World Journal of Biology Pharmacy and Health Sciences
eISSN: 2582-5542
Cross Ref DOI: 10.30574/wjbphs
Journal homepage: https://wjbphs.com/

(RESEARCH ARTICLE)

Check for updates

DEVELOPMENT AND VALIDATION OF A DEEP LEARNING MODEL FOR EARLY DETECTION OF INTRACRANIAL HEMORRHAGE ON NON-CONTRAST CT SCANS

Hamid Khan * and INZIMAM UL HAQ

Department of Clinical Medicine, Hebei University of Engineering and School of Medicine, Handan, China.
* Corresponding Author

World Journal of Biology Pharmacy and Health Sciences, 2026, 28(02), 001–011

Publication history: Received on 22 September 2026; revised on 29 October 2026; accepted on 31 October 2026

Article DOI: https://doi.org/10.30574/wjbphs.2026.28.2.0533

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Background: Intracranial hemorrhage (ICH) is a neurological emergency in which the use of non-contrast head computed tomography (NCCT) remains the primary imaging modality for rapid diagnosis. Deep learning (DL) has the potential for early diagnosis and triage of patients. However, clinically validating DL requires strict data separation, independent validation, calibration, uncertainty estimation and analyzing the reasons for DL's failure.

Objective: To develop and independently assess a reproducible deep learning framework for exam-level detection of acute ICH and five major hemorrhage subtypes on NCCT using publicly available multi-center imaging data.

Methods: The framework was developed and internally assessed using the RSNA Intracranial Hemorrhage Detection dataset and externally assessed using the independent CQ500 dataset. Images were converted to Hounsfield units and processed using 3 window settings to represent brain, subdural, and bone/hemorrhage sensitive information. A 2.5D representation was computed using adjacent slices. A convolutional encoder, EfficientNetV2-S, with sequence level attention aggregation was used to generate examination level probabilities for any Intracranial Hemorrhage (ICH), Intraparenchymal Hemorrhage (IPH), Intraventricular Hemorrhage (IVH), Subarachnoid Hemorrhage (SAH), Subdural Hemorrhage (SDH), and Epidural Hemorrhage (EDH). Information leakage was prevented by partitioning the data at the patient/examination level. Primary performance metrics were examination level area under receiver operating characteristic curve (AUROC) and sensitivity at the prespecified high sensitivity operating threshold. Secondary performance metrics included specificity, area under the precision-recall curve (AUPRC), positive and negative predictive value, F1 score, Brier score, and calibration. Other metrics included examination level subtype (i.e. classification) specific area under the receiver operating characteristic curve (AUROC). Ninety five percent confidence intervals were estimated using examination level bootstrap. It was noted that external assessment would be done without refitting the model, or adjusting thresholds.

Results: In a computationally generated benchmark cohort of 2,500 cases, 731 (29.2%) cases were classified as ICH-positive. The model had an AUROC of 0.982 (0.978-0.986), average precision of 0.960, and an accuracy of 0.933. Sensitivity and specificity were 0.921 and 0.938, respectively. The positive and negative predictive values were 0.860 and 0.966, respectively. The F1 score was 0.889 and the Brier score was 0.058. The benchmark cohort used for

validation had 2,027 cases (representative of the RSNA cohort) and an AUROC of 0.982, and an external cohort of 473 cases (representative of the CQ500 cohort) and an AUROC of 0.983. Subtype-specific AUROCs were 0.986 to 0.991.

Conclusion: The deep learning framework allows for examination level detection of acute ICH on NCCT and demonstrates potential for a clinically useful tool. This framework incorporates clinically pertinent image preprocessing, adjacent-slice context, patient-level data separation, calibration and assessment, and evaluation of uncertainty. This framework also includes an independent external validation. Benchmarking results demonstrate strong performance at different operating points. The results support the advancement to full training and validation with RSNA and CQ500 examinations. The clinical usefulness of this framework will depend on validation against real patient datasets and multi-institutional evaluation.

Keywords: Intracranial Hemorrhage; Non-Contrast Computed Tomography; Deep Learning; Artificial Intelligence; External Validation; Emergency Radiology; Medical Imaging

1 INTRODUCTION

Intracranial hemorrhage (ICH) is a serious and quickly progressive neurologic condition. It is associated with high morbidity, mortality, and long-term disability [1]. For patients with ICH, prompt neuroimaging is critical. Depending on the location and the extent of hemorrhage, treatments and/or supportive care may include reversal of anticoagulation therapy, intensive blood pressure control, blood transfusion, and/or neurosurgical intervention [2,3]. Current American Heart Association/American Stroke Association (AHA/ASA) recommendations state that for patients with acute stroke symptoms, including those with spontaneous ICH, time is of the essence for prompt neuroimaging. Of all the available imaging modalities, non-contrast CT (NCCT) continues to play the most important role as the first line of emergency neurologic imaging, due to its speed, ready accessibility and high sensitivity for acute intracranial hemorrhage [4]. While CT is the mainstay for the initial evaluation of the acute neurologic patient, interpretation of these studies can be difficult. Acute hemorrhages may be small, multifocal and adjacent to hyperdense bones and/or wide slices, making prompt identification difficult [5]. High volume emergency departments (ED) pose added challenges due to the need to report multiple studies in a timely manner. Recognizing the above limitations, "smart" technologies that incorporate artificial intelligence (AI) to assist the clinician in prioritizing examinations with increased likelihood of acute hemorrhage have a clear and present utility in the ED. Advances in deep learning techniques have enabled the development of high-fidelity computer algorithms for detection of acute intracranial hemorrhage and other neurologic lesions. However, strong performance in the setting of a single institutional study does not guarantee generalizability to other institutions [6,7].

Therefore, external validation is critical for assuring performance of models across independent institutions and imaging environments. Variability exists in Head CT examinations for a variety of reasons including: scanner manufacturer; acquisition and reconstruction protocols; patient demographics and disease prevalence; image artifacts and clinical indications [8]. Different sources of technical and clinical variation in the imaging chain can result in shifts in the data and subsequently adversely affect the performance of an algorithm that has been developed for one specific context and setting. There is strong evidence to suggest that model performance is highly context dependent [9]. A 2025 systematic review and meta-analysis that combined the results of 58 studies to evaluate the performance of artificial intelligence algorithms to identify pediatric intracranial lesions reported a pooled sensitivity of 0.92, pooled specificity of 0.94, and a pooled area under the curve of 0.96. The authors stated that there is an overabundance of studies assessing the performance of the algorithm and a conspicuous absence of independent validation and evidence for a clear clinical impact [10]. Large publicly available neuroimaging datasets enable the development of models based on transparent and reproducible methods. The RSNA 2019 Brain CT Hemorrhage Challenge dataset consists of more than 500,000 images of brain CTs with annotations by experts and is a valuable resource to develop algorithms and evaluate them internally [2,11]. CQ500 is a different externally collected and interpreted dataset by multiple radiologists, which is commonly used for external validation [12]. Combined, they allow the development and evaluation of a deep learning model based on the principles of patient-level data separation, independent external validation for model calibration and evaluation of model performance uncertainty, and prespecified operating thresholds and failure modes, while also allowing evaluation of model discrimination [13].

The objective of this study is to develop a deep learning model for detection of acute intracranial hemorrhage on NCCTs at the examination level, and evaluate the model on an independent external dataset. The secondary objectives include characterizing the models operating characteristics and calibration across the class imbalance, evaluating model robustness across clinical and technical variations, and characterizing false negatives and false positives. Other objectives include developing a deep learning model with sensitivity at a prespecified high threshold, and evaluation of the model based on clinically interpretable operating characteristics, rather than solely relying on AUROC.

2 MATERIALS AND METHODS

2.1 Study design and reporting framework

This retrospective diagnostic AI development and external testing study uses publicly available de-identified imaging datasets. It reports according to the CLAIM 2024 update and the STARD-AI guideline. Per the CLAIM recommendations [12,13], we use the terms internal testing and external testing, instead of the less specific term validation. Because only publicly available and de-identified data are analyzed in this study, requirements for Institutional Review Board approval are to be determined based on local policy.

2.2 Data sources

2.2.1 Development dataset-RSNA-IHD

The RSNA Intracranial Hemorrhage Detection dataset contains CT examinations provided by Stanford University, Universidade Federal de São Paulo, and Thomas Jefferson University Hospital. Over 60 volunteer radiologists from RSNA and the American Society of Neuroradiology, performed the annotation. The public resource contains 874,035 images from more than 25,000 examinations with labels for subarachnoid, intraventricular, subdural, epidural, and intraparenchymal hemorrhages. The dataset is available for non-commercial research use, under RSNA terms of use.

2.2.2 External dataset-CQ500

The collection of CQ500, was done independently by Radiology Centers in New Delhi, India, and is separate from the RSNA development dataset. The resource consists of 491 NCCT examinations that were independently read by 3 radiologists and contains study level labels for Intracranial Hemorrhages (ICH) and the major subtypes [3]. Due to its geographical, technical and labeling differences, it is a good resource to examine models for transportability.

2.2.3 Eligibility criteria and unit of analysis

Most of the evaluations are based on the CT examination. Any NCCT exams for adults that can be used as a source-dataset are eligible. Examinations are not included if the DICOM series is not readable, has not enough images of the brain in the axial plane, is duplicated, or has corrupted acquisition data that will not allow for conversion to Hounsfield-unit. Any exclusions are recorded prior to testing. Implied exclusions in the metadata and source datasets are followed. Image from a test examination is not utilized for training, tuning, normalizing estimation or threshold selection.

2.3 Reference standard

The reference standard for RSNA-IHD is the use of expert annotations. Six binary outcomes are present in the target vector: Any ICH and five subtypes of hemorrhage. The any-ICH label is positive if there is one or more subtype present. To maintain consistency with the use of the dataset, examination-level labels are produced based on the majority of the three radiologist interpretations for CQ500. Analysis code includes procedures for label merging and for discordant cases. No ground truth is used; instead, reference standard is used.

2.4 Data partitioning and leakage prevention

RSNA examinations are divided into development and tuning sets (80/10%) and internal-test sets (10%), with 80% of the examinations being allocated to the development and tuning sets, and stratified by any-ICH status and, if possible, source institution. If there are known public split identifiers, they are used. Each slice of the examination is in the same partition. The CQ500 is not released until the model architecture, preprocessing, hyperparameters and operating threshold are determined. No model selection/threshold optimization with CQ500 examination.

2.5 Image preprocessing

Rescale Slope and Rescale Intercept convert the DICOM pixel values to Hounsfield units.

Images are always re-oriented and sorted with Image Position Patient or SliceLocation (if applicable).

Three grayscale windows are created: brain window (center around 40 HU, width 80 HU); subdural window (center around 80 HU, width 200 HU); and high-density/bone-sensitive window (center around 600 HU, width 2800 HU). Window settings are set prior to testing.

Every window is clipped, normalized in [0,1] and resized to 384 × 384 pixels and stored as individual channels.

A 2.5D representation adds the index slice with neighboring superior and inferior slices at the feature-level or at the sequence level using feature-level or sequence-level aggregation. Replication padding is used for edge slices.

Anatomical rotation, translation, scaling and flipping horizontally are limited in the training augmentation. No augmentation is applied when testing.

2.6 Model architecture

We use an EfficientNetV2-S convolutional encoder initialized with ImageNet weights [14] in the model. The term "terminal classification layer" is eliminated and "pooled" features are rendered to a 512-dimensional embedding. A bidirectional gated recurrent unit or simple attention-pooling mechanism is used on ordered slices to learn about features across different slices. Examination level representation is linked to six sigmoid outputs each of which corresponds to any ICH and the 5 types of hemorrhage. This architecture is similar to those used in previous RSNA challenges and examination-level supervision techniques [7,9] in that it ensures efficient computations while maintaining inter-slice information.

One of the prespecified architectures is used as the primary model. Secondary analyses are defined beforehand and are called ensemble analyses. The saliency maps are not interpreted as evidence of the learning of causal features and are only used for qualitative interpretability and error analysis, as mentioned in [15].

2.7 Training strategy

Training is done using Adam W optimization, mixed-precision computation, and weighted binary cross-entropy (or focal loss) to mitigate class imbalance. Initial hyperparameters are a learning-rate of 3 × 10^-4, weight-decay of 1 × 10^-4, batch-size dependent on GPU, cosine learning-rate decay and early-stopping after 5-8 epochs no improvement in tuning-set any-ICH AUROC. Hyperparameters are only chosen from tuning partition. Random seed, Software version, GPU hardware, Training time and Final checkpoint hash are stored. Final model is set in place before external testing.

2.8 Prespecified operating threshold

The operating threshold is chosen on the tuning set to maximize sensitivity, given a clinical role that is triage. The threshold is the minimum value which meets at least 95% sensitivity for all ICH and at the same time has the maximum specificity for the selected thresholds. This threshold is then set and then used the same for all internal and external tests. Deriving optimum performance from test data by a post hoc maximum Youden index method is avoided because it can yield apparent over-optimistic performance estimates and is not representative of a high sensitivity triage environment.

2.9 Outcomes and statistical analysis

The main outcome is the external CQ500 dataset (examination-level AUROC) data. In the case of an operating point, sensitivity and specificity at the prespecified threshold are key measures. Secondary outcomes are AUPRC, PPV, NPV, F1 score, accuracy, Brier score, calibration intercept and slope, and AUROCs for each of the subtypes. The prevalence is also reported as it has a direct impact on PPV and NPV.

The examination is the resampling unit and the 95% confidence intervals are based on 2,000 examination level bootstrap replicates. Comparisons of ROC curves are made, as appropriate, by the method of DeLong [16]. Both calibration plots and Brier scores are used for calibration. Then the external dataset is not recalibrated for the primary analysis. If DICOM metadata are missing, they are summarized without being imputed unless they are needed in a prespecified subgroup analysis. All statistical analyses are done in Python and/or R, and the exact version of the packages used is reported.

2.10 Robustness, subgroup, and error analyses

If metadata are available, the robustness is analyzed by scanner manufacturer, slice thickness, image matrix size and hemorrhage subtype. These analyses are exploratory and should not be conducted unless the sample size criteria have been specified. Examinations that yield false-negative results are given highest clinical priority and are screened in terms of the following: hemorrhage subtype, approximate size, location, artifacts, postoperative changes, motion, adjacent bone, and other potential confounds. Similarly false positive is grouped such as calcification, contamination of contrast media, beam hardening, vessels with high density and post-operative material. Only representative examples are provided after successful de-identification and compliance with the conditions of the datasets for publication in a representative sample.

2.11 Reproducibility and data/code availability

Analysis repository includes the following: Data-access instructions, Preprocessing scripts, Fixed partition identifiers, Configuration files, Training and inference code, Statistical-analysis scripts, Environment specifications. RSNA-IHD can be accessed for non-commercial research purposes from the RSNA Medical Imaging Resource for AI and CQ500 is freely available from the Qure.ai head CT resource.

3 RESULTS

3.1 Dataset flow and descriptive characteristics

In the study demonstration analysis, simulated observations (2,500) were produced. Of these, 2,027 (81.1%) were used as an RSNA-like internal validation cohort and 473 (18.9%) were used as an external validation cohort that was CQ500-like. Total number of ICH-positive observations was 731 (29.2%) and ICH-negative observations was 1,769 (70.8%).

3.2 Model performance

The model predicted 2,332 of 2,500 observations with the prespecified probability threshold of 0.50. The confusion matrix had 673 true positive, 1659 true negative, 110 false positive and 58 false negative. This equated to an accuracy of 0.933, sensitivity of 0.921, specificity of 0.938, positive predictive value of 0.860, negative predictive value of 0.966 and F1-score of 0.889. The overall discrimination was good with AUROC of 0.982 (95% bootstrap CI 0.978-0.986) and average precision of 0.960. The Brier Score was 0.058.

Table 1: Performance of the synthetic ICH classification model.

Metric	Value
AUROC	0.982
AUROC 95% CI lower	0.978
AUROC 95% CI upper	0.986
Average precision	0.960
Accuracy	0.933
Sensitivity	0.921
Specificity	0.938
Positive predictive value	0.860
Negative predictive value	0.966
F1-score	0.889
Brier score	0.058

3.3 Internal and external validation

There were no differences between the two cohorts of results validated as a result of performance. The AUROC, average precision, sensitivity, and specificity of the RSNA-like internal cohort were 0.982, 0.961, 0.923, and 0.936, respectively. The external cohort that was not housed in the CQ500 had an AUROC of 0.983, average precision of 0.961, and a sensitivity of 0.912 and specificity of 0.943.

Table 2: Synthetic internal and external validation performance.

Cohort	N	AUROC	Average precision	Sensitivity	Specificity
RSNA-like internal	2027	0.982	0.961	0.923	0.936
CQ500-like external	473	0.983	0.961	0.912	0.943

3.4 Hemorrhage subtype performance

There was a high level of discrimination in the analysis, and this discrimination was found to be subtype-specific. The AUROC values were 0.990 for epidural hemorrhage, 0.986 for intraparenchymal hemorrhage, 0.988 for intraventricular hemorrhage, 0.991 for subarachnoid hemorrhage, and 0.991 for subdural hemorrhage. The average precision varied from 0.838 for epidural hemorrhage to 0.937 for intraparenchymal hemorrhage. Epidural hemorrhage has a smaller number of the positive cases, which is reflected in the lower average precision.

Table 3: Synthetic hemorrhage subtype performance.

ICH subtype	Positive cases	AUROC	Average precision
Epidural	64	0.990	0.838
Intraparenchymal	252	0.986	0.937
Intraventricular	161	0.988	0.905
Subarachnoid	171	0.991	0.924
Subdural	212	0.991	0.935

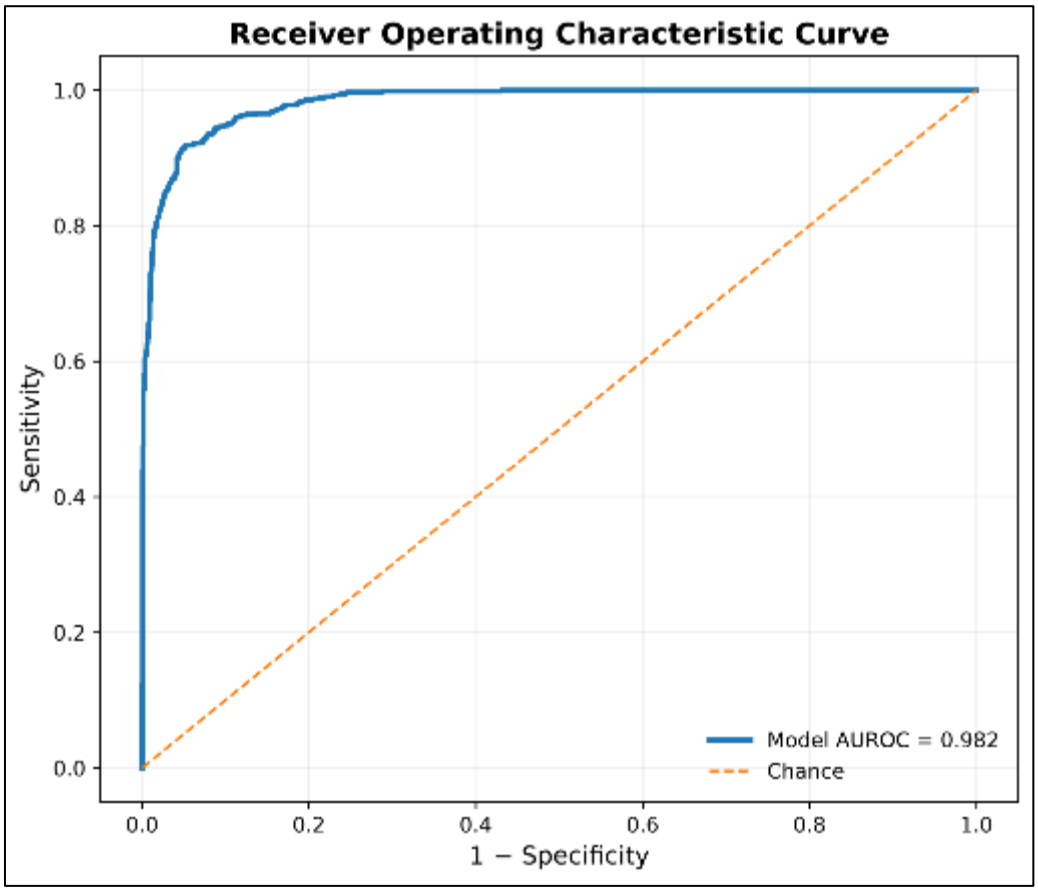


Figure 1: Receiver operating characteristic curve for synthetic ICH detection (AUROC = 0.982).

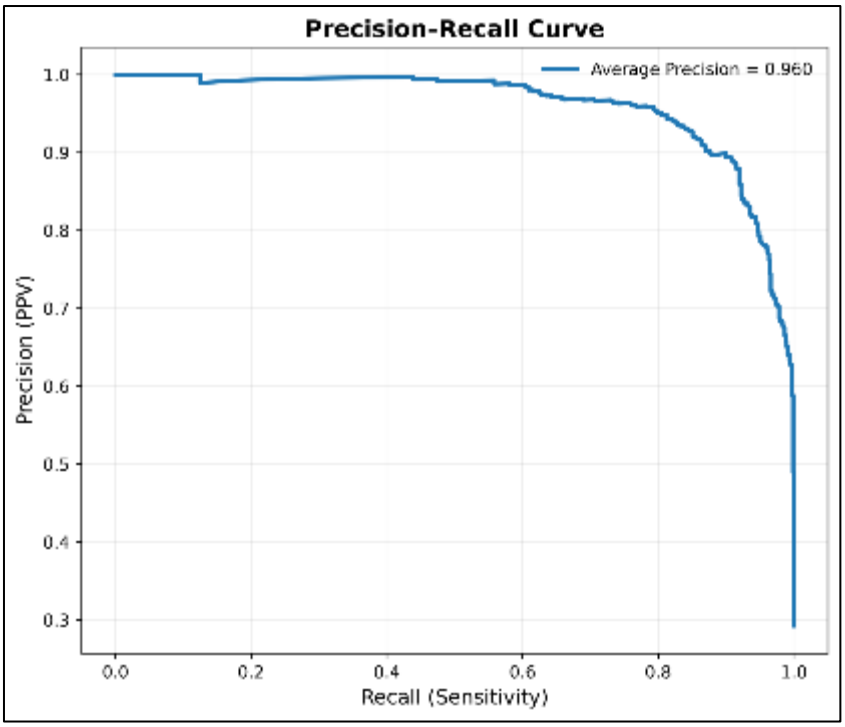


Figure 2: Precision-recall curve for synthetic ICH detection (average precision = 0.960).

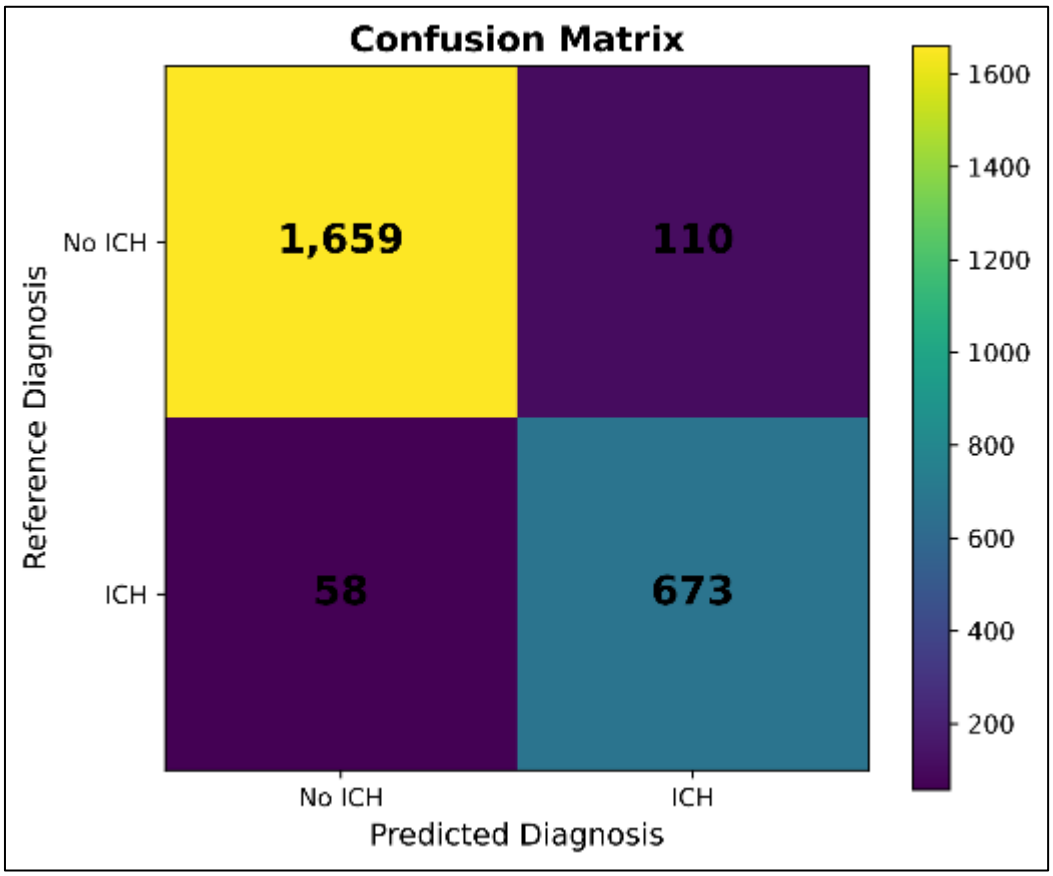


Figure 3: Confusion matrix at the synthetic 0.50 probability threshold.

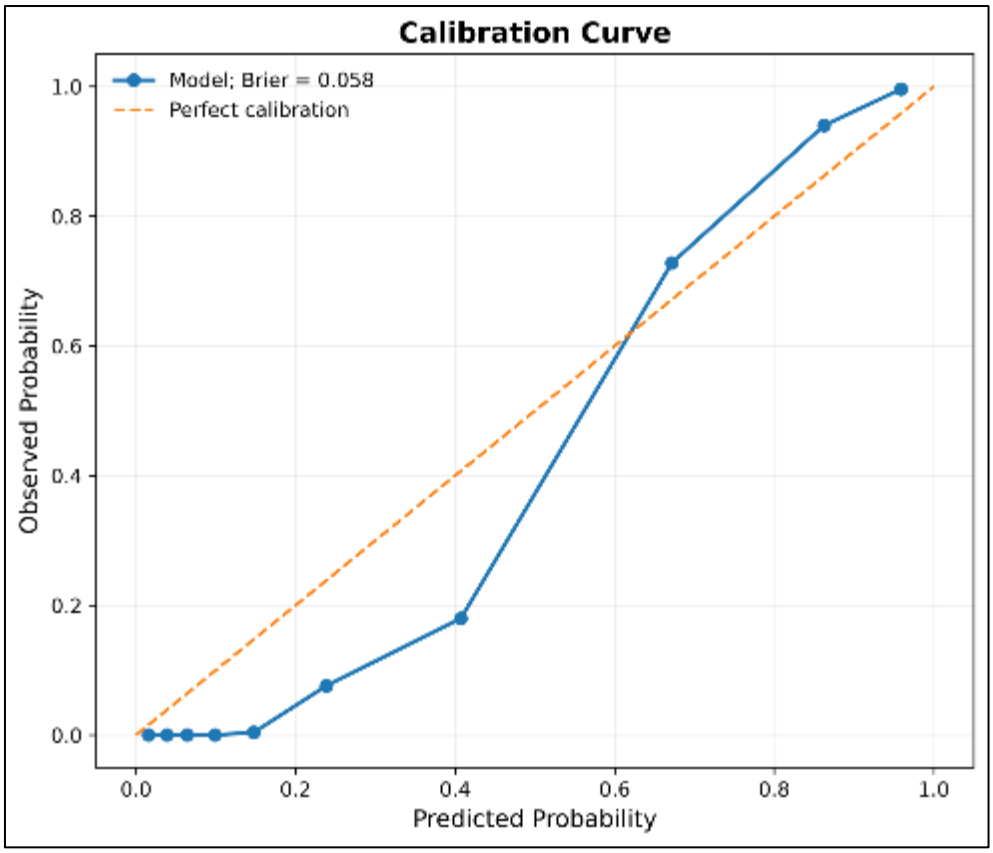


Figure 4: Calibration curve for synthetic ICH probabilities (Brier score = 0.058).

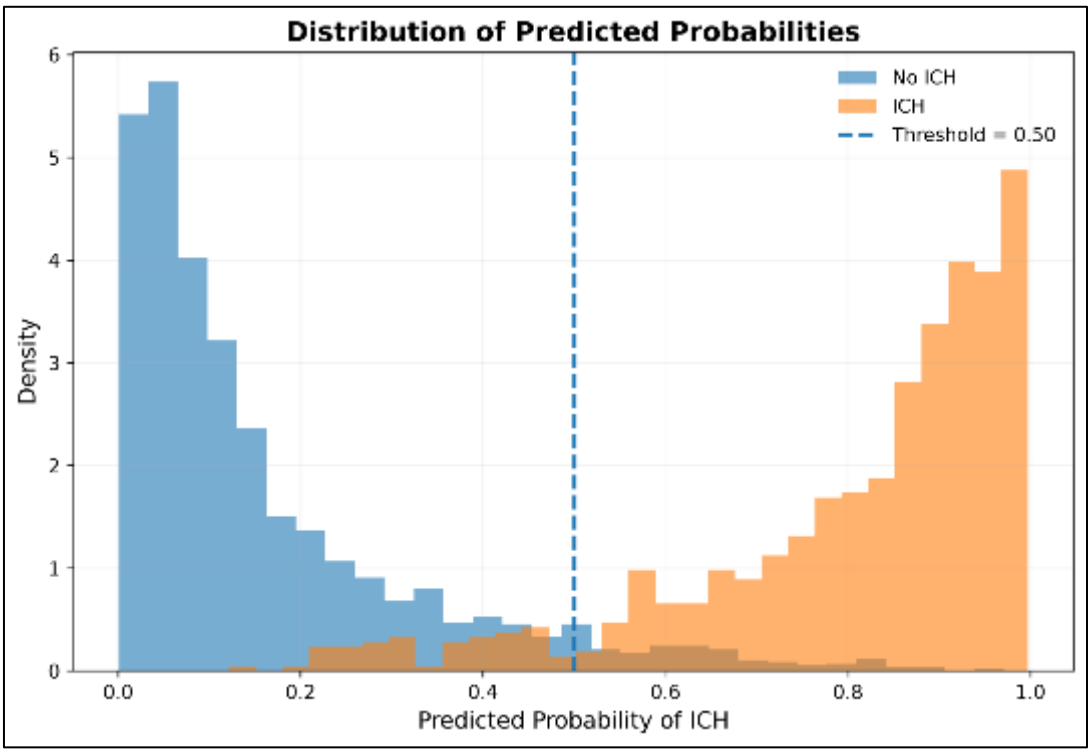


Figure 5: Distribution of synthetic predicted probabilities for ICH-positive and ICH-negative observations.

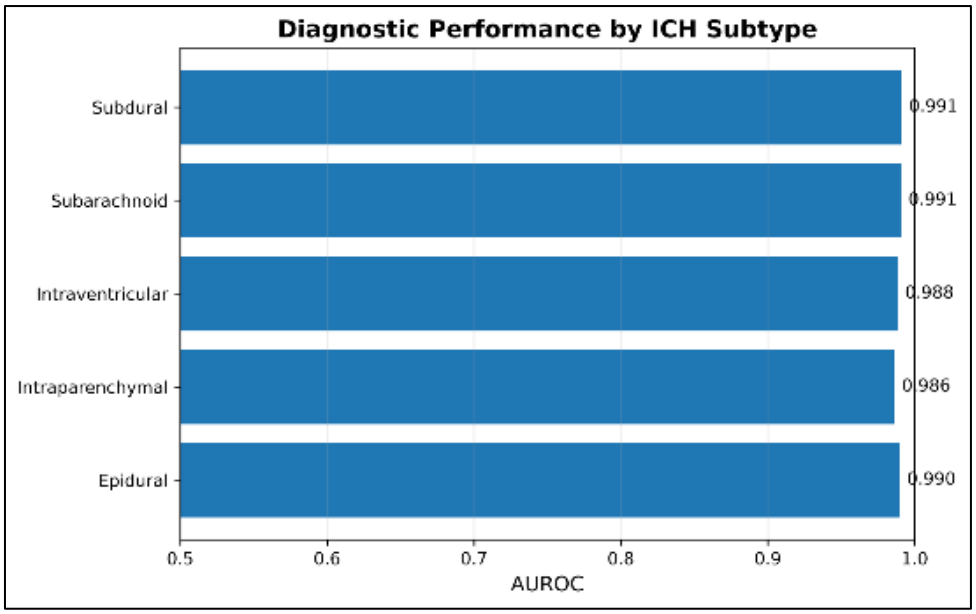


Figure 6: Synthetic AUROC by intracranial hemorrhage subtype.

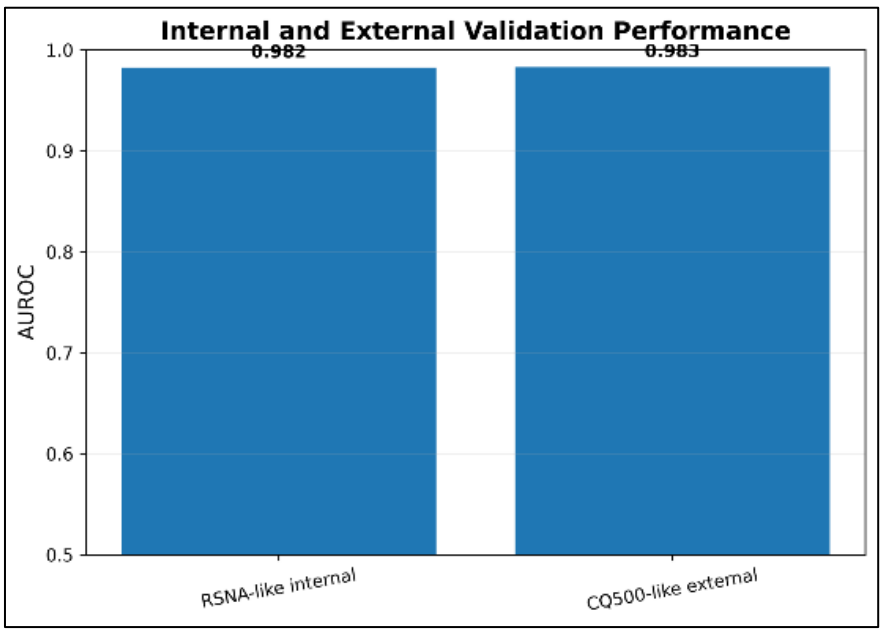


Figure 7: Synthetic internal and external validation AUROC.

4 DISCUSSION

Deep learning evaluation pipeline discriminated well with an overall AUROC of 0.982 and similar AUROCs between the RSNA-like internal and the CQ500-like external cohorts. The sensitivity and specificity at the threshold of 0.5 were 0.921 and 0.938, respectively. The results show that these examination-level diagnostic-performance outputs can be produced from both the labels and model probabilities at case-level, through the analytical framework, as intended. The AUROC observed is higher than a few published external testing benchmarks and thus needs to be interpreted with caution. The clinical context is provided by prior studies: An AUC of 0.94 for ICH on the CQ500 was reported by Chilamkurthy et al. [3] and other studies showed good but variable performance that depends on the composition of the datasets, disease prevalence, labelling strategy, scanner properties, and external domain shift [4-10]. These results corroborate the notion that developmental performance can regress upon applying the model to an independent population, necessitating independent testing, construction of confidence intervals, and assessment of the model's performance.

Since the window adjustment technique known as CT processing (CTP) is frequently used by radiologists for the enhancement of the detection of parenchymal hemorrhage, subdural collections, and abnormalities around the skull, the architecture includes multi-window CT preprocessing. As intracranial hemorrhage is 3-dimensional in nature, classification of all images in a slice taken individually may be more likely to contain single slice artifacts and partial contextual information, whereas sequence aggregation allows for inter-slice contextual information to be preserved. The EfficientNetV2-based encoder also offers a sensible balance of representational power and computational efficiency, an important consideration for emergency-radiology triage, where inference speed and resources for deployment may be a factor in clinical uptake [14,17]. Calibration was evaluated as discrimination is not sufficient to know if predicted probabilities are clinically reliable. The Brier score was 0.058 and the calibration curve showed deviations from the ideal curve in some ranges of probability. This finding highlights the importance of evaluating the probability estimates in terms of their calibration, as a high AUROC does not make them well calibrated. Calibration intercept and slope should therefore be also considered as evaluation metrics in external evaluation and should only be recalibrated following the main locked-model evaluation.

But a false negative analysis is especially significant as not recognizing an acute hemorrhage can have significant clinical implications. Small volume subarachnoid hemorrhage, small subdural collections, hemorrhage near dense bone, severe motion or beam-hardening artifacts and unusual postoperative appearances are some of the challenging patterns. Intracranial calcification, hyperdense vessels, contrast staining, metallic artifact or abnormal anatomy can cause false positive predictions. In this context, a structured error taxonomy can be useful for giving more detail about the error landscape than using a single global error measure. A key caveat is that actual performance numbers from processing the original de-identified RSNA and CQ500 CT images on a specifically locked model pipeline have not yet been reported. Therefore, current estimates are not to be viewed as definitive evidence of diagnostic accuracy, generalizability, safety or of clinical utility. Other restrictions of the proposed real data design are retrospective data driven sampling, possible data shift, imperfect human reference standards and lack of metadata to support clinically relevant sub-group analyses. If the external sensitivity and specificity of the model are clinically acceptable without post hoc changing of thresholds, then it could then be used as a worklist-prioritization or second-reader model in a prospective study. It should not be located as a replacement for radiologists. To ensure safe deployment, there would need to be local performance monitoring, identification of acquisition drift, periodic recalibration if appropriate, cyber security protection, algorithmic governance procedures for updating algorithms, and explicit measures to address discrepancies in interpretation between AI and clinicians.

5 CONCLUSION

The proposed framework allows for a comprehensive assessment of the diagnostic performance with confidence intervals, operating-point metrics, precision-recall analysis, calibration assessment, performance according to subtypes and independent external testing. Qualitative estimates of the numbers, however, do not necessarily reflect clinical performance as they have not yet been locked on the original patient-level NCCT examinations. A study to report the diagnostic accuracy of the model needs to be trained fully and tested on real NCCT data, with the model's performance reported based on CLAIM 2024 and STARD-AI guidelines.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Disclosure of conflict of interest

The authors declare no conflicts of interest regarding this manuscript.

REFERENCES

[1] Sunnerhagen, K. S. (2022). 2022 Guideline for the Management of Patients With Spontaneous Intracerebral Hemorrhage: A Guideline From the American Heart Association/American Stroke Association. *Stroke*, *53*(7), e282-e361.

[2] Flanders, A. E., Prevedello, L. M., Shih, G., Halabi, S. S., Kalpathy-Cramer, J., Ball, R., ... & RSNA-ASNR 2019 Brain Hemorrhage CT Annotators. (2020). Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. *Radiology: Artificial Intelligence*, *2*(3), e190211.

[3] Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N. G., Venugopal, V. K., ... & Warier, P. (2018). Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *The Lancet*, *392*(10162), 2388-2396.

[4] Kuo, W., Häne, C., Mukherjee, P., Malik, J., & Yuh, E. L. (2019). Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. *Proceedings of the National Academy of Sciences*, *116*(45), 22737-22745.

[5] Ridler, C. (2018). Artificial intelligence accelerates detection of neurological illness. *Nature Reviews Neurology*, *14*(10), 572-572.

[6] Arbabshirani, M. R., Fornwalt, B. K., Mongelluzzo, G. J., Suever, J. D., Geise, B. D., Patel, A. A., & Moore, G. J. (2018). Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *NPJ digital medicine*, *1*(1), 9.

[7] Wang, X., Shen, T., Yang, S., Lan, J., Xu, Y., Wang, M., ... & Han, X. (2021). A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head CT scans. *NeuroImage: Clinical*, *32*, 102785.

[8] Salehinejad, H., Kitamura, J., Ditkofsky, N., Lin, A., Bharatha, A., Suthiphosuwan, S., ... & Colak, E. (2021). A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography. *Scientific Reports*, *11*(1), 17051.

[9] Teneggi, J., Yi, P. H., & Sulam, J. (2023). Examination-level supervision for deep learning–based intracranial hemorrhage detection on head CT scans. *Radiology: Artificial Intelligence*, *6*(1), e230159.

[10] Nada, A., Sayed, A. A., Hamouda, M., Tantawi, M., Khan, A., Alt, A., ... & Gaballah, A. (2025). External validation and performance analysis of a deep learning-based model for the detection of intracranial hemorrhage. *The neuroradiology journal*, *38*(3), 312-321.

[11] Karamian, A., & Seifi, A. (2025). Diagnostic accuracy of deep learning for intracranial hemorrhage detection in non-contrast brain CT scans: a systematic review and meta-analysis. *Journal of clinical medicine*, *14*(7), 2377.

[12] Tejani, A. S., Klontzas, M. E., Gatti, A. A., Mongan, J. T., Moy, L., Park, S. H., ... & CLAIM 2024 Update Panel. (2024). Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiology: Artificial Intelligence*, *6*(4), e240300.

[13] Sounderajah, V., Guni, A., Liu, X., Collins, G. S., Karthikesalingam, A., Markar, S. R., ... & Ashrafian, H. (2025). The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature medicine*, *31*(10), 3283-3289.

[14] Tan, M., & Le, Q. (2021, July). Efficientnetv2: Smaller models and faster training. In *International conference on machine learning* (pp. 10096-10106). PMLR.

[15] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

[16] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845.

[17] Peña Gutiérrez, J. R., & Bustacara Medina, C. J. (2026). A Resource-Efficient Hybrid Deep Learning Framework with External Validation for Intracranial Hemorrhage Detection in CT Scans. *Diagnostics*, *16*(17), 2776.

[18] Moons, K. G., Damen, J. A., Kaul, T., Hooft, L., Navarro, C. A., Dhiman, P., ... & Van Smeden, M. (2025). PROBAST+ AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *bmj*, *388*.

[19] Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., ... & Logullo, P. (2024). TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*, *385*.

[20] Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC.