

Strengthening Deep-learning Models for Intracranial Hemorrhage

Detection: Strongly Annotated Computed Tomography Images and Model

Ensembles

Dong-Wan Kang^{1,2,3}, Gi-Hun Park⁴, Wi-Sun Ryu⁴, Dawid Schellingerhout⁵, Museong Kim^{6,7},
Yong Soo Kim⁸, Chan-Young Park⁹, Keon-Joo Lee¹⁰, Moon-Ku Han³, Han-Gil Jeong^{3,6},
Dong-Eog Kim^{11,12}

¹ Headquarters for Public Health Care, Seoul National University Bundang Hospital,
Seongnam, Republic of Korea

² Department of Neurology, Gyeonggi Provincial Medical Center, Icheon Hospital, Icheon,
Republic of Korea

³ Department of Neurology, Seoul National University Bundang Hospital, Seoul National
University College of Medicine, Seongnam, Republic of Korea

⁴ JLK Inc., Artificial Intelligence Research Center, Seoul, Republic of Korea

⁵ Department of Neuroradiology and Imaging Physics, The University of Texas M.D.
Anderson Cancer Center, Houston, USA

⁶ Department of Neurosurgery, Seoul National University Bundang Hospital, Seoul National
University College of Medicine, Seongnam, Republic of Korea

⁷ Hospital Medicine Center, Seoul National University Bundang Hospital, Seoul National
University College of Medicine, Seongnam, Republic of Korea

⁸ Department of Neurology, Nowon Eulji Medical Center, Eulji University School of
Medicine, Seoul, Republic of Korea

⁹ Department of Neurology, Chung-Ang University Hospital, Seoul, Republic of Korea

¹⁰ Department of Neurology, Korea University Guro Hospital, Seoul, Republic of Korea

¹¹ Department of Neurology, Dongguk University Ilsan Hospital, Goyang, Republic of Korea

¹² National Priority Research Center for Stroke, Goyang, Republic of Korea

Corresponding authors:

Han-Gil Jeong, MD, MSc

Department of Neurosurgery and Neurology

Seoul National University Bundang Hospital

82, Gumi-ro 173beon-gil, Bundang-gu, Seongnam 13620, Republic of Korea

Tel: +82-31-787-7460, Fax: +82-31-787-4059

E-mail: han.g.jeong@gmail.com

Dong-Eog Kim, MD, PhD

Department of Neurology

Dongguk University Ilsan Hospital

27 Dongguk-ro, Ilsandong-gu, Goyang 10326, Republic of Korea

Tel: +82-31-961-7207 Fax: +82-31-961-7212

E-mail: kdongeog@duih.org

Abstract

Multiple attempts at intracranial hemorrhage (ICH) detection using deep-learning techniques have been made and plagued with clinical failures. Most studies for ICH detection have insufficient data or weak annotations. We sought to determine whether a deep-learning algorithm for ICH detection trained on a strongly annotated dataset outperforms that trained on a weakly annotated dataset, and whether a weighted ensemble model that integrates separate models trained using datasets with different ICH subtypes is more accurate. We used publicly available brain CT scans from the Radiological Society of North America (27,861 CT scans, 3,528 ICHs) and AI-Hub (53,045 CT scans, 7,013 ICHs) for training datasets. For external testing, 600 CT scans (327 with ICH) from Dongguk University Medical Center and 386 CT scans (160 with ICH) from Qure.ai were used. DenseNet121, InceptionResNetV2, MobileNetV2, and VGG19 were trained on strongly and weakly annotated datasets and compared. We then developed a weighted ensemble model combining separate models trained on all ICH, subdural hemorrhage (SDH), subarachnoid hemorrhage (SAH), and small-lesion ICH cases. The final weighted ensemble model was compared to four well-known deep-learning models. Six neurologists reviewed difficult ICH cases after external testing. InceptionResNetV2, MobileNetV2, and VGG19 models outperformed when trained on strongly annotated datasets. A weighted ensemble model combining models trained on SDH, SAH, and small-lesion ICH had a higher AUC than a model only trained on all ICH cases. This model outperformed four well-known deep-learning models in terms of sensitivity, specificity, and AUC. Strongly annotated data are superior to weakly annotated data for training deep-learning algorithms. Since no model can capture all aspects of a complex task well, we developed a weighted ensemble model for ICH detection after training with large-scale strongly annotated CT scans. We also showed that a better understanding and

management of cases challenging for AI and human is required to facilitate clinical use of ICH detection algorithms.

Key Points

Question Can a weighted ensemble method and strongly annotated training datasets develop a deep-learning model with high accuracy to detect intracranial hemorrhage?

Findings A deep-learning algorithm for detecting ICH trained with a strongly annotated dataset outperformed models trained with a weakly annotated dataset. After ensembling separate models that were trained with only SDH, SAH, and small-lesion ICH, a weighted ensemble model had a higher AUC.

Meaning This study suggests that to enhance the performance of deep-learning models, researchers should consider the distinct imaging characteristics of each hemorrhage subtype and use strongly annotated training datasets.

Introduction

Intracranial hemorrhage (ICH) occurs in the intracranial space and encompasses the following six types: epidural hemorrhage (EDH), subdural hemorrhage (SDH), subarachnoid hemorrhage (SAH), intraparenchymal hemorrhage (IPH), intraventricular hemorrhage (IVH), and mixed hemorrhage. A timely and accurate diagnosis of ICH and its subtypes is critical for treatment, because of the high mortality and morbidity. In addition, assessing the location and extent of ICH is important for outcome prediction. However, neuroradiology training requires a significant investment of time and resources; accordingly, neuroradiologists are scarce in many countries.¹ Without neuroradiologists' assistance, doctors who see ICH patients often misdiagnose.²

Deep-learning algorithms have recently made progress in accurately detecting ICH on CT scans.³ However, their clinical use is limited due to challenges in identifying SDH or SAH.⁴ In contrast to other ICH subtypes, SDH is more likely to present in the subacute stage and has reduced CT attenuation, similar to that of brain tissue. SAH may also appear isoattenuating if there is only a small amount of blood mixed with cerebrospinal fluid.⁵ Moreover, a pseudo-SAH is not an uncommon finding; it manifests as high-attenuation areas along the basal cisterns, Sylvian fissure, tentorium cerebelli, or cortical sulci. These point out the importance of lesion location information for a more accurate diagnosis of ICH, particularly when the hemorrhage is difficult to detect due to its tiny volume or faint attenuation.

One of the challenges to increasing the accuracy of ICH detection algorithms is the lack of a large dataset with expert annotations, which would take a lot of effort and resources to produce.⁶ Thus, insufficient data or weak annotations have been employed in the majority of published research,^{3,7,8} In the weakly-supervised settings such as classification-based deep

learning using images with annotations that are relatively easy to obtain (presence vs. absence of ICH), saliency maps could not locate the exact location of lesions.⁹

In this study, we hypothesized that a deep-learning model performs better if it is trained on a large annotated dataset with slice-wise manual segmentation, compared to the one trained on a weakly annotated dataset. To improve performance and robustness across all ICH subtypes and sizes, we also designed a weighted ensemble model to integrate multiple models trained on distinct strongly annotated datasets reflecting ICH features to minimize the prediction errors of each individual model. In addition, six experts reviewed challenging ICH cases after external testing of the final model.

Methods

1. Model development with weakly and strongly annotated datasets

1.1. Datasets

Weakly annotated dataset We used open data from the Radiological Society of North America (RSNA) comprising 27,861 brain CT scans (3,528 hemorrhages). Per slice, the neuroradiologists labeled the presence/absence of a hemorrhage without spatial annotation.⁷

Strongly annotated dataset We used 53,045 brain CT scans (7,013 with and 46,032 without ICH) from the AI-Hub directed by the Korean National Information Society Agency (<https://aihub.or.kr/aidata/34101>). The AI-Hub dataset was collected from six Korean university hospitals in 2020 as part of a large-scale data collection initiative for cerebrovascular disease. Each hospital's neuroradiologist interpreted the CT scans, labeled the presence of hemorrhage per slice, and manually segmented the outline of the hemorrhage. A total of 7,013 CT scans with hemorrhages included 2,424 SAHs, 2,738 SDHs, 371 EDHs, 1,266 IVHs, 3,367 IPHs, and 2,833 mixed hemorrhage.

1.2. Training and validation dataset

To compare the performance of the deep-learning models trained on weakly and strongly annotated datasets, and to account for different data sizes, we randomly selected the same number of slices with and without hemorrhage ($n = 6,500$ each) from the RSNA and AI-Hub datasets. For CT scans with hemorrhage, the same number of slices as those of IPH, IVH, SDH, EDH, and SAH ($n = 1,300$) were included, i.e. the hemorrhagic types were balanced. All training dataset images were pre-processed into four-channel input data (eMethods and eFigure 1).

1.3. External test dataset

Two datasets were used for the external testing of the deep-learning models. The first dataset comprised 600 brain CT scans (327 with and 273 without ICH) from a tertiary hospital in Korea (Dongguk University Medical Center (DUMC)). The second was an open dataset (Qure.ai) comprising 386 CTs (160 with and 226 without ICH). A vascular neurologist and neuroradiologist performed consensus labeling for the DUMC dataset, and three neuroradiologists labeled the Qure.ai dataset using a majority vote. This study was approved by the institutional review board of DUMC and JLK Inc. (No. DUIH 2018-03-018 and 20220407-01).

1.4. Comparison of models trained with weakly and strongly annotated datasets

DenseNet121,¹⁰ InceptionResNetV2,¹¹ MobileNetV2,¹² and VGG19¹³ were used for model development to compare deep-learning models trained on datasets with weak and strong annotations. For models using weakly annotated datasets, we used slice-wise hemorrhage labeling for both the RSNA and AI-Hub datasets (eFigure 2A-B). For classification loss, we compared the slice-wise model output and ground-truth labeling. The same input image was fed into the deep-learning model trained on the strongly annotated AI-Hub dataset (eFigure 2C). We extracted the saliency map from the last convolutional layer of each deep-learning model, compared it to the ground-truth hemorrhage segmentation, and computed the segmentation loss in addition to the classification loss to train the hemorrhage location.

We tested each model on an external test dataset and calculated sensitivity, specificity, and AUC and the threshold of 0.5 as the model's performance. We used 500 bootstrap replications to calculate 95% confidence intervals. We used the DeLong test for AUC comparison.¹⁴

2. Ensemble model

2.1. Training and test dataset

Considering class imbalance, we randomly selected an equal number of brain CT scans with and without ICH from the AI-Hub dataset (6,963 each). We trained five U-net based segmentation models (eFigure 3): Lesion segmentation model using all training datasets (Model 1), lesion subtype pre-trained segmentation model using all training datasets (Model 2), SDH model (Model 3), SAH model (Model 4), and small lesion (≤ 5 mL) model (Model 5). A summary of each model and the training dataset is shown in Figure 1. The DUMC and Qure.ai datasets were used for external testing.

2.2. Ensemble base models

Five deep-learning models were trained using 2D U-net with the Inception module (eFigure 3).^{15,16} For the lesion subtype pre-trained segmentation model (Model 2), a pre-trained model in which down-sampling layers of U-net were pre-trained using hemorrhage subtype labeling was used. The Dice loss function, Adam optimizer, and a learning rate of $1e-4$ were used for model training.

To determine whether the ensemble of base models (Models 1–5) improved the performance of hemorrhage detection in SAH, SDH, and small hemorrhage cases, we combined the base models and evaluated the performance of each combination model. From the two external test datasets, we extracted SAH, SDH, and small hemorrhage cases (with the same number of normal CT scans) to test the combination model.

2.3. Weighted ensemble model

To ensemble the five base models using distinct datasets, their outputs needed to be assigned appropriate weight values according to the input data. Hence, we developed an additional

weight model using input data comprising five-channel segmentation results from five base models, ranging from zero to one. Using random initiative weight values, the model was trained to select the weight values that minimized the Dice loss between the predicted segmentation at the pixel probability threshold of 0.5 and ground-truth segmentation (eMethods and eFigure 4 and 5).

3. Review of “difficult” ICH cases

After external testing of the weighted ensemble model, we defined the difficult ICH cases for expert reviews. “Difficult-for-AI” cases were chosen from the DUMC dataset when a) the probability of lesion ≤ 0.6 among cases annotated as hemorrhage or b) the probability of lesion ≥ 0.4 among cases annotated as no hemorrhage. “Difficult-for-humans” cases were selected from the Qure.ai dataset if the three annotators had not unanimously agreed with the ground truth during the initial labeling process.

Six neurology experts with six to eighteen years of clinical experience re-annotated the presence vs. absence of ICH in these two types of difficult images. The sensitivity, specificity, and accuracy of the weighted ensemble model and each expert were calculated. The inter-rater agreement among the experts was also calculated. If the ground truth and the opinions of six experts did not concur, a consensus meeting was held to amend the ground truth with a majority (4 or higher) vote. After the consensus meeting, the sensitivity, specificity, and accuracy of the weighted ensemble model and those of the six experts were re-calculated. Cases where the weighted ensemble model made incorrect predictions were subjected to qualitative assessments.

Results

Baseline characteristics of the datasets

The AI-Hub dataset, in which the presence of each hemorrhage subtype was labeled per slice and the hemorrhage was manually segmented, had 53,045 cases (mean 57.5 years, 47.5% female), and 13.2% (7,013) had ICH. In the RSNA dataset with the presence of each ICH subtype being labeled per slice, 39.6% (7,449) of 18,938 cases had ICH. Table 1 shows the proportion of each ICH subtype in the AI-Hub and RSNA datasets. The baseline characteristics of the external datasets from DUMC (n = 600) and Qure.ai (n = 386) are presented in Table 1.

Comparison of models trained with weakly vs. strongly annotated datasets

A dataset with strong annotations (AI-Hub dataset with location information) and two datasets with weak annotations (AI-Hub dataset without location information and RSNA dataset) were utilized for the training of four well-known deep-learning networks: DenseNet121, InceptionResNetV2, MobileNetV2, and VGG19. We tested four trained models on a composite of the DUMC and Qure.ai datasets. When trained using the RSNA dataset, the accuracies of DenseNet121, InceptionResNetV2, MobileNetV2, and VGG19 was 0.771 (95% C.I. 0.767–0.775), 0.770 (95% C.I. 0.766–0.774), 0.649 (95% C.I. 0.645–0.653), and 0.708 (95% C.I. 0.704–0.712), respectively. When trained using the AI-Hub dataset without location information, the accuracies were 0.812 (95% C.I. 0.809–0.816), 0.810 (95% C.I. 0.807–0.814), 0.645 (95% C.I. 0.641–0.650), and 0.707 (95% C.I. 0.705–0.711), respectively. When trained using the AI-Hub dataset with location information, the accuracies of all deep-learning networks except for DenseNet121 improved significantly, with the values being 0.756 (95% C.I. 0.753–0.761), 0.818 (95% C.I. 0.812–0.820), 0.658 (95% C.I. 0.655–0.664), and 0.862 (95% C.I. 0.859–0.865), respectively (Table 2 and eFigure 6).

Development of a weighted ensemble model

To improve the detection of SDH, SAH, and tiny lesions, which is regarded to be challenging, we designed a weighted ensemble model using multiple distinct datasets that not only had strong annotations but also reflected the various ICH features.

We first generated “Ensemble basic”, a weighted ensemble of two models, a lesion segmentation model (Model 1) and a subtype classification/segmentation model (Model 2), which were trained using CT images encompassing all ICH subtypes. Overall ICH detection accuracy measured by using a composite dataset of DUMC and Qure.ai, was 0.938 (95% C.I. 0.922–0.953, Table 3). The accuracy for SDH and SAH cases was respectively 0.893 (95% C.I. 0.865–0.919) and 0.944 (95% C.I. 0.942–0.945). Next, to further increase the accuracy in the diagnosis of SDH, SAH, and small lesions, we additionally developed models 3, 4, and 5, which were respectively trained on only SDH cases, only SAH cases, and only small lesions ≤ 5 mL. We then investigated whether the model performance was improved by combining Models 3, 4, or 5 with the weighted ensemble models for Models 1 and 2. “Ensemble SDH” showed an increased accuracy for SDH compared to “Ensemble basic”, from 0.893 (95% C.I. 0.865–0.919) to 0.927 (95% C.I. 0.903–0.948; P for AUC difference = 0.0002). For SAH, “Ensemble SAH” showed a comparable accuracy to “Ensemble basic” (0.952 [95% C.I. 0.951–0.954] vs. 0.944 [95% C.I. 0.942–0.945], P for AUC difference = 0.2439). “Ensemble small lesions” showed a comparable accuracy for total ICH compared to “Ensemble basic” (0.946 [95% C.I. 0.931–0.960] vs. 0.938 [95% C.I. 0.922–0.953], P for AUC difference = 0.1180). Finally, we developed a final weighted ensemble model that ensembles all models 1 to 5 and showed a significantly higher accuracy for total ICHs (0.951 [95% C.I. 0.937–0.964], P for AUC difference = 0.0379), compared to “Ensemble basic”.

Comparison of the final weighted ensemble model with AI models that were previously built

We compared the performance of the final weighted ensemble model with that of DenseNet121, InceptionResNetV2, MobileNetV2, and VGG19 by using a test dataset that combines the DUMC and Qure.ai datasets. The weighted ensemble model significantly outperformed the other models in terms of sensitivity, specificity, and AUC (Figure 2 and eTable 1). Additional tests using either the DUMC or Qure.ai dataset showed similar results (eFigure 7).

Review of “difficult” 161 ICH cases

A total of 91 cases from the DUMC dataset were selected as difficult-for-AI cases: 17 cases from those classified as ICH that AI identified with a lesion probability of ≤ 0.6 and 74 cases from those classified as normal that AI identified with a lesion probability of ≥ 0.4 in external testing. A total of 70 cases from the Qure.ai datasets were selected as difficult-for-humans based on the three annotators' disagreement. Six experts re-annotated these 161 cases for the presence vs. absence of ICH; there was complete agreement among the six experts for 81 cases, whereas there was at least one disagreement for 80 cases. For all 161 cases, the final weighted ensemble model showed an accuracy of 0.441, sensitivity of 0.431, and specificity of 0.462. The accuracies of the six experts were 0.671, 0.764, 0.708, 0.640, 0.683, and 0.714, with the interrater agreement (the Fleiss' kappa value) being 0.536.

Among 95 cases where one or more experts disagreed with the initial ground truth, 80 were unanimous, five were disagreed upon by one expert, seven were disagreed upon by two experts, and three were disagreed upon by three experts. In 48 cases, the ground truth was changed following discussion and majority voting in the consensus meeting. The accuracy, sensitivity, and specificity of the weighted ensemble model for the revised ground truth was

respectively 0.491, 0.457, and 0.525, without showing significant differences when compared with those for the original ground truth. eTable 2 shows the qualitative assessments of the cases where the weighted ensemble model predicted incorrectly.

Discussion

This study demonstrated that a) a deep-learning algorithm for detecting ICH trained with a strongly annotated dataset outperformed models trained with a weakly annotated dataset, and b) a weighted ensemble model that integrated separate models trained using SDH, SAH, or small-lesion (≤ 5 mL) ICH datasets achieved a higher AUC than four previous deep-learning models on external testing.

Medical image segmentation requires a lot of labor and resources.¹⁷ Although many transfer learning methods for weakly or partially annotated data have been developed,¹⁸ there is still a need for large-scale annotated data. To the best of our knowledge, no deep-learning algorithm for ICH detection has been developed using large-scale CT scans with segmentation annotation. We found that the accuracy of three of the four previously reported deep-learning models improved after training with strongly annotated datasets, compared to weakly annotated datasets.

Compared with magnetic resonance imaging, CT is less expensive, faster, and better at detecting ICH. However, some ICHs are more likely to be misdiagnosed due to their varied location and shape, small lesion size, and similar attenuation to adjacent tissue.¹⁹ For example, SDH and SAH are often difficult to distinguish from adjacent tissue,⁴ despite their distinct locations and shapes. Although there are classification methods for small lesions,²⁰ a new deep learning strategy based on more comprehensive feature data may improve ICH detection performance. We achieved higher ICH detection accuracies using ensemble models that combined multiple separate models trained with datasets specialized for SDH, SAH, or small lesions. Employing the ensemble models, we also observed an increasing trend in specificity.

Despite recent development of many deep-learning algorithms for imaging diagnosis of ICH, their clinical application has yet to be accomplished. In addition to technical challenges such as the domain shift problem and the shortcut problem, there are also

instances where determining the ground truth is difficult or inter-physician agreement is limited.²¹⁻²³ After our expert meeting, as high as 30% (48/161) of difficult ICH cases, which however represented only 4.9% of the total test dataset cases (n = 986), required re-labeling of their ground truth. This may explain why it is challenging for AI to learn medical images that are difficult for experienced clinicians. Moreover, the re-labeling did not improve the accuracy of ICH detection by our weighted ensemble model trained with a total of 13,926 strongly annotated CT data. Further studies are required to investigate if fine-tuning a model after training with a larger high-quality training dataset, where difficult data are augmented, could increase the robustness, generalization, and discriminative power of the deep-learning algorithm.

Noisy labels have a negligible effect on the model performance in large datasets. In a handwritten number dataset, increasing the accuracy of random labels by only 1% significantly improved the model performance.^{24,25} However, if difficult cases are mixed at a low frequency across a dataset, noisy labels may affect ICH detection. There is a high demand in the medical profession for a model that can accurately diagnose both easy and complex cases. False positives and negatives may result in unnecessary and missed therapy, respectively. Future research should investigate if a) using ensemble models and b) expert review of difficult cases and adding them to a training dataset could overcome these challenges.

Our study has limitations. First, except for age and sex, no clinical information was available. Second, the classification and segmentation of some training data may not be accurate, because of the inclusion of difficult cases. Third, the proportions of mixed hemorrhages were high in the RSNA and AI-Hub datasets.

In conclusion, we developed a weighted ensemble model for ICH detection by training with strongly annotated CT scans obtained from multiple centers. Although

challenging cases existed, external testing with a dataset from different ethnic origins demonstrated excellent performance of our model. We also showed that a better understanding and management of cases that are challenging for AI and humans is required to facilitate clinical use of ICH detection algorithms.

Acknowledgment

This research was supported by the National Research Foundation of Korea and funded by the Ministry of Science and ICT (Grant NRF-2020M3E5D9079768). Wi-Sun Ryu had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Wi-Sun Ryu and Gi-Hun Park are employed by JLK Inc. Dong-Eog Kim is a stockholder of JLK Inc.

References

1. Wardlaw JM, Mair G, von Kummer R, et al. Accuracy of Automated Computer-Aided Diagnosis for Stroke Imaging: A Critical Evaluation of Current Evidence. *Stroke*. Jul 2022;53(7):2393-2403. doi:10.1161/STROKEAHA.121.036204
2. Vermeulen MJ, Schull MJ. Missed diagnosis of subarachnoid hemorrhage in the emergency department. *Stroke*. Apr 2007;38(4):1216-21. doi:10.1161/01.STR.0000259661.05525.9a
3. Kuo W, Hone C, Mukherjee P, Malik J, Yuh EL. Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. *Proc Natl Acad Sci U S A*. Nov 5 2019;116(45):22737-22745. doi:10.1073/pnas.1908021116
4. Seyam M, Weikert T, Sauter A, Brehm A, Psychogios MN, Blackham KA. Utilization of Artificial Intelligence-based Intracranial Hemorrhage Detection on Emergent Noncontrast CT Images in Clinical Workflow. *Radiol Artif Intell*. Mar 2022;4(2):e210168. doi:10.1148/ryai.210168
5. Tetsuka S, Matsumoto E. Diagnosis of a subarachnoid hemorrhage with only mild symptoms using computed tomography in Japan. *BMC Neurol*. Oct 18 2016;16(1):196. doi:10.1186/s12883-016-0726-9
6. Malhotra P, Gupta S, Koundal D, Zaguia A, Enbeyle W. Deep Neural Networks for Medical Image Segmentation. *J Healthc Eng*. 2022;2022:9580991. doi:10.1155/2022/9580991
7. Flanders AE, Prevedello LM, Shih G, et al. Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. *Radiol Artif Intell*. May 2020;2(3):e190211. doi:10.1148/ryai.2020190211
8. Majumdar A, Brattain L, Telfer B, Farris C, Scalera J. Detecting Intracranial Hemorrhage with Deep Learning. *Annu Int Conf IEEE Eng Med Biol Soc*. Jul

- 2018;2018:583-587. doi:10.1109/EMBC.2018.8512336
9. Arun N, Gaw N, Singh P, et al. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiol Artif Intell.* Nov 2021;3(6):e200267. doi:10.1148/ryai.2021200267
 10. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. 2016:arXiv:1608.06993. doi:10.48550/arXiv.1608.06993 Accessed August 01, 2016. <https://ui.adsabs.harvard.edu/abs/2016arXiv160806993H>
 11. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. 2016:arXiv:1602.07261. doi:10.48550/arXiv.1602.07261 Accessed February 01, 2016. <https://ui.adsabs.harvard.edu/abs/2016arXiv160207261S>
 12. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018:arXiv:1801.04381. doi:10.48550/arXiv.1801.04381 Accessed January 01, 2018. <https://ui.adsabs.harvard.edu/abs/2018arXiv180104381S>
 13. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014:arXiv:1409.1556. doi:10.48550/arXiv.1409.1556 Accessed September 01, 2014. <https://ui.adsabs.harvard.edu/abs/2014arXiv1409.1556S>
 14. Demler OV, Pencina MJ, D'Agostino RB, Sr. Misuse of DeLong test to compare AUCs for nested models. *Stat Med.* Oct 15 2012;31(23):2577-87. doi:10.1002/sim.5328
 15. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.* 2015:234-241:chap Chapter 28. *Lecture Notes in Computer Science.*
 16. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. presented at: 2016 IEEE Conference on Computer

- Vision and Pattern Recognition (CVPR); 2016;
17. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Med Image Anal.* Jul 2020;63:101693. doi:10.1016/j.media.2020.101693
 18. Ardalan Z, Subbian V. Transfer Learning Approaches for Neuroimaging Analysis: A Scoping Review. *Front Artif Intell.* 2022;5:780405. doi:10.3389/frai.2022.780405
 19. Wang X, Shen T, Yang S, et al. A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head CT scans. *Neuroimage Clin.* 2021;32:102785. doi:10.1016/j.nicl.2021.102785
 20. Dubost F, Adams H, Yilmaz P, et al. Weakly supervised object detection with 2D and 3D regression neural networks. *Med Image Anal.* Oct 2020;65:101767. doi:10.1016/j.media.2020.101767
 21. Geirhos R, Jacobsen J-H, Michaelis C, et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence.* 2020;2(11):665-673. doi:10.1038/s42256-020-00257-z
 22. DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence.* 2021;3(7):610-619. doi:10.1038/s42256-021-00338-7
 23. Guan H, Liu M. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Trans Biomed Eng.* Mar 2022;69(3):1173-1185. doi:10.1109/TBME.2021.3117407
 24. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis.* 2020;65doi:10.1016/j.media.2020.101759
 25. Rolnick D, Veit A, Belongie S, Shavit N. Deep Learning is Robust to Massive Label Noise. 2017:arXiv:1705.10694. doi:10.48550/arXiv.1705.10694 Accessed May 01, 2017. <https://ui.adsabs.harvard.edu/abs/2017arXiv170510694R>

Figure Legends

Figure 1. Summary of each deep-learning model and the training dataset used. EDH, epidural hemorrhage; IPH, intraparenchymal hemorrhage; IVH, intraventricular hemorrhage; SAH, subarachnoid hemorrhage; SDH, subdural hemorrhage.

Figure 2. Receiver Operating Characteristic (ROC) curves representing the performance of the deep-learning models. DenseNet121, InceptionResNetV2, MobileNetV2, VGG19 trained on strongly annotated datasets and the final weighted ensemble model were applied to a test dataset combining the DUMC and Qure.ai datasets.

Tables

| | Training dataset | | Test dataset | |
|--------------------------|----------------------------|----------------------------|--------------------------|--------------------------|
| | AI-Hub | RSNA | DUMC | Qure.ai |
| Country | South Korea | Unites States | South Korea | India |
| Number of cases | 53,045 | 18,938 | 600 | 386 |
| Age (mean ± S.D.) | 57.5 ± 19.9 | N/A | 65.7 ± 14.9 | N/A |
| Female Sex | 25,185 (47.5%) | N/A | 213 (35.5%) | N/A |
| Normal | 46,032 (86.8%) | 11,439 (60.4%) | 273 (45.5%) | 226 (58.5%) |
| ICH | 7,013 (13.2%) | 7,499 (39.6%) | 327 (54.5%) | 160 (41.5%) |
| IPH | 1,744 (24.9%) [†] | 1,008 (13.4%) [†] | 221 (67.6%) [†] | 45 (28.1%) [†] |
| IVH | 135 (1.9%) [†] | 239 (3.2%) [†] | 8 (2.4%) [†] | 0 (0%) [†] |
| EDH | 159 (2.3%) [†] | 73 (1.0%) [†] | 1 (0.3%) [†] | 0 (0%) [†] |
| SDH | 1,442 (20.6%) [†] | 1,298 (17.3%) [†] | 15 (4.6%) [†] | 0 (0%) [†] |
| SAH | 700 (10.0%) [†] | 456 (6.1%) [†] | 82 (25.1%) [†] | 0 (0%) [†] |
| Mixed | 2,833 (40.4%) [†] | 4,425 (59.0%) [†] | 0 (0%) [†] | 115 (71.9%) [†] |

Table 1. Baseline characteristics of the datasets. [†]The percentages indicate the proportion of each subtype to the total number of lesions. RSNA, Radiological Society of North America; DUMC, Dongguk University Medical Center; ICH, intracranial hemorrhage; IPH, intraparenchymal hemorrhage; IVH, intraventricular hemorrhage; EDH, epidural hemorrhage; SDH, subdural hemorrhage; SAH, subarachnoid hemorrhage.

| 95% C.I. | RSNA dataset | | | AI-Hub without location dataset | | | AI-Hub with location dataset | | |
|--------------------------|--------------------------|--------------------------|--------------------------|---------------------------------|--------------------------|--------------------------|------------------------------|--------------------------|--------------------------|
| | AUC | Sensitivity | Specificity | AUC | Sensitivity | Specificity | AUC | Sensitivity | Specificity |
| DenseNet121 | 0.771 (0.767 - 0.775) | 0.616 (0.611 - 0.620) | 0.754 (0.745 - 0.765) | 0.812 (0.809 - 0.816) | 0.554 (0.548 - 0.558) | 0.887 (0.880 - 0.895) | 0.756 (0.753 - 0.761) | 0.693 (0.688 - 0.698) | 0.657 (0.647 - 0.668) |
| P for AUC difference | Reference | | | $P = 0.0003$ (*0.0039) | | | $P < 0.0001$ (*0.0030) | | |
| InceptionResNetV2 | 0.770 (0.766 - 0.774) | 0.623 (0.618 - 0.628) | 0.724 (0.716 - 0.736) | 0.810 (0.807 - 0.814) | 0.658 (0.654 - 0.663) | 0.825 (0.817 - 0.834) | 0.818 (0.812 - 0.820) | 0.745 (0.741 - 0.750) | 0.725 (0.715 - 0.735) |
| P for AUC difference | Reference | | | $P < 0.0001$ (*0.0031) | | | $P < 0.0001$ (*0.0031) | | |
| MobileNetV2 | 0.649 (0.645 - 0.653) | 0.599 (0.595 - 0.605) | 0.600 (0.589 - 0.611) | 0.645 (0.641 - 0.65) | 0.599 (0.595 - 0.605) | 0.608 (0.598 - 0.620) | 0.658 (0.655 - 0.664) | 0.708 (0.704 - 0.713) | 0.499 (0.488 - 0.510) |
| P for AUC difference | Reference | | | $P = 0.4221$ (*0.0047) | | | $P = 0.0227$ (*0.0040) | | |
| VGG19 | 0.708 (0.704 - 0.712) | 0.569 (0.564 - 0.574) | 0.754 (0.745 - 0.764) | 0.707 (0.705 - 0.711) | 0.460 (0.455 - 0.465) | 0.876 (0.869 - 0.884) | 0.862 (0.859 - 0.865) | 0.816 (0.812 - 0.820) | 0.731 (0.721 - 0.741) |
| P for AUC difference | Reference | | | $P = 0.6526$ (*0.0032) | | | $P < 0.0001$ (*0.0029) | | |

Table 2. Comparison of models trained with weakly and strongly annotated datasets. Area under the curve (AUC), sensitivity, and specificity of four deep-learning networks trained on RSNA dataset, AI-Hub dataset without location information, and AI-Hub dataset with location information were shown. The AUC of each of the four deep-learning networks trained on the RSNA dataset was set as a reference, and the AUCs of the remaining models were compared using the DeLong test. C.I., confidence interval. *Standard error.

| Combination of the models | All (492 ICH, 494 normal) | | | SDH (43 SDH, 494 normal) | | | SAH (117 SAH, 494 normal) | | |
|---|---------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|---------------------------|--------------------------|--------------------------|
| | AUC | Sensitivity | Specificity | AUC | Sensitivity | Specificity | AUC | Sensitivity | Specificity |
| Ensemble basic 1 + 2 | 0.938 (0.922 - 0.953) | 0.982 (0.966 - 0.992) | 0.702 (0.662 - 0.744) | 0.893 (0.865 - 0.919) | 0.977 (0.877 - 0.999) | 0.702 (0.662 - 0.744) | 0.944 (0.942 - 0.945) | 0.992 (0.953 - 1.000) | 0.702 (0.662 - 0.744) |
| P for AUC difference | Reference | | | Reference | | | Reference | | |
| Ensemble SDH 1 + 2 + 3 | 0.936 (0.920 - 0.951) | 0.957 (0.935 - 0.973) | 0.781 (0.757 - 0.830) | 0.927 (0.903 - 0.948) | 0.977 (0.877 - 0.999) | 0.781 (0.757 - 0.830) | 0.941 (0.940 - 0.943) | 0.992 (0.953 - 1.000) | 0.781 (0.757 - 0.830) |
| P for AUC difference | $P = 0.7612$ (*0.0060) | | | $P = 0.0002$ (*0.0092) | | | $P = 0.6537$ (*0.0071) | | |
| Ensemble SAH 1 + 2 + 4 | 0.944 (0.928 - 0.958) | 0.963 (0.943 - 0.978) | 0.777 (0.753 - 0.826) | 0.868 (0.837 - 0.896) | 0.860 (0.721 - 0.947) | 0.777 (0.753 - 0.826) | 0.952 (0.951 - 0.954) | 0.983 (0.940 - 0.998) | 0.777 (0.753 - 0.826) |
| P for AUC difference | $P = 0.3337$ (*0.0057) | | | $P = 0.1695$ (*0.0186) | | | $P = 0.2439$ (*0.0069) | | |
| Ensemble small lesions 1 + 2 + 5 | 0.946 (0.931 - 0.960) | 0.970 (0.950 - 0.983) | 0.775 (0.738 - 0.813) | 0.874 (0.843 - 0.901) | 0.860 (0.721 - 0.947) | 0.775 (0.738 - 0.813) | 0.950 (0.948 - 0.951) | 0.983 (0.940 - 0.998) | 0.775 (0.738 - 0.813) |
| P for AUC difference | $P = 0.1180$ (*0.0053) | | | $P = 0.2023$ (*0.0155) | | | $P = 0.3621$ (*0.0063) | | |
| Ensemble all 1 + 2 + 3 + 4 + 5 | 0.951 (0.937 - 0.964) | 0.943 (0.919 - 0.962) | 0.826 (0.794 - 0.862) | 0.893 (0.864 - 0.918) | 0.884 (0.749 - 0.961) | 0.826 (0.794 - 0.862) | 0.958 (0.958 - 0.960) | 0.974 (0.927 - 0.995) | 0.826 (0.794 - 0.862) |
| P for AUC difference | $P = 0.0379$ (*0.0064) | | | $P = 0.9845$ (*0.0194) | | | $P = 0.0644$ (*0.0077) | | |

Table 3. The structures of five weighted ensemble models and accuracies for all cases, SDH cases, and SAH cases. The area under the curves (AUCs) of “Ensemble basic” for total ICH, SDH, and SAH were set as references, and the AUCs of the remaining models were compared using the DeLong test. ICH, intracranial hemorrhage; SDH, subdural hemorrhage; SAH, subarachnoid hemorrhage. *Standard error.



