JAMA Network Open

Original Investigation | Health Informatics

Deep Learning–Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model

Allison Park, BA; Chris Chute, BS; Pranav Rajpurkar, MS; Joe Lou; Robyn L. Ball, PhD; Katie Shpanskaya, BS; Rashad Jabarkheel, BS; Lily H. Kim, BS; Emily McKenna, BS; Joe Tseng, MD; Jason Ni, MD; Fidaa Wishah, MD; Fred Wittber, MD; David S. Hong, MD; Thomas J. Wilson, MD; Safwan Halabi, MD; Sanjay Basu, MD, PhD; Bhavik N. Patel, MD, MBA; Matthew P. Lungren, MD, MPH; Andrew Y. Ng, PhD; Kristen W. Yeom, MD

Abstract

IMPORTANCE Deep learning has the potential to augment clinician performance in medical imaging interpretation and reduce time to diagnosis through automated segmentation. Few studies to date have explored this topic.

OBJECTIVE To develop and apply a neural network segmentation model (the HeadXNet model) capable of generating precise voxel-by-voxel predictions of intracranial aneurysms on head computed tomographic angiography (CTA) imaging to augment clinicians' intracranial aneurysm diagnostic performance.

DESIGN, SETTING, AND PARTICIPANTS In this diagnostic study, a 3-dimensional convolutional neural network architecture was developed using a training set of 611 head CTA examinations to generate aneurysm segmentations. Segmentation outputs from this support model on a test set of 115 examinations were provided to clinicians. Between August 13, 2018, and October 4, 2018, 8 clinicians diagnosed the presence of aneurysm on the test set, both with and without model augmentation, in a crossover design using randomized order and a 14-day washout period. Head and neck examinations performed between January 3, 2003, and May 31, 2017, at a single academic medical center were used to train, validate, and test the model. Examinations positive for aneurysm had at least 1 clinically significant, nonruptured intracranial aneurysm. Examinations with hemorrhage, ruptured aneurysm, posttraumatic or infectious pseudoaneurysm, arteriovenous malformation, surgical clips, coils, catheters, or other surgical hardware were excluded. All other CTA examinations were considered controls.

MAIN OUTCOMES AND MEASURES Sensitivity, specificity, accuracy, time, and interrater agreement were measured. Metrics for clinician performance with and without model augmentation were compared.

RESULTS The data set contained 818 examinations from 662 unique patients with 328 CTA examinations (40.1%) containing at least 1 intracranial aneurysm and 490 examinations (59.9%) without intracranial aneurysms. The 8 clinicians reading the test set ranged in experience from 2 to 12 years. Augmenting clinicians with artificial intelligence–produced segmentation predictions resulted in clinicians achieving statistically significant improvements in sensitivity, accuracy, and interrater agreement when compared with no augmentation. The clinicians' mean sensitivity increased by 0.059 (95% CI, 0.028-0.091; adjusted P = .01), mean accuracy increased by 0.038 (95% CI, 0.014-0.062; adjusted P = .02), and mean interrater agreement (Fleiss κ) increased by 0.060, from 0.799 to 0.859 (adjusted P = .05). There was no statistically significant change in mean

(continued)

Key Points

Question How does augmentation with a deep learning segmentation model influence the performance of clinicians in identifying intracranial aneurysms from computed tomographic angiography examinations?

Findings In this diagnostic study of intracranial aneurysms, a test set of 115 examinations was reviewed once with model augmentation and once without in a randomized order by 8 clinicians. The clinicians showed significant increases in sensitivity, accuracy, and interrater agreement when augmented with neural network model–generated segmentations.

Meaning This study suggests that the performance of clinicians in the detection of intracranial aneurysms can be improved by augmentation using deep learning segmentation models.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

specificity (0.016; 95% CI, −0.010 to 0.041; adjusted *P* = .16) and time to diagnosis (5.71 seconds; 95% CI, 7.22-18.63 seconds; adjusted *P* = .19).

CONCLUSIONS AND RELEVANCE The deep learning model developed successfully detected clinically significant intracranial aneurysms on CTA. This suggests that integration of an artificial intelligence–assisted diagnostic model may augment clinician performance with dependable and accurate predictions and thereby optimize patient care.

JAMA Network Open. 2019;2(6):e195600. doi:10.1001/jamanetworkopen.2019.5600

Introduction

Diagnosis of unruptured aneurysms is a critically important clinical task: intracranial aneurysms occur in 1% to 3% of the population and account for more than 80% of nontraumatic life-threatening subarachnoid hemorrhages.[1] Computed tomographic angiography (CTA) is the primary, minimally invasive imaging modality currently used for diagnosis, surveillance, and presurgical planning of intracranial aneurysms,[2,3] but interpretation is time consuming even for subspecialty-trained neuroradiologists. Low interrater agreement poses an additional challenge for reliable diagnosis.[4-7]

Deep learning has recently shown significant potential in accurately performing diagnostic tasks on medical imaging.[8] Specifically, convolutional neural networks (CNNs) have demonstrated excellent performance on a range of visual tasks, including medical image analysis.[9] Moreover, the ability of deep learning systems to augment clinician workflow remains relatively unexplored.[10] The development of an accurate deep learning model to help clinicians reliably identify clinically significant aneurysms in CTA has the potential to provide radiologists, neurosurgeons, and other clinicians an easily accessible and immediately applicable diagnostic support tool.

In this study, a deep learning model to automatically detect intracranial aneurysms on CTA and produce segmentations specifying regions of interest was developed to assist clinicians in the interpretation of CTA examinations for the diagnosis of intracranial aneurysms. Sensitivity, specificity, accuracy, time to diagnosis, and interrater agreement for clinicians with and without model augmentation were compared.

Methods

The Stanford University institutional review board approved this study. Owing to the retrospective nature of the study, patient consent or assent was waived. The Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline was used for the reporting of this study.

Data

A total of 9455 consecutive CTA examination reports of the head or head and neck performed between January 3, 2003, and May 31, 2017, at Stanford University Medical Center were retrospectively reviewed. Examinations with parenchymal hemorrhage, subarachnoid hemorrhage, posttraumatic or infectious pseudoaneurysm, arteriovenous malformation, ischemic stroke, nonspecific or chronic vascular findings such as intracranial atherosclerosis or other vasculopathies, surgical clips, coils, catheters, or other surgical hardware were excluded. Examinations of injuries that resulted from trauma or contained images degraded by motion were also excluded on visual review by a board-certified neuroradiologist with 12 years of experience. Examinations with nonruptured clinically significant aneurysms (>3 mm) were included.[11]

Radiologist Annotations

The reference standard for all examinations in the test set was determined by a board-certified neuroradiologist at a large academic practice with 12 years of experience who determined the presence of aneurysm by review of the original radiology report, double review of the CTA examination, and further confirmation of the aneurysm by diagnostic cerebral angiograms, if available. The neuroradiologist had access to all of the Digital Imaging and Communications in Medicine (DICOM) series, original reports, and clinical histories, as well as previous and follow-up examinations during interpretation to establish the best possible reference standard for the labels. For each of the aneurysm examinations, the radiologist also identified the location of each of the aneurysms. Using the open-source annotation software ITK-SNAP,[12] the identified aneurysms were manually segmented on each slice.

Model Development

In this study, we developed a 3-dimensional (3-D) CNN called HeadXNet for segmentation of intracranial aneurysms from CT scans. Neural networks are functions with parameters structured as a sequence of layers to learn different levels of abstraction. Convolutional neural networks are a type of neural network designed to process image data, and 3-D CNNs are particularly well suited to handle sequences of images, or volumes.

HeadXNet is a CNN with an encoder-decoder structure (eFigure 1 in the Supplement), where the encoder maps a volume to an abstract low-resolution encoding, and the decoder expands this encoding to a full-resolution segmentation volume. The segmentation volume is of the same size as the corresponding study and specifies the probability of aneurysm for each voxel, which is the atomic unit of a 3-D volume, analogous to a pixel in a 2-D image. The encoder is adapted from a 50-layer SE-ResNeXt network,[13-15] and the decoder is a sequence of 3 × 3 transposed convolutions. Similar to UNet,[16] skip connections are used in 3 layers of the encoder to transmit outputs directly to the decoder. The encoder was pretrained on the Kinetics-600 data set,[17] a large collection of YouTube videos labeled with human actions; after pretraining the encoder, the final 3 convolutional blocks and the 600-way softmax output layer were removed. In their place, an atrous spatial pyramid pooling[18] layer and the decoder were added.

Training Procedure

Subvolumes of 16 slices were randomly sampled from volumes during training. The data set was preprocessed to find contours of the skull, and each volume was cropped around the skull in the axial plane before resizing each slice to 208 × 208 pixels. The slices were then cropped to 192 × 192 pixels (using random crops during training and centered crops during testing), resulting in a final input of size 16 × 192 × 192 per example; the same transformations were applied to the segmentation label. The segmentation output was trained to optimize a weighted combination of the voxelwise binary cross-entropy and Dice losses.[19]

Before reaching the model, inputs were clipped to [−300, 700] Hounsfield units, normalized to [−1, 1], and zero-centered. The model was trained on 3 Titan Xp graphical processing units (GPUs) (NVIDIA) using a minibatch of 2 examples per GPU. The parameters of the model were optimized using a stochastic gradient descent optimizer with momentum of 0.9 and a peak learning rate of 0.1 for randomly initialized weights and 0.01 for pretrained weights. The learning rate was scheduled with a linear warm-up from 0 to the peak learning rate for 10 000 iterations, followed by cosine annealing[20] over 300 000 iterations. Additionally, the learning rate was fixed at 0 for the first 10 000 iterations for the pretrained encoder. For regularization, L2 weight decay of 0.001 was added to the loss for all trainable parameters and stochastic depth dropout[21] was used in the encoder blocks. Standard dropout was not used.

To control for class imbalance, 3 methods were used. First, an auxiliary loss was added after the encoder and focal loss was used to encourage larger parameter updates on misclassified positive examples. Second, abnormal training examples were sampled more frequently than normal

examples such that abnormal examples made up 30% of training iterations. Third, parameters of the decoder were not updated on training iterations where the segmentation label consisted of purely background (normal) voxels.

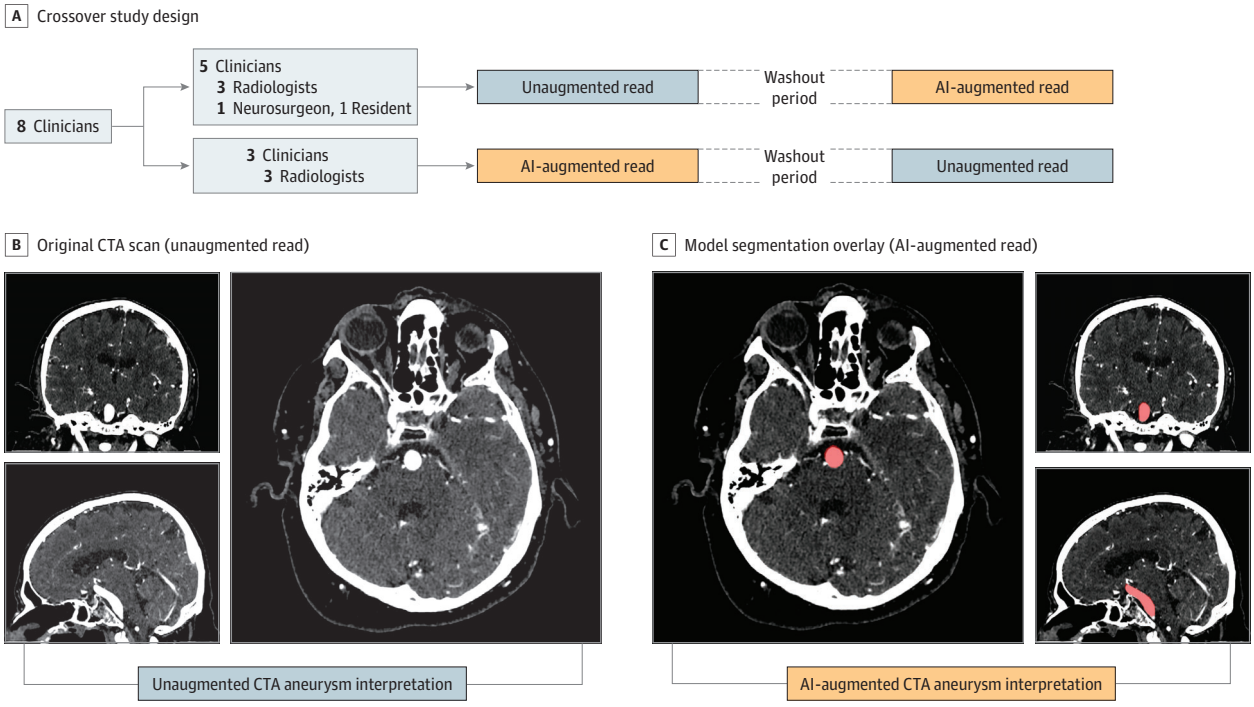
To produce a segmentation prediction for the entire volume, the segmentation outputs for sequential 16-slice subvolumes were simply concatenated. If the number of slices was not divisible by 16, the last input volume was padded with 0s and the corresponding output volume was truncated back to the original size.

Study Design

We performed a diagnostic accuracy study comparing performance metrics of clinicians with and without model augmentation. Each of the 8 clinicians participating in the study diagnosed a test set of 115 examinations, once with and once without assistance of the model. The clinicians were blinded to the original reports, clinical histories, and follow-up imaging examinations. Using a crossover design, the clinicians were randomly and equally divided into 2 groups. Within each group, examinations were sorted in a fixed random order for half of the group and sorted in reverse order for the other half. Group 1 first read the examinations without model augmentation, and group 2 first read the examinations with model augmentation. After a washout period of 14 days, the augmentation arrangement was reversed such that group 1 performed reads with model augmentation and group 2 read the examinations without model augmentation (**Figure 1**A).

Clinicians were instructed to assign a binary label for the presence or absence of at least 1 clinically significant aneurysm, defined as having a diameter greater than 3 mm. Clinicians read alone in a diagnostic reading room, all using the same high-definition monitor (3840 × 2160 pixels) displaying CTA examinations on a standard open-source DICOM viewer (Horos).[22] Clinicians entered

Figure 1. Study Design



A, Crossover study design. Clinicians were divided into 2 groups to perform reads with and without model augmentation in random order, with a 2-week washout period between. B, Unaugmented read, with original CTA scan in axial, coronal, and sagittal view. C, Augmented read, with model segmentation overlay on CTA in axial, coronal, and sagittal view. Readers had the option to toggle overlays off and view the scan as shown in B. AI indicates artificial intelligence; CTA, computed tomographic angiography.

their labels into a data entry software application that automatically logged the time difference between labeling of the previous examination and the current examination.

When reading with model augmentation, clinicians were provided the model's predictions in the form of region of interest (ROI) segmentations directly overlaid on top of CTA examinations. To ensure an image display interface that was familiar to all clinicians, the model's predictions were presented as ROIs in a standard DICOM viewing software. At every voxel where the model predicted a probability greater than 0.5, readers saw a semiopaque red overlay on the axial, sagittal, and coronal series (Figure 1C). Readers had access to the ROIs immediately on loading the examinations, and the ROIs could be toggled off to reveal the unaltered CTA images (Figure 1B). The red overlays were the only indication that was given whether a particular CTA examination had been predicted by the model to contain an aneurysm. Given these model results, readers had the option to take it into consideration or disregard it based on clinical judgment. When readers performed diagnoses without augmentation, no ROIs were present on any of the examinations. Otherwise, the diagnostic tools were identical for augmented and nonaugmented reads.

Statistical Analysis

On the binary task of determining whether an examination contained an aneurysm, sensitivity, specificity, and accuracy were used to assess the performance of clinicians with and without model augmentation. Sensitivity denotes the number of true-positive results over total aneurysm-positive cases, specificity denotes the number of true-negative results over total aneurysm-negative cases, and accuracy denotes the number of true-positive and true-negative results over all test cases. The microaverage of these statistics across all clinicians was also computed by measuring each statistic pertaining to the total number of true-positive, false-negative, and false-positive results. In addition, to convert the models' segmentation output of the model into a binary prediction, a prediction was considered positive if the model predicted at least 1 voxel as belonging to an aneurysm and negative otherwise. The 95% Wilson score confidence intervals were used to assess the variability in the estimates for sensitivity, specificity, and accuracy.[23]

To assess whether the clinicians achieved significant increases in performance with model augmentation, a 1-tailed *t* test was performed on the differences in sensitivity, specificity, and accuracy across all 8 clinicians. To determine the robustness of the findings and whether results were due to inclusion of the resident radiologist and neurosurgeon, we performed a sensitivity analysis: we computed the *t* test on the differences in sensitivity, specificity, and accuracy across board-certified radiologists only.

The average time to diagnosis for the clinicians with and without augmentation was computed as the difference between the mean entry times into the spreadsheet of consecutive diagnoses; 95% *t* score confidence intervals were used to assess the variability in the estimates. To account for interruptions in the clinical read or time logging errors, the 5 longest and 5 shortest time to diagnosis for each clinician in each reading were excluded. To assess whether model augmentation significantly decreased the time to diagnosis, a 1-tailed *t* test was performed on the difference in average time with and without augmentation across all 8 clinicians.

The interrater agreement of clinicians and for the radiologist subset was computed using the exact Fleiss κ.[24] To assess whether model augmentation increased interrater agreement, a 1-tailed permutation test was performed on the difference between the interrater agreement of clinicians on the test set with and without augmentation. The permutation procedure consisted of randomly swapping clinician annotations with and without augmentation so that a random subset of the test set that had previously been labeled as *read with augmentation* was now labeled as being *read without augmentation*, and vice versa; the exact Fleiss κ values (and the difference) were computed on the test set with permuted labels. This permutation procedure was repeated 10 000 times to generate the null distribution of the Fleiss κ difference (the interrater agreement of clinician annotations with augmentation is not higher than without augmentation) and the unadjusted *P* value

calculated as the proportion of Fleiss κ differences that were higher than the observed Fleiss κ difference.

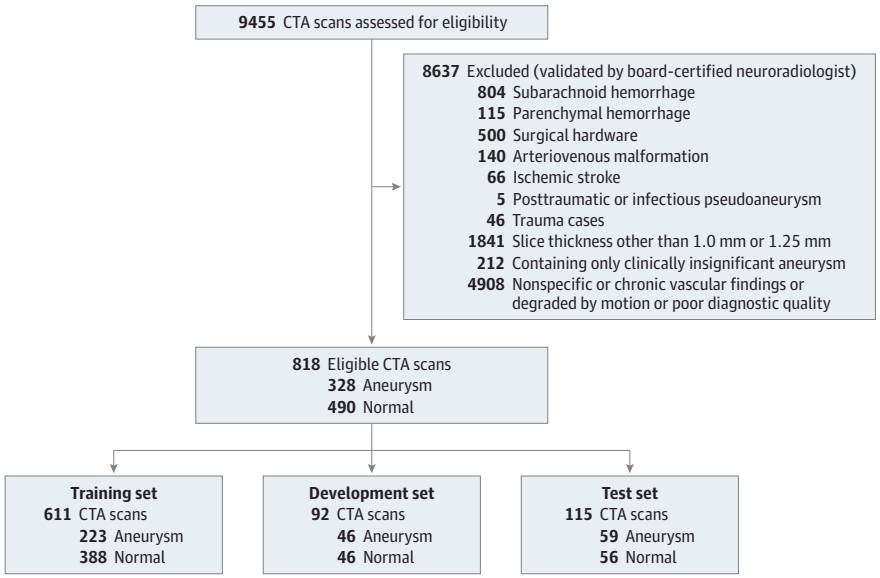
To control the familywise error rate, the Benjamini-Hochberg correction was applied to account for multiple hypothesis testing; a Benjamini-Hochberg–adjusted $P \leq .05$ indicated statistical significance. All tests were 1-tailed.[25]

Results

The data set contained 818 examinations from 662 unique patients with 328 CTA examinations (40.1%) containing at least 1 intracranial aneurysm and 490 examinations (59.9%) without intracranial aneurysms (**Figure 2**). Of the 328 aneurysm cases, 20 cases from 15 unique patients contained 2 or more aneurysms. One hundred forty-eight aneurysm cases contained aneurysms between 3 mm and 7 mm, 108 cases had aneurysms between 7 mm and 12 mm, 61 cases had aneurysms between 12 mm and 24 mm, and 11 cases had aneurysms 24 mm or greater. The location of the aneurysms varied according to the following distribution: 99 were located in the internal carotid artery, 78 were in the middle cerebral artery, 50 were cavernous internal carotid artery aneurysms, 44 were basilar tip aneurysms, 41 were in the anterior communicating artery, 18 were in the posterior communicating artery, 16 were in the vertebrobasilar system, and 12 were in the anterior cerebral artery. All examinations were performed either on a GE Discovery, GE LightSpeed, GE Revolution, Siemens Definition, Siemens Sensation, or a Siemens Force scanner, with slice thicknesses of 1.0 mm or 1.25 mm, using standard clinical protocols for head angiogram or head/neck angiogram. There was no difference between the protocols or slice thicknesses between the aneurysm and nonaneurysm examinations. For this study, axial series were extracted from each examination and a segmentation label was produced on every axial slice containing an aneurysm. The number of images per examination ranged from 113 to 802 (mean [SD], 373 [157]).

The examinations were split into a training set of 611 examinations (494 patients; mean [SD] age, 55.8 [18.1] years; 372 [60.9%] female) used to train the model, a development set of 92 examinations (86 patients; mean [SD] age, 61.6 [16.7] years; 59 [64.1%] female) used for model selection, and a test set of 115 examinations (82 patients; mean [SD] age, 57.8 [18.3] years; 74 [64.4%] female) to evaluate the performance of the clinicians when augmented with the model (Figure 2). Using stratified random sampling, the development and test sets were formed to include

Figure 2. Data Set Selection Flow Diagram and Patient Demographics



Of 9455 computed tomography angiogram (CTA) examinations performed between 2003 and 2017 at Stanford University Medical Center, 818 were selected according to an exclusion criteria validated by a board-certified neuroradiologist. These examinations were split into the training set, development set, and test set to be used for training models, selecting the best model, and assessing the selected model, respectively.

50% aneurysm examinations and 50% normal examinations; the remaining examinations composed the training set, of which 36.5% were aneurysm examinations. Forty-three patients had multiple examinations in the data set due to examinations performed for follow-up of the aneurysm. To account for these repeat patients, examinations were split so that there was no patient overlap between the different sets. Figure 2 contains pathology and patient demographic characteristics for each set.

A total of 8 clinicians, including 6 board-certified practicing radiologists, 1 practicing neurosurgeon, and 1 radiology resident, participated as readers in the study. The radiologists' years of experience ranged from 3 to 12 years, the neurosurgeon had 2 years of experience as attending, and the resident was in the second year of training at Stanford University Medical Center. Groups 1 and 2 consisted of 3 radiologists each; the resident and neurosurgeon were both in group 1. None of the clinicians were involved in establishing the reference standard for the examinations.

Without augmentation, clinicians achieved a microaveraged sensitivity of 0.831 (95% CI, 0.794-0.862), specificity of 0.960 (95% CI, 0.937-0.974), and an accuracy of 0.893 (95% CI, 0.872-0.912). With augmentation, the clinicians achieved a microaveraged sensitivity of 0.890 (95% CI, 0.858-0.915), specificity of 0.975 (95% CI, 0.957-0.986), and an accuracy of 0.932 (95% CI, 0.913-0.946). The underlying model had a sensitivity of 0.949 (95% CI, 0.861-0.983), specificity of 0.661 (95% CI, 0.530-0.771), and accuracy of 0.809 (95% CI, 0.727-0.870). The performances of the model, individual clinicians, and their microaverages are reported in eTable 1 in the Supplement.

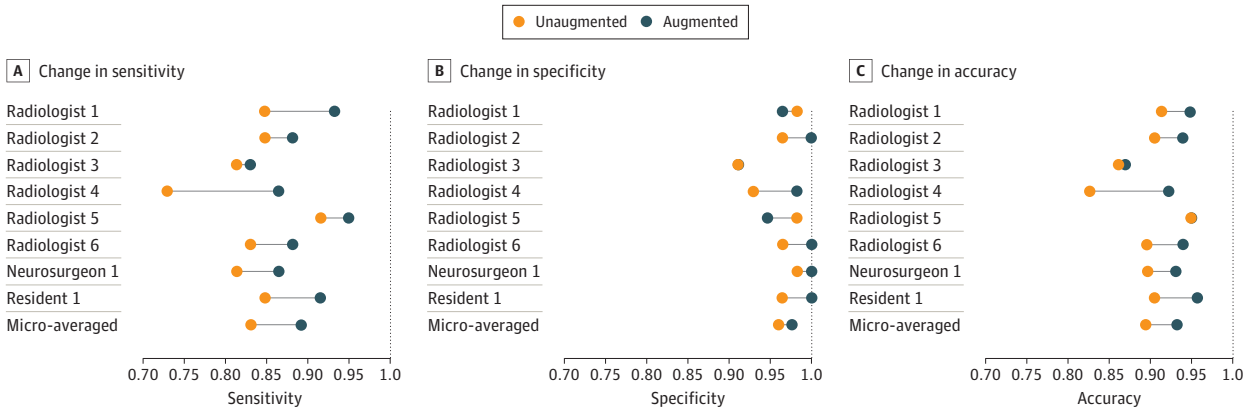
With augmentation, there was a statistically significant increase in the mean sensitivity (0.059; 95% CI, 0.028-0.091; adjusted P = .01) and mean accuracy (0.038; 95% CI, 0.014-0.062; adjusted P = .02) of the clinicians as a group. There was no statistically significant change in mean specificity (0.016; 95% CI, −0.010 to 0.041; adjusted P = .16). Performance improvements across clinicians are detailed in the **Table**, and individual clinician improvement in **Figure 3**. Individual performances with and without model augmentation are shown in eTable 1 in the Supplement. The sensitivity analysis confirmed that even among board-certified radiologists, there was a statistically significant increase

Table. Clinician Performance Metrics With and Without Augmentation

	Microaverage (95% CI)			*P* Value	
Metric	Without Augmentation	With Augmentation	Mean Increase (95% CI)	Unadjusted	Adjusted[a]
Sensitivity	0.831 (0.794 to 0.862)	0.890 (0.858 to 0.915)	0.059 (0.028 to 0.091)	.001	.01
Specificity	0.960 (0.937 to 0.974)	0.975 (0.957 to 0.986)	0.016 (−0.010 to 0.041)	.10	.16
Accuracy	0.893 (0.782 to 0.912)	0.932 (0.913 to 0.946)	0.038 (0.014 to 0.062)	.004	.02

[a] *P* values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg correction.

Figure 3. Change in Individual Clinicians' Performance Metric



Horizontal lines depict the change in performance metric for each clinician with and without model augmentation. The orange dot represents performance without model, and the blue dot represents performance with model augmentation.

in mean sensitivity (0.059; 95% CI, 0.013-0.105; adjusted P = .04) and accuracy (0.036; 95% CI, 0.001-0.072; adjusted P = .05). Performance improvements of board-certified radiologists as a group are shown in eTable 2 in the Supplement.

The mean diagnosis time per examination without augmentation microaveraged across clinicians was 57.04 seconds (95% CI, 54.58-59.50 seconds). The times for individual clinicians are detailed in eTable 3 in the Supplement, and individual time changes are shown in eFigure 2 in the Supplement. With augmentation, there was no statistically significant decrease in mean diagnosis time (5.71 seconds; 95% CI, −7.22 to 18.63 seconds; adjusted P = .19). The model took a mean of 7.58 seconds (95% CI, 6.92-8.25 seconds) to process an examination and output its segmentation map.

Confusion matrices, which are tables reporting true- and false-positive results and true- and false-negative results of each clinician with and without model augmentation, are shown in eTable 4 in the Supplement.

There was a statistically significant increase of 0.060 (adjusted P = .05) in the interrater agreement among the clinicians, with an exact Fleiss κ of 0.799 without augmentation and 0.859 with augmentation. For the board-certified radiologists, there was an increase of 0.063 in their interrater agreement, with an exact Fleiss κ of 0.783 without augmentation and 0.847 with augmentation.

Discussion

In this study, the ability of a deep learning model to augment clinician performance in detecting cerebral aneurysms using CTA was investigated with a crossover study design. With model augmentation, clinicians' sensitivity, accuracy, and interrater agreement significantly increased. There was no statistical change in specificity and time to diagnosis.

Given the potential catastrophic outcome of a missed aneurysm at risk of rupture, an automated detection tool that reliably detects and enhances clinicians' performance is highly desirable. Aneurysm rupture is fatal in 40% of patients and leads to irreversible neurological disability in two-thirds of those who survive; therefore, an accurate and timely detection is of paramount importance. In addition to significantly improving accuracy across clinicians while interpreting CTA examinations, an automated aneurysm detection tool, such as the one presented in this study, could also be used to prioritize workflow so that those examinations more likely to be positive could receive timely expert review, potentially leading to a shorter time to treatment and more favorable outcomes.

The significant variability among clinicians in the diagnosis of aneurysms has been well documented and is typically attributed to lack of experience or subspecialty neuroradiology training, complex neurovascular anatomy, or the labor-intensive nature of identifying aneurysms. Studies have shown that interrater agreement of CTA-based aneurysm detection is highly variable, with interrater reliability metrics ranging from 0.37 to 0.85,[6,7,26-28] and performance levels that vary depending on aneurysm size and individual radiologist experience.[4,6] In addition to significantly increasing sensitivity and accuracy, augmenting clinicians with the model also significantly improved interrater reliability from 0.799 to 0.859. This implies that augmenting clinicians with varying levels of experience and specialties with models could lead to more accurate and more consistent radiological interpretations.

Currently, tools to improve clinician aneurysm detection on CTA include bone subtraction,[29] as well as 3-D rendering of intracranial vasculature,[30-32] which rely on application of contrast threshold settings to better delineate cerebral vasculature and create a 3-D–rendered reconstruction to assist aneurysm detection. However, using these tools is labor- and time-intensive for clinicians; in some institutions, this process is outsourced to a 3-D lab at additional costs. The tool developed in this study, integrated directly in a standard DICOM viewer, produces a segmentation map on a new examination in only a few seconds. If integrated into the standard workflow, this diagnostic tool

could substantially decrease both cost and time to diagnosis, potentially leading to more efficient treatment and more favorable patient outcomes.

Deep learning has recently shown success in various clinical image-based recognition tasks. In particular, studies have shown strong performance of 2-D CNNs in detecting intracranial hemorrhage and other acute brain findings, such as mass effect or skull fractures, on CT head examinations.[33-36] Recently, one study[10] examined the potential role for deep learning in magnetic resonance angiogram–based detection of cerebral aneurysms, and another study[37] showed that providing deep learning model predictions to clinicians when interpreting knee magnetic resonance studies increased specificity in detecting anterior cruciate ligament tears. To our knowledge, prior to this study, deep learning had not been applied to CTA, which is the first-line imaging modality for detecting cerebral aneurysms. Our results demonstrate that deep learning segmentation models may produce dependable and interpretable predictions that augment clinicians and improve their diagnostic performance. The model implemented and tested in this study significantly increased sensitivity, accuracy, and interrater reliability of clinicians with varied experience and specialties in detecting cerebral aneurysms using CTA.

Limitations

This study has limitations. First, because the study focused only on nonruptured aneurysms, model performance on aneurysm detection after aneurysm rupture, lesion recurrence after coil or surgical clipping, or aneurysms associated with arteriovenous malformations has not been investigated. Second, since examinations containing surgical hardware or devices were excluded, model performance in their presence is unknown. In a clinical environment, CTA is typically used to evaluate for many types of vascular diseases, not just for aneurysm detection. Therefore, the high prevalence of aneurysm in the test set and the clinician's binary task could have introduced bias in interpretation. Also, this study was performed on data from a single tertiary care academic institution and may not reflect performance when applied to data from other institutions with different scanners and imaging protocols, such as different slice thicknesses.

Conclusions

A deep learning model was developed to automatically detect clinically significant intracranial aneurysms on CTA. We found that the augmentation significantly improved clinicians' sensitivity, accuracy, and interrater reliability. Future work should investigate the performance of this model prospectively and in application of data from other institutions and hospitals.

ARTICLE INFORMATION

Accepted for Publication: April 23, 2019.

Published: June 7, 2019. doi:10.1001/jamanetworkopen.2019.5600

Corresponding Author: Kristen W. Yeom, MD, School of Medicine, Department of Radiology, Stanford University, 725 Welch Rd, Ste G516, Palo Alto, CA 94304 (kyeom@stanford.edu).

Author Affiliations: Department of Computer Science, Stanford University, Stanford, California (Park, Chute, Rajpurkar, Lou, Ng); AIMI Center, Stanford University, Stanford, California (Ball); Roam Analytics, San Mateo, California (Ball); School of Medicine, Stanford University, Stanford, California (Shpanskaya, Jabarkheel, Kim); School of Medicine, Department of Radiology, Stanford University, Stanford, California (McKenna, Tseng, Ni, Wishah, Wittber, Halabi, Basu, Patel, Lungren, Yeom); School of Medicine, Department of Neurosurgery, Stanford University, Stanford, California (Hong, Wilson).

Author Contributions: Ms Park and Dr Yeom had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Ms Park and Messrs Chute and Rajpurkar are co-first authors. Drs Ng and Yeom are co-senior authors.

Concept and design: Park, Chute, Rajpurkar, Lou, Shpanskaya, Ni, Basu, Lungren, Ng, Yeom.

Acquisition, analysis, or interpretation of data: Park, Chute, Rajpurkar, Lou, Ball, Shpanskaya, Jabarkheel, Kim, McKenna, Tseng, Ni, Wishah, Wittber, Hong, Wilson, Halabi, Patel, Lungren, Yeom.

Drafting of the manuscript: Park, Chute, Rajpurkar, Lou, Ball, Jabarkheel, Kim, McKenna, Hong, Halabi, Lungren, Yeom.

Critical revision of the manuscript for important intellectual content: Park, Chute, Rajpurkar, Ball, Shpanskaya, Jabarkheel, Kim, Tseng, Ni, Wishah, Wittber, Wilson, Basu, Patel, Lungren, Ng, Yeom.

Statistical analysis: Park, Chute, Rajpurkar, Lou, Ball, Lungren.

Administrative, technical, or material support: Park, Chute, Shpanskaya, Jabarkheel, Kim, McKenna, Tseng, Wittber, Hong, Wilson, Lungren, Ng, Yeom.

Supervision: Park, Ball, Tseng, Halabi, Basu, Lungren, Ng, Yeom.

Conflict of Interest Disclosures: Drs Wishah and Patel reported grants from GE and Siemens outside the submitted work. Dr Patel reported participation in the speakers bureau for GE. Dr Lungren reported personal fees from Nines Inc outside the submitted work. Dr Yeom reported grants from Philips outside the submitted work. No other disclosures were reported.

Funding/Support: This work was supported by National Institutes of Health National Center for Advancing Translational Science Clinical and Translational Science Award UL1TR001085.

Role of the Funder/Sponsor: The National Institutes of Health had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

1. Jaja BN, Cusimano MD, Etminan N, et al. Clinical prediction models for aneurysmal subarachnoid hemorrhage: a systematic review. *Neurocrit Care*. 2013;18(1):143-153. doi:10.1007/s12028-012-9792-z

2. Turan N, Heider RA, Roy AK, et al. Current perspectives in imaging modalities for the assessment of unruptured intracranial aneurysms: a comparative analysis and review. *World Neurosurg*. 2018;113:280-292. doi:10.1016/j.wneu.2018.01.054

3. Yoon NK, McNally S, Taussky P, Park MS. Imaging of cerebral aneurysms: a clinical perspective. *Neurovasc Imaging*. 2016;2(1):6. doi:10.1186/s40809-016-0016-3

4. Jayaraman MV, Mayo-Smith WW, Tung GA, et al. Detection of intracranial aneurysms: multi-detector row CT angiography compared with DSA. *Radiology*. 2004;230(2):510-518. doi:10.1148/radiol.2302021465

5. Bharatha A, Yeung R, Durant D, et al. Comparison of computed tomography angiography with digital subtraction angiography in the assessment of clipped intracranial aneurysms. *J Comput Assist Tomogr*. 2010;34(3):440-445. doi:10.1097/RCT.0b013e3181d27393

6. Lubicz B, Levivier M, François O, et al. Sixty-four-row multisection CT angiography for detection and evaluation of ruptured intracranial aneurysms: interobserver and intertechnique reproducibility. *AJNR Am J Neuroradiol*. 2007;28(10):1949-1955. doi:10.3174/ajnr.A0699

7. White PM, Teasdale EM, Wardlaw JM, Easton V. Intracranial aneurysms: CT angiography and MR angiography for detection prospective blinded comparison in a large patient cohort. *Radiology*. 2001;219(3):739-749. doi:10.1148/radiology.219.3.r01ma16739

8. Suzuki K. Overview of deep learning in medical imaging. *Radiol Phys Technol*. 2017;10(3):257-273. doi:10.1007/s12194-017-0406-5

9. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med*. 2018;15(11):e1002686. doi:10.1371/journal.pmed.1002686

10. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med*. 2018;15(11):e1002699. doi:10.1371/journal.pmed.1002699

11. Morita A, Kirino T, Hashi K, et al; UCAS Japan Investigators. The natural course of unruptured cerebral aneurysms in a Japanese cohort. *N Engl J Med*. 2012;366(26):2474-2482. doi:10.1056/NEJMoa1113260

12. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*. 2006;31(3):1116-1128. doi:10.1016/j.neuroimage.2006.01.015

13. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; June 27, 2016; Las Vegas, NV.

14. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. Paper presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 25, 2017; Honolulu, HI.

15. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Paper presented at: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 21, 2018; Salt Lake City, Utah.

16. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Basel, Switzerland: Springer International; 2015:234–241.

17. Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. Paper presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 25, 2017; Honolulu, HI.

18. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. https://arxiv.org/abs/1706.05587. Published June 17, 2017. Accessed May 7, 2019.

19. Milletari F, Navab N, Ahmadi S-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. Paper presented at: 2016 IEEE Fourth International Conference on 3D Vision (3DV); October 26-28, 2016; Stanford, CA.

20. Loshchilov I, Hutter F. Sgdr: Stochastic gradient descent with warm restarts. Paper presented at: 2017 Fifth International Conference on Learning Representations; April 24-26, 2017; Toulon, France.

21. Huang G, Sun Y, Liu Z, Sedra D, Weinberger KQ. Deep networks with stochastic depth. *European Conference on Computer Vision*. Basel, Switzerland: Springer International; 2016:646–661.

22. Horos. https://horosproject.org. Accessed May 1, 2019.

23. Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc*. 1927;22(158):209-212. doi:10.1080/01621459.1927.10502953

24. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas*. 1973;33(3):613-619. doi:10.1177/001316447303300309

25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995;57(1):289-300.

26. Maldaner N, Stienen MN, Bijlenga P, et al. Interrater agreement in the radiologic characterization of ruptured intracranial aneurysms based on computed tomography angiography. *World Neurosurg*. 2017;103:876-882.e1. doi:10.1016/j.wneu.2017.04.131

27. Wang Y, Gao X, Lu A, et al. Residual aneurysm after metal coils treatment detected by spectral CT. *Quant Imaging Med Surg*. 2012;2(2):137-138.

28. Yoon YW, Park S, Lee SH, et al. Post-traumatic myocardial infarction complicated with left ventricular aneurysm and pericardial effusion. *J Trauma*. 2007;63(3):E73-E75. doi:10.1097/01.ta.0000246896.89156.70

29. Tomandl BF, Hammen T, Klotz E, Ditt H, Stemper B, Lell M. Bone-subtraction CT angiography for the evaluation of intracranial aneurysms. *AJNR Am J Neuroradiol*. 2006;27(1):55-59.

30. Shi W-Y, Li Y-D, Li M-H, et al. 3D rotational angiography with volume rendering: the utility in the detection of intracranial aneurysms. *Neurol India*. 2010;58(6):908-913. doi:10.4103/0028-3886.73743

31. Lin N, Ho A, Gross BA, et al. Differences in simple morphological variables in ruptured and unruptured middle cerebral artery aneurysms. *J Neurosurg*. 2012;117(5):913-919. doi:10.3171/2012.7.JNS111766

32. Villablanca JP, Jahan R, Hooshi P, et al. Detection and characterization of very small cerebral aneurysms by using 2D and 3D helical CT angiography. *AJNR Am J Neuroradiol*. 2002;23(7):1187-1198.

33. Chang PD, Kuoy E, Grinband J, et al. Hybrid 3D/2D convolutional neural network for hemorrhage evaluation on head CT. *AJNR Am J Neuroradiol*. 2018;39(9):1609-1616. doi:10.3174/ajnr.A5742

34. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet*. 2018;392(10162):2388-2396. doi:10.1016/S0140-6736(18)31645-3

35. Jnawali K, Arbabshirani MR, Rao N, Patel AA. Deep 3D convolution neural network for CT brain hemorrhage classification. Paper presented at: Medical Imaging 2018: Computer-Aided Diagnosis. February 27, 2018; Houston, TX. doi:10.1117/12.2293725

36. Titano JJ, Badgeley M, Schefflein J, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med*. 2018;24(9):1337-1341. doi:10.1038/s41591-018-0147-y

37. Ueda D, Yamamoto A, Nishimori M, et al. Deep learning for MR angiography: automated detection of cerebral aneurysms. *Radiology*. 2019;290(1):187-194.

SUPPLEMENT.

eFigure 1. Diagram of Model Architecture
eFigure 2. Individual Changes in Time Spent per Case
eTable 1. Comparison of Individual Unaugmented and Augmented Clinicians in Aneurysm Detection on the Test Set
eTable 2. Mean Increase in Board-Certified Radiologists' Metrics as a Group
eTable 3. Comparison of Individual Unaugmented and Augmented Clinicians in Time Spent on Aneurysm Detection on the Test Set
eTable 4. Comparison of Confusion Matrices of Individual Unaugmented and Augmented Clinicians on Aneurysm Detection on the Test Set