nature medicine

LETTERS

https://doi.org/10.1038/s41591-018-0147-y

Automated deep-neural-network surveillance of cranial images for acute neurologic events

Joseph J. Titano[1,5], Marcus Badgeley[2,5], Javin Schefflein[1], Margaret Pain[2], Andres Su[1], Michael Cai[1], Nathaniel Swinburne[1], John Zech[1], Jun Kim[3], Joshua Bederson[2], J. Mocco[2], Burton Drayer[1], Joseph Lehar[4], Samuel Cho[2,3], Anthony Costa[2] and Eric K. Oermann[2]*

Rapid diagnosis and treatment of acute neurological illnesses such as stroke, hemorrhage, and hydrocephalus are critical to achieving positive outcomes and preserving neurologic function—'time is brain'[1–5]. Although these disorders are often recognizable by their symptoms, the critical means of their diagnosis is rapid imaging[6–10]. Computer-aided surveillance of acute neurologic events in cranial imaging has the potential to triage radiology workflow, thus decreasing time to treatment and improving outcomes. Substantial clinical work has focused on computer-assisted diagnosis (CAD), whereas technical work in volumetric image analysis has focused primarily on segmentation. 3D convolutional neural networks (3D-CNNs) have primarily been used for supervised classification on 3D modeling and light detection and ranging (LiDAR) data[11–15]. Here, we demonstrate a 3D-CNN architecture that performs weakly supervised classification to screen head CT images for acute neurologic events. Features were automatically learned from a clinical radiology dataset comprising 37,236 head CTs and were annotated with a semisupervised natural-language processing (NLP) framework[16]. We demonstrate the effectiveness of our approach to triage radiology workflow and accelerate the time to diagnosis from minutes to seconds through a randomized, double-blinded, prospective trial in a simulated clinical environment.

Current studies of CAD in radiology have shown a positive effect on radiologists' ability to detect lesions in a variety of settings from mammography to magnetic resonance angiography. In each of these contexts, the sensitivity for lesion detection is quite impressive, exceeding 95% in most instances, but is often accompanied by large numbers of false positives. Many of these errors stem from the challenge of automatically featurizing the images, which remains an open problem despite the successes of deep learning at automated featurization in domains outside of radiology[17,18]. Part of the reason for this problem is that medical-imaging labels often describe an entire image rather than a specific region containing pathology, thus forcing the classification problem to be the more challenging problem of weakly supervised learning[19]. Some groups have proposed combining manual featurization with CAD-based decision support to achieve an effective semiautomated hybrid solution[20]. Previous studies of CAD, however, suggest that in a triage framework, radiologists can control for false-positive classifications through manual inspection, thereby leaving the triaging system to prioritize review and benefit workflow even without an improvement in accuracy[21,22]. The lack of large, curated medical-imaging databases as well as tools to easily build them has been cited as a major hurdle in improving CAD technology. Efforts to train CNNs on large, curated image databases offer an opportunity to improve radiologists' ability to detect and classify lesions as well as to improve existing CAD technologies[21]. The present study sought to avoid the challenges of CAD entirely by focusing on optimizing the triage of studies to radiologists by using a fully automated approach and leaving the question of diagnosis itself up to the physician reviewing the case.

Owing to the recent successes of deep-learning algorithms on computer-vision problems, there is substantial interest in applying these techniques to medical imaging[17,23]. Existing work has focused on 2D imaging, in which highly successful algorithms developed for image classification may be easily deployed on new medical use cases[24–27]. Many of these studies exploit transfer learning, whereby 2D-CNNs are trained on millions of nonmedical 2D images, and the pretrained CNN classifiers are fine-tuned with the smaller medical datasets[24,25,27]. However, much of medical imaging data are volumetric and require 3D-CNNs or 2D projections of the 3D space to be processed. Large-scale pretraining datasets are not widely available for 3D-CNNs, and applications to date have largely focused on CAD or LiDAR data, or segmentation, in which these methods have had particular success[11–13,15,28–32]. In these cases, both true 3D-CNNs with voxels (volumetric pixels) as well as orthogonal 2D projections of the 3D space have been used with 3D-CNNs and have yielded better results than projection-based approaches[13–15].

Our dataset of 37,236 studies and radiology reports for a further 96,303 studies was accumulated as part of the ICAHNC project, an institutional review board–approved computer-vision initiative within the Department of Radiology and part of the Icahn School of Medicine AI Consortium (AISINAI). A detailed description of data acquisition, preprocessing, and labeling can be found in the Methods and in the Nature Research Reporting Summary online. In brief, images and their accompanying reports were standardized and processed with a crowdsourcing platform and an NLP pipeline (Fig. 1a). To differentiate between labels obtained by expert review of the images themselves and the labels obtained by the NLP model's inferring labels from the study reports, we adopted the terms 'gold-standard labels' and 'silver-standard labels', after Agarwal and Halpern[33,34]. Gold-standard labels refer to the studies that were labeled through manual physician review of the patient medical record, including images and follow-up studies, whereas silver-standard labels refer to the noisy labels obtained after application of an NLP algorithm to the report for each study. The NLP algorithm was trained on

[1]Department of Radiology, Icahn School of Medicine, New York, NY, USA. [2]Department of Neurological Surgery, Icahn School of Medicine, New York, NY, USA. [3]Department of Orthopedic Surgery, Icahn School of Medicine, New York, NY, USA. [4]Bioengineering and Bioinformatics, Boston University, Boston, MA, USA. [5]These authors contributed equally: Joseph J. Titano, Marcus Badgeley. *e-mail: eric.oermann@mountsinai.org

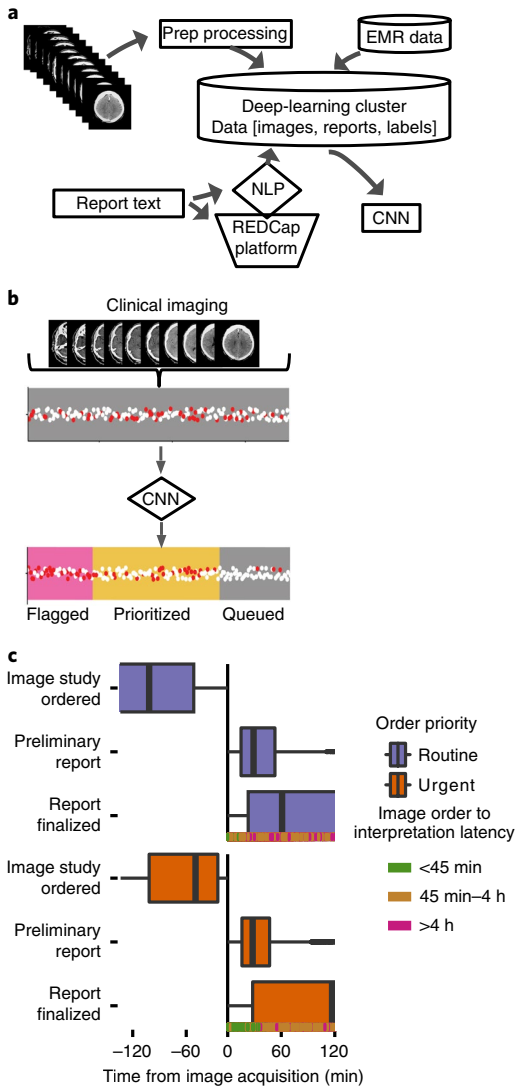

Fig. 1 | Data pipeline, CNN triage concept, and analysis of current CT head workflow. a, A multistage preprocessing pipeline with a crowdsourcing component facilitates the training of deep neural networks on radiology data from a hospital system. **b**, 180 test images and corresponding clinical information were used to build sample work queues in a simulated clinical environment to test the ability of our deep-learning-based image-triage system to practically affect radiology workflow. **c**, Distribution of durations between an image study being ordered and final interpretation. The boxes indicate the median, twenty-fifth and seventy-fifth quantiles, and whiskers extend to 1.58 × the interquartile range (IQR)/$\sqrt{n}$. Images ordered urgently (orange, $n = 60{,}254$) took an average time of 174 min (s.d. 216 min) until a preliminary report was published, whereas routine studies (purple, $n = 35{,}876$) took 241 min (s.d. 250 min) (two-sided t test, $P < 2 \times 10^{-16}$). For urgent studies, the time from being ordered to acquisition was 74 min versus 169 min (t test, $P < \times 10^{-16}$). For all studies, 77.6% were completely interpreted within 4 h of an initial order being placed.

the 96,303 reports, and all 37,236 imaging studies were annotated with silver-standard labels based on diagnostic terms derived from the Universal Medical Language System (UMLS) concept universal identifiers. The NLP labels for UMLS concept universal identifiers were pooled into higher-level disease types, and radiologists manually assigned each disease's acuity as being critical or noncritical, on the basis of institutional protocols (full mapping in Supplementary Table 1).

We trained a 3D-CNN modeled after the ResNet-50 architecture[23] to identify whether an image contained acute neurological illnesses or noncritical findings for the purpose of triaging studies, so that images were seen in the order of potential acuity rather than the order in which they were acquired (Fig. 1b). With this schema and the CNN's findings, studies were flagged as critical or noncritical, and ordered in a work queue according to the probability of a critical finding (training details in Methods). To obtain a clinically meaningful measure of our deep-learning-based triage system, we performed a small double-blinded randomized controlled trial (RCT) of image interpretation in a simulated clinical environment. In the trial, both the model and radiologists were assessed regarding how quickly they recognized and provided notification of a critical finding (Methods and Supplementary Fig. 1). Model performance was assessed with standard metrics of classification performance, including the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and accuracy (ACC). Because of our focus on queue prioritization and rapid assessment, the model was also assessed for its runtime. Statistical comparisons were calculated with appropriate nonparametric statistical tests. The Matthews correlation coefficient was calculated to assess the correlation between binary variables. All P values are reported at $\alpha = 0.05$. Data are reported as mean ± 1 s.d.

According to a survey of 96,303 radiology reports and their metadata, the average time for an initial report to be generated from the time of acquisition was 83 min (s.d. 191 min). For studies labeled as urgent, the average time was 87 min (s.d. 200 min) compared with 75 min (s.d. 171 min) for studies labeled as routine; 77.6% of all studies were completely interpreted within 4 h of an initial order being placed in the electronic medical record (EMR) (Fig. 1c). Being flagged as urgent in the EMR system itself was not associated with diagnosis of a critical finding (ACC = 0.55, Matthews correlation coefficient = –0.12), thus suggesting a need for a smarter CNN-based approach of prioritization. Two datasets were subsequently constructed as previously described: a training-validation dataset with silver-standard labels and a test dataset with gold-standard labels. For the test dataset, 180 images were randomly queried from the dataset of silver-standard labels to achieve an approximately 1:1 split of noncritical to critical studies (Fig. 2a). Gold-standard labels were obtained for the studies through manual review of patient medical records, and the remaining studies were pooled into silver-standard-label training and validation sets in an 80:20 ratio. The prevalence of critical findings in the silver-standard dataset was 7.6%, 0.8%, and 0.7% at thresholds of 4, 8, and 10 UMLS critical findings per note, respectively (Supplementary Table 2).

For patients included in the test cohort, the average age was 59.7 years (s.d. 22.7), and 51% ($n = 92$) of the patients were female. Studies were evenly distributed among acquisition settings: 36% came from the emergency department, 33% came from an inpatient unit, and 31% came from an outpatient setting. The patient symptoms were mostly nonspecific, and the top three symptoms were headache (27%), altered mental status (17%), and ataxia/dizziness (9%). There was an association between a clinical presentation suggestive of neurological disease and positive findings on imaging (47% versus 24%, $P = 0.016$) (diagnoses, symptoms, and findings in Supplementary Tables 3–5).

We constructed a deep neural network based on the ResNet architecture and trained it to predict critical versus noncritical findings (Fig. 2b and Supplementary Fig. 2). For predicting the silver-standard labels of 'critical finding' on the training set, the 3D-CNN had an AUC of 0.88 (Fig. 3a). The neural network's performance was subsequently tested against that of two radiologists (J.S. and M.C.) and a neurosurgeon (E.K.O.) in predicting the gold-standard labels. For predicting gold-standard labels, EMR prioritization performed poorly (ACC = 0.55, sensitivity = 0.16, specificity = 0.73), whereas the NLP model's best results based on the radiology reports were

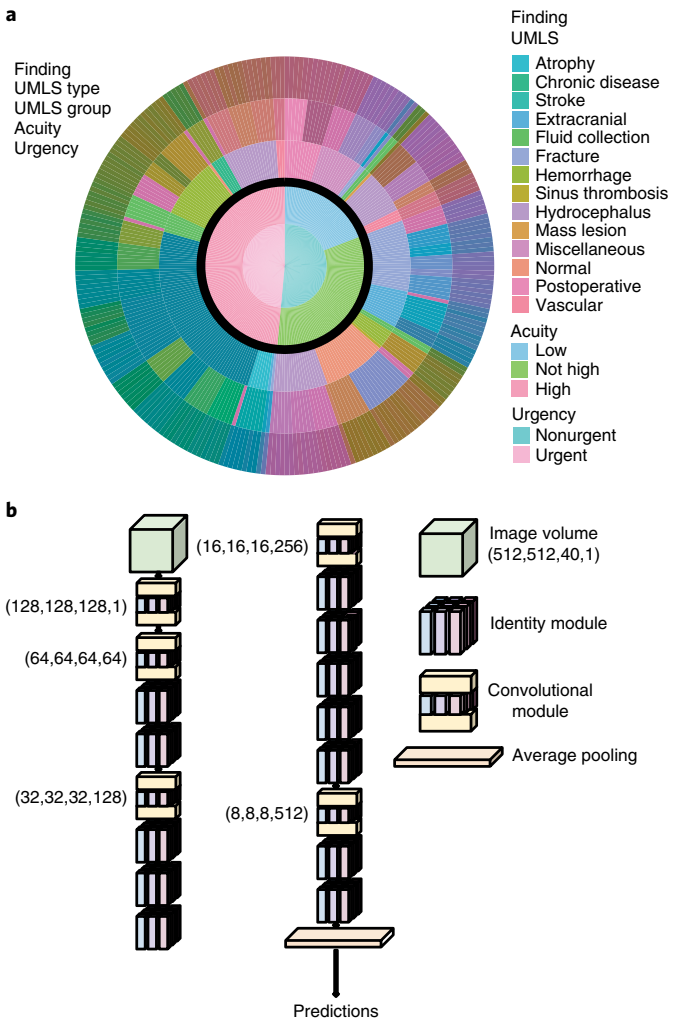

Fig. 2 | Systems for encoding images and mapping diseases to urgency. **a**, Hundreds of potential head CT diagnoses were designated urgent or nonurgent through manual mapping of 14 UMLS groupings. The definition of critical, made on an institutional level, is largely standardized and reflects illnesses that require prompt intervention. **b**, A 3D-CNN was modeled after the ResNet-50 (ref. [23]).

better (ACC = 0.71, sensitivity = 0.18, specificity = 0.95). Here, the 3D-CNN had an AUC of 0.73 (Fig. 3b). The sensitivity and specificity of the model were 0.79 and 0.48, relative to an average sensitivity and specificity of 0.79 (s.d. 0.04) and 0.85 (s.d. 0.07), respectively, for the physicians. Notably, for several of the labels, many cases contained degenerate or highly variable visual features (Fig. 3c). The average inference time of the algorithm was 134 ms (s.d. 900 ms). According to these results, we decided that a triage system could be constructed that functions at a human level of sensitivity and could theoretically alert physicians in 50% of critical cases ($n = 21$) with a 21% false-alarm rate ($n = 85$).

To assess whether a deep-learning-based technology could meaningfully triage studies, we performed an RCT in a simulated clinical environment of our system as an alarm mechanism as well as a triage/prioritization system. Notably, in our simulated clinical environment, the average time for the algorithm to preprocess an image, run its inference method, and potentially raise an alarm was 150 times faster than that for humans (1.2 s versus 177 s; $P = 2 \times 10^{-17}$) (Fig. 4a). When generating work queues, we found a significant difference in queue position between the urgent and routine cases in the two cohorts: more urgent studies appeared earlier in the queue in the prioritized list ($P = 0.01$) (Fig. 4b). An example work queue demonstrates the CNN-mediated triage of images with decreased time to recognition even with false-positive results (Fig. 4c).

Here, we demonstrated our creation of a machine-learning framework to triage radiology studies for critical findings. Our analysis was performed on actual clinical cases from an academic medical center and validated in an RCT through a simulated work environment with actual radiologists. The 3D-CNN algorithm augments human performance by prioritizing the normal radiology work queue. This choice of clinical task, triage rather than diagnosis, was motivated by our clinical findings regarding current imaging workflow as well as our technical findings that triage is a more tractable problem to solve with a weakly supervised deep-learning classifier. Furthermore, the critical versus noncritical class labels are mutually exclusive, thus allowing for an easier classification problem, in contrast to the non-mutually exclusive, multilabel nature of diagnostic medical imaging. Consequently, by choosing to target prioritization rather than classification, we alleviated the need for extremely accurate models and emphasized one of our method's strengths: speed. Our results demonstrate that formal CAD is still challenging with weakly supervised classifiers and silver-standard labels (further examples of model failure in Supplementary Fig. 3). Although some studies have reported accurate CAD results by using weakly supervised models, they have often focused on single diagnoses (e.g., hemorrhage, stroke, and hydrocephalus). Furthermore, we believe that such reports must be viewed skeptically in the absence of rigorous validation, because our results on actual diagnostic accuracy were poorer than those of human performance[35]. Addressing triage rather than CAD is a better pairing of clinical task with technical solution. With a total processing and interpretation time of approximately 1 s, such a triage system can alert physicians to a critical finding that might otherwise remain in a queue for minutes to hours.

Notably, to use images alone to make diagnoses is to work with an impoverished input. Medical decisions are not made in isolation solely on the basis of imaging findings, and the diagnostic work of radiologists requires careful tailoring of models to the prior probabilities of disease. During our review of the test cases, we observed that arriving at a clinically meaningful diagnosis in one-third of the images would be impossible without access to prior imaging or clinical information. Our results are based on the studies and reports from a single hospital system, and will require validation externally. There are also clear ways to improve this system; for example, building a visualization layer over the CNN classifier to label areas of activation might further facilitate human interpretation. An additional opportunity for improvement is found in the quality and nature of the labels. Image-level labels require the problem to be approached as a case of weakly supervised learning; however, doing so is particularly challenging in the medical context, in which pathological findings often exist within a small region of an otherwise normal study, and a single pathological entity can vary considerably among cases[19] (Supplementary Fig. 4). Strongly supervised approaches involving segmentation or object detection are likely to obtain better results and to more accurately reflect the radiological task. Even higher-quality weak labels would lead to improved classifier performance and consequently to superior triage results. We also were unable to fully explore the data space available with modern CT imaging. Different types of image acquisition may possibly affect model performance. For example, volume averaging from a 5-mm slice thickness might obscure more subtle findings that a radiologist would be able to delineate by referencing the thinnest slices available[36–39].

Although our trial provides an initial assessment of a CNN-directed triage system, there are several deviations from actual practice that bear mentioning. For patients with concerning symptoms, radiologists are often called by ordering clinicians, and the

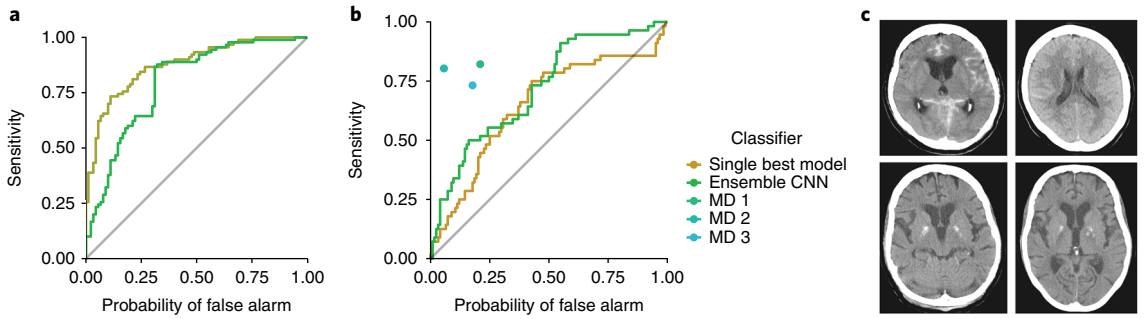

Fig. 3 | Human and algorithm classifier performance and diagnosis heterogeneity. a,**b**, Receiver operating characteristic curves showing AUC values of different classifiers. For predicting the silver-standard labels, the 3D-CNN had an AUC of 0.88 (**a**), whereas for the gold-standard labels, the 3D-CNN had an AUC of 0.73 (**b**). The accuracy of the model at an optimal threshold for screening (negative predictive value = 0.90) was 0.56. **c**, Four CT scans demonstrating the challenging nature of a weakly supervised classification task. The top-left and top-right images show subarachnoid hemorrhages occurring in two different locations with different visual features. On the bottom left and bottom right, there are similarly appearing benign basal ganglia calcifications (bottom left) and a basal ganglia hemorrhage (bottom right).

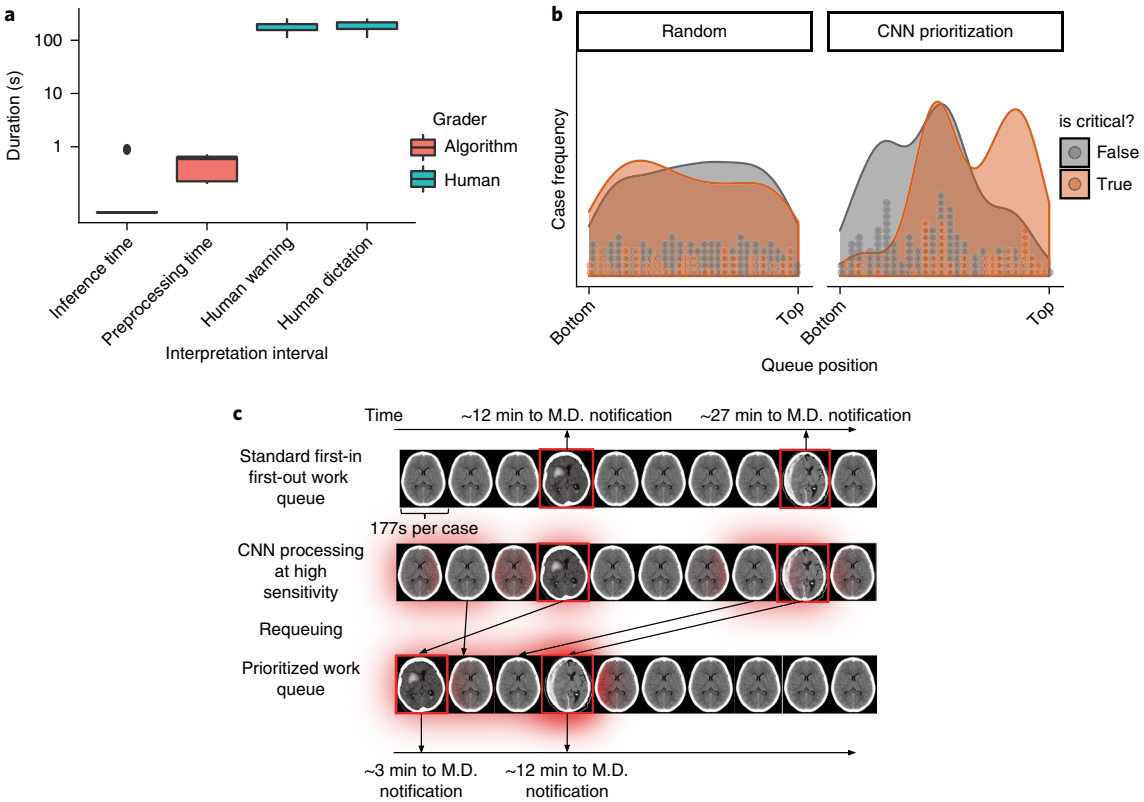

Fig. 4 | Interpretation speed and CNN queue triage. a, Box plots from an RCT of interpretation speed in a simulated clinical environment, showing that a deep-learning-based system interprets images 150 times faster than humans (1.2 s versus 177 s; $P = 4 \times 10^{-12}$, paired t test, $n = 17$). The boxes map to the median, twenty-fifth and seventy-fifth quantiles, and whiskers extend to $1.58 \times \mathrm{IQR}/\sqrt{n}$. **b**, The deep-learning-triaged queue had a significantly lower distribution of queue positions for critical studies compared with noncritical studies (two-sided Wilcoxon rank-sum test, $P = 0.01$, $n = 180$). **c**, A sample work queue containing ten cases demonstrating the system in deployment.

ordering physicians may even accompany patients to the department of radiology. Our lack of a notification mechanism such as this is inconsistent with actual practice and probably works against our radiologists. However, because this was a simulation, there were no unrelated interruptions or distractions, nor were there other studies requiring the attention of our physicians. As in most simulations, the present study is idealized and will require validation in a formal multicenter clinical trial, which our group intends to perform, using our home institution as our initial site.

Further research in applying modern deep-learning and computer-vision techniques to radiological imaging is underway and is a clear imperative for medical care in the twenty-first century. For the near term, optimizing study-label generation, specifying pathological regions in medical images to build strongly supervised models rather than weakly supervised ones, and tailoring algorithms for deployment in picture archiving and communication system (PACS) platforms will provide avenues for algorithm improvement and entrance into the clinical setting. Most importantly, though,

will be the active involvement of the medical community to ensure that developed algorithms are intelligently integrated into existing medical practice to solve actual medical problems and that techniques are rigorously validated for clinical efficacy.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41591-018-0107-6.

Received: 11 October 2017; Accepted: 23 May 2018;
Published online: 13 August 2018

References

1. Furlan, A. J. Time is brain. *Stroke* **37**, 2863–2864 (2006).
2. Del Zoppo, G. J., Saver, J. L., Jauch, E. C., Adams, H. P. Jr & American Heart Association Stroke Council. Expansion of the time window for treatment of acute ischemic stroke with intravenous tissue plasminogen activator: a science advisory from the American Heart Association/American Stroke Association. *Stroke* **40**, 2945–2948 (2009).
3. Jovin, T. G. et al. Thrombectomy within 8h after symptom onset in ischemic stroke. *N. Engl. J. Med.* **372**, 2296–2306 (2015).
4. Saver, J. L. Time is brain: quantified. *Stroke* **37**, 263–266 (2006).
5. Seelig, J. M. et al. Traumatic acute subdural hematoma: major mortality reduction in comatose patients treated within four hours. *N. Engl. J. Med.* **304**, 1511–1518 (1981).
6. National Collaborating Centre for Chronic Conditions (UK). *Stroke: National Clinical Guideline for Diagnosis and Initial Management of Acute Stroke and Transient Ischaemic Attack (TIA)*. (Royal College of Physicians, London, 2011).
7. Broderick, J. P. et al. Guidelines for the management of spontaneous intracerebral hemorrhage: a statement for healthcare professionals from a special writing group of the Stroke Council, American Heart Association. *Stroke* 30, (905–915 (1999).
8. Ferro, J. M. et al. Diagnosis of stroke by the nonneurologist: a validation study. *Stroke* **29**, 1106–1109 (1998).
9. Mullins, M. E. et al. CT and conventional and diffusion-weighted MR imaging in acute stroke: study in 691 patients at presentation to the emergency department. *Radiology* **224**, 353–360 (2002).
10. Navi, B. B. et al. The use of neuroimaging studies and neurological consultation to evaluate dizzy patients in the emergency department. *Neurohospitalist* **3**, 7–14 (2013).
11. Sedaghat, N., Zolfaghari, M. & Brox, T. Orientation-boosted voxel nets for 3D object recognition. Preprint at https://arxiv.org/abs/1604.03351/ (2016).
12. Maturana, D. & Scherer, S. VoxNet: a 3D convolutional neural network for real-time object recognition. in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 922–928 (IEEE, Piscataway, NJ, USA, 2015).
13. Hegde, V. & Zadeh, R. FusionNet: 3D object classification using multiple data representations. Preprint at https://arxiv.org/abs/1607.05695/ (2016).
14. Sinha, A., Bai, J. & Ramani, K. Deep learning 3D shape surfaces using geometry images. in *Computer Vision – ECCV 2016* 223–240 (Springer, Cham, Switzerland, 2016).
15. Brock, A., Lim, T., Ritchie, J. M. & Weston, N. Generative and discriminative voxel modeling with convolutional neural networks. Preprint at https://arxiv.org/abs/1608.04236/ (2016).
16. Shin, H.-C. et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **35**, 1285–1298 (2016).
17. Deng, J. et al. ImageNet: a large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (IEEE, Piscataway, NJ, USA, 2009).
18. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. Preprint at https://arxiv.org/abs/1512.03385/ (2015).
19. Durand, T., Mordan, T., Thome, N. & Cord, M. Wildcat: weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*. (IEEE, Piscataway, NJ, USA, 2017).
20. Chen, P.-H., Botzolakis, E., Mohan, S., Nick Bryan, R. & Cook, T. Feasibility of streamlining an interactive Bayesian-based diagnostic support tool designed for clinical practice. in M*edical Imaging2016: PACS and Imaging Informatics: Next Generation and Innovations* Vol. 9789, 97890C (International Society for Optics and Photonics, Bellingham, WA, USA, 2016).
21. Doi, K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput. Med. Imaging Graph.* **31**, 198–211 (2007).
22. Lehman, C. D. et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern. Med.* **175**, 1828–1837 (2015).
23. Wen, W., Wu, C., Wang, Y., Chen, Y. & Li, H. Learning structured sparsity in deep neural networks. Preprint at http://arxiv.org/abs/1608.03665/ (2016).
24. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
25. Lakhani, P. & Sundaram, B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* **284**, 574–582 (2017).
26. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *J. Am. Med. Assoc.* **316**, 2402–2410 (2016).
27. Liu, Y. et al. Detecting cancer metastases on gigapixel pathology images. Preprint at https://arxiv.org/abs/1703.02442/ (2017).
28. Oktay, O. et al. Anatomically constrained neural networks (ACNN): application to cardiac image enhancement and segmentation. *IEEE Trans. Med. Imaging* **37**, 384–395 (2017).
29. Li, X. et al. H-DenseUNet: hybrid densely connected UNet for liver and liver tumor segmentation from CT volumes. Preprint at https://arxiv.org/abs/1709.07330/ (2017).
30. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. Preprint at https://arxiv.org/abs/1606.06650/ (2016).
31. Brosch, T. et al. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Trans. Med. Imaging* **35**, 1229–1239 (2016).
32. Kamnitsas, K. et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **36**, 61–78 (2017).
33. Agarwal, V. et al. Learning statistical models of phenotypes using noisy labeled training data. *J. Am. Med. Inform. Assoc.* **23**, 1166–1173 (2016).
34. Halpern, Y., Choi, Y., Horng, S. & Sontag, D. Using anchors to estimate clinical state without labeled data. A*MIA Annu. Symp. Proc.* **2014**, 606–615 (2014).
35. Merkow, J. et al. DeepRadiologyNet: radiologist level pathology detection in CT head images. Preprint at https://arxiv.org/abs/1711.09313/ (2017).
36. Riedel, C. H. et al. Assessment of thrombus in acute middle cerebral artery occlusion using thin-slice nonenhanced computed tomography reconstructions. *Stroke* **41**, 1659–1664 (2010).
37. Kim, E. Y. et al. Detection of thrombus in acute ischemic stroke: value of thin-section noncontrast-computed tomography. *Stroke* **36**, 2745–2747 (2005).
38. Rubinstein, D., Escott, E. J. & Mestek, M. F. Computed tomographic scans of minimally displaced type II odontoid fractures. *J. Trauma* **40**, 204–210 (1996).
39. Bush, K., Huikeshoven, M. & Wong, N. Nasofrontal outflow tract visibility in computed tomography imaging of frontal sinus fractures. *Craniomaxillofac. Trauma Reconstr.* **6**, 237–240 (2013).

Acknowledgements

We thank E. Gordon for assistance with data collection for the purposes of building the NLP pipeline. We also thank the National Library of Medicine for making the UMLS Metathesaurus available.

Author contributions

A.C. and E.K.O. designed and built the computing environment for this project. J.J.T., J.M., and E.K.O. conceived the study. A.C., J.L., and E.K.O. developed the computer-vision algorithms. J.Z., E.K.O., J.S., and M.P. developed the NLP algorithms. J.J.T., M.P., A.S., J.K., and S.C. built the crowdsourcing platform for obtaining ground-truth labels. J.J.T., J.S., M.C., N.S., and J.Z. generated gold-standard labels and assembled the datasets. M.B. and E.K.O. designed and oversaw the clinical simulation. M.B. performed the statistical analysis and designed the figures. J.J.T., J.B., B.D., and E.K.O. supervised the project. All authors contributed to writing and editing the manuscript.

Competing interests

J.L. currently works for Merck in addition to his role as an adjunct professor at Boston University. M.B. currently works for Verily Life Sciences in addition to his role as a medical student at Mount Sinai. All of the present work was performed within the Mount Sinai Health System, and Merck and Verily played no role in the research and have no commercial interest in it.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0147-y.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to E.K.O.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Image acquisition and preprocessing. Basic demographic information, the location of the patient at time of analysis, and other clinical variables that would be available to a radiologist or clinician at the time of image interpretation were abstracted from the hospital's electronic medical record (EMR) system. Per-protocol head CT imaging was defined as a DICOM image series acquired in the axial plane with 5-mm slice thickness. Series meeting those criteria were extracted and subsequently stored as arrays with isotropic dimensions (1 mm^3) and intervoxel spacing of 1 mm, 1 mm, and 1 mm along the axial, coronal, and sagittal dimensions, respectively.

All relevant ethical regulations were followed as part of the conduct of this study. The Icahn School of Medicine Institutional Review Board approved the present study and granted a waiver of consent for the use of retrospective imaging data used in this study. Patient-level data and imaging were gathered for 37,236 studies and from radiology reports for a further 96,303 patients who underwent head CT imaging between January 2006 and January 2016. Head CT imaging for identified patients was extracted from the hospital PACS system and linked to clinical data through a unique identifying code. CT imaging was acquired randomly between the assigned years and screened for completed studies. Studies that were either not completed or not completed per the standard protocol, such as studies performed after contrast administration or those obtained with a low-radiation-dose protocol were excluded, thus resulting in a final cohort of 37,236 studies. The reports describing the image findings from an attending neuroradiologist were stored separately for later use in generating clinical labels. All data were stored in an encrypted, dual-token-authenticated, high-performance computing cluster specifically built for the handling of large amounts of protected health information and executing deep-learning graphs.

Silver language and image labels: NLP-based radiology note abstraction. An NLP tool was developed as a scalable but imperfect approach to obtain image labels from the radiologists' original clinical notes[40]. We considered the NLP image labels as a silver standard to contrast with the manually collected gold-standard labels discussed below. The machine-learning-based NLP algorithm itself has labels, referred to as language labels, to distinguish them from the image labels that were the focus of the current study. We briefly summarize the method below; a full report on the NLP tool and methods can be found in ref. [40].

Critical findings were defined on the basis of Viertel et al.[41], in accordance with our own institutional protocols. 1,004 of the 96,303 total head CT reports, which were uniformly sampled over the entire 2010–2016 time period, were manually reviewed for critical findings by three physicians (two senior radiology residents and a senior neurological surgery resident). Findings were decided by a majority vote. Manual annotation of these 1,004 studies was performed with a collaborative setup on the REDCap platform (Vanderbilt University). A hierarchical-labeling schema was implemented, in which users could specify key diagnoses and were then prompted to further detail the findings. For example, users could select 'subdural', and then the system would prompt them for acuity (acute, subacute, or chronic) as well as laterality and lobe (Supplementary Table 1).

Lasso logistic regression was performed to predict the binary presence of a primary UMLS term by using an indicator 'bag of words' vector based on the corresponding head CT report. For example, a report containing only the words 'findings: unremarkable exam' would be converted into a vector with 1 at three locations corresponding to 'findings', 'unremarkable', and 'exam', and a 0 at every other location in the vector corresponding to all other words in the vocabulary. This vector would be fed as input into a Lasso logistic regression with a binary target consisting of one of the primary UMLS terms identified in the report by the physician labelers. A separate regression was performed for each UMLS term of interest. Once fitted, these regression models were used to infer the presence or absence of terms in the remaining 95,299 unlabeled reports.

After primary UMLS labels were assigned to all available studies through the previously described NLP approach, UMLS labels were binned into critical or noncritical categories. The binning schema (Supplementary Table 1) included specific acute vascular events or other potentially surgical or medically emergent conditions as critical, and chronic or other nonspecific conditions as noncritical. Of the 1,004 reports manually labeled with UMLS terms, 477 were determined to contain one or more critical findings on the basis of the binning schema. This schema can be directly mapped back to the original UMLS categories (Supplementary Table 1), and it resulted in a wide distribution of labels across diagnoses and symptoms (Supplementary Tables 3 and 4). The number of critical labels to define a critical scan was optimized as a hyperparameter at training time and was found to be four positive labels per patient. For the NLP model, when trained on 60% of the 1,004 manually labeled reports ($n=602$), the model achieved AUC for the presence of any critical finding of 0.95 (sensitivity = 0.88; specificity = 0.87) on the remaining 40% validation reports ($n=402$).

Gold image-label manual chart review. A subset of images ($n=180$) was held out from the training and validation of the model to obtain a test set for estimating the out-of-sample error of the model as well as assessing outcomes in a simulated clinical setting. Test-image reports written by a neuroradiologist with full access to the patients' charts were obtained, and the reports, the images, and the clinical charts were reviewed by a radiologist and a neurosurgeon.
A consensus was reached among the diagnoses noted by the neuroradiologist in the report, the reviewing radiologist, and the reviewing neurosurgeon to define gold-standard UMLS terms for diagnosis for each image. After primary UMLS terms were assigned to all available studies, UMLS terms were binned into critical or noncritical categories, similarly to the NLP-pipeline procedure. Notably, this ground truth was made with full knowledge of the patient's clinical status in the EMR as well as with knowledge of prior imaging. To provide a more fair comparison, further labels were generated by having two radiologists and a neurosurgeon review the images alone without access to the EMR or to prior imaging. The second set of human labels constituted the human benchmarks for classifier performance. For the purposes of grouping symptoms for analysis, patients with focal deficits, ataxia, seizures, or a prior diagnosis of neurological disease were considered to have a clinical presentation suggestive of primary neurologic disease.

For all patients included in this test cohort, a manual chart review of the clinical record was also performed to ascertain symptoms and clinical disposition at the time of imaging (Supplementary Table 4). The patients' symptoms were mostly nonspecific, and the top three symptoms recorded included headache (27%), altered mental status (17%), and ataxia/dizziness (9%). Patients with symptoms likely to be associated with intracranial disease, focal neurological deficits, ataxia, seizures, and surveillance for known neurological conditions accounted for 29% of the cohort. For the 29% of patients with positive symptoms, approximately 47% had positive imaging findings. In comparison, patients without concerning symptoms had positive findings in only 24% of cases ($P=0.016$).

CNN image model design, training, and evaluation. We developed a novel architecture for the purposes of this study, termed VolResNet50, which was closely modeled after ResNet50 from Wen et al.[23] A deep neural network (50 layers) with residual connections was chosen to maximize the ability of the 3D-CNN to learn features as well as extensive use of batch normalization and dropout to regularize the network and encourage generalizability[42]. The ability of residual connections to minimize the number of free parameters while maximizing network depth is particularly critical for volumetric imaging with the addition of a third spatial dimension. We experimented with the use of batch normalization followed by rectified linear units as well as exponential linear units and found mild gains in training performance but not overall accuracy on the current dataset (Supplementary Fig. 3). As a means of comparison, we tried shallower architectures but consistently were unable to obtain a good fit. The combination of noisy, high-dimensional volumetric images makes for a challenging learning problem that requires substantially deeper representations to extract meaningful features for classification.

All CNN models were trained on a pair of nVidia GTX 1080 s by using Tensorflow 1.0.2 and Keras 2.0.1 (https://github.com/fchollet/keras/). Models were trained with a categorical cross-entropy loss with a stochastic-gradient-descent optimizer and nesterov momentum of 0.9. The learning rate was initially set at 0.001 and decayed 0.0005 per epoch. Simple data augmentation was performed to encourage the model to learn representations that would be invariant to rotation. At training time, a single axis was chosen at random, and then the volume was randomly rotated from 0° to 180° around that axis. Although training the algorithm required on average 3 d of computing time on a pair of GTX 1080 s (nVidia), inference on the trained net could be readily implemented at local clinics with minimal computing power.

In addition to training the single models, we trained an ensemble of CNNs to produce a more robust classifier. Using an elastic net, we combined the predictions of three different primary CNNs to predict the gold-standard labels. Cross-validation is a machine-learning technique for assessing generalization in which for each 'fold', the dataset is partitioned into two subsets (training and validation). A model is fit to the training partition and tested on the validation partition, and then the dataset is repartitioned for the next fold. Tenfold cross-validation was used to tune a grid of alpha and lambda penalization terms, and the model was evaluated by combining the 10% of held-out data from each of ten training folds to compute the ROC and other performance metrics.

The NLP was performed in Python. All deep-learning preprocessing, training, and inference were performed in Python. The ROC curves, sensitivity, specificity, and other statistical analyses were performed in R.

Validation-subset image-only human interpretation. For the purposes of comparing the VolResNet50 results to the human results, a held-out test set of 180 images was drawn from the hospital PACS on the basis of the silver-standard labels to obtain a 2:1 split of noncritical to critical images. Reviewers for the human comparison were selected to include both neurosurgeons and radiologists. A neurosurgery instructor (E.K.O.) and a PGY5 radiology house officer (J.S.) were given a shuffled set of unique identifying numbers for each imaging study. For each study, the clinical PACS viewer was set up to disable prior notification and windowing capability. Windowing capability was restricted to a level of 50 and a width of 90 to account for the CNN having access to only standard cranial windows, as compared with bone or stroke windows. Prior notifications were disabled despite being available in standard practice to better assess the model's

performance relative to humans on a similar task. Although it is standard clinical workflow to have an appreciation for the context of a given study, most importantly regarding whether any prior imaging studies have already been obtained, for the purposes of this test of model accuracy, we deemed it best to compare human performance with the model by using a similar set of images.

Imaging latency analysis. To characterize the latency in radiograph interpretation, we performed a large retrospective analysis on 95,201 CTH orders in addition to a prospective trial described below. We retrieved time stamps for image order, acquisition, and interpretation for 95,201 CTH orders from the MSH PACS system. We excluded studies that were ordered over 24 h before acquisition and images with order time stamps after the scan was acquired. Similarly, we excluded scans that were interpreted over 24 h after acquisition and images interpreted before the acquisition timestamp. Because the algorithm is designed to run at image acquisition, we used the time of image acquisition as $t=0$ and calculated the latency in the image workflow and associated the ordering clinician's indication of whether the order was urgent or high priority.

Radiologist-interpretation RCT. To obtain the most clinically meaningful metric as to whether a deep-learning-based image-triage system can improve radiologist performance and decrease the time to notification of urgent findings, we performed a prospective trial involving a simulated clinical workflow with our enqueuing system. The delay between image acquisition and radiologist documentation includes the time before the radiologist looked at the image, image interpretation, and image documentation; however, when critical findings are found, typically a call is placed to the appropriate clinical team. Therefore we assessed how long radiologists spent on the image-interpretation step itself, and we attempted to quantify the delay between when the radiologists would contact the ordering clinician for critical findings, rather than being finished with documentation (Supplementary Fig. 1).

The 180-image test set was randomly split into two sets of work queues with ten images per queue. One set of work queues was ordered randomly, to reflect obtaining scans in chronological order, and the other was prioritized by the deep-learning system's predicted probability of critical findings. The queues were masked with md5 hashes for a double-blinded image interpretation and analysis. Two radiologists were provided with the instructions included in Supplementary Fig. 1, and they each spent a maximum of 30 min interpreting images from each queue. The radiologists were instructed to open the next study in the queue and to interpret it as they normally would with access to prior imaging, clinical information, and the study images themselves. As soon as they noticed a finding that they deemed critical, they recorded the time stamp when they would have placed a call. Once they had 'called' the referring physician, the radiologists were asked to complete a dictation of the current study, and a time stamp was again recorded when dictation was complete. After completion of their dictation, they proceeded to the next study. After they completed a session, they took a short break and then performed another session on a simulated queue. Using these three time points gathered across a total of ten sessions, we calculated the time interval required for a radiologist to interpret a study and arrive at a clinical finding, and we evaluated the effect of deep-learning-based enqueuing on the overall time to notification for critical findings on enqueued studies.

Code availability. All code related to this project was written in Python. Custom code related to the image extraction, preprocessing pipeline, deep-learning model builder, data provider, and experimenter driver will be made available at https://github.com/aisinai/.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability. The imaging studies used for algorithm development are not publicly available, because they contain protected patient health information. Derived data supporting the findings of this study are available from the corresponding author on reasonable request.

References

40. Zech, J. et al. Natural language-based machine learning models for the annotation of clinical radiology reports. *Radiology* **287**, 570–580 (2018).
41. Viertel, V. G. et al. Reporting of critical findings in neuroradiology. *AJR Am. J. Roentgenol.* **200**, 1132–1137 (2013).
42. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. in *2015 IEEE International Conference on Computer Vision (ICCV)* (IEEE, Piscataway, NJ, USA, 2015)

nature research

Corresponding author(s): Eric K Oermann, M.D.

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   Sample sizes were not pre-determined, but represented the largest samples that we could pull from our PACs system utilizing the tools available to use when the query was made. We elected to obtain as large a dataset as possible under the general premise that deep learning systems perform better with more data.

   The sample of non-contrast CT head studies utilized for training and testing of the algorithms described in the present study represents the largest labeled collection of such studies compiled to date. The studies were acquired following search for relevant study descriptions of the picture archiving and communication system at our institution within the IRB-approved dates.

   For the natural language processing (NLP) algorithm verification, 1004 reports were sampled randomly from each year of the 10 year period approved for inclusion in the study and were annotated by three physicians with majority vote deciding discrepancies. Details regarding the NLP pipeline have been published in Zech et al, Radiology, 2018.

   For convolution neural network (CNN) testing, 180 studies were randomly sampled from the 37,236 imaging studies that had been previously labeled by the NLP algorithm until an approximate two-to-one split of critical to non-critical studies were obtained. These cases were subsequently review by physicians to obtain gold-standard labels of the imaging pathology. This review included the entire medical record including note review, prior imaging, and follow-up imaging studies.

2. Data exclusions

   Describe any data exclusions.

   None

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   The CNN demonstrated consistent performance in the prediction of gold-standard critical findings with an area under the curve of 0.73. These results were consistent across multiple folds of training. When tested prospectively in a simulated clinical trial, the CNN was able to alert the interpreting radiologist to 50% of critical cases with a false alarm rate of 21%. The clinical simulation also demonstrated that the CNN was able to significantly reduce time to notification from minutes to seconds.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   For natural language processing (NLP) algorithm verification, 1004 reports were sampled randomly from each year of the 10 year period approved for inclusion in the study. This was done to ensure that variations in institutional reporting standards and personnel over the study period were included in the test set.

   For convolution neural network (CNN) testing, 180 studies were randomly sampled from the 37,236 imaging studies that had been previously labeled by the NLP algorithm until an approximate two-to-one split of critical to non-critical studies were obtained.

For CNN training and validation, CT studies were allocated to the validation and training data sets randomly. We repeated this randomization five times for each of the folds and subsequently integrated them into the ensemble model.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

When comparing neural network performance to physicians, blinding was performed by restricting the physicians ability to study the clinical scenario associated with a study and by restricting image manipulation capabilities normally available to physicians working in practice. A description of blinding procedures is included in the supplemental methods section.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	☐ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

The NLP algorithms utilized for silver-standard label generation are described in detail in Zech et al, Radiology, 2018. In brief, we tested multiple NLP approaches to CT head report label generation and utilized the best performing model (BOW with unigrams, bigrams, and trigrams plus average word embeddings vector) to generate the silver-standard labels utilized for subsequent CNN training and validation.

We developed a novel 3D CNN intended to identify critical findings on noncontrast CT head studies. The architectures tested and the ensemble of CNNs trained are described in detail in the supplemental methods section. All deep learning preprocessing, training, and inference were performed in C++ and Python. All models were trained on a pair of nVidia GTX 1080s using Tensorflow 1.0.2 and Keras 2.0.1.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	The imaging studies used to train and test the neural network described in the manuscript are subject to privacy regulations and cannot be made available in totality. Subsets of the imaging studies can be made available following verification of appropriate deidentification. If accepted for publication, our group would make the code discussed in the manuscript publicly available.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No eukaryotic cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animal were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Demographics related to the 37,236 included patients are as follows: mean (standard deviation) age 61.5 ± 21.0, female sex 52%, male sex 48%. The noncontrast CT Head studies included normal studies as well as a broad range of pathologies including intracranial hemorrhage, stroke, hydrocephalus, mass lesions, fractures, and post-operative changes. The revised manuscript contains detailed tables documenting the schema for critical finding identification as well as demographic data and clinical presentation descriptions for the patients included in the gold-standard set of studies.