Deep Learning-based Detection of Neonatal Intracranial Hemorrhage Using Cranial Ultrasound

Mike Weaver, PhD, Gary Ge, PhD, Alexander Alsalihi, Nanditha George, MD, Adrian Dawkins, MD, Jie Zhang, PhD
Department of Radiology, University of Kentucky, Lexington, KY 40513

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

ABSTRACT

Purpose:
Intracranial hemorrhage in premature infants is a frequent and clinically significant finding, with cranial ultrasound serving as the first-line imaging modality. Interpretation can be challenging due to subtle and variable imaging appearances. This study aimed to develop and evaluate a deep learning model for detection of neonatal intracranial hemorrhage on cranial ultrasound.

Methods:
Ultrasound images were retrospectively collected from 2,127 neonatal patients, including 124 hemorrhage-positive cases (5.8% prevalence). A total of 1,488 examinations were used for model development. A ResNet-50–based deep learning model was pretrained on public ultrasound datasets and fine-tuned on the study cohort. Data were split into training, validation, and test sets (85/15) with patient-level separation and five-fold cross-validation. A multiple instance learning framework was applied, representing each examination as a set of ultrasound frames to enable patient-level classification. Model performance was evaluated using area under the receiver operating characteristic curve (AUROC), precision–recall area under the curve (PR-AUC), and sensitivity and specificity at selected decision thresholds.

Results:
Across five cross-validation folds, AUROC ranged from 0.68 to 0.75 (mean 0.72 ± 0.03), and PR-AUC ranged from 0.12 to 0.24 (mean 0.18 ± 0.06). At thresholds determined by Youden's J statistic, mean sensitivity was 0.64 ± 0.18 and specificity was 0.66 ± 0.16. When thresholds were adjusted to prioritize high sensitivity (target 0.90), mean sensitivity increased to 0.77 ± 0.16 with a corresponding decrease in specificity to 0.48 ± 0.21.

Conclusion:
Despite substantial class imbalance, the deep learning model demonstrated meaningful discrimination for detection of neonatal intracranial hemorrhage on cranial ultrasound. Performance reflected expected sensitivity–specificity trade-offs when prioritizing hemorrhage detection. These results support the feasibility of deep learning–based decision support for neonatal cranial ultrasound and motivate further optimization to improve clinical utility.

CONTACT

Mike Weaver, PhD
University of Kentucky
Lexington, KY 40513
mike.weaver@uky.edu
859-218-0404

INTRODUCTION

Neonatal germinal matrix-intraventricular hemorrhage is a clinically important complication of prematurity, associated with mortality, ventricular dilation/hydrocephalus, and adverse neurodevelopmental outcome [1]. Cranial ultrasound is the first-line tool for routine bedside surveillance in high-risk preterm infants because it is portable, repeatable, inexpensive, and avoids ionizing radiation [2,3]. Interpretation remains challenging because early hemorrhage may be subtle, appearance evolves over serial examinations, and image quality depends on acoustic windows, acquisition technique, and lesion location; MRI remains complementary for selected subtle, posterior fossa, or cerebellar abnormalities [4].

Recent neonatal cranial-ultrasound AI studies have demonstrated feasibility for germinal matrix hemorrhage detection/grading and broader severe-lesion screening [5-7]. However, patient-level hemorrhage detection from routine multi-frame cranial ultrasound examinations remains less developed. Multiple instance learning is well suited to this setting because an examination can be treated as a bag of frames with a single patient-level label, avoiding the need for exhaustive frame-level hemorrhage annotation [8]. In this study, we evaluated a ResNet-50 model pretrained with radiology-domain weights and adapted to a MIL framework for patient-level detection of neonatal intracranial hemorrhage on cranial ultrasound [9,10].

METHODS AND MATERIALS

This retrospective study included cranial ultrasound examinations from 2,127 neonatal patients, including 124 hemorrhage-positive patients (5.8% prevalence). A total of 1,488 examinations were used for model development. All data partitioning was performed at the patient level to avoid leakage across splits.
DICOM ultrasound images were filtered to exclude invalid images, and retained grayscale frames were converted to 3-channel tensors, intensity-normalized, and resized to a fixed input matrix. Each examination was represented as a bag of frames. During training, a fixed number of frames was sampled from each exam to construct a bag; during evaluation, frames were sampled deterministically across the exam to provide consistent patient-level inference. Representative final settings used 288 × 288 input images, 72 frames per bag during training, and 100 frames per bag during evaluation.

A ResNet-50 backbone initialized with RadImageNet weights was used as the frame encoder [10]. Frame-level logits were aggregated with a MIL pooling operation to generate a single patient-level hemorrhage probability. Training used five-fold patient-level cross-validation, AdamW optimization, mixed-precision training, and progressive fine-tuning of the backbone. Performance was evaluated with AUROC, precision-recall AUC (PR-AUC), and sensitivity/specificity at prespecified operating points. PR-AUC was emphasized because of the low positive prevalence [11].

RESULTS

Across five cross-validation folds, AUROC ranged from 0.68 to 0.75, with a mean AUROC of 0.72 ± 0.03 (Fig. 1). PR-AUC ranged from 0.12 to 0.24, with a mean PR-AUC of 0.18 ± 0.06 (Fig. 2). The cohort hemorrhage prevalence was 5.8%, corresponding to the no-skill baseline for the precision-recall curve.

At thresholds selected using Youden's J statistic, mean sensitivity was 0.64 ± 0.18 and mean specificity was 0.66 ± 0.16. At the high-sensitivity operating point, mean sensitivity was 0.77 ± 0.16 and mean specificity was 0.48 ± 0.21.

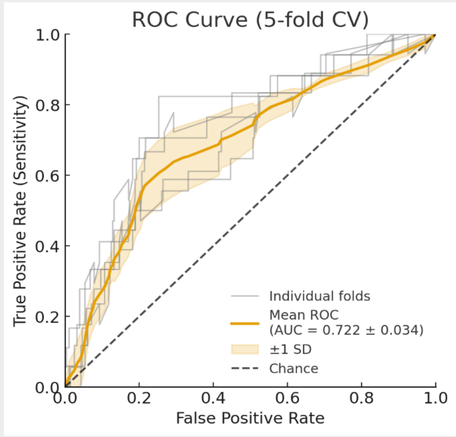

Figure 1. Mean receiver operating characteristic (ROC) curve across five cross-validation folds. Shaded region indicates ±1 standard deviation. Individual fold curves are shown in gray.

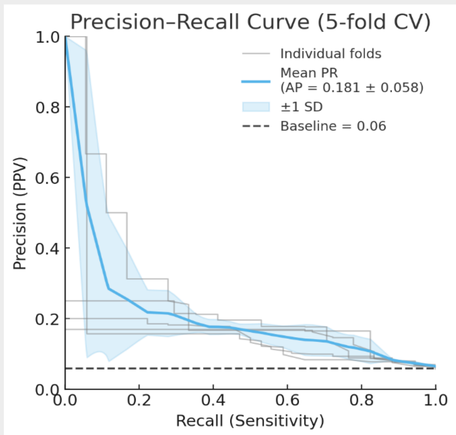

Figure 2. Mean precision–recall (PR) curve across five cross-validation folds. Shaded region indicates ±1 standard deviation. The dashed line denotes baseline performance corresponding to hemorrhage prevalence

DISCUSSION

This study supports the feasibility of weakly supervised, patient-level hemorrhage detection from routine neonatal cranial ultrasound. The MIL design is appropriate for this task because hemorrhage-related signal may be present in only a small subset of frames, while labels are available only at the examination or patient level [8]. The observed AUROC and PR-AUC suggest that the model learned discriminative signal despite severe class imbalance.

DISCUSSION (CONT.)

The results should be framed as decision-support feasibility, not autonomous diagnosis. The PR-AUC was approximately three times the prevalence baseline, but absolute precision remains constrained by the low positive rate. Similarly, threshold selection strongly affected the balance between sensitivity and specificity. A sensitivity-prioritized operating point may be more appropriate for a screening-support workflow, but the associated reduction in specificity would increase the number of false-positive exams requiring review.

Key limitations include single-institution retrospective data, a small positive cohort, weak patient-level labels, no frame-level localization, and no external validation. Future work will focus on increasing positive case representation, improving frame sampling and MIL aggregation, calibrating model outputs, identifying which frames drive patient-level predictions, and validating performance on independent data.

CONCLUSION

A ResNet-50–based multiple instance learning model demonstrated feasible patient-level detection of neonatal intracranial hemorrhage on cranial ultrasound, achieving a mean AUROC of 0.72 and a mean PR-AUC of 0.18 across five cross-validation folds. These findings support continued development of AI-assisted decision support for neonatal cranial ultrasound.

REFERENCES

1. Papile LA, Burstein J, Burstein R, Koffler H. Incidence and evolution of subependymal and intraventricular hemorrhage: a study of infants with birth weights less than 1,500 gm. J Pediatr. 1978;92(4):529-534. doi:10.1016/S0022-3476(78)80282-0.
2. Hand IL, Shellhaas RA, Milla SS, et al. Routine neuroimaging of the preterm brain. Pediatrics. 2020;146(5):e2020029082. doi:10.1542/peds.2020-029082.
3. van Wezel-Meijler G, Steggerda SJ, Leijser LM. Cranial ultrasonography in neonates: role and limitations. Semin Perinatol. 2010;34(1):28-38. doi:10.1053/j.semperi.2009.10.002.
4. Burkitt K, Kang O, Jyoti R, et al. Comparison of cranial ultrasound and MRI for detecting brain injury in extremely preterm infants and correlation with neurological outcomes at 1 and 3 years. Pediatr Res. 2019;86(3):379-386.
5. Kim KY, Wang K, Yang W, et al. Assessment of germinal matrix hemorrhage on head ultrasounds with convolutional neural network. J Ultrasound Med. 2022. PMID:35064324.
6. Ibrahim NM, Alanize H, Alqahtani L, et al. Deep Learning Approaches for the Assessment of Germinal Matrix Hemorrhage Using Neonatal Head Ultrasound. Sensors (Basel). 2024;24(21):7052. doi:10.3390/s24217052.
7. Lin Z, et al. Deep learning approach for screening neonatal cerebral lesions on ultrasound in China. Nat Commun. 2025;16:63096. doi:10.1038/s41467-025-63096-9.
8. Quellec G, Cazuguel G, Cochener B, Lamard M. Multiple-instance learning for medical image and video analysis. IEEE Rev Biomed Eng. 2017;10:213-234. doi:10.1109/RBME.2017.2651164.
9. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.
10. Mei X, Liu Z, Robson PM, et al. RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. Radiol Artif Intell. 2022;4(5):e210315. doi:10.1148/ryai.210315.
11. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432.