches, stacking multiple adjacent axial slices as input channels, substantially improved localization of small mediastinal and abdominal nodes compared to purely 2D approaches. This 2.5D strategy became widely adopted because it captures inter-slice context without the prohibitive GPU memory demands of full 3D convolutions. Subsequent work by Wang et al. [13] extended contextual embedding learning to enrich 2D networks with explicit volumetric information from adjacent slices, further improving segmentation performance on volumetric CT data. Kumar et al. [14] later formalized this with a flexible 2.5D framework incorporating both in-slice and cross-slice attention, achieving strong performance across multiple volumetric segmentation and detection benchmarks. Deep learning has also been applied beyond detection to predict sentinel lymph node metastasis status directly from staging CT images, with Yang et al. [15] demonstrating that CT-based signatures can achieve clinically relevant preoperative prediction accuracy in breast cancer.

On the detection architecture side, the YOLO family [6] has emerged as a dominant paradigm for real-time medical image detection due to its speed and competitive accuracy. Wang et al. [16] applied YOLOv7 to cervical lymphoma CT scans and reported 96.4% mAP, demonstrating that YOLO-based detectors generalize effectively to lymphoid tissue localization tasks. The newer YOLOv8 [7] and YOLOv11 [8] releases further improve small-object detection through anchor-free detection heads and enhanced feature pyramid networks, making them particularly suited to lymph node detection where targets are small, numerous, and variably distributed across anatomical regions.

Transformer-based detection architectures have also gained traction. DETR [17] reformulated object detection as a direct set-prediction problem using a transformer encoder-decoder, eliminating hand-crafted pipeline components such as anchor generation and non-maximum suppression entirely. Although DETR typically requires longer training convergence and larger annotated datasets than CNN-based detectors, it offers strong localization accuracy at stricter IoU thresholds and end-to-end training simplicity. Subsequent work has tailored

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

transformer detection specifically to lymph nodes: Yu et al. [18] addressed positional bias in transformer-based detectors by introducing location-debiased query selection and contrastive query representation, achieving effective lymph node detection across heterogeneous CT datasets. These advances demonstrate that transformers can learn discriminative spatial representations for small and morphologically variable targets such as lymph nodes when attention mechanisms are appropriately adapted.

The impact of training data composition on detection performance has received comparatively little systematic attention. Most prior studies either pool heterogeneous multi-region datasets or focus on a single anatomical station, making it difficult to isolate the effect of anatomical homogeneity from architectural differences. The present work directly addresses this gap by evaluating Faster R-CNN, YOLOv8-m, MULAN, YOLOv11-m, and RT-DETR-L under identical conditions on both a heterogeneous multi-region dataset and a mediastinal-focused subset, enabling a controlled quantitative assessment of how anatomical consistency affects detection learning across fundamentally different architectural paradigms.

Methods and Materials

This section describes the experimental design, dataset characteristics, preprocessing pipeline, and detection model architectures evaluated in this study.

Dataset

This study employed two datasets with contrasting anatomical compositions. Ver1 [19] contains 21,906 annotated CT slices from 689 patients across multiple institutions, covering mediastinal, abdominal, and pelvic lymph nodes. Ver2 [20] contains 15,056 slices from 603 patients, restricted to mediastinal lymph nodes only. Both datasets used patient-level splitting of approximately 70%, 15%, and 15% for training, validation, and test sets respectively, to prevent data leakage. The split ratios apply to patient counts; due to variation in the number of slices per patient, the resulting slice proportions differ marginally from 70/15/15. Ver1 splits comprise 482 training patients (14,876 slices), 103 validation (3,196 slices), and 104 test (3,834 slices). Ver2 splits comprise 422 training patients (10,699 slices), 90 validation (2,052 slices), and 91 test (2,305 slices). All CT scans were acquired using standard clinical protocols with slice thickness 1.0–5.0 mm and image dimensions 512×512 to 768×768 pixels. Ground truth annotations were produced by experienced radiologists. Bounding box annotations for detection models were derived directly from the pixel-level segmentation masks by computing the tightest enclosing rectangle around each annotated lymph node region. Figure 1 provides an overview of the complete experimental pipeline, from dataset preparation and 2.5D preprocessing through to the five detection architectures evaluated in this study.

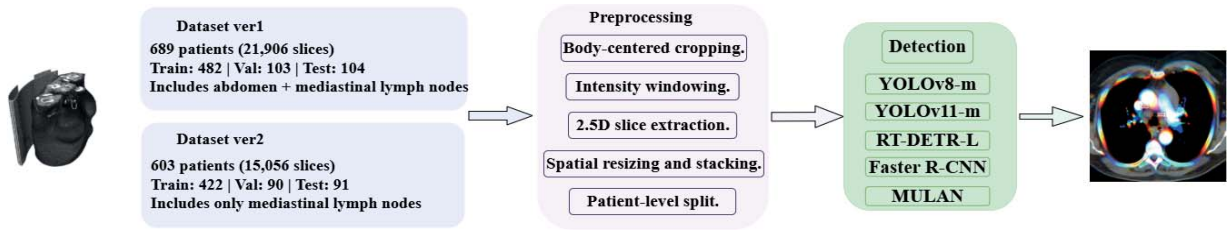

Figure 1. Overview of the experimental pipeline for lymph node detection, illustrating dataset preparation, 2.5D preprocessing, and the five detection model architectures

Preprocessing

All CT images underwent a standardized preprocessing pipeline to ensure consistency across the heterogeneous datasets. The pipeline comprises four sequential stages.

The first stage is body-centered region of interest extraction. Raw CT volumes contain substantial non-anatomical background regions that provide no diagnostic information. An automated body center detection algorithm operates on CT Hounsfield Unit (HU) values. For each

DOI: 10.37943/SJEN9984
© Beibit Abdikenov, Zhaidar Kairat,
Temirlan Karibekov, Tomiris Zhaksylyk.

axial slice, a binary body mask is generated by applying an intensity threshold at −200 HU, which effectively separates tissue from air. The geometric center of the body mask is computed as the mean of all body pixel coordinates, and a 384×384 pixel region is extracted around this center to focus computational resources on anatomically relevant areas while maintaining sufficient context.

The second stage is intensity windowing and normalization. To enhance lymph node visibility and improve contrast, a single fixed window was applied to all slices within each dataset, following clinical radiology standards for soft tissue visualization. For the mediastinal dataset (Ver2), a window level (WL) of 40 HU and window width (WW) of 400 HU were applied uniformly to all slices. For the multi-region dataset (Ver1), which covers mediastinal, abdominal, and pelvic lymph nodes, a window level of 50 HU and window width of 350 HU were applied uniformly to all slices. This single-window approach for Ver1 is appropriate because abdominal and pelvic lymph nodes share similar soft tissue attenuation characteristics, and a unified window avoids the need for per-slice anatomical region classification, which would introduce additional complexity and potential errors. Intensities outside the window range were clipped and linearly rescaled to the range [0, 255].

The third stage is memory-efficient 2.5D representation. For each target slice at index i, three consecutive axial slices are extracted at positions (i−1), (i), and (i+1), forming the input tensor:

$$X = [I_{i-1}, I_i, I_{i+1}] \qquad (1)$$

For boundary slices where (i−1) or (i+1) are unavailable, the target slice is replicated to maintain consistent input dimensions. This approach maintains inter-slice context while remaining computationally tractable on consumer-grade GPUs. Images were stored using PNG-based lossless compression, providing faster data loading during training, reduced storage requirements, and compatibility with standard DICOM-to-PNG export workflows used in clinical imaging systems.

The fourth stage is spatial resizing and channel stacking. Image patches of 384×384 pixels were resized to 256×256 pixels using bilinear interpolation to balance spatial resolution with computational efficiency. The three consecutive slices were stacked along the channel dimension to form the final 3-channel 256×256 input fed to all detection models. For detection models, bounding box coordinates were scaled and transformed consistently with the spatial resizing applied to the images. Figure 2 summarizes the four-stage preprocessing pipeline described above.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

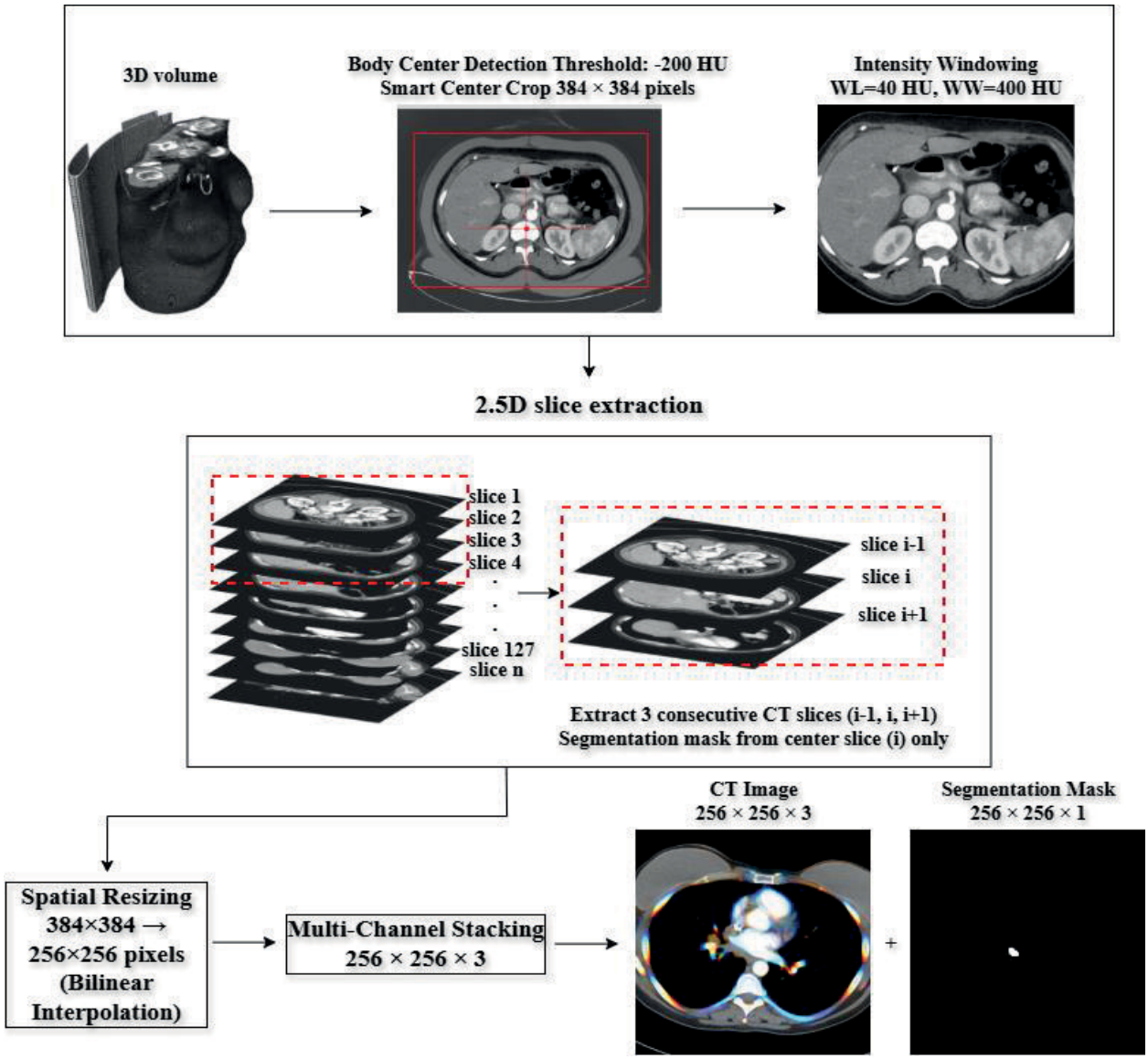

Figure 2. Preprocessing pipeline showing body-centered ROI extraction, intensity windowing, multi-slice 2.5D extraction, and spatial resizing to 256×256 pixels.

Model architecture

Five state-of-the-art object detection architectures were evaluated, spanning two-stage CNN-based detectors, single-stage anchor-free detectors, real-time transformer-based detectors, and a multi-task lesion detection framework.

YOLO Detectors. YOLOv8 [7] and YOLOv11 [8] represent state-of-the-art single-stage detection architectures optimized for real-time performance. Both employ anchor-free detection heads that directly predict bounding box coordinates and class probabilities without requiring predefined anchor boxes. The architectures feature CSPNet (Cross Stage Partial Network) backbones for efficient multi-scale feature extraction and PANet (Path Aggregation Network) necks for feature pyramid fusion, enabling robust detection across varying node sizes. The medium variants (YOLOv8-m and YOLOv11-m) were used in this study, with YOLOv8-m having approximately 25.9 million parameters (12.7 GFLOPs) and YOLOv11-m approximately 20.1 million parameters (11.0 GFLOPs) at the 256×256 input resolution used in this study, selected for their balance of capacity and computational efficiency. Both models were trained using Complete IoU (CIoU) loss for bounding box regression combined with binary cross-entropy (BCE) for classification:

DOI: 10.37943/SJEN9984
© Beibit Abdikenov, Zhaidar Kairat,
Temirlan Karibekov, Tomiris Zhaksylyk.

$$L_{YOLO} = L_{CIoU} + L_{BCE} \quad (2)$$

RT-DETR-L. RT-DETR-L [10] is a real-time end-to-end object detector that builds on the transformer detection paradigm with an efficient hybrid encoder decoupling intra-scale interaction and cross-scale fusion. Unlike the original DETR [17], RT-DETR eliminates non-maximum suppression through direct set prediction while achieving real-time inference speeds. A ResNet-50 backbone extracts multi-scale feature maps that are processed by the hybrid encoder and passed to a transformer decoder with uncertainty-minimal query selection. The total training loss combines classification cross-entropy, L1 bounding box regression, and generalized IoU loss:

$$L_{DETR} = \lambda_{cls} L_{CE} + \lambda_{L1} L_{L1} + \lambda_{giou} L_{GIoU} \quad (3)$$

We employed RT-DETR-L [10] with a ResNet-50 backbone. In Equation (3), λ_cls, λ_L1, and λ_giou are the loss weighting coefficients; following the RT-DETR defaults [10], these were set to λ_cls = 1, λ_L1 = 5, and λ_giou = 2. Faster R-CNN. Faster R-CNN [9] is a two-stage detector that first generates region proposals via a Region Proposal Network (RPN) sharing convolutional features with the detection head, then classifies and refines these proposals. It uses a ResNet backbone with a Feature Pyramid Network (FPN) for multi-scale feature extraction. We employ the standard Faster R-CNN with a ResNet-50-FPN backbone, totaling approximately 41.8 million parameters and 45.3 GFLOPs. MULAN. MULAN [11] is a universal lesion detection framework built on an improved Mask R-CNN backbone with three task-specific head branches for detection, tagging, and segmentation. It employs a 3D feature fusion strategy to incorporate inter-slice context and achieves strong performance on the large-scale DeepLesion dataset covering lesions across multiple body parts. The model used in this study has approximately 43.7 million parameters and 52.0 GFLOPs.

Experimental Design

To investigate the effect of anatomical composition on lymph node detection performance, we adopt a dual-dataset experimental design. Two dataset configurations with contrasting anatomical coverage are used: Ver1, a heterogeneous multi-region dataset covering mediastinal, abdominal, and pelvic lymph nodes, and Ver2, a mediastinal-focused subset. Each of the five detection models (Faster R-CNN, YOLOv8-m, MULAN, YOLOv11-m, and RT-DETR-L) is trained and evaluated independently on both dataset configurations, yielding ten trained models in total. This setup allows direct quantitative comparison of how anatomical homogeneity influences detection learning across architecturally distinct frameworks. All experiments were repeated with three random seeds (42, 123, 456) to ensure robustness and statistical stability.

Training Configuration

All experiments were conducted on a workstation equipped with two NVIDIA GeForce RTX 2080 Ti GPUs, each with 11 GB of VRAM, running Ubuntu 22.04.5 LTS with CUDA 12.2. Detection models were implemented using PyTorch, with YOLOv8-m sourced from the Ultralytics library [7] and YOLOv11-m from the Ultralytics YOLO11 release [8], RT-DETR-L from the Ultralytics RT-DETR implementation [10], Faster R-CNN from the torchvision detection library [9], and MULAN following the original codebase [11].

All detection models (Faster R-CNN, YOLOv8-m, MULAN, YOLOv11-m, and RT-DETR-L) were initialized with COCO-pretrained weights. A uniform input resolution of 256×256 was used across all experiments, with 3-channel 2.5D input representation. Training was conducted with a batch size of 8 for 100 epochs, and early stopping with a patience of 15 epochs was applied to prevent overfitting. Optimization was performed using AdamW with an initial learning rate of 2×10^{-4} and a weight decay of 1×10^{-4} for all models. For YOLO-based and RT-DETR models, a cosine learning rate schedule was applied with a final learning rate ratio of 0.05 and momentum of 0.937. Faster R-CNN and MULAN used cosine annealing scheduling. Data

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

augmentation included horizontal flipping with probability 0.5, random rotation within ±15°, and scaling within ±10% for YOLOv8-m, YOLOv11-m, and RT-DETR-L. Gradient clipping was applied only in MULAN with a maximum norm of 5.0. Mixed precision training (AMP) was enabled for YOLOv8-m, YOLOv11-m, and RT-DETR-L.

Regularization and Augmentation

Data augmentation was applied online during training to enhance model robustness. For YOLOv8-m, YOLOv11-m, and RT-DETR-L, geometric transformations included horizontal flipping with probability 0.5, random rotation within ±15°, and scaling within ±10%. Gradient clipping with a maximum norm of 5.0 was applied exclusively to MULAN. All bounding box annotations were transformed consistently with their corresponding image augmentations to maintain annotation integrity. These settings were kept consistent across corresponding model groups to ensure that performance differences arose solely from dataset composition and architectural characteristics rather than training variability.

Evaluation Metrics

Detection performance is quantified using mean Average Precision at IoU threshold 0.5 (mAP50) as the primary metric, computed as the area under the precision-recall curve with detections considered correct when IoU $\geq$ 0.5. mAP50-95 is additionally reported, averaging performance over IoU thresholds from 0.5 to 0.95 in steps of 0.05 for stricter localization evaluation. Precision, Recall, and F1 score are reported at IoU threshold 0.5, with F1 computed as the harmonic mean of the tabulated Precision and Recall.

Results

Table 1 presents detection performance across all five evaluated architectures on both dataset configurations. All models demonstrated consistent performance gains when trained and evaluated on the mediastinal-focused dataset Ver2 compared to the multi-region dataset Ver1. This pattern mirrors the anatomical homogeneity effect observed consistently across all experiments.

On the mediastinal dataset Ver2, Faster R-CNN and MULAN achieved the highest mAP50 of 0.641, reflecting the advantage of two-stage and anatomy-aware detection pipelines on this focused task. Faster R-CNN showed the highest precision of 0.719, while MULAN demonstrated similar precision at 0.714. However, their recall values (0.450 and 0.429 respectively) and F1 scores (0.554 and 0.536) are lower than those of the YOLO and RT-DETR-L models, which range from 0.619 to 0.630 in F1, reflecting the trade-off between precision-focused two-stage detection and recall-oriented single-stage approaches. RT-DETR-L achieved the highest recall of 0.646 and the best F1 score of 0.630 on Ver2, indicating the most balanced trade-off between sensitivity and specificity among all models. Among the YOLO-based models, YOLOv11-m slightly outperformed YOLOv8-m on Ver2 (mAP50: 0.621 vs. 0.616), with YOLOv11-m achieving the highest precision among YOLO variants (0.656 on Ver2). Statistical analysis using paired t-tests (n=3 seeds) revealed that on Ver2, MULAN significantly outperformed YOLOv11-m (Δ=+0.020, p=0.007) and YOLOv8-m (Δ=+0.025, p=0.019). Faster R-CNN showed a trend toward outperforming YOLOv8-m (Δ=+0.025, p=0.065). Note that all deltas are computed from per-seed means and may differ marginally from the rounded values shown in Table 1. On Ver1, no significant differences were observed between any pair of models (all p>0.58), indicating performance convergence under greater anatomical diversity.

On the multi-region dataset Ver1, performance differences between models were smaller, with mAP50 values ranging from 0.517 (YOLOv8-m) to 0.534 (MULAN). Faster R-CNN and MULAN demonstrated markedly higher precision on Ver1 (0.750 and 0.721 respectively) but at the cost of substantially lower recall (0.299 and 0.295), resulting in lower F1 scores. RT-DETR-L showed the best recall-precision balance on Ver1 (recall: 0.570, F1: 0.549), while YOLO models occupied an intermediate position. These patterns suggest that two-stage and anatomy-aware detectors are conservative on heterogeneous data, generating fewer but more precise detections.

DOI: 10.37943/SJEN9984
© Beibit Abdikenov, Zhaidar Kairat,
Temirlan Karibekov, Tomiris Zhaksylyk.

Table 1. Object detection performance on test sets (IoU threshold 0.5). Best per dataset in bold.

Model	Dataset	mAP50	mAP50-95	Precision	Recall	F1
YOLOv8-m	Ver1	0.517	0.245	0.561	0.529	0.545
YOLOv8-m	Ver2	0.616	**0.307**	0.654	0.587	0.619
YOLOv11-m	Ver1	0.523	0.251	0.571	0.514	0.540
YOLOv11-m	Ver2	0.621	0.305	**0.656**	0.591	0.621
RT-DETR-L	Ver1	0.533	0.244	0.531	0.570	0.549
RT-DETR-L	Ver2	0.630	0.298	0.615	0.646	0.630
Faster R-CNN	Ver1	0.530	0.238	0.750	0.299	0.428
Faster R-CNN	Ver2	0.641	0.236	0.719	0.450	0.554
MULAN	Ver1	0.534	0.231	0.721	0.295	0.419
MULAN	Ver2	0.641	0.239	0.714	0.429	0.536

Across all models, the mAP50-95 values were considerably lower than mAP50, ranging from 0.231 (MULAN Ver1) to 0.307 (YOLOv8-m Ver2), reflecting the inherent difficulty of precise bounding box localization in lymph node detection. Notably, Faster R-CNN is the only model whose mAP50-95 does not increase on Ver2, showing a marginal change from Ver1 (0.238) to Ver2 (0.236). Given the small magnitude of this difference relative to seed-level variation, it should be interpreted cautiously, but it may reflect the two-stage detector producing higher-confidence yet slightly less precisely localized boxes on the mediastinal-focused data, where the region proposal network can be less well-calibrated at strict IoU thresholds despite achieving higher mAP50. Representative qualitative detection results on the Ver2 test set are shown in Figure 3.

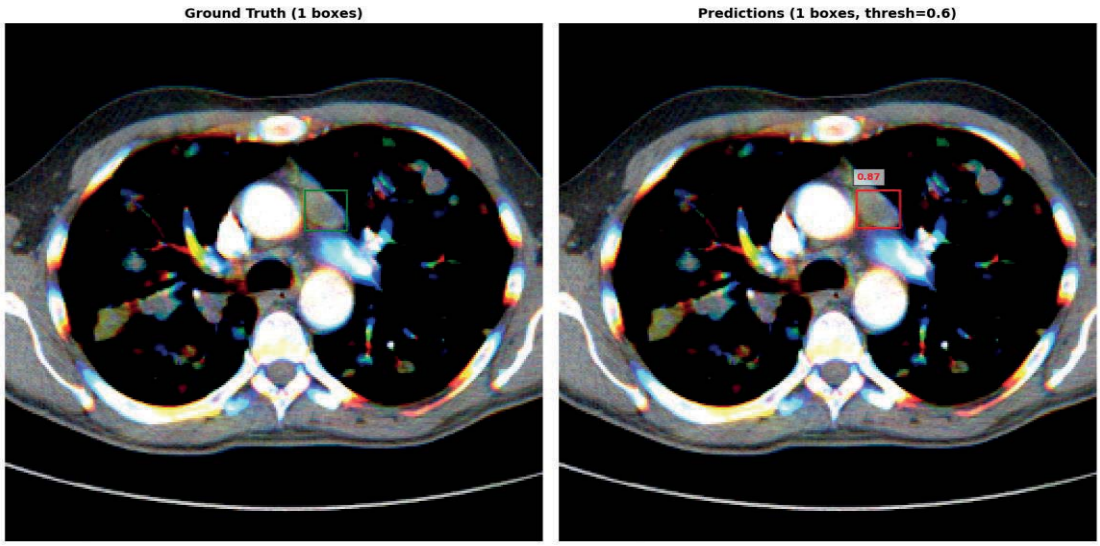

Figure 3. Representative detection outputs on Ver2 test cases. Green boxes indicate true positive detections, red boxes indicate false positives, and yellow boxes show ground truth annotations (best viewed in color).

A quantitative comparison of the four core detection metrics across all architectures and both datasets is presented in Figure 4.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

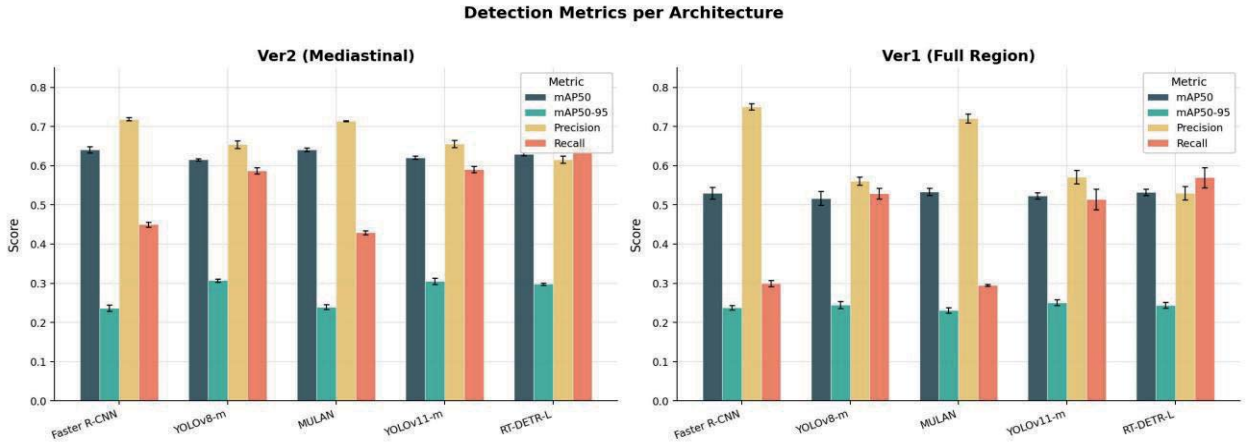

Figure 4. Detection metrics (mAP50, mAP50-95, Precision, Recall) per architecture on Ver2 (Mediastinal) and Ver1 (Full Region). Error bars show standard deviation across three random seeds.

Figure 5 presents the parameter and computational efficiency analysis. YOLOv11-m achieves competitive accuracy with only 20.1M parameters and 11.0 GFLOPs, offering the best accuracy-efficiency trade-off, while MULAN requires 43.7M parameters and 52.0 GFLOPs for comparable mAP50. All GFLOPs are measured at the 256×256 input resolution used in this study; values at standard 640×640 resolution would be higher.

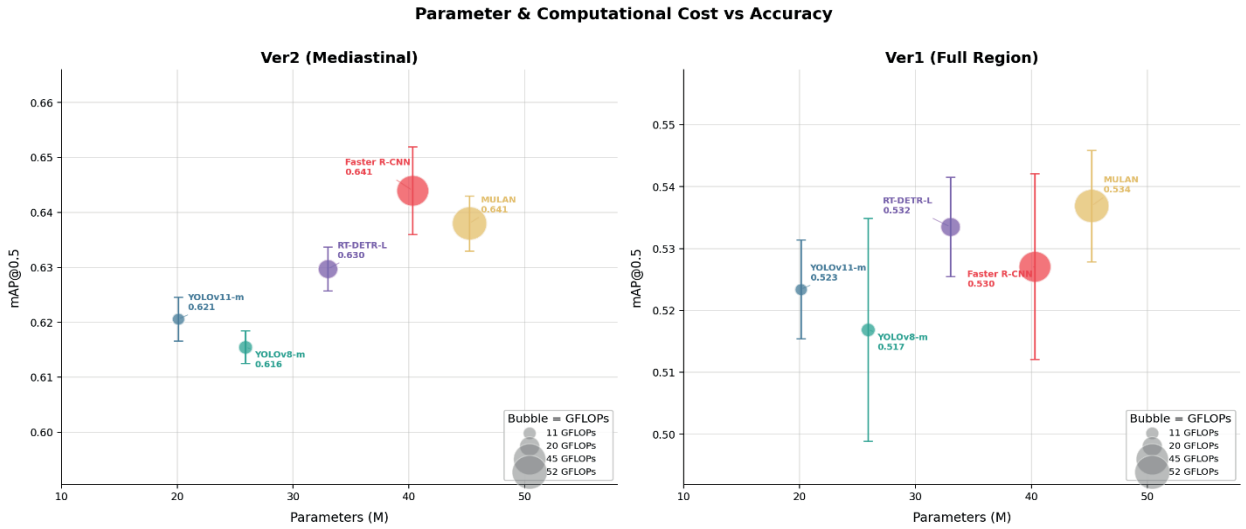

Figure 5. Parameter and computational cost (GFLOPs, bubble size) vs. mAP@0.5 for all five models on Ver2 (Mediastinal, left) and Ver1 (Full Region, right). Error bars indicate standard deviation across seeds.

Figure 6 provides an integrated visual comparison of mAP50, precision, recall, and F1 across all models on both datasets.

DOI: 10.37943/SJEN9984
© Beibit Abdikenov, Zhaidar Kairat,
Temirlan Karibekov, Tomiris Zhaksylyk.

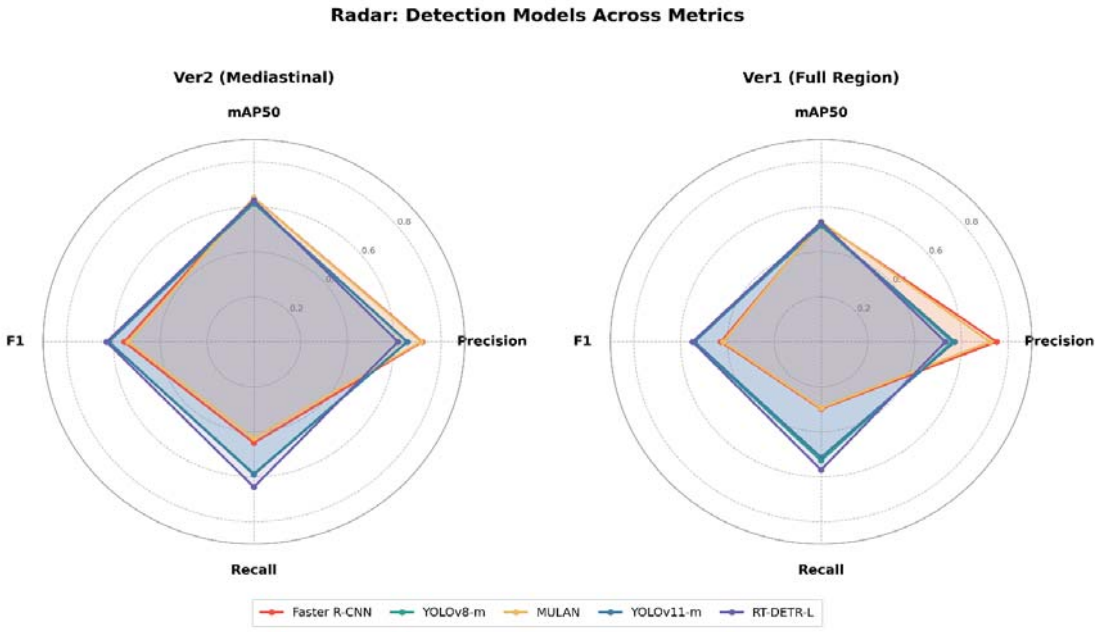

Figure 6. Radar charts comparing all five detection models across mAP50, Precision, Recall, and F1 on Ver2 (Mediastinal, left) and Ver1 (Full Region, right).

The consistent cross-dataset improvements across all five architecturally distinct detection frameworks, spanning two-stage detectors, anchor-free YOLO variants, anatomy-aware transformers, and real-time transformer detectors, demonstrate that the benefits of anatomical homogeneity are not specific to any particular detection paradigm but reflect a fundamental property of the learning problem. Training on anatomically focused data reduces intra-class variability in node appearance and surrounding tissue context, allowing all models to learn more discriminative and consistent detection boundaries. This finding has direct practical implications for clinical dataset curation, suggesting that anatomically stratified training datasets may yield better detection performance than maximally diverse multi-region collections.

Inference Efficiency

Table 2 reports inference latency and throughput for all five models measured at 256×256 input resolution (batch size 1) on an NVIDIA H200 GPU. YOLOv8-m achieves the highest throughput at 238.7 FPS (4.19 ms per image), followed by YOLOv11-m at 163.7 FPS (6.11 ms). RT-DETR-L is the slowest at 13.3 FPS (75.09 ms), reflecting the cost of its transformer decoder. Faster R-CNN and MULAN share a throughput of 95.4 FPS (10.48 ms), consistent with their similar two-stage architectures. For real-time clinical screening applications, the YOLO variants offer a clear latency advantage, while Faster R-CNN and MULAN represent a reasonable accuracy-speed trade-off for batch-processing workflows.

Table 2. Inference efficiency at 256×256 input resolution, batch size 1, NVIDIA H200 GPU. GFLOPs for Faster R-CNN and MULAN are input-dependent (region proposals vary per image) and are therefore approximate averages.

Model	Params (M)	GFLOPs	ms / image	FPS
Faster R-CNN	41.8	45.3	10.48	95.4
YOLOv8-m	25.9	12.7	4.19	238.7
MULAN	43.7	52.0	10.48	95.4
YOLOv11-m	20.1	11.0	6.11	163.7
RT-DETR-L	33.0	20.2	75.09	13.3

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

Discussion

This study systematically evaluated five state-of-the-art object detection architectures (Faster R-CNN, YOLOv8-m, MULAN, YOLOv11-m, and RT-DETR-L) for automated lymph node detection in CT imaging across two datasets with contrasting anatomical compositions. The results reveal two consistent and clinically meaningful findings: first, that anatomical homogeneity in training data substantially improves detection performance across all evaluated frameworks; and second, that the evaluated architectures exhibit distinct precision-recall trade-offs with direct implications for clinical workflow design.

Effect of Anatomical Homogeneity

The most consistent finding across all experiments is that models trained on the mediastinal-focused dataset Ver2 outperformed their counterparts trained on the heterogeneous multi-region dataset Ver1 in mAP50. This improvement was observed uniformly across all five architectures despite their fundamental differences in detection philosophy (two-stage proposal-based detection in Faster R-CNN, anchor-free single-stage prediction in the YOLO variants, anatomy-aware transformer detection in MULAN, and real-time transformer detection in RT-DETR-L), suggesting that the benefit is not tied to any specific design choice but reflects a fundamental property of the detection learning problem itself.

The underlying mechanism is likely the reduction in intra-class appearance variability. When a model is trained on lymph nodes spanning mediastinal, abdominal, and pelvic regions simultaneously, it must learn detection boundaries across nodes embedded in highly variable tissue contexts, attenuation profiles, and surrounding anatomical structures. Restricting training to the mediastinal region reduces this variability substantially, allowing the model to learn more focused and discriminative feature representations. This interpretation is consistent with the statistical finding that differences between models become significant on the focused Ver2 dataset, while performance converges on the more challenging Ver1 dataset (all $p>0.58$), suggesting that anatomical homogeneity amplifies architectural advantages that are otherwise masked by distribution complexity.

These findings have important practical implications for clinical dataset curation. Rather than always pursuing maximal anatomical diversity in training collections, dataset curators and clinical AI developers should consider anatomically stratified training strategies, particularly when deploying models for specific anatomical stations such as mediastinal staging in lung cancer or axillary assessment in breast cancer. The consistent performance advantage of Ver2 across all architectures provides strong empirical support for this recommendation.

YOLO Variants and RT-DETR: Efficiency-Accuracy Trade-offs

Among the YOLO-based models, YOLOv11-m achieved slightly higher mAP50 on Ver2 (0.621) compared to YOLOv8-m (0.616), and demonstrated higher precision (0.656 vs. 0.654). YOLOv11-m also requires fewer parameters (20.1M vs. 25.9M) and fewer GFLOPs (11.0 vs. 12.7), making it the more efficient choice for deployment in resource-constrained clinical environments or real-time screening applications. RT-DETR-L offered the best F1 score and recall on Ver2 (recall: 0.646, F1: 0.630), beneficial for high-sensitivity detection scenarios where missing nodes carries greater clinical consequences than generating reviewable false positives.

Faster R-CNN demonstrated the highest precision on both datasets (0.750 on Ver1, 0.719 on Ver2), reflecting its two-stage proposal-refinement approach that favors conservative, high-confidence detections. However, its lower recall values (0.299 on Ver1, 0.450 on Ver2) suggest it misses a substantial fraction of nodes, which may be a limiting factor in screening applications. MULAN exhibited a similar precision-recall profile to Faster R-CNN, consistent with its anatomy-aware design prioritizing specificity over sensitivity. For multi-region protocols where minimizing false positives is critical, Faster R-CNN or MULAN may be preferred, while RT-DETR-L offers the best balance for mediastinal screening where recall matters most.

Faster R-CNN and MULAN: Precision-Focused Two-Stage and Anatomy-Aware Detectors

Faster R-CNN and MULAN both achieved the highest mAP50 on Ver2 (0.641 each), tied for top performance on the mediastinal dataset. Their high precision (0.719 and 0.714 respectively) indicates conservative detection strategies that produce fewer but more reliable detections. This behavior aligns with the design philosophy of two-stage detectors like Faster R-CNN, where region proposals are refined before classification, and with MULAN's anatomy-aware representations that encode spatial priors about lesion appearance. Both models incur higher computational costs: Faster R-CNN requires 45.3 GFLOPs and MULAN 52.0 GFLOPs, compared to 11.0 GFLOPs for YOLOv11-m, which may be a practical consideration for real-time deployment.

Notably, the relatively lower recall of Faster R-CNN (0.450) and MULAN (0.429) on Ver2 compared to RT-DETR-L (0.646) suggests that in clinical screening applications where missing a metastatic node carries significant consequences, RT-DETR-L or YOLO variants may be preferred. A complementary deployment strategy could combine high-recall RT-DETR-L or YOLO models for initial screening with Faster R-CNN or MULAN for false-positive reduction, leveraging the complementary strengths of each architecture paradigm.

Together, these architectural comparisons highlight that no single model dominates across all clinical criteria; the optimal choice depends on whether precision, recall, efficiency, or balanced F1 performance is prioritized in a given deployment context.

Limitations

Several limitations should be acknowledged. The datasets, while clinically realistic, originate from a limited number of institutions, and model performance on fully external multi-institutional data remains to be validated. Ground truth annotations carry inherent inter-observer variability of 15–35%, introducing label noise that likely suppresses absolute performance across all models. Detection was evaluated at the slice level using 2.5D inputs rather than in full 3D, which means that volumetric consistency of detections across adjacent slices has not been explicitly enforced or evaluated. Furthermore, all models were evaluated at a fixed confidence threshold, and threshold optimization for specific clinical operating points such as maximum sensitivity or maximum F1 was not explored. Finally, inference benchmarks were conducted on a single GPU (NVIDIA H200) and may not generalize across hardware configurations.

Clinical Translation

For clinical deployment, the present results suggest a task-specific model selection strategy. RT-DETR-L is recommended for high-throughput screening applications where sensitivity (recall) is the primary concern, such as initial nodal staging in oncology workflows, given its leading recall and F1 on Ver2. YOLOv11-m offers the best efficiency-accuracy trade-off and is suitable for resource-constrained deployment or settings requiring real-time inference. Faster R-CNN and MULAN are more appropriate when precision is paramount, such as reducing radiologist review burden in multi-region protocols. In all cases, real-world deployment will require prospective clinical validation, integration with picture archiving and communication systems (PACS), uncertainty quantification mechanisms, and formal regulatory approval before clinical adoption.

Future Directions

Promising directions for future research include federated learning to leverage multi-institutional data without centralizing sensitive patient information, self-supervised pretraining on large unannotated CT archives to reduce dependence on expensive expert annotations, and longitudinal detection modeling to track nodal changes across treatment cycles. Extending the 2.5D pipeline toward full 3D detection frameworks as GPU memory constraints continue to ease could further improve volumetric consistency of detections. Integration of detection outputs with

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

radiomics features and clinical metadata may enhance the clinical utility of automated nodal assessment beyond simple localization toward prognostic prediction.

Conclusion

This study compared five detection architectures (Faster R-CNN, YOLOv8-m, MULAN, YOLOv11-m, and RT-DETR-L) for automated lymph node detection in CT imaging across two datasets with contrasting anatomical compositions. Results consistently showed that anatomically focused training improves detection performance across all architectures, providing strong empirical support for anatomically stratified dataset curation in clinical AI development. Among the evaluated models, Faster R-CNN and MULAN achieved the highest mAP50 on the mediastinal dataset (0.641 each), while RT-DETR-L led in recall and F1 score, making it well suited for high-sensitivity screening applications. YOLOv11-m demonstrated the best parameter efficiency. Statistical analysis confirmed that MULAN significantly outperforms the YOLO variants on the focused mediastinal task, with Faster R-CNN showing a comparable trend, while performance converges on the more challenging multi-region dataset.

Absolute performance remains limited for small lymph nodes, reflecting the fundamental challenges of class imbalance and low tissue contrast that persist across all architectures. Future work should focus on larger multi-institutional datasets, full 3D detection frameworks, and prospective clinical validation to advance automated lymph node detection toward routine deployment in oncologic imaging workflows.

References

[1] Ji, H., Hu, C., Yang, X., Liu, Y., Ji, G., Ge, S., Wang, X., & Wang, M. (2023). Lymph node metastasis in cancer progression: Molecular mechanisms, clinical significance and therapeutic interventions. *Signal Transduction and Targeted Therapy*, 8, Article 367. https://doi.org/10.1038/s41392-023-01576-4

[2] Terán, M. D., & Brock, M. V. (2014). Staging lymph node metastases from lung cancer in the mediastinum. *Journal of Thoracic Disease*, 6(3), 230–236. https://doi.org/10.3978/j.issn.2072-1439.2013.12.18

[3] Matsuda, S., Takeuchi, M., Kawakubo, H., & Kitagawa, Y. (2023). Lymph node metastatic patterns and the development of multidisciplinary treatment for esophageal cancer. *Diseases of the Esophagus*, 36(4), Article doad006. https://doi.org/10.1093/dote/doad006

[4] Schwartz, L. H., Litière, S., de Vries, E. G. E., Ford, R., Gwyther, S., Mandrekar, S., Shankar, L., Bogaerts, J., Chen, A., Dancey, J., Hayes, W., Hodi, F. S., Hoekstra, O. S., Huang, E. P., Lin, N., Liu, Y., Therasse, P., Wolchok, J. D., & Seymour, L. (2016). RECIST 1.1—Update and clarification: From the RECIST committee. *European Journal of Cancer*, 62, 132–137. https://doi.org/10.1016/j.ejca.2016.03.081

[5] Sobocińska, M., Białecki, M., Sobociński, B., Martynowska, I., & Cieściński, J. (2022). The influence of contrast enhancement and experience of observers on the assessment of mediastinal lymph nodes in sarcoidosis patients. *Polish Journal of Radiology*, 87, e392–e396. https://doi.org/10.5114/pjr.2022.118303

[6] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 779–788). IEEE. https://doi.org/10.1109/CVPR.2016.91

[7] Jocher, G., Chaurasia, A., & Qiu, J. (2023). *Ultralytics YOLOv8* [Computer software]. https://github.com/ultralytics/ultralytics

[8] Jocher, G., & Qiu, J. (2024). *Ultralytics YOLO11* [Computer software]. https://github.com/ultralytics/ultralytics

[9] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

[10] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., & Chen, J. (2024). DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 16965–16974). IEEE. https://doi.org/10.1109/CVPR52733.2024.01605

[11] Yan, K., Tang, Y., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., & Summers, R. M. (2019). MULAN: Multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation. In D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, & A. Khan (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (Lecture Notes in Computer Science, Vol. 11769, pp. 194–202). Springer. https://doi.org/10.1007/978-3-030-32226-7_22

[12] Roth, H. R., Lu, L., Seff, A., Cherry, K. M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., & Summers, R. M. (2014). A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In P. Golland, N. Hata, C. Barillot, J. Hornegger, & R. Howe (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014* (Lecture Notes in Computer Science, Vol. 8673, pp. 520–527). Springer. https://doi.org/10.1007/978-3-319-10404-1_65

[13] Wang, Z., Sun, D., Zeng, X., Wu, R., & Wang, Y. (2024). Contextual embedding learning to enhance 2D networks for volumetric image segmentation. *Expert Systems with Applications*, 253, Article 124279. https://doi.org/10.1016/j.eswa.2024.124279

[14] Kumar, A., Jiang, H., Imran, M., Valdes, C., Leon, G., Kang, D., Nataraj, P., Zhou, Y., Weiss, M. D., & Shao, W. (2024). A flexible 2.5D medical image segmentation approach with in-slice and cross-slice attention. *Computers in Biology and Medicine*, 182, Article 109173. https://doi.org/10.1016/j.compbiomed.2024.109173

[15] Yang, X., Wu, L., Ye, W., Zhao, K., Wang, Y., Liu, W., Li, J., Li, H., Liu, Z., & Liang, C. (2020). Deep learning signature based on staging CT for preoperative prediction of sentinel lymph node metastasis in breast cancer. *Academic Radiology*, 27(9), 1226–1233. https://doi.org/10.1016/j.acra.2019.11.007

[16] Wang, Y., Yang, C., Yang, Q., Zhong, R., Wang, K., & Shen, H. (2024). Diagnosis of cervical lymphoma using a YOLO-v7-based model with transfer learning. *Scientific Reports*, 14, Article 11073. https://doi.org/10.1038/s41598-024-61955-x

[17] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (Lecture Notes in Computer Science, Vol. 12346, pp. 213–229). Springer. https://doi.org/10.1007/978-3-030-58452-8_13

[18] Yu, Q., Wang, Y., Yan, K., Li, H., Guo, D., Zhang, L., Shen, N., Wang, Q., Ding, X., Lu, L., Ye, X., & Jin, D. (2024). Effective lymph nodes detection in CT scans using location debiased query selection and contrastive query representation in transformer. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer Vision – ECCV 2024* (Lecture Notes in Computer Science, Vol. 15100, pp. 180–198). Springer. https://doi.org/10.1007/978-3-031-72946-1_11

[19] Roth, H. R., Lu, L., Seff, A., Cherry, K. M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., & Summers, R. M. (2015). A new 2.5D representation for lymph node detection in CT [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.AQIIDCNM

[20] Idris, T., Somarouthu, S., Jacene, H., LaCasce, A., Ziegler, E., Pieper, S., Khajavi, R., Dorent, R., Pujol, S., Kikinis, R., & Harris, G. (2024). Mediastinal lymph node quantification (LNQ): Segmentation of heterogeneous CT data (Version 1) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/QVAZ-JA09