

Ultrasound Deep Learning for Distinguishing Benign Reactive from Hematologic Malignancy-Associated Cervical Lymphadenopathy in Children: Development and Validation

Zhiqing Wu

Children's Hospital of Soochow University

Yali Yang

Children's Hospital of Soochow University

Shijie Zhang

Children's Hospital of Soochow University

Zhiming Chen

Children's Hospital of Soochow University

Shibei Du

Children's Hospital of Soochow University

Bin Wu

Children's Hospital of Soochow University

Xiaobo Liu

Children's Hospital of Soochow University

Ran Zhuo

rannyzhuo@163.com

Children's Hospital of Soochow University

Research Article

Keywords: pediatric cervical lymphadenopathy, ultrasonography, deep learning, leukemia, lymphoma, reactive hyperplasia

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11182960/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Cervical lymphadenopathy is a common reason for pediatric outpatient visits. It is usually reactive but may also accompany leukemia or lymphoma. Because these conditions may have overlapping ultrasound appearances, we developed a deep-learning model to distinguish benign reactive from hematologic malignancy-associated cervical lymphadenopathy. The model may provide valuable support for diagnosing hematologic malignancies and guiding differential diagnosis in pediatric outpatient practice.

Methods

This single-center retrospective study included 653 cervical lymph-node annotations after four duplicate images with discordant diagnoses were excluded. The benign reactive group comprised 283 annotations, and the hematologic malignancy-associated group comprised 370. Related images were kept in the same data fold, yielding 559 non-redundant image groups for analysis. An ImageNet-pretrained EfficientNetV2-S model using consensus lesion masks was evaluated with grouped nested five-fold cross-validation. Model probabilities were calibrated with Platt scaling. The primary endpoint was the area under the receiver operating characteristic curve (AUC); accuracy, sensitivity, specificity, predictive values, and the Brier score were also assessed.

Results

The model achieved an AUC of 0.863 (95% confidence interval, 0.834 to 0.893) and an average precision of 0.865. At the Youden operating point, accuracy was 0.784, sensitivity was 0.773, specificity was 0.796, and positive predictive value was 0.824. The calibrated Brier score was 0.154. AUCs for the four diagnosis-specific comparisons ranged from 0.839 to 0.883.

Conclusions

Ultrasound image analysis showed good internal discrimination between benign reactive and hematologic malignancy-associated cervical lymphadenopathy in children. These findings support multicenter, patient-level evaluation of the model as an adjunct to ultrasound assessment. This model could provide radiologists and other clinicians with decision support for the non-invasive ultrasound assessment of cervical lymphadenopathy in children.

1. Background

Cervical lymphadenopathy is a frequent reason for pediatric consultation. Most enlarged lymph nodes are related to infection or reactive hyperplasia, but a small and clinically important proportion are associated with leukemia, lymphoma, or another malignant disorder [1]. Ultrasound is the preferred first-line imaging examination in children because it is readily available, does not use ionizing radiation, and can assess nodal shape, echogenicity, hilum, margins, and vascularity. Rounded morphology, loss of the echogenic hilum, heterogeneous echotexture, and abnormal vascular patterns have all been linked to malignancy; however, these findings overlap across diseases and no single feature is sufficiently reliable in isolation [2-5].

Previous pediatric studies have mainly combined radiologist-assessed ultrasound findings with clinical or laboratory variables. Decision trees and conventional machine-learning models based on these features have shown useful diagnostic performance [4-6]. Magnetic resonance and computed tomography radiomics have also been explored in children, but these approaches require additional imaging or manually structured features [7, 8]. They do not directly address whether routine ultrasound images alone contain enough information to support the distinction that matters most at the first examination.

Deep learning and radiomics have recently been applied to cervical lymph-node ultrasound. Multicenter work in adult or mixed-age populations has reported high diagnostic performance using grayscale and color Doppler images [9], while other studies have shown that convolutional neural networks and ultrasound radiomics can capture image patterns associated with nodal etiology [10-12]. Pediatric evidence remains limited, particularly for the clinically focused distinction between benign reactive lymphadenopathy and lymphadenopathy associated with hematologic malignancy [13].

We therefore developed a deep-learning model that uses routine cervical lymph-node ultrasound images and lesion contours to distinguish benign reactive from hematologic malignancy-associated lymphadenopathy. We evaluated discrimination, probability calibration, and clinically relevant operating points using nested internal validation. We also examined performance across the four underlying diagnoses: acute lymphoblastic leukemia, lymphoma, Epstein-Barr virus (EBV)-associated lymphadenitis, and nonspecific reactive hyperplasia.

2. Methods

2.1. Study Design and Reference Standards

This retrospective study was conducted at Children's Hospital of Soochow University and was approved by the institutional review board (No. 2026CS095). The requirement for informed consent was waived because de-identified data from a retrospective image archive were used. The manuscript was prepared with reference to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) 2024 Update [14].

The source diagnosis associated with each image was established before model development. Acute lymphoblastic leukemia was diagnosed by bone marrow morphology and immunophenotyping.

Lymphoma was confirmed by histopathologic examination of an excisional or core-needle biopsy. Epstein-Barr virus-associated lymphadenitis was diagnosed using EBV-DNA and serologic testing. Nonspecific reactive hyperplasia was defined by clinical follow-up showing regression of the enlarged lymph node without subsequent evidence of malignancy.

Epstein-Barr virus-associated lymphadenitis and nonspecific reactive hyperplasia were combined as the benign reactive group. Acute lymphoblastic leukemia and lymphoma were combined as the hematologic malignancy-associated group. The hematologic malignancy-associated group was defined as the positive class and the benign reactive group as the negative class throughout the analysis.

The initial archive contained 657 cervical lymph-node annotations. Four annotations corresponded to exact duplicate images with discordant diagnoses and were excluded before model training. The final dataset comprised 653 annotations. Images that were identical or highly similar were assigned to the same cross-validation fold, resulting in 559 non-redundant image groups. The model used ultrasound images and lesion masks only; age, sex, symptoms, laboratory tests, and other clinical variables were not included. The study sample by diagnosis and clinical group is summarized in Table 1.

Table 1
Study sample by diagnosis and clinical group.

Diagnosis	Clinical group	Lesion annotations	Non-redundant image groups
EBV-associated lymphadenitis	Benign reactive	123	108
Nonspecific reactive hyperplasia	Benign reactive	160	142
Acute lymphoblastic leukemia	HM-associated	171	152
Lymphoma	HM-associated	199	157
Total	Both groups	653	559

Four exact duplicate images with discordant diagnoses were excluded before model training. Related images were kept in the same cross-validation fold. EBV, Epstein-Barr virus; HM, hematologic malignancy.

2.2. Image Annotation and Preprocessing

Three ultrasound physicians at attending level or above independently delineated the lymph nodes. They were blinded to disease group and to one another's contours. Contours that differed were jointly reviewed by all three physicians, and a consensus contour was agreed upon before model development.

For each image, a rectangular region centered on the annotated lymph node was expanded to include the nodal margin and a small amount of adjacent tissue, then resized to 384 x 384 pixels. Pixels outside the lesion mask were replaced with a neutral gray value to reduce the influence of screen text, device interface elements, and distant anatomy. The red, green and blue (RGB) image was normalized using ImageNet parameters, and the binary lesion mask was supplied as a fourth input channel. The same manually agreed consensus masks defined the image crop and fourth input channel during both training and outer-fold testing; automatic segmentation was not evaluated in this study.

During training, image augmentation included horizontal flipping, small rotations and translations, modest scaling, and limited brightness and contrast adjustment. Geometric transformations were applied identically to the image and mask.

2.3. Deep-Learning Model

We used an EfficientNetV2-S network initialized with ImageNet weights [15]. The first convolutional layer was adapted to accept the RGB image and lesion mask, and the final classifier was replaced with a two-class output layer. Models were trained with the AdamW optimizer and cross-entropy loss, with early stopping based on validation loss. Mild label smoothing and horizontal-flip test-time augmentation were considered within the inner validation loop and were retained only when they improved validation performance.

When a non-redundant image group contained more than one related annotation, logits were averaged within the group before the loss and performance measures were calculated. This prevented repeated views or annotations of the same image from receiving disproportionate weight.

For visual interpretation, gradient-weighted class activation mapping (Grad-CAM) was applied to the final convolutional block of one representative refit. One correctly classified, single-annotation outer-fold test image was selected for each underlying diagnosis. The attention maps were descriptive and were not used for model selection or performance estimation. De-identified wider-field images of the same examples are provided in Additional file 1 (Figure S2).

2.4. Nested Internal Validation

Model development and testing used grouped nested five-fold cross-validation [16, 17]. In each outer iteration, one fold was held out for testing and the remaining four folds were used for training and inner model selection. The outer test fold was not used to choose the model, calibrator, or classification threshold. Identical or highly similar images remained in the same fold throughout the analysis. Patient identifiers were not available for partitioning; folds were therefore disjoint at the non-redundant image-group level.

The inner loop selected between the base model and the label-smoothed model and determined whether horizontal-flip test-time augmentation should be used. To reduce variability from neural-network training, each selected outer-fold model was refitted 10 times and the resulting probabilities were averaged. The

combined outer-fold predictions therefore represent internal test predictions for all 559 non-redundant image groups. Outer-fold composition and selected models are detailed in Additional file 1 (Table S1).

2.5. Probability Calibration and Statistical Analysis

Model probabilities were calibrated using Platt scaling [18, 19]. For each outer test fold, the calibration parameters and operating thresholds were estimated only from the corresponding development data. The main operating point was selected using the Youden index. A second operating point placed greater emphasis on sensitivity for hematologic malignancy. Fold-specific calibration parameters and thresholds are provided in Additional file 1 (Table S2).

The primary endpoint was the area under the receiver operating characteristic curve (ROC AUC). Secondary measures were average precision, accuracy, balanced accuracy, sensitivity, specificity, F1 score, positive predictive value (PPV), negative predictive value (NPV), Brier score, calibration slope, and expected calibration error. Ninety-five percent confidence intervals were calculated with 2000 stratified bootstrap samples. All performance measures were calculated at the non-redundant image-group level. AUCs were also calculated separately for acute lymphoblastic leukemia versus each benign diagnosis and for lymphoma versus each benign diagnosis.

Analyses were performed with Python 3.11.9, PyTorch 2.12.0.dev20260408 + cu128, and scikit-learn 1.6.1.

3. Results

3.1. Study Sample

The final dataset included 653 lymph-node annotations: 283 in the benign reactive group and 370 in the hematologic malignancy group. The benign group comprised 123 EBV-associated lymphadenitis annotations and 160 nonspecific reactive hyperplasia annotations. The malignant group comprised 171 acute lymphoblastic leukemia annotations and 199 lymphoma annotations. After related images were grouped, 250 benign and 309 malignant non-redundant image groups remained. The study workflow is shown in Fig. 1, and representative images and model attention maps are shown in Fig. 2.

3.2. Model Selection and Discrimination

Inner validation selected the mildly label-smoothed model in 3 of the five outer iterations and the base model in the remaining 2. Horizontal-flip test-time augmentation did not provide a consistent benefit and was not retained.

The internally tested model achieved an AUC of 0.863 (95% confidence interval (CI), 0.834 to 0.893). AUCs in the five outer test folds ranged from 0.801 to 0.916. The pooled cross-fitted calibrated AUC was 0.854 (95% CI, 0.823 to 0.883), and average precision was 0.865 (95% CI, 0.832 to 0.897).

The ROC and precision-recall curves, outer-fold AUCs, and diagnosis-specific comparisons are presented in Fig. 3. The principal discrimination and calibration measures are summarized in Table 2.

Table 2
Discrimination and calibration of the final model.

Metric	Estimate (95% CI)
ROC AUC	0.863 (95% CI, 0.834 to 0.893)
Pooled calibrated ROC AUC	0.854 (95% CI, 0.823 to 0.883)
Average precision	0.865 (95% CI, 0.832 to 0.897)
Brier score	0.154 (95% CI, 0.138 to 0.171)
Calibration intercept	0.193 (95% CI, 0.070 to 0.323)
Calibration slope	0.971 (95% CI, 0.829 to 1.164)
Expected calibration error	0.046 (95% CI, 0.039 to 0.083)

AUC, area under the receiver operating characteristic curve; CI, confidence interval. Confidence intervals were obtained from 2000 stratified bootstrap samples.

3.3. Calibration and Operating Points

After calibration, the Brier score was 0.154 (95% CI, 0.138 to 0.171), the calibration intercept was 0.193 (95% CI, 0.070 to 0.323), the calibration slope was 0.971 (95% CI, 0.829 to 1.164), and the expected calibration error was 0.046 (95% CI, 0.039 to 0.083). At the Youden operating point, accuracy was 0.784 (95% CI, 0.751 to 0.818), sensitivity was 0.773 (95% CI, 0.728 to 0.819), specificity was 0.796 (95% CI, 0.748 to 0.844), PPV was 0.824 (95% CI, 0.788 to 0.862), and NPV was 0.740 (95% CI, 0.700 to 0.781). At the higher-sensitivity operating point, accuracy was 0.791 (95% CI, 0.758 to 0.823), sensitivity was 0.809 (95% CI, 0.767 to 0.851), and specificity was 0.768 (95% CI, 0.716 to 0.820).

Table 3
Classification performance at two clinically relevant operating points.

Operating point	Accuracy	Sensitivity	Specificity	PPV	NPV
Youden operating point	0.784 (0.751–0.818)	0.773 (0.728–0.819)	0.796 (0.748–0.844)	0.824 (0.788–0.862)	0.740 (0.700–0.781)
Higher-sensitivity operating point	0.791 (0.758–0.823)	0.809 (0.767–0.851)	0.768 (0.716–0.820)	0.812 (0.778–0.847)	0.765 (0.723–0.807)

Hematologic malignancy-associated lymphadenopathy was the positive class. Values are estimates with 95% bootstrap confidence intervals. PPV, positive predictive value; NPV, negative predictive value.

3.4. Performance across Underlying Diagnoses

The AUC was 0.839 for acute lymphoblastic leukemia versus EBV-associated lymphadenitis, 0.852 for acute lymphoblastic leukemia versus nonspecific reactive hyperplasia, 0.876 for lymphoma versus EBV-associated lymphadenitis, and 0.883 for lymphoma versus nonspecific reactive hyperplasia. Discrimination was therefore slightly higher for lymphoma than for acute lymphoblastic leukemia in both benign comparisons. The corresponding comparisons and diagnosis-specific probability distributions are provided in Additional file 1 (Table S3 and Figure S1).

4. Discussion

4.1. Principal Findings

In this pediatric cervical lymph-node ultrasound study, the deep-learning model achieved an internally validated AUC of 0.863 and an average precision of 0.865. At the Youden operating point, sensitivity was 0.773 and specificity was 0.796. These findings show that routine ultrasound images contain information for separating benign reactive from hematologic malignancy-associated lymphadenopathy, even without clinical or laboratory variables.

The model was developed for a practical pediatric ultrasound question: whether an enlarged cervical lymph node is more consistent with a benign reactive process or warrants concern for leukemia or lymphoma. It used only the ultrasound image and the delineated lesion, without clinical or laboratory variables. Performance remained above an AUC of 0.80 in every outer test fold, indicating that the result was not driven by a single favorable split. The calibrated model also provided a balanced combination of sensitivity and specificity at the Youden operating point.

4.2. Comparison with Previous Studies

Previous pediatric studies have established that nodal morphology and vascularity can help identify malignancy. Park et al. reported validation accuracies of 83.3% to 86.1% for a diagnostic tree combining ultrasound and clinical findings [4], and a 500-child cohort reported an AUC of 0.92 for a regression model that included structured ultrasound assessment [5]. These approaches depend on manually recorded imaging features or clinical information. In contrast, our model analyzed the image and lesion contour directly and achieved an AUC of 0.863 without additional clinical variables.

In pediatric radiomics, a magnetic resonance imaging (MRI)-based model for benign and malignant cervical lymph nodes achieved a test AUC of 0.80 [7]. In predominantly adult cervical lymph-node cohorts, Yuan et al. reported a maximum AUC of 0.81 for grayscale ultrasound deep learning [10], whereas Zhu et al. reported AUCs of 0.896 to 0.963 in multicenter test cohorts using both grayscale and color Doppler images [9]. Direct numerical comparison is limited by differences in age, disease spectrum, imaging protocol, and validation design. Nevertheless, these studies and our findings support the premise that routine ultrasound images contain quantitative patterns that are relevant to nodal etiology.

4.3. Clinical Implications

At the Youden operating point, the PPV exceeded 0.82, indicating that higher model scores were associated with hematologic malignancy in this study cohort. Such an output could provide a quantitative cue for closer ultrasound review when interpreted alongside the clinical assessment. Because predictive values depend on disease prevalence and cohort composition, they should not be transferred directly to general pediatric practice. A higher-sensitivity operating point was also evaluated for settings in which missing hematologic malignancy would be of greater concern.

The AUC was slightly higher for lymphoma than for acute lymphoblastic leukemia when each was compared with the two benign diagnoses. Lymphoma may more often show a rounded shape, marked hypoechogenicity, and hilar abnormality, whereas cervical nodes in children with acute lymphoblastic leukemia may share more nonspecific appearances with viral or reactive lymphadenopathy. This finding suggests that the most difficult cases for future model development may be those separating leukemia-associated from reactive or EBV-associated lymphadenitis.

The intended role of the model is to provide an objective second reading for the ultrasound physician. It is not designed to replace clinical assessment, laboratory investigation, bone marrow examination, or histopathology.

4.4. Strengths and Limitations

Strengths of this study include diagnosis-specific reference standards, independent blinded contouring by three experienced ultrasound physicians, grouped nested cross-validation, repeated model fitting, and explicit probability calibration. The study was retrospective and conducted at a single center; external testing was not performed. Lesion masks were also provided manually. Multicenter studies with patient-level partitioning should assess automatic segmentation and prospective reader assistance across different ultrasound systems and institutions.

5. Conclusions

The proposed ultrasound deep-learning model showed good internal discrimination (AUC 0.863) and a balanced sensitivity-specificity profile for distinguishing benign reactive from hematologic malignancy-associated cervical lymphadenopathy in children. The findings support further multicenter, patient-level validation and prospective reader studies of the model as an adjunct to ultrasound assessment.

Abbreviations

AUC

area under the receiver operating characteristic curve

CI

confidence interval

CLAIM

Checklist for Artificial Intelligence in Medical Imaging

DNA
deoxyribonucleic acid
EBV
Epstein-Barr virus
Grad-CAM
gradient-weighted class activation mapping
HM
hematologic malignancy
MRI
magnetic resonance imaging
NPV
negative predictive value
PPV
positive predictive value
RGB
red,green and blue
ROC
receiver operating characteristic.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Children's Hospital of Soochow University (approval No. 2026CS095). The requirement for informed consent was waived by the Institutional Review Board because of the retrospective study design and the use of de-identified data.

Consent for publication

All ultrasound images presented in the manuscript and additional files were de-identified before figure preparation.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Suzhou Science, Education and Strong Health Project (QNXM2024026), the Youth Talents in Science and Education Program of Suzhou (KJXW2022019), the Suzhou Applied

Author Contribution

Conceptualization, Z.W., Z.C., B.W. and R.Z.; methodology, Z.W., S.Z., Z.C. and S.D.; software and formal analysis, Z.W., S.Z. and S.D.; ultrasound image curation and annotation, Z.C.; clinical validation and interpretation, Y.Y., B.W. and X.L.; writing - original draft, Z.W., Y.Y. and Z.C.; writing - review and editing, all authors; supervision and funding acquisition, R.Z. All authors have read and agreed to the published version of the manuscript.

Acknowledgement

The authors thank the ultrasound and hematology teams at Children's Hospital of Soochow University for their support in image review and diagnostic verification.

Data Availability

The de-identified ultrasound images are not publicly available because of institutional privacy restrictions. Derived aggregate results and the analysis code are available from the corresponding author on reasonable request, subject to institutional approval.

References

1. Deosthali A, Donches K, DelVecchio M, Aronoff S. Etiologies of pediatric cervical lymphadenopathy: a systematic review of 2687 subjects. *Glob Pediatr Health*. 2019;6:2333794X19865440. <https://doi.org/10.1177/2333794X19865440>.
2. Moon J, Hwang J, Kim PH. Diagnostic performance of ultrasound for differentiating malignant from benign cervical lymphadenopathy in children: a systematic review and meta-analysis. *J Ultrasound Med*. 2025;44:1329–41. <https://doi.org/10.1002/jum.16686>.
3. Kartal Ö, Ataş E, Gürsel O. Differentiation of benign from malignant cervical lymphadenopathy by ultrasonography in children. *Arch Argent Pediatr*. 2020;118:11–7. <https://doi.org/10.5546/aap.2020.eng.11>.
4. Park JE, Ryu YJ, Kim JY, Kim YH, Park JY, Lee H, et al. Cervical lymphadenopathy in children: a diagnostic tree analysis model based on ultrasonographic and clinical findings. *Eur Radiol*. 2020;30:4475–85. <https://doi.org/10.1007/s00330-020-06794-w>.
5. Çalışkan Kamış Ş, Yağcı B. Childhood lymphadenopathy: ultrasonographic predictors of malignancy in a retrospective cohort of 500 patients. *Front Pediatr*. 2025;13:1663515. <https://doi.org/10.3389/fped.2025.1663515>.

6. Zijtregtop EAM, Winterswijk LA, Beishuizen TPA, Zwaan CM, Nievelstein RAJ, Meyer-Wentrup FAG, et al. Machine learning logistic regression model for early decision making in referral of children with cervical lymphadenopathy suspected of lymphoma. *Cancers (Basel)*. 2023;15:1178. <https://doi.org/10.3390/cancers15041178>.
7. Xu Y, Chu C, Wang Q, Xiang L, Lu M, Yan W, et al. Using T2-weighted magnetic resonance imaging-derived radiomics to classify cervical lymphadenopathy in children. *Pediatr Radiol*. 2024;54:1302–14. <https://doi.org/10.1007/s00247-024-05954-0>.
8. Al Bulushi Y, Saint-Martin C, Muthukrishnan N, Maleki F, Reinhold C, Forghani R. Radiomics and machine learning for the diagnosis of pediatric cervical non-tuberculous mycobacterial lymphadenitis. *Sci Rep*. 2022;12:2962. <https://doi.org/10.1038/s41598-022-06884-3>.
9. Zhu Y, Meng Z, Fan X, Duan Y, Jia Y, Dong T, et al. Deep learning radiomics of dual-modality ultrasound images for hierarchical diagnosis of unexplained cervical lymphadenopathy. *BMC Med*. 2022;20:269. <https://doi.org/10.1186/s12916-022-02469-z>.
10. Yuan H, Ruan J, Wen W, Liu J, Peng Y. Development and validation of deep learning models for qualitative classification of benign and malignant enlarged cervical lymph nodes based on ultrasound images. *Gland Surg*. 2025;14:1649–59. <https://doi.org/10.21037/gs-2024-576>.
11. Liu Y, Chen J, Zhang C, Li Q, Zhou H, Zeng Y, et al. Ultrasound-based radiomics can classify the etiology of cervical lymphadenopathy: a multi-center retrospective study. *Front Oncol*. 2022;12:856605. <https://doi.org/10.3389/fonc.2022.856605>.
12. Gong W, Li M, Wang S, Jiang Y, Wu J, Li X, et al. An improved diagnostic deep learning model for cervical lymphadenopathy characterization. *Ultrasound Med Biol*. 2025;51:1814–20. <https://doi.org/10.1016/j.ultrasmedbio.2025.06.025>.
13. Mezher M, Elgamal M, Schoeman S, Al-Hasani M, Otero HJ, Sultan LR. Ultrasound radiomics in pediatric imaging: current applications, challenges, and future directions toward clinical implementation. *Diagnostics (Basel)*. 2026;16:1669. <https://doi.org/10.3390/diagnostics16111669>.
14. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiol Artif Intell*. 2024;6:e240300. <https://doi.org/10.1148/ryai.240300>.
15. Tan M, Le QV. EfficientNetV2: smaller models and faster training. In: Meila M, Zhang T, editors. *Proceedings of the 38th International Conference on Machine Learning*. PMLR; 2021. pp. 10096–10106. (Proceedings of Machine Learning Research; vol. 139). <https://proceedings.mlr.press/v139/tan21a.html>
16. Rouzrokh P, Khosravi B, Faghani S, Moassefi M, Vera Garcia DV, Singh Y, et al. Mitigating bias in radiology machine learning: 1. Data handling. *Radiol Artif Intell*. 2022;4:e210290. <https://doi.org/10.1148/ryai.210290>.
17. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7:91. <https://doi.org/10.1186/1471-2105-7-91>.

18. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning. PMLR; 2017. pp. 1321–1330. (Proceedings of Machine Learning Research; vol. 70). <https://proceedings.mlr.press/v70/guo17a.html>

19. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Topic Group ‘Evaluating diagnostic tests and prediction models’ of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. <https://doi.org/10.1186/s12916-019-1466-7>.

Figures

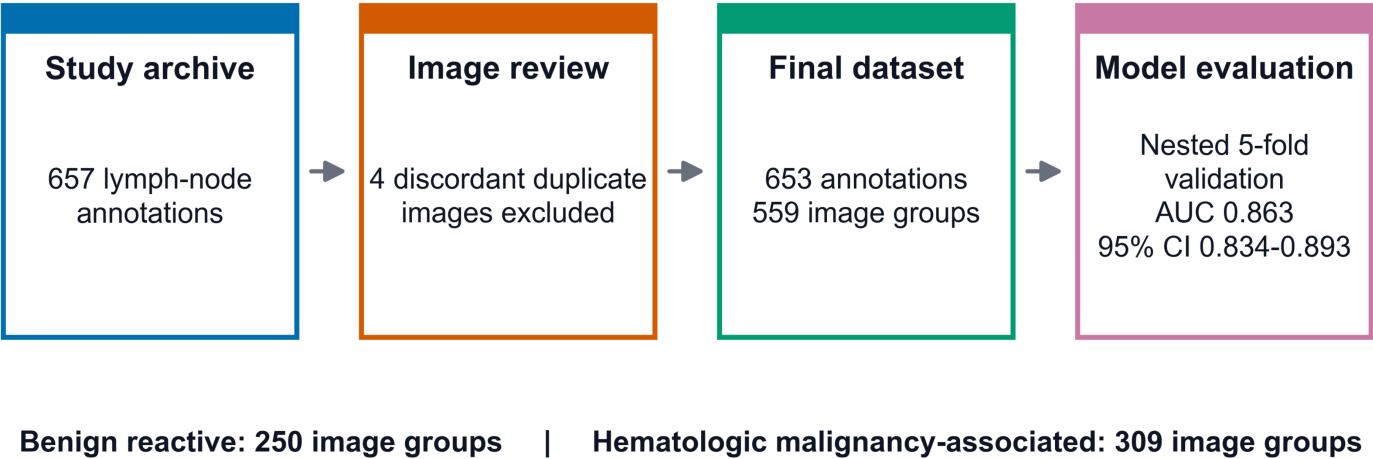


Figure 1

Study workflow. After removal of four duplicate annotations with conflicting diagnoses, 653 lymph-node annotations were retained and organized into 559 non-redundant image groups. Related images remained in the same fold throughout nested cross-validation. AUC, area under the receiver operating characteristic curve; CI, confidence interval.

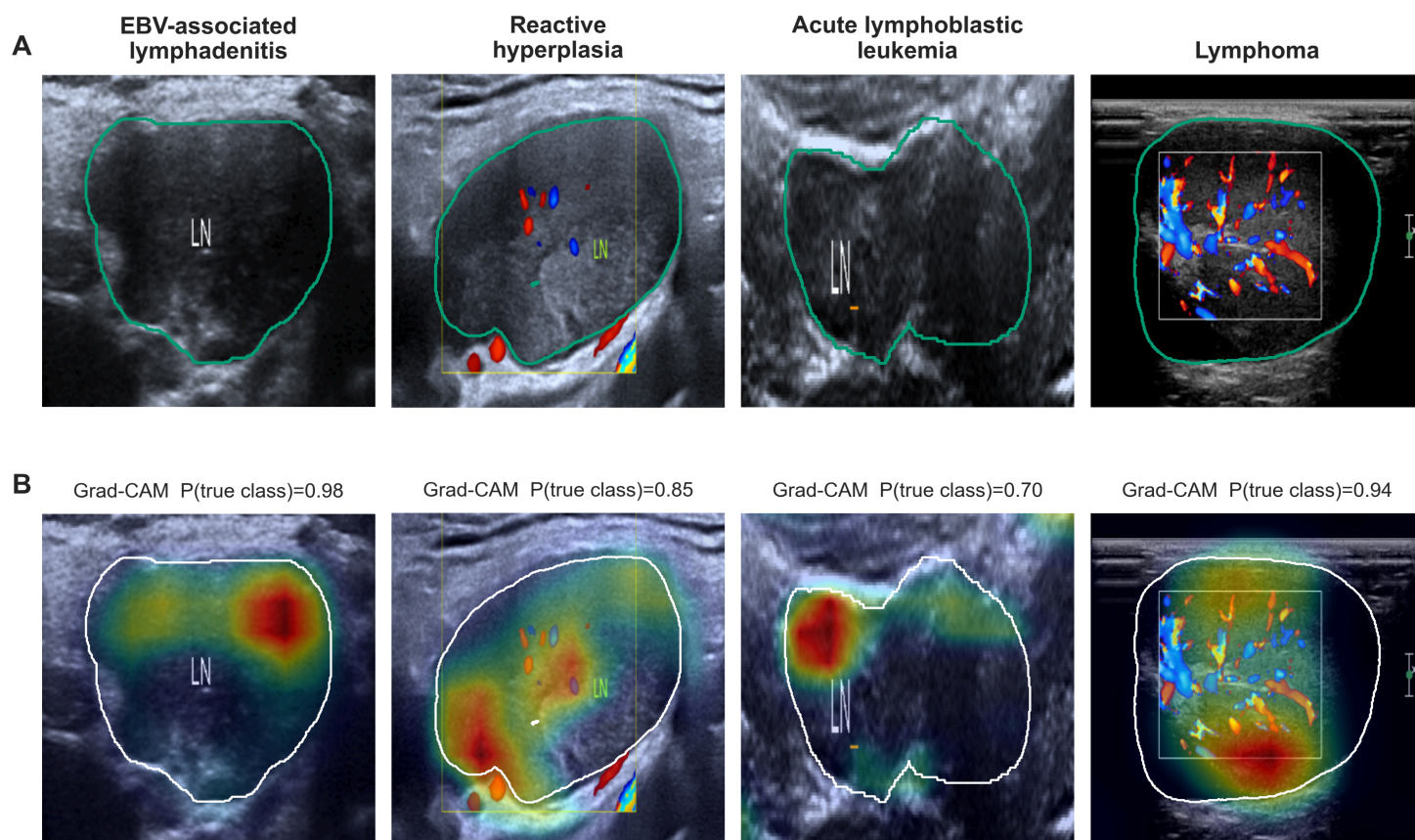


Figure 2

Representative ultrasound regions and Grad-CAM attention maps from correctly classified outer-fold test images. Columns show EBV-associated lymphadenitis, nonspecific reactive hyperplasia, acute lymphoblastic leukemia, and lymphoma. The top row shows the de-identified lesion region with the consensus contour; the bottom row shows the corresponding attention map for the true binary class, generated from one representative refit of the selected outer-fold model. EBV, Epstein-Barr virus; Grad-CAM, gradient-weighted class activation mapping; LN, lymph node.

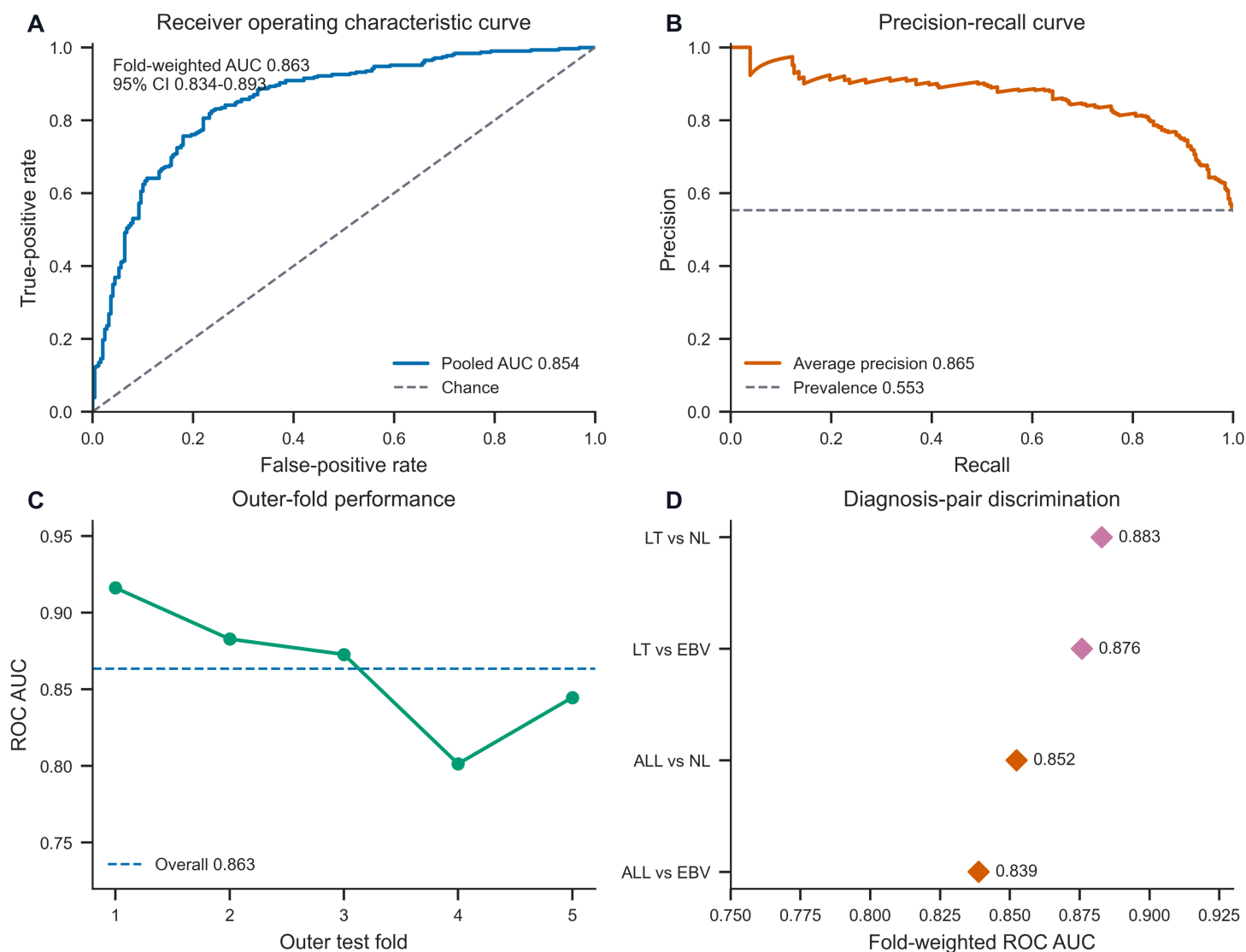


Figure 3

Discrimination of the final deep-learning model. (A) Receiver operating characteristic curves; (B) precision-recall curve; (C) AUC in each outer test fold; and (D) AUC for the four diagnosis-specific comparisons. In panel D, ALL denotes acute lymphoblastic leukemia, LT lymphoma, NL nonspecific reactive hyperplasia, and EBV Epstein-Barr virus-associated lymphadenitis. AUC, area under the receiver operating characteristic curve; CI, confidence interval; ROC, receiver operating characteristic.

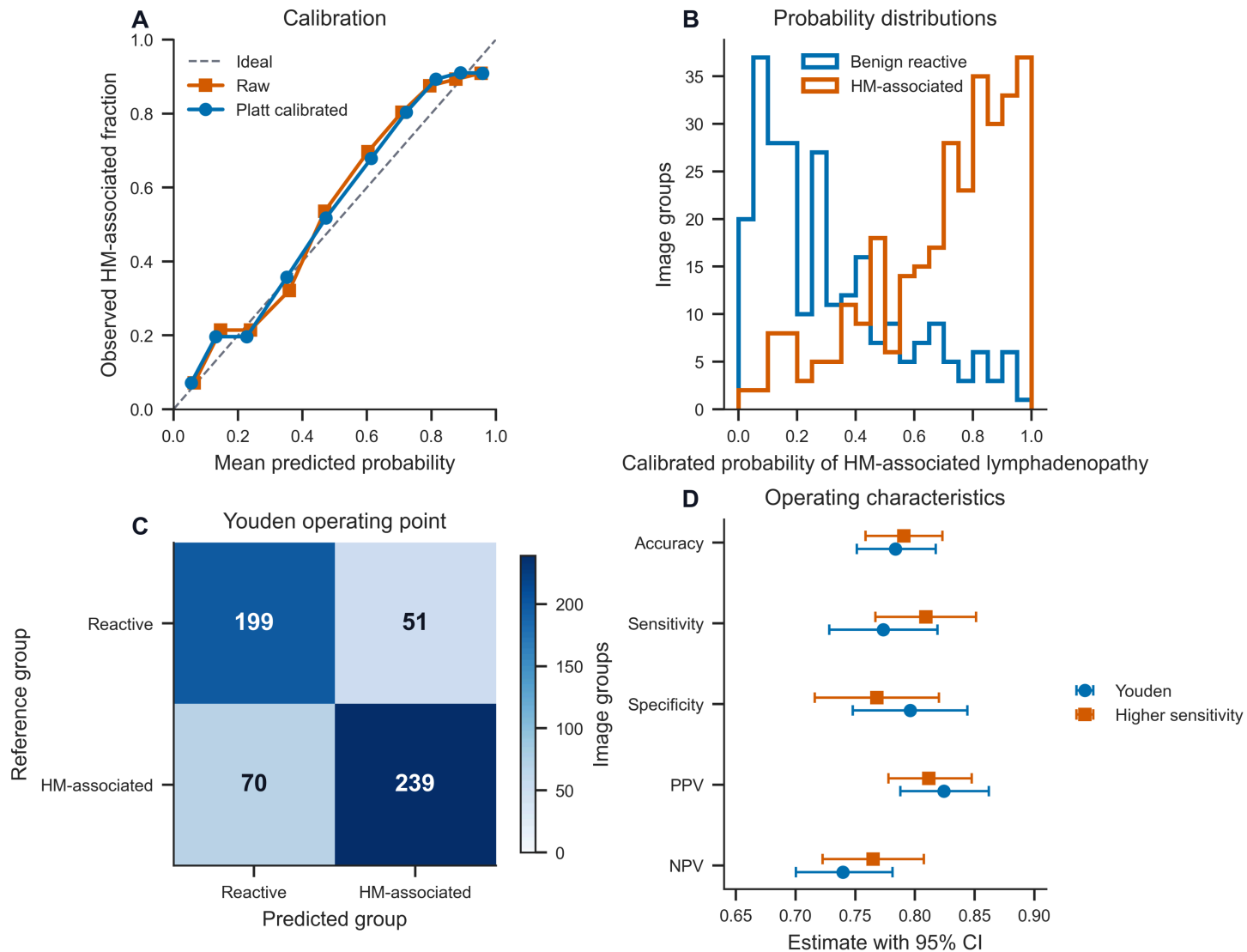


Figure 4

Calibration and classification performance. (A) Calibration curve; (B) calibrated probability distributions; (C) confusion matrix at the Youden operating point; and (D) accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. HM, hematologic malignancy; PPV, positive predictive value; NPV, negative predictive value; CI, confidence interval.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterial.docx](#)