공학석사 학위논문

응급 뇌 CT 트리아제를 위한 생성 모델 기반 이상치 검출 알고리즘

Emergency triage of brain computed tomography via anomaly detection with a deep generative model

울산대학교대학원

의공학과

이승준

Emergency triage of brain computed tomography via anomaly detection with a deep generative model

지도교수 김남국

이 논문을 공학 석사 학위 논문으로 제출함

2022년 08월

울산대학교대학원

의공학과

이승준

이승준의 공학석사학위 논문을 인준함

심사위원 홍 길 선 인

심사위원 김 남 국 인

심사위원 이 준 구 인

울 산 대 학 교 대 학 원

2022 년 08 월

감사의 글

많은 분들의 도움으로 석사 과정을 무사히 마칠 수 있었습니다. 이 글을 통해 감사 인사를 드리고자 합니다.

연구의 큰 그림을 그릴 수 있게 방향을 지도해주시고 연구를 아낌없이 지원해주신 김남국 교수님과 홍길선 교수님께 감사드립니다. 항상 성심성의껏 지도해주셔서 연구자로서 교수님들 덕분에 좋은 연구를 할 수 있었고 많이 배울 수 있었습니다.

바쁜 연구 일정에도 저에게 많은 격려와 도움을 주었던 연구실 선생님들께 모두 감사드립니다. 인생 선배로서 그리고 같은 연구실 동료로서 항상 옆에서 많은 격려를 해주고 조언을 주었던 김동원 박사님 고맙습니다.

항상 저를 믿어주시고 묵묵히 응원해주는 우리 가족 모두 사랑하고 고맙습니다. 우리 어머니, 지혜자, 우리 아버지, 이태진, 그리고 우리 형, 이승재, 모두 고맙습니다.

지면으로 미처 언급하지 못했지만, 저를 아끼고 격려해주셨던 모든 분들께 진심으로 감사하다는 말씀을 전합니다. 더욱 정진하며 바른 모습으로 한층 성장해 공헌할 수 있는 존재가 되도록 노력하겠습니다.

국문요약

트리아제(triage)는 신경학적 응급 상황의 조기 진단과 보고에 필수적이다. 본 연구는 건강한 개인들의 뇌 컴퓨터 단층 촬영 (CT) 영상 데이터를 훈련 데이터로, 응급 뇌 CT 영상을 판독 리스트에서 재정렬하고 이상 영역을 나타내는 심층 생성 모델 기반의 이상 검출 알고리즘을 개발하였다. 응급 환자 감지 성능에 대한 내부 및 외부 검증으로 AUC (95% 신뢰 구간)는 각각 0.85 (0.81–0.89)과 0.87 (0.85–0.89)이었다. 응급실 코호트 임상 시뮬레이션 테스트 결과, 트리아제 시스템 적용 전후로 응급 환자의 대기 시간 중위값 294 초 (422.5 초 [사분위 범위 299] - 70.5 초 [사분위 범위 168])만큼 유의하게 줄었으며, 방사선 전문의의 보고 처리 시간 중위값은 297.5 초 (445.0 초 [사분위 범위 298] - 88.5 초 [사분위 범위 179])로 유의하게 빨라졌다 ($p < 0.001$).

차례

그림목차

서론

Neurological emergencies should be diagnosed and treated as soon as possible to reduce mortality and morbidity rates and to enhance functional outcomes[1–3]. For the initial screening and diagnosis of neurological conditions, non-contrast brain computed tomography (CT) is the current standard imaging modality. In this regard, radiology worklist reprioritization based on image findings is critical in the emergency department (ED).

With the excellent achievements of deep learning in various radiological tasks, several studies have demonstrated that deep learning-based radiological triage can improve radiology workflow efficiency, accelerate radiology reporting, and enable timely management of patients with critical findings (e.g., intracranial hemorrhage or large vessel occlusion on brain images) [4–7]. However, data-related problems have restricted the broad clinical application of deep learning. The construction of large-scale annotated training datasets across diverse populations, disease entities from common to rare, medical centers, and acquisition protocols has remained a significant obstacle to developing a deep learning system in medicine. In addition, the clinical efficacy of supervised deep learning models has been validated only in selected patients with the risk of having a single disease or a few specific diseases. Therefore, this approach cannot guarantee that deep learning can cope with new or previously unseen conditions. As a result, the clinical applicability of supervised deep learning with a narrow clinical focus has been limited.

Recently, pilot studies have shown that deep generative models trained on normal data can detect anomalies[8–13]. Deep generative models learn to capture target data distribution; hence, they can detect anomalous data that deviate from the target distribution without prior knowledge of anomalies. Moreover, the anomaly detection framework based on deep generative models can visually highlight the model's prediction using reconstruction error. Although previous studies using this framework have attracted considerable attention, they have two limitations: 1) lack of external and clinical validation tests (hence, whether the model can be generalized to real-world situations cannot be guaranteed) and 2) no clinical utility test of them.

This study aimed to develop and validate an anomaly detection algorithm (ADA) based on a deep generative model trained only with normal brain CT images and investigate the clinical impact of an ADA-based triage system on ED radiology workflow using a randomized crossover clinical simulation test. Importantly, this study aimed to assess the real-world performance of the ADA using brain CT images in internal and external ED screening cohorts.

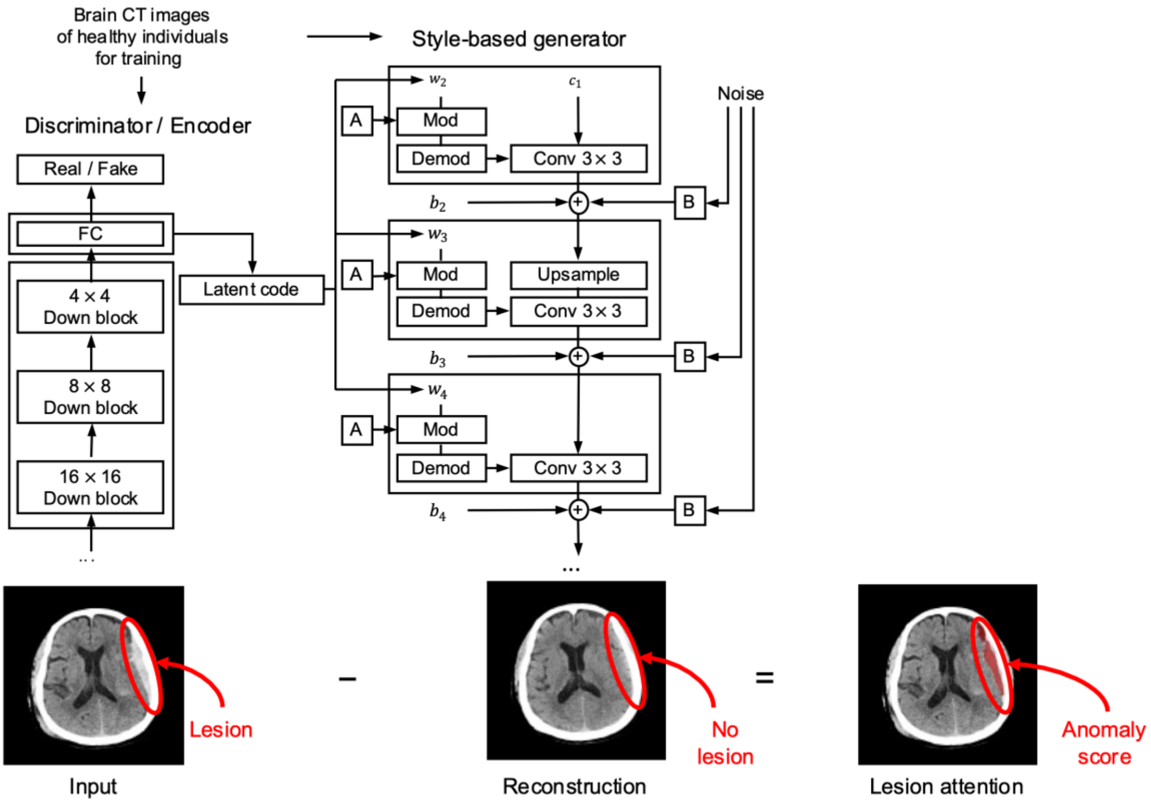


Fig. 1: Our proposed anomaly detection framework based on a deep generative model— closest normal style-based generative adversarial network (CN-StyleGAN)—trained on normal brain CT images. The model reconstructs a brain CT image as its closest normal-style brain CT image. Based on the density error, the anomaly score of the scan is determined and used to identify emergency cases, followed by visualization of the detected lesion upon overlay.

We developed an ADA based on a deep generative model called the closest normal style-based generative adversarial network (CN-StyleGAN). CN-StyleGAN was closely modeled after StyleGAN2. CN-StyleGAN comprised three deep neural networks: a style-based generator (G), discriminator (D), and style-based encoder (E). We used the same architecture as StyleGAN2 for G and D; E followed the architecture of D, although the last fully connected layer was modified to output an 8192-dimensional latent code, $\mathbf{w} \in \mathbf{W}^{+}$, followed by a leaky ReLU of $\alpha = 0.2$[14]. Given a brain CT image as an input, E encodes the image into the closest normal-style latent code, and G generates the closest normal-style brain CT image from the latent code, trying to fool D by making the generated image indistinguishable from the true image. Then, D tries to discriminate the generated image from the true image. Using brain CT images from healthy individuals, CN-StyleGAN was trained to reconstruct a scan into the closest normal-style scan. The density error between the actual scan and the reconstructed scan was used to determine the anomaly score of the scan to identify emergency cases. Cases identified as emergency cases were reprioritized based on their anomaly scores in the radiology worklist as well as the visualization of the predicted lesions (Fig. 1).

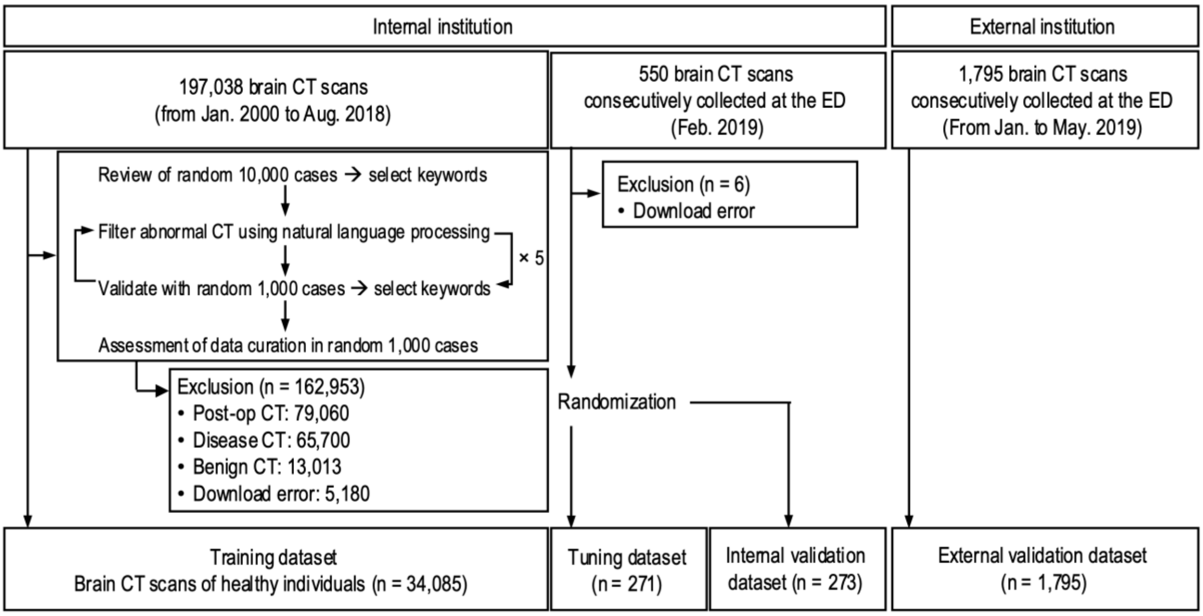


Fig. 2: Data flow diagram of the collection and curation process of the training, tuning, internal validation, and external validation datasets. The training dataset was collected and curated to include brain CT scans from healthy individuals by reviewing and applying NLP algorithms to radiological reports. In addition, consecutive brain CT scans from individuals who underwent emergency screening for suspected neurological conditions in the EDs of the internal and external institutions were independently and retrospectively collected. The internal dataset was randomly divided into two parts: a tuning dataset and an internal validation dataset. The external validation dataset included brain CT scans from 1,795 consecutive individuals who had visited the ED of the external institution for five months.

Table 1. Baseline characteristics of the patients and image acquisition information according to the training, tuning, and validation sets. Data are presented as the mean ± standard deviation or number of cases (%). Abbreviations: CT, computed tomography

Characteristics	**Training**	**Tuning**	**Internal validation**	**External validation**
Sample size	34,085	271	273	1,795
Age (year)	42.9 ± 19.6	58.1 ± 18.0	59.1 ± 17.6	60.3 ± 19.3
Sex				
Female	18,232 (53.5%)	143 (52.8%)	137 (50.2%)	875 (48.7%)
Male	15,853 (46.5%)	128 (47.2%)	136 (49.8%)	920 (51.3%)
Emergency				

severity				
Immediate		18 (6.6%)	18 (6.6%)	80 (4.5%)
Urgent		22 (8.1%)	23 (8.4%)	117 (6.5%)
Indeterminate		10 (3.7%)	10 (3.7%)	50 (2.8%)
Benign		59 (21.8%)	60 (22.0%)	436 (24.3%)
Normal		162 (59.8%)	162 (59.3%)	1,112 (61.9%)
Diseases in the emergency group		40	41	197
Brain mass-like lesion		10 (25%)	16 (39.0%)	20 (10.2%)
Acute infarction		6 (15%)	3 (7.3%)	39 (19.8%)
Hemorrhage		19 (47.5%)	18 (43.9%)	128 (65.0%)
Hydrocephalus		4 (10%)	2 (4.9%)	6 (3.0%)
Other diseases		1 (2.5%)	2 (4.9%)	4 (2.0%)
CT scanner	Siemens Healthcare (n = 19, 420) · Definition · SOMATOM Definition · SOMATOM Definition Flash · Definition AS · SOMATOM Definition Edge · SOMATOM Force · SOMATOM Definition AS+ · SOMATOM Definition AS · Sensation 16 GE Healthcare (n = 14,656) · LifeSpeed Plus	Siemens Healthcare (n = 268) · SOMATOM Definition Edge · SOMATOM Definition Flash · SOMATOM Definition AS+ GE Healthcare (n = 3) · Discovery CT750 HD · LightSpeed VCT	Siemens Healthcare (n = 270) · SOMATOM Definition Edge · SOMATOM Force · SOMATOM Definition AS+ GE Healthcare (n = 3) · Discovery CT750 HD · LightSpeed VCT	Siemens Healthcare (n = 1751) · SOMATOM Definition Edge · SOMATOM Scope GE Healthcare (n = 44) · LightSpeed 16 · LightSpeed VCT

	· LightSpeed QX/i · HiSpeed CT/i · Optima CT660 · Discovery CT750 HD · LightSpeed VCT · LightSpeed16 Neurologica (n = 9) · CereTom			
Slice thickness (mm)	· 4.8 (n = 14,316) · 5 (n = 19,676) · 10 (n = 93)	· 5 (n = 271)	· 5 (n = 273)	· 4.8 (n = 1,065) · 5 (n = 730)

Fig. 2 and Table 1 summarize the data collection, baseline characteristics, and image acquisition information for the datasets. For the development of CN-StyleGAN, a total of 197,038 non-contrast brain CT scans and paired radiology reports were retrospectively collected from patients who visited an urban, tertiary, academic hospital between January 1, 2000, and August 31, 2018. After iterations of the data curation process, the training dataset comprised 34,085 normal brain CT scans from healthy patients. In detail, the data curation process included three steps. First, we reviewed the radiology reports from 10,000 randomly sampled CT scans and selected keywords for anomalous CT findings such as positive pathological findings, benign lesions, and postoperative changes. Second, a natural language processing (NLP) algorithm (PyConTextNLP[15]) was used to exclude anomalous brain CT scans based on these keywords. Finally, two radiologists (with 14 years and four years of experience in reading brain CT images, respectively) randomly selected 1,000 CT scans and reviewed their radiology reports. If anomalous CT scans were found during this step, additional keywords were added. This data curation cycle was repeated five times to obtain completely normal CT scans. A total of 79,060 postoperative CT scans and 78,713 abnormal CT scans were excluded. Finally, the NLP-based data curation was assessed by manually reviewing the radiology reports of 1,000 randomly selected cases. Of the 39,265 potentially eligible cases, CT scans from 5,180 cases were not available for automatic downloading using the in-house system. Finally, non-contrast brain CT scans from 34,085 healthy individuals (mean age ± standard deviation [SD]: 42.9 ± 19.6 years; female: 18,232 [53.5%]) were retrospectively collected from a tertiary academic hospital for the training dataset.

Furthermore, the brain CT scans were collected independently and retrospectively from consecutive individuals who underwent emergency screening for suspected neurological conditions in the EDs of an internal and an external institution. For the tuning and internal validation test, after six cases were excluded due to download errors, 544 non-contrast brain CT scans of ED patients (mean age ± SD: 58.6 ± 17.8 years; women: 280 [51.5%]) were consecutively collected from Asan Medical Center in February 2019. The internal dataset was subsequently randomly divided into two parts: a tuning dataset and an internal validation dataset, and the ratio of each emergency severity group was preserved. For the external validation test, 1,795 non-contrast brain CT scans from ED patients (mean age ± SD: 60.3 ± 19.3 years; female: 875 [48.7%]) were consecutively collected from Gangneung Asan Hospital from January 1, 2019, to May 31, 2019. A board-certified emergency radiologist (with 14 years of experience reading brain CT images) reviewed all CT images in the internal and external validation datasets and classified the cases according to the category system for emergency severity[16–18]. This system categorized the cases into the following categories based on the urgency of treatment: normal, benign, indeterminate, urgent, and immediate (Table 2). Subsequently, both urgent and immediate cases were defined as emergency cases that required emergency intervention, regardless of the neurological entity. Cases of a critical, life-threatening condition that required immediate medical or surgical treatment were defined as immediate cases. Cases that were not life-threatening currently but required rapid treatment because they could deteriorate were defined as urgent cases. The disease entities in the emergency cases were categorized as brain mass-like lesions, acute infarctions, intracranial hemorrhages, hydrocephalus, and other diseases. A brain mass-like lesion was defined as a volumetric space-occupying lesion (e.g., brain tumor, brain abscess, tumefactive demyelinating disease, or encephalitis) distinct from the brain parenchyma with a normal appearance. The emergency cases accounted for 15.0% (41 of 273) and 11.0% (197 of 1,795) of the internal and external validation datasets, respectively. Disease entities from the internal and external validation datasets included brain mass-like lesions (39.0% [16 of 41] vs. 10.2% [20 of 197]), acute infarctions (7.3% [3 of 41] vs. 19.8% [39 of 197]), intracranial hemorrhage (43.9% [18 of 41] vs. 65.0% [128 of 197]), hydrocephalus (4.9% [2 of 41] vs. 3.0% [6 of 197]), and other diseases (4.9% [2 of 41]] vs. 2.0% [4 of 197]).

Table 2. Emergency severity categories according to brain CT findings

Categories		Definition
Emergency	**Immediate**	**CT findings suggest a critical, life-threatening condition that requires immediate medical or surgical treatment.**
		· Brain tumor with a mass effect resulting in midline shift and herniation · Intracranial hemorrhage with a mass effect resulting in midline shift and herniation · Extensive subarachnoid hemorrhage · Hypoxic encephalopathy · Large territorial or malignant acute infarction
	Urgent	**CT findings suggest a current non-life-threatening condition that requires rapid treatment to prevent deterioration.**
		· Intracranial hemorrhage without a mass effect · Focal acute infarction · Tumor without a mass effect, such as midline shift and herniation · Marked hydrocephalus (Evans' index > 0.4) · Unruptured giant aneurysm
Non-emergency	**Indeterminate**	**CT findings suggest that prompt treatment is not required but further workup or follow-up is required.**
		· Indeterminate small hypodense cerebral lesions · Incidental pituitary adenoma, small meningioma, or suspected small aneurysm · Hydrocephalus (0.4 ≥ Evans' index > 0.34)
	Benign	**CT findings suggest that no further workup is required in the emergency department.**
		· Severe brain atrophy, arachnoid cyst, encephalomalacia, leukoaraiosis, or postoperative change
	Normal	**Normal**

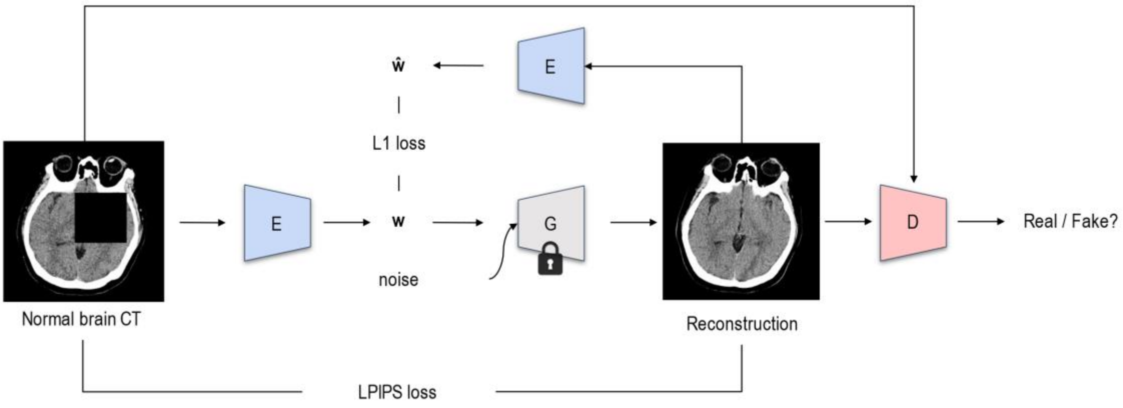


Fig. 3: Training process of CN-StyleGAN.

The architecture of CN-StyleGAN includes a style-based generator (**G)**, discriminator (**D)**, and style-based encoder neural network (**E**). The training dataset containing normal brain CT axial slices was used to jointly train **D** and **E,** while the pre-trained weights of **G** were kept constant. The model learned to reconstruct a query brain CT image as the closest-normal brain CT image. Note that the input brain CT image was randomly erased for the model to learn the context of normal brain CT images by filling in the missing region.

Training. Fig. 3 illustrates the training process of CN-StyleGAN. We trained CN-StyleGAN using normal brain CT images and several training processes for the model to encode the style of normal brain CT images. First, we trained G and D for 160,000 iterations following the original training process of StyleGAN2. Subsequently, we trained E and D but not G with loss functions including VGG16-based learned perceptual image patch similarity (LPIPS) loss[19,20], domain-guided loss[21], and adversarial loss functions from StyleGAN2. LPIPS loss measured the discrepancy between real images $(\mathbf{x})$ and reconstructed images $\left(\mathrm{G}\big(\mathrm{E}(\mathbf{x})\big)\right)$ in the feature space of VGG16. To improve the performance and increase the stability, we downsampled the images to a resolution of 256 × 256 pixels before computing the LPIPS distance. The domain-guided loss measured the L1 distance between $\mathrm{E}(\mathbf{x})$ and $\mathrm{E}\left(\mathrm{G}\big(\mathrm{E}(\mathbf{x})\big)\right)$ for the in-domain property, regularizing the latent code to be inside the latent space of the normal brain CT data distribution. For adversarial loss, non-saturating loss[22] was used with R1-regularization[23] at every 16th step to stabilize the training of D. After adversarial training, the reconstructed images were indistinguishable from the normal brain CT images. Furthermore, random erasing of brain CT images[24] was used so that E could learn the semantics of normal brain CT images by filling in the missing region. We trained the model in PyTorch[25] with the Adam optimizer[26] for 200,000 iterations with hyper-parameters ($\beta_1 = 0$, $\beta_2 = 0.99$, $\varepsilon = 10^{-8}$, and minibatch = 32). The learning rate was 10^{-5} for the E and 10^{-6} for the D.

Gaussianized latent space. Previous studies on StyleGAN have indicated that data distribution can be explicitly modeled as a normal distribution in the intermediate latent space of StyleGAN[27,28]. Similarly, we explicitly modeled the data distribution of normal brain CT images in the intermediate latent space. We used E to map each normal brain CT image, slice-by-slice, from the training data to the latent space and used the latent codes to estimate the sample statistics for each slice order. Thus, the empirical covariance matrices, $\sum$, and means, $\boldsymbol{\mu}$, were accumulated for each layer of the intermediate latent space.

Inference. Fig. 4 illustrates the inference method and anomaly scoring system of CN-StyleGAN. A CT scan included up to 32 axial slices from the bottom to the top. We initialized the latent code, $\boldsymbol{w_{init}}$, for each axial slice $(\mathbf{x})$ of the scan as $\mathrm{E}(\mathbf{x})$ and the noise maps $(\mathbf{n})$ from a normal distribution. We Gaussianized and optimized the latent code $(\mathbf{w})$ with L1, LPIPS, and the in-domain loss functions using the Adam optimizer for 100 epochs. Furthermore, the in-domain loss was modified to regularize the latent vector in the Gaussianized latent space only when the latent code deviated from the mean of the data distribution of normal brain CT images in the latent space compared with the in-domain latent code, $\mathrm{E}\big(\mathrm{G}(\mathbf{x})\big)$. After the latent code was optimized as $\mathbf{w}^*$, we optimized the noise maps with the L1 loss function for 100 iterations. Noise maps can be optimized to generate out-of-domain images[29]; therefore, we proposed a masked noise optimization that forced the model to reconstruct the normal region alone. At each optimization step, a binary mask, $\mathbf{M}$, was defined to predict the lesion area in the scan. To calculate $\mathbf{M}$, the residual difference between an image $(\mathbf{x})$ and the reconstructed image $\big(\mathrm{G}(\mathbf{w}^*, \mathbf{n})\big)$ was brain-extracted[30], median-filtered with a window size of 17, and thresholded by 5 Hounsfield units. Moreover, the number of false positives in $\mathbf{M}$ decreased because of the intersections of binary masks at the previous optimization steps. Consequently, $\mathbf{M}$ was used to set a target image for optimization:

$$\mathbf{x}_{\mathrm{target}} = \mathbf{M} \odot \mathrm{G}(\mathbf{w}^*, \mathbf{n}_{\mathrm{init}}) + (\mathbf{1} - \mathbf{M}) \odot \mathbf{x} \qquad (1)$$

where $\odot$ denotes a pointwise multiplication. At the last optimization step, the binary mask was used as the lesion attention map for prediction.

Anomaly score. The anomaly score was calculated as follows: first, reconstruction error for a slice $\mathbf{x}_i$ of a scan was defined as:

$$\mathrm{R}(\mathbf{x_i}) = \left\|\mathbf{M} \odot \big(\, \mathbf{x}_{\mathrm{i}} - \mathrm{G}(\mathbf{w}, \mathbf{n})\big)\right\| \qquad (2)$$

which is the binary masked density error between the slice, $\mathbf{x}_i$, and the reconstructed slice, $\mathrm{G}(\mathbf{w}, \mathbf{n})$. Second, this reconstruction error was normalized, slice-by-slice, based on the slice order, using the reconstruction error statistics of the mean, $\mathbf{R}_{\boldsymbol{\mu}}$, and SD, $\mathbf{R}_{\boldsymbol{\sigma}}$, of the normal brain CT images. A total of 1,000 scans were randomly selected from the training dataset for the normal brain reconstruction error statistics. Finally, this normalized per-slice reconstruction error of 32 slices for the scan was summed to obtain the anomaly score:

$$\text{Anomaly score} = \sum_{i=1}^{32} \frac{\mathbf{R(x}_i\mathbf{)} - \mathbf{R}_{\mu_i}}{\mathbf{R}_{\sigma_i}} \tag{3}$$

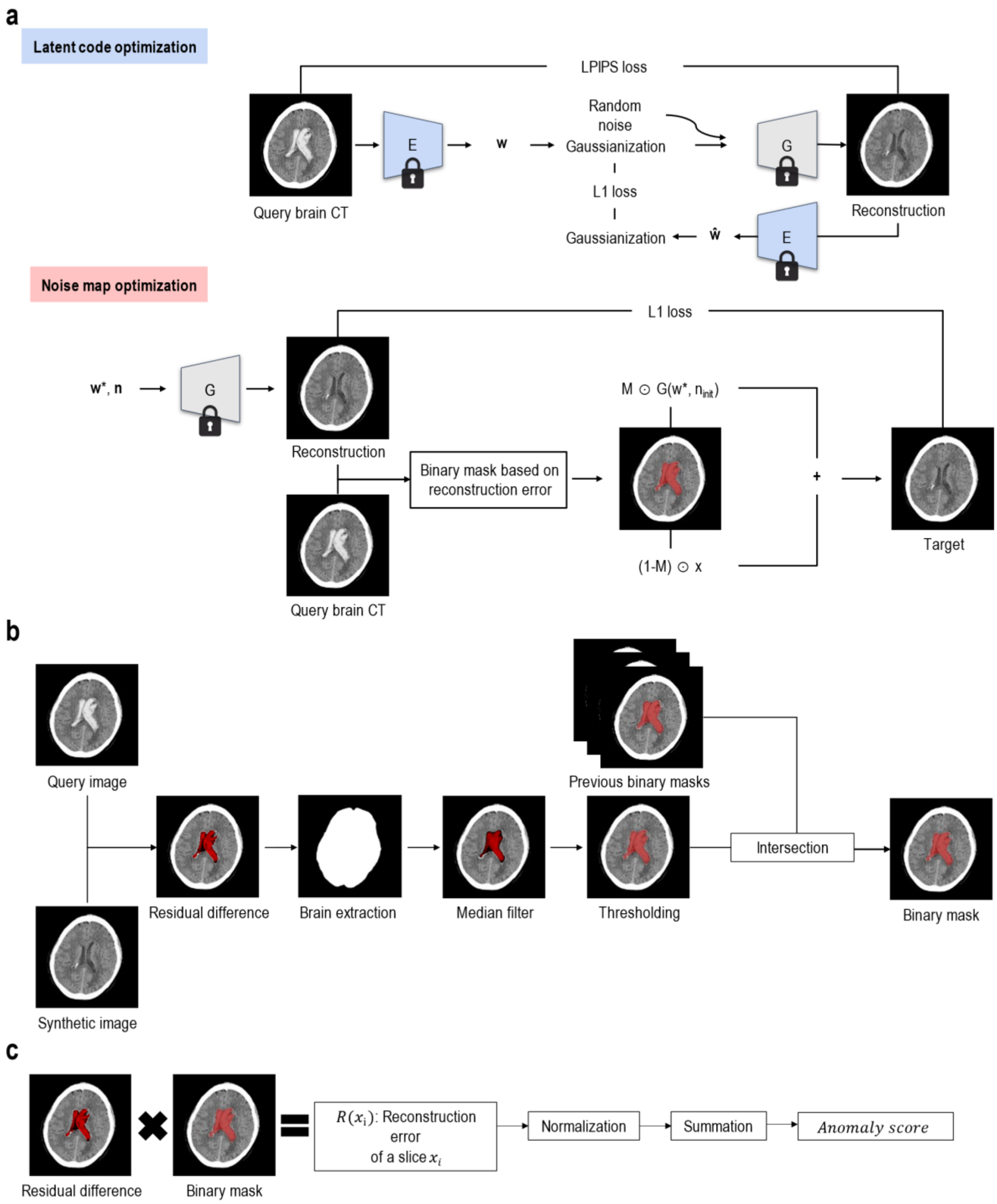


Fig. 4: Inference method and anomaly score system of CN-StyleGAN

This figure demonstrates the inference method and anomaly scoring system of CN-StyleGAN. **a,** The inference method of CN-StyleGAN. Given a brain CT slice, $\boldsymbol{x}$, the latent vector, $\boldsymbol{w}$, was initialized as $E(\boldsymbol{x})$, and the noise maps, $\boldsymbol{n}$, were initialized from the unit normal distribution. After the latent vector optimization, the noise maps were optimized. Masked noise optimization was proposed for the noise map optimization. **b,** Derivation of the binary mask in the masked noise optimization process. The residual difference between an input image and the reconstructed image was brain-extracted, median-filtered, and thresholded. Moreover, the false positives in the binary mask were reduced because of the intersections between the binary masks in the previous optimization steps. **c**, The calculation process of the anomaly scoring system. The reconstruction error of a slice was derived as the binary masked density error between the input slice and the reconstructed slice. This reconstruction error was normalized, slice by slice, based on the slice order, using the reconstruction error statistics (mean and SD of normal brain CT images) from the training dataset. Finally, this normalized per-slice reconstruction error of 32 slices for a CT scan was summed to determine the anomaly score.

Clinical simulation test. To investigate the clinical efficacy of the ADA-based triage system for radiology workflow, a randomized crossover study was performed in two sessions using the external validation dataset by referring to the existing study[33] (Fig. 11a). Two radiologists (each with ≥ 14 years of experience in reading brain CT images) independently and retrospectively performed a clinical simulation test using a washout period and varying reading orders in a crossover design to assess brain CT scans with and without the help of the triage system. Specifically, a total of 1,795 brain CT scans from the external validation dataset were randomized to two groups (group A [898 brain CT scans] and group B [897 radiographs]). Each block enrolled 23 brain CT scans, except for one block in group A that enrolled 24 brain CT scans, as the number of imaging studies (n = 878) in group A could not be divided evenly by 23. In the first session, each reader assessed the brain CT scans in group A without the help of the triage system and those in group B with the help of the triage system. In the second session, each reader assessed the brain CT scans in group A with the help of the triage system and those in group B without the help of the triage system. The first and second sessions were separated by at least two weeks, and the reading order of the blocks was randomized and different for each reading session. Our triage system reprioritized emergency cases based on their anomaly scores and labeled them in red in the worklist to attract the readers' attention. The readers were able to overlay the segmentation mask (lesion attention) predicted by CN-StyleGAN on the brain CT image. The readers interpreted the brain CT images and determined the presence of critical findings in the CT scans using an in-house user interface that provided the radiology worklists of the brain CT scans and their images (Fig. 5). The readers were blinded to the clinical information, imaging reports, and number of emergency cases included in the study.

The clinical efficacy of the ADA was analyzed according to three radiological time metrics based on previous studies[6,31,32]: wait time (WT; the time required to open a CT for image review from the beginning of one block), radiology report turnaround time (TAT; the time required to report a critical CT finding from the beginning of one block), and reading time (RT; the time between opening and closing a CT) for each case in each block. These time metrics were calculated based on the timepoints in the CT interpretation process, which were automatically recorded by the software. The metrics were calculated for each case in each block and were defined as follows:

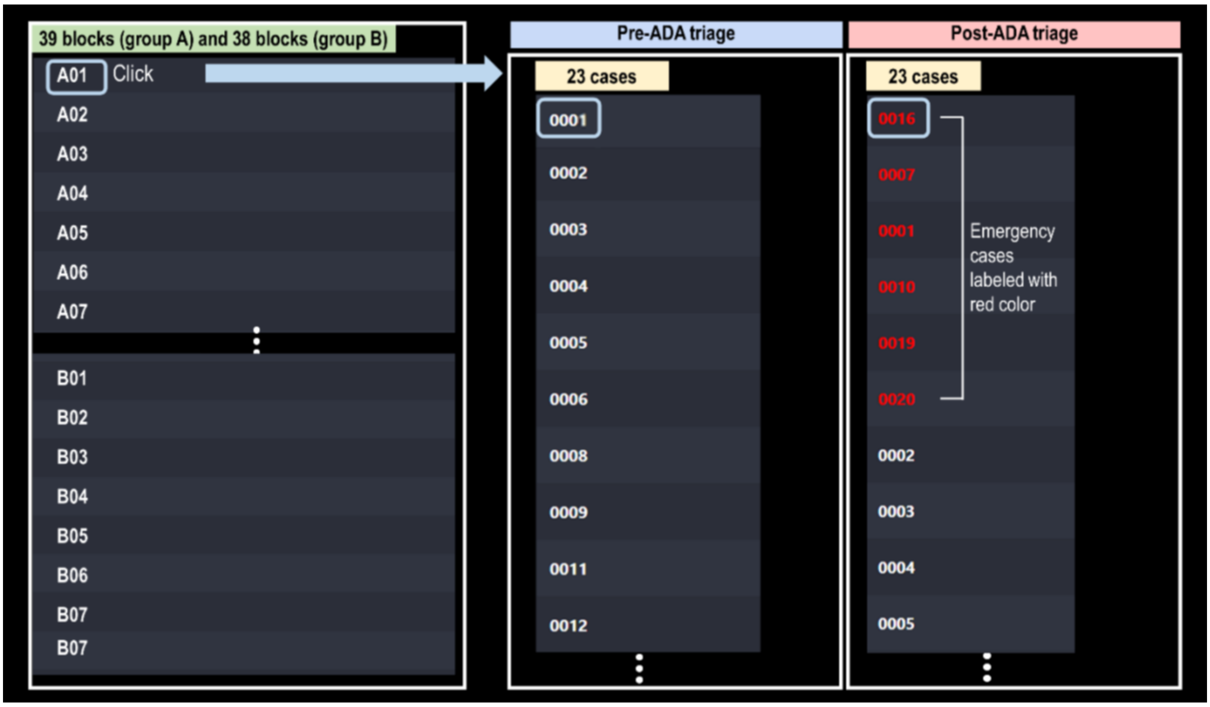
$$\mathrm{WT}(i) = (\text{Timestamp of opening CT}_i) - (\text{Timestamp of opening a block}) \tag{4}$$

$$\mathrm{TAT}(i) = (\text{Timestamp of reporting image findings in CT}_i) - (\text{Timestamp of opening a block}) \tag{5}$$

$$\mathrm{RT}(i) = (\text{Timestamp of closing CT}_i) - (\text{Timestamp of opening CT}_i), \tag{6}$$

where CT_i is the i th CT in a block.

Statistical analyses. The mean values of the anomaly scores between emergency and non-emergency cases were compared using independent *t*-tests. The emergency case detection performance of CN-StyleGAN was analyzed by calculating the AUC, sensitivity, specificity, and accuracy using the internal and external validation datasets. The optimal anomaly score cutoff value was determined from the maximum value of Youden's index for the ROC curve analysis using the tuning dataset. The bootstrap method (10,000 iterations) was used to calculate 95% CIs. The median values of the time factors in the clinical simulation test were compared using the Wilcoxon signed-rank test and Wilcoxon rank-sum test. Analyses were performed using Python version 3.8.5 (sklearn 0.23.2; Python Software Foundation), R version 4.1.0 (R Foundation for Statistical Computing), and ggplot2 version 3.6.3. All statistical tests were two-sided, and the statistical significance was set at $p = 0.05$.



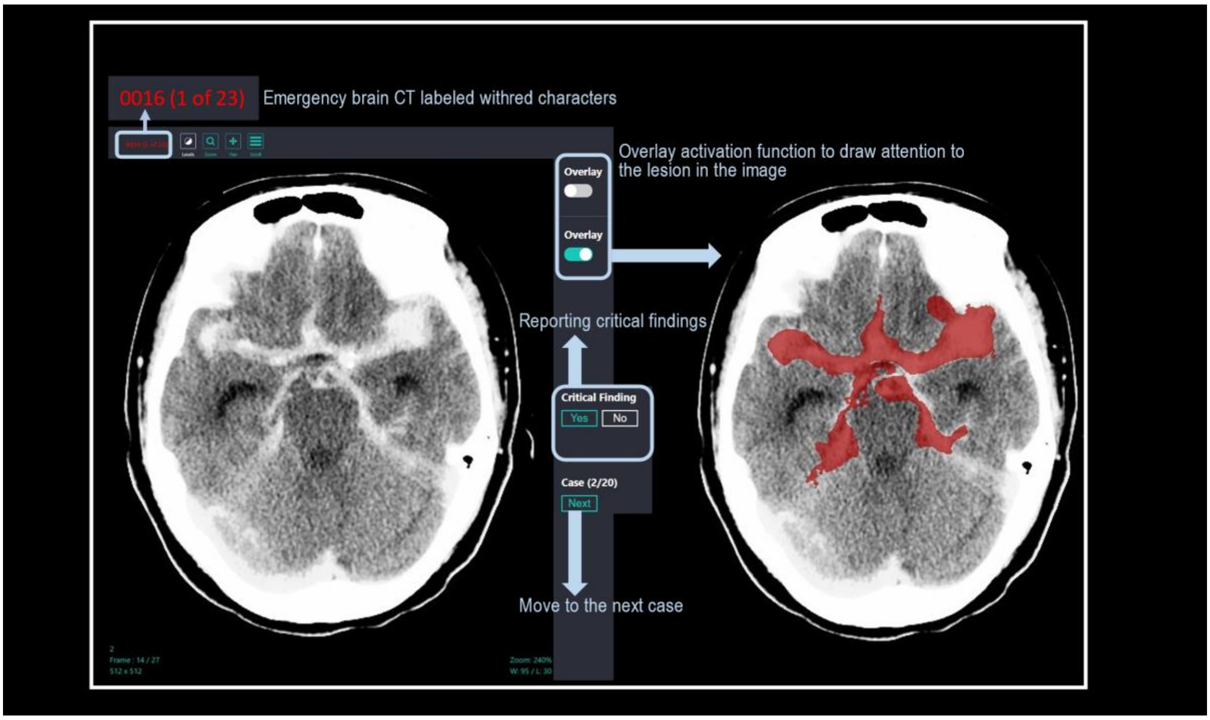

0016 (1 of 23)
Emergency brain CT labeled withred characters
Overlay activation function to draw attention to the lesion in the image
Overlay
Overlay
Reporting critical findings
Critical Finding
Yes
No
Case (2/20)
Next
Move to the next case
2
Frame : 14 / 27
512 x 512
Zoom: 240%
W: 95 / L: 30

Fig. 5: **User interface for the clinical simulation test**

The web-based user interface shown here provides the radiology worklists of brain CT images and displays Digital Imaging and Communications in Medicine (DICOM) images. In the first screen, the readers can select a block in a top-to-bottom order. In the next screen, the worklists of brain CT images in the selected block are observed. The readers can open each brain CT image according to the assigned order. The opening times of each block and each brain CT image are automatically recorded. The readers can adjust the window level of the images and can zoom in to magnify the images. After the readers determine the presence or absence of emergency CT findings, they click the "Critical Finding" button to report the CT findings. If the readers click the button "Next," the user interface will automatically move to the next case. The time is automatically recorded upon clicking the buttons. Pre- ADA triage, the user interface provides worklists of randomly ordered brain CT images. Post-ADA triage, the ADA reprioritizes brain CT images in the worklists and labels emergency cases with red color. The user interface provides the overlay activation function. The readers can see the lesion attention (mask overlay) predicted by the ADA by clicking the button "Overlay."

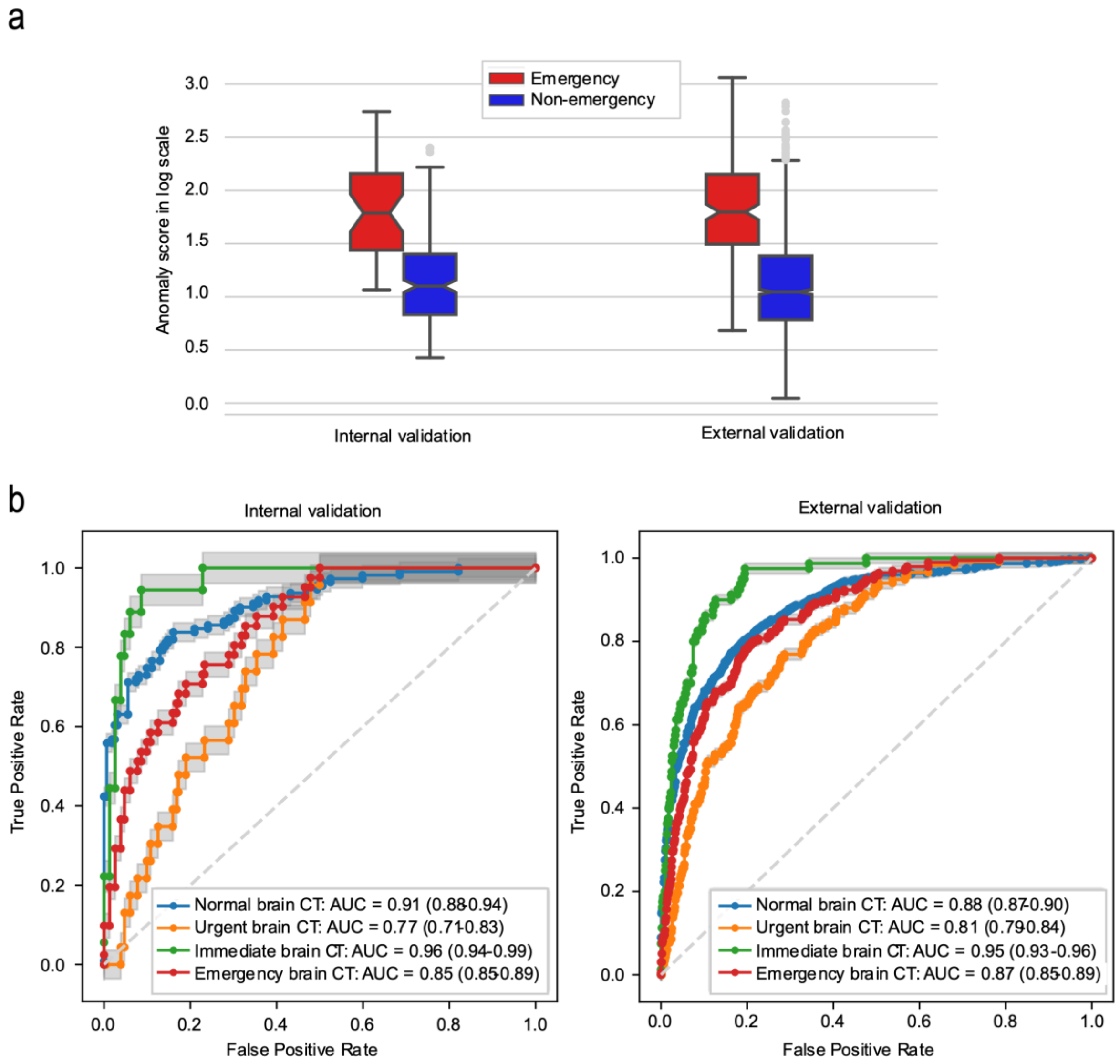


Fig. 6: Detection performance of the ADA for brain CT triage. a In both the internal and external validation tests, the anomaly scores differed significantly between non-emergency (n = 232 for internal validation; n = 1,598 for external validation) and emergency cases (n = 41 for internal validation; n = 197 for external validation) (all p < 0.001). Box plots show the median (center line), first and third quartiles (box edges), and whiskers 1.5 times the IQR. Data points outside the whiskers are considered outliers. Two-sided p value was calculated using independent t-tests. **b** ROC curve analysis for assessing the performance of the ADA according to different target groups in the internal and external validation tests. Date are presented as mean AUC values with 95% CI.

Emergency case detection performance of the ADA. The mean ± SD of the anomaly score was significantly different between the non-emergency and emergency groups in the internal and external validation tests (14.8 ± 36.9 vs. 98.6 ± 119.7, $p < 0.001$, and 14.5 ± 47.3 vs. 118.5 ± 177.3, $p < 0.001$, respectively) (Fig. 6a). The emergency case detection performance of the ADA was analyzed by calculating the area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, specificity, and accuracy with 95% confidence intervals (CIs). The maximum value of Youden's index for the ROC curve analysis using the tuning dataset revealed the optimal anomaly score cutoff value. In the internal and external validation datasets, no data were excluded to reflect real data without sampling bias. Consequently, the AUC, sensitivity, specificity, and accuracy with 95% CIs were 0.85 (0.81–0.89), 0.71 (0.60–0.82), 0.78 (0.74–0.82), and 0.77 (0.73–0.80), respectively, in the internal validation test and 0.87 (0.85–0.89), 0.78 (0.74–0.82), 0.81 (0.80–0.83), and 0.81 (0.80–0.82), respectively, in the external validation test (Fig. 6, Fig. 7, and Table 3). The false negative rates were 29.3% (12 of 41 in the internal validation dataset) and 22.3% (44 of 197 in the external validation dataset). The false positive rates were 22.4% (52 of 232 in the internal validation dataset) and 19.1% (305 of 1,598 in the external validation dataset). For the detection of immediate cases, the ADA achieved the AUC values of 0.96 (0.94–0.99) and 0.95 (0.93–0.96) in the internal and external validation tests, respectively. According to disease entity, the AUC values with 95% CIs in the internal and external validation tests were as follows: brain mass-like lesions, 0.92 (0.88–0.96) vs. 0.92 (0.88–0.96); acute infarctions, 0.91 (0.86–0.95) vs. 0.87 (0.83–0.91); intracranial hemorrhages, 0.78 (0.70–0.85) vs. 0.86 (0.83–0.88); hydrocephalus, 0.82 (0.64–0.97) vs. 0.94 (0.92–0.97); and other diseases, 0.95 (0.91–0.99) vs. 0.80 (0.65–0.94) (Fig. 8). Fig. 9 shows representative cases of various diseases detected as emergency cases by the ADA and lesion attention maps provided by the ADA (see Fig. 10 for representative false-positive and false-negative cases).

Furthermore, sensitivities and specificities (95% CI) were calculated, with the thresholds derived using the tuning dataset at high sensitivity levels of 0.95 and 1.00. At a sensitivity level of 0.95 for the tuning dataset, the sensitivity and specificity were 0.90 (0.83–0.97) and 0.60 (0.56–0.65), respectively, in the internal validation set and 0.89 (0.86–0.92) and 0.63 (0.62–0.65), respectively, in the external validation set. At a sensitivity level of 1.00 for the tuning dataset, the sensitivity and specificity were 1.00 (1.00–1.00) and 0.42 (0.37–0.47), respectively, in the internal validation set and 0.96 (0.95–0.98) and 0.47 (0.45–0.49), respectively, in the external validation set.

Table 3. **Detection performance of CN-StyleGAN according to target severity.** The results are presented as the area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, specificity, and accuracy with 95% confidence intervals. The threshold was derived using the maximum value of Youden's index for the ROC curve using the tuning dataset.

Target group	Performance of CN-StyleGAN			
Internal validation	**AUC**	**Sensitivity**	**Specificity**	**Accuracy**
Normal brain CT	**0.91 (0.88–0.94)**	**0.70 (0.63–0.77)**	**0.94 (0.92–0.97)**	**0.85 (0.82–0.88)**
Urgent brain CT	**0.77 (0.71–0.83)**	**0.52 (0.36–0.69)**	**0.78 (0.74–0.82)**	**0.75 (0.71–0.79)**
Immediate brain CT	**0.96 (0.94–0.99)**	**0.94 (0.88–1.00)**	**0.78 (0.74–0.82)**	**0.79 (0.75–0.83)**
Emergency brain CT (urgent and immediate)	**0.85 (0.81–0.89)**	**0.71 (0.60–0.82)**	**0.78 (0.74–0.82)**	**0.77 (0.73–0.80)**
External validation				
Normal brain CT	**0.88 (0.87–0.90)**	**0.56 (0.54–0.59)**	**0.94 (0.93–0.95)**	**0.80 (0.78–0.81)**
Urgent brain CT	**0.81 (0.79–0.84)**	**0.68 (0.61–0.75)**	**0.81 (0.80–0.83)**	**0.80 (0.79–0.82)**
Immediate brain CT	**0.95 (0.93–0.96)**	**0.93 (0.90–0.99)**	**0.81 (0.80–0.83)**	**0.81 (0.80–0.82)**
Emergency brain CT (urgent and immediate)	**0.87 (0.85–0.89)**	**0.78 (0.74–0.82)**	**0.81 (0.80–0.83)**	**0.81 (0.80–0.82)**

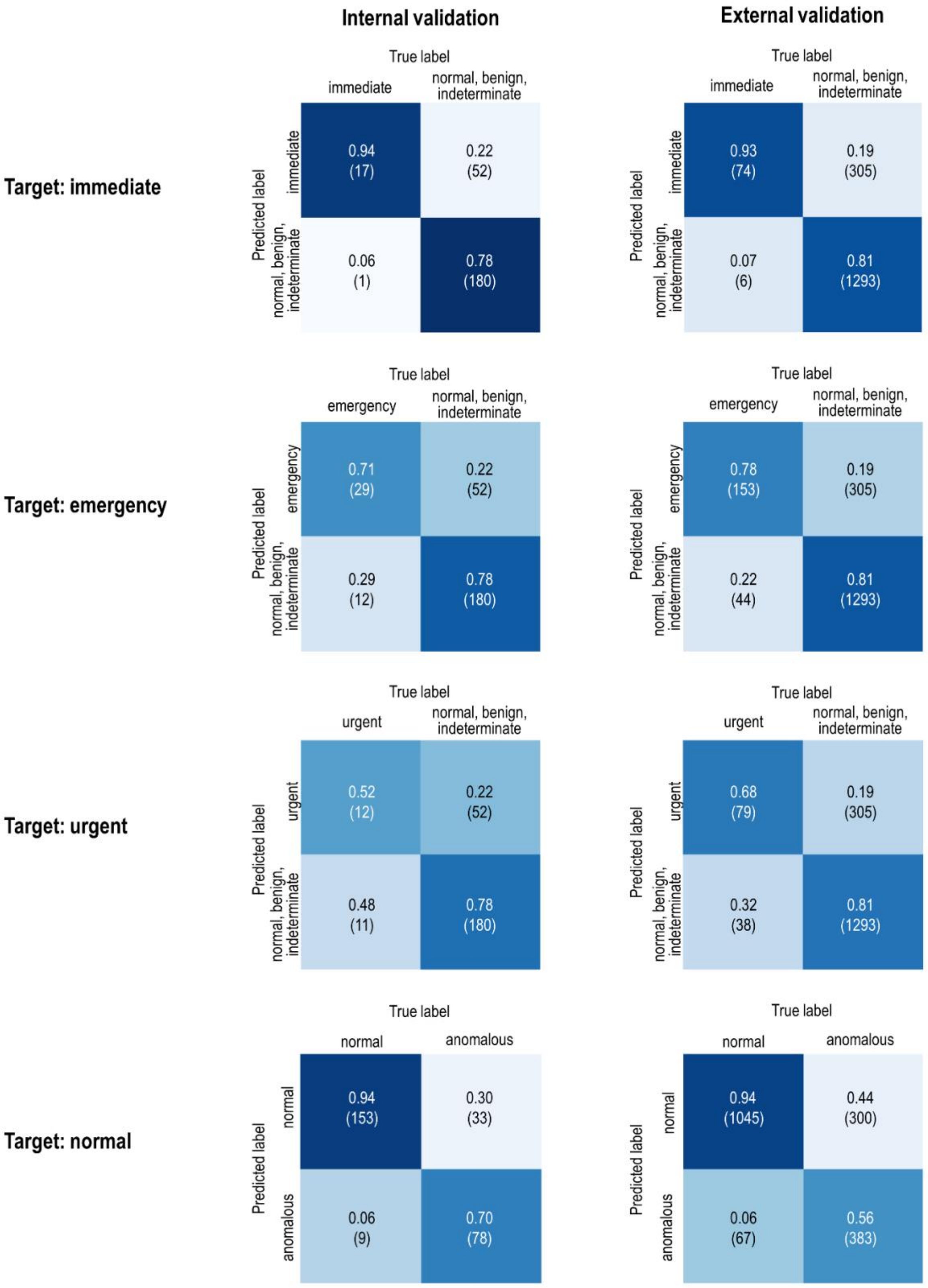


Fig. 7: Confusion matrices of the classification of brain CTs based on emergency severity for assessing the performance of CN-StyleGAN.

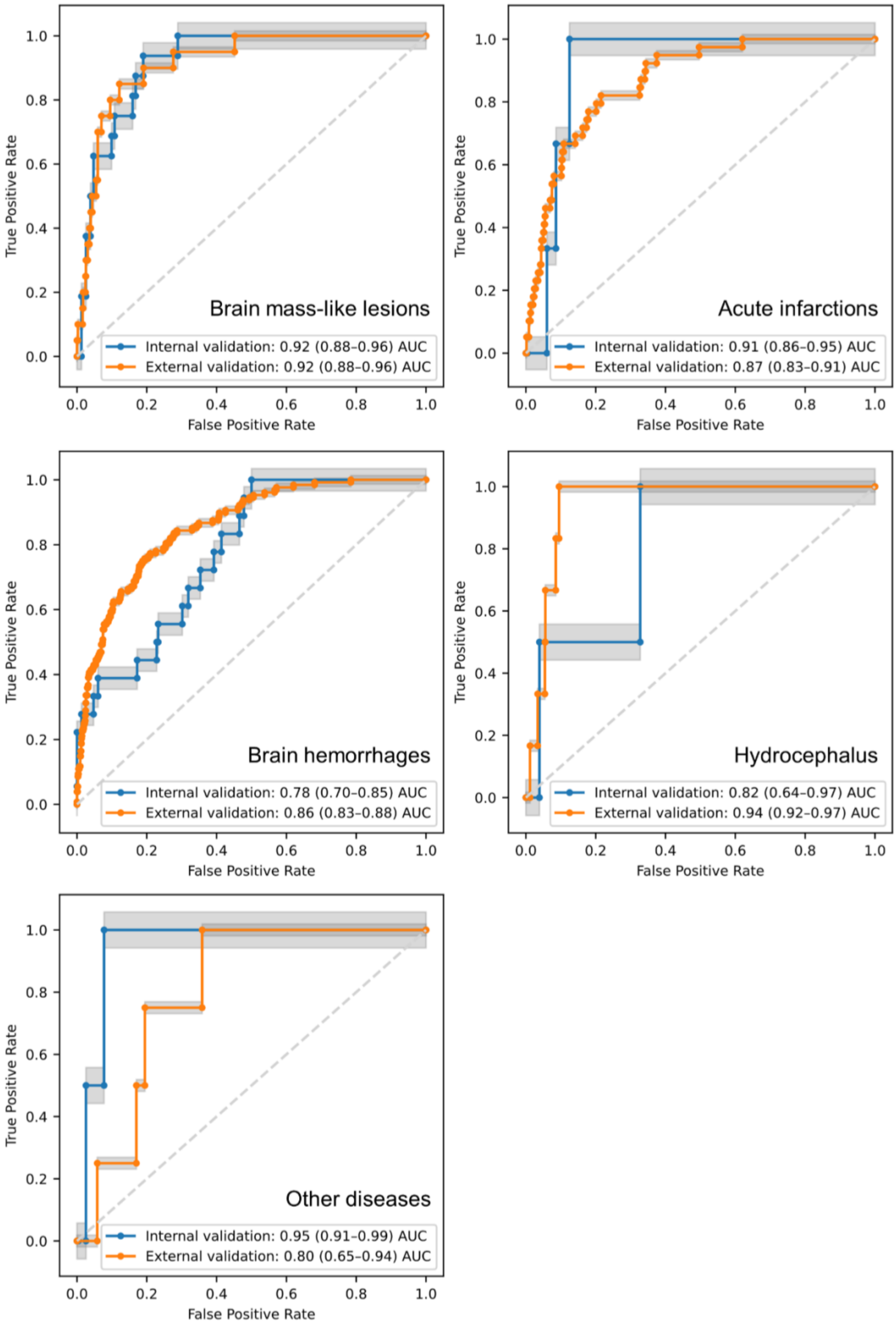


Fig. 8: ROC curve analysis for assessing the performance of CN-StyleGAN for the detection of emergency cases by disease entity. Data are presented as mean AUC values with 95% CI.

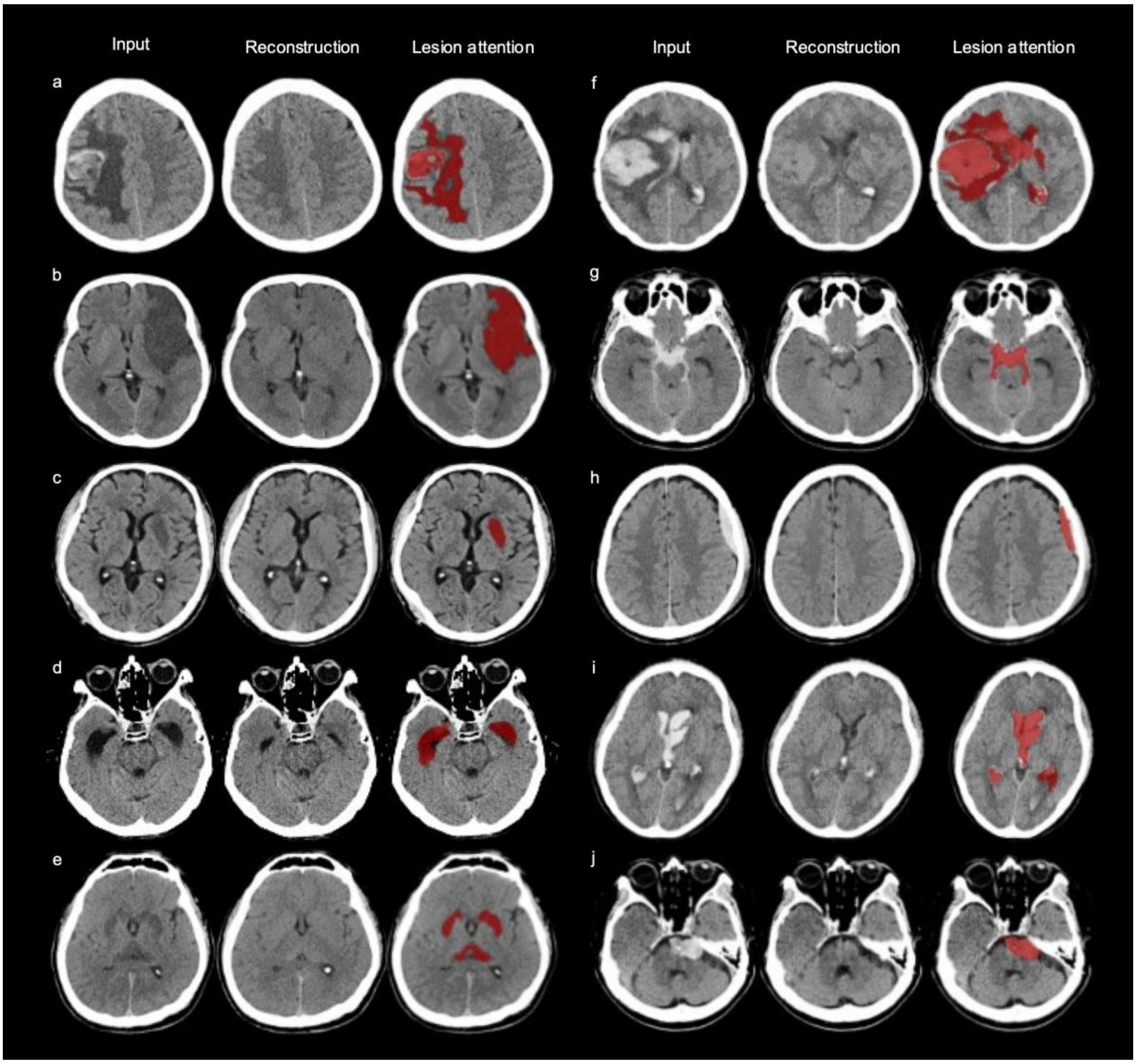


Fig. 9: Localization of the predicted lesion on emergency brain CT images from patients with various diseases. The columns, from the left to right, of each case represent input images, reconstructed images, and lesion attention. The attention maps localize anomalies related to secondary brain changes such as midline shift or perilesional edema as well as space-occupying brain lesions. **a** brain mass-like lesions, **b** acute territory infarction, **c** acute basal ganglionic infarction, **d** hydrocephalus, **e** hypoxic encephalopathy, **f** intracerebral hemorrhage (ICH), **g** subarachnoid hemorrhage (SAH), **h** subdural hemorrhage (SDH), **i** intraventricular hemorrhage (IVH), and **j** unruptured aneurysm.

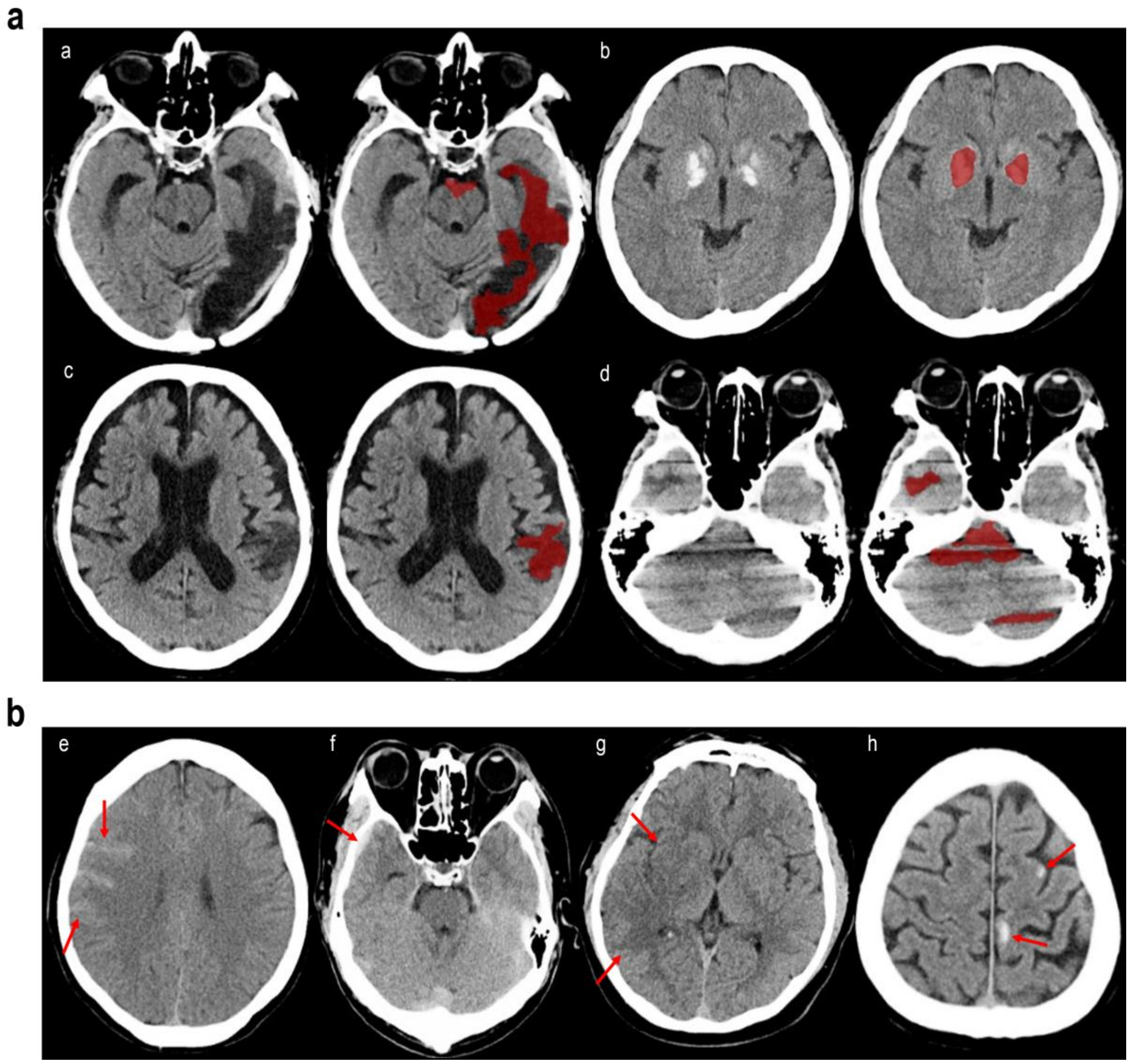


Fig. 10: False-positive and false-negative cases predicted by CN-StyleGAN

a The false-positive cases include anomalous cases that do not require urgent or immediate treatment. In each example, the left-sided image represents the original input image, and the right-sided image represents the image with predicted abnormal regions. **b** Most false-negative cases consisted of brain lesions with a relatively small volume or a subtle attenuation change (arrows). a, encephalomalacia (old infarction); b, intracranial calcification not related to normal aging; c, normal age-related prominent sulci; d, motion artifact; e, traumatic subarachnoid hemorrhage; f, subdural hemorrhage; g, early-stage acute infarction; and h, small, calcified metastases.

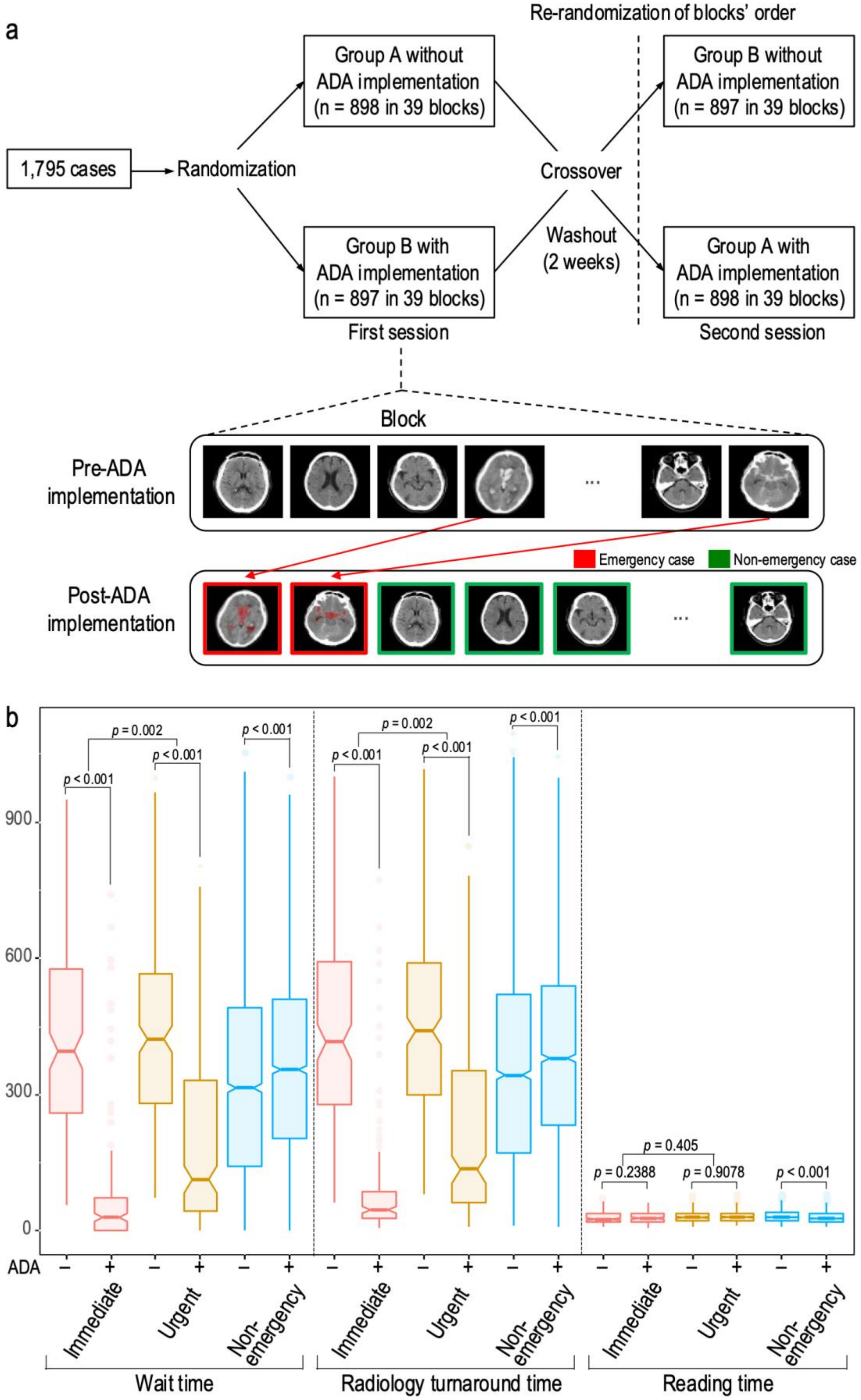

a
Re-randomization of blocks' order
1,795 cases
Randomization
Group A without ADA implementation (n = 898 in 39 blocks)
Group B with ADA implementation (n = 897 in 39 blocks)
First session
Crossover
Washout (2 weeks)
Group B without ADA implementation (n = 897 in 39 blocks)
Group A with ADA implementation (n = 898 in 39 blocks)
Second session
Block
Pre-ADA implementation
...
🟥 Emergency case
🟩 Non-emergency case
Post-ADA implementation
...
b
p = 0.002
p < 0.001
p < 0.001
p < 0.001
p = 0.002
p < 0.001
p < 0.001
p < 0.001
900
600
300
0
p = 0.405
p = 0.2388
p = 0.9078
p < 0.001
ADA
–
+
–
+
–
+
–
+
–
+
–
+
–
+
–
+
–
+
Immediate
Urgent
Non-emergency
Immediate
Urgent
Non-emergency
Immediate
Urgent
Non-emergency
Wait time
Radiology turnaround time
Reading time

Fig. 11: Clinical simulation test. a Randomized crossover study design. **b** Comparison of outcomes in subgroups pre- and post-ADA triage (immediate [n = 80], urgent [n = 117], and non-emergency cases [n = 1,598]). Data are reported as the median ± IQR. Box plots show the median (center line), first and third quartiles (box edges), and whiskers 1.5 times the IQR. Data points outside the whiskers are considered outliers. Two-sided *p* values were calculated using the Wilcoxon signed-rank test for comparison between pre- and post-ADA triage, and the Wilcoxon rank-sum test was used for comparison between immediate and urgent cases.

Clinical simulation test for emergency case prioritization. Table 4 summarizes the outcomes before and after ADA implementation and presents them as median values in seconds (interquartile range [IQR]). In the emergency group, the median WT was significantly shorter post-ADA triage by 294 s (70.5 s [IQR 168]) than pre-ADA triage (422.5 s [IQR 299]) ($p < 0.001$). The median TAT was significantly faster post-ADA triage by 297.5 s (88.5 s [IQR 179]) than pre-ADA triage (445.0 s [IQR 298]) ($p < 0.001$). There was no significant difference in RT between pre-ADA and post-ADA triage (29.0 s [IQR 12.5] vs. 30.0 s [IQR 11.0], $p = 0.38$). As expected, in the non-emergency group, there was a significant delay in the WT and TAT when the ADA was implemented. However, the absolute difference in the WT and TAT between pre-ADA and post-ADA triage was significantly smaller in the non-emergency group (79.3 s [IQR 197.9] and 72.8 s [IQR 202.3]) than in the emergency group (-294.0 s [IQR 352] and -297.5 s [IQR 347]) ($p < 0.001$). The RT was significantly shorter post-ADA triage by 1.5 s (31.00 s [11.5]) than pre-ADA triage (28.00 [11.5]) ($p < 0.001$). In the false negatives, the median WT and TAT were significantly delayed by 71 s and 70.3 s, respectively, post-ADA triage compared with pre-ADA triage (358.0 [IQR 291.5] to 449.8 s [IQR 199.3], $p = 0.009$ and 471.0 s [IQR 205] to 384.3 [IQR 300.9], respectively; $p = 0.02$) (Table 5). Fig. 11b shows the significant reduction in the WT and TAT in the subgroups of emergency cases. Note that the WT and TAT were significantly shorter in the immediate group (350 s [260.3] and 355 s [266.6], respectively) than in the urgent group (245.5 s [422.5] and 245.5 s [439.5], respectively) (all $p = 0.002$).

Our study proposed an anomaly detection approach based on a deep generative model trained only with normal brain CT images from healthy individuals. Although the proposed model did not reach the level of the supervised learning-based model performance, our study showed that the ADA has a clear advantage in terms of covering a diversity of diseases seen in the ED. In particular, our research demonstrated the potential clinical applicability of the ADA as a triage system for patients with emergency conditions.

Our research demonstrated the moderate but consistent performance of the ADA based on a deep generative model for internal and external validation datasets. Our external validation dataset represents real-world data that were consecutively collected from ED patients with neurologic symptoms and acquired from diverse CT machines and scanning protocols. Our results are supported by the findings of previous related studies in terms of the acceptable performance by an anomaly detection model and good generalizability. Han et al.[34] reported on a GAN-based anomaly detection model with an AUC of 0.727–0.894 for detecting Alzheimer's disease and an AUC of 0.921 for detecting brain metastases from MRI. Choi et al.[9] reported on a deep learning model trained only using normal brain images to identify brain abnormalities (AUC of 0.74) in on brain positron emission tomography-CT (PET-CT) images. Fujioka et al.[10] proposed a GAN-based anomaly detection model with an AUC of 0.936 for distinguishing normal tissue from benign and malignant masses based on breast ultrasound imaging. These prior studies are valuable in that they demonstrated the capability of anomaly detection models in various medical images. However, the previous studies lacked external clinical validation tests; thus, whether these models can be generalized to real-world situations cannot be guaranteed. Therefore, further evidence with real-world data is warranted. Our study serves this purpose.

The other critical point of our study is that our research demonstrated the feasibility of our ADA as a triage system for brain CT scans in the ED. Our study revealed that ADA implementation significantly reduced the WT and TAT in emergency cases. Our results are comparable to those of previous studies regarding the clinical feasibility of patient triage by supervised anomaly detection models. Titano et al.[35] reported that their supervised model potentially raised the alarm 150 times faster than humans for urgent cases in brain CT scans. Wood et al. [36] demonstrated that the supervised anomaly detection model significantly reduced the mean reporting time for abnormal MRI examinations from 28 days to 14 days and from 9 days to 5 days for two hospital networks. Notably, in the detailed subgroup analysis of our study, ADA implementation led to a significant reduction in the WT and TAT in immediate (more urgent) cases than in urgent cases. This is because ADA-based classification is based on anomaly scores. A higher anomaly score for a limited intracranial space likely reflects a correspondent urgency on an emergency brain CT scan. Unexpectedly, the increase in the WT and TAT in non-emergency cases was significantly smaller than the decrease in the WT and TAT in emergency cases. This finding is likely due to the small percentage of emergency cases and shorter RT following ADA implementation in non-emergency cases. Although the emergency cases led to a radiology workflow delay in the non-emergency cases, the faster RT in the relatively larger non-emergency cases seemed to offset these effects. Given our study design with a clinical simulation test, the shorter RT in the non-emergency cases may be due to the change in the radiologists' confidence or behavior for image interpretation in the normal brain CT scans predicted by ADA rather than due to recall bias or a learning effect. However, this issue needs further study.

The unresolved problem for anomaly detection models is the relatively high false-positive and false-negative rates. In the randomized controlled study conducted by Titano et al.[35], their supervised model for the triage of urgent brain CT scans could alert physicians in 50% of critical cases, with a 21% false-alarm rate. Our model had a high false-negative rate (22.3%) and false-positive rate (19.1%). In our clinical simulation test, the ADA implementation caused a significant delay in the median WT and TAT in the false negatives compared with the pre-ADA group. Therefore, the triage system with the anomaly detection model posed a risk of undermining the timely management of patients with critical CT findings. For false positives, a false alarm can reduce physicians' faith in a model and negatively affect emergency patients who need fast treatment. Although these problems could be solved using technical advances, this will be an ongoing issue unless the triage algorithm achieves perfect accuracy. Therefore, it is important that interpreting radiologists understand the optimization strategy and are prepared to deal with false positives or negatives.

This study has several limitations. First, our current system relies on a single brain CT scan and does not refer to prior imaging examinations or clinical information. This could result in mis-triage of some less urgent cases as high priority cases. For example, even if a previously diagnosed infarction has already been treated, it could be detected as an emergency case. Furthermore, anomaly cases of benign conditions (e.g., an arachnoid cyst or encephalomalacia with an old infarction) may also be incorrectly classified as emergency conditions. In addition, brain shrinkage is a normal part of the aging process but can indicate early-onset neurodegenerative diseases in younger patients. Therefore, generating brain images that are the closest to normal without age information is challenging. Age information could be a prerequisite for correct classification in our anomaly detection model. These problems can be mitigated by training the model on benign conditions and incorporating meta-information regarding factors that affect clinical diagnosis. Third, we used clinical and radiological diagnoses as reference standards. However, many neurological ED cases (e.g., small traumatic intracranial hemorrhage, minor stroke, or transient ischemic attack) do not require surgical treatment or aggressive intervention because of their low risk of rapid exacerbation. Therefore, this may be an unavoidable limitation in an emergency screening cohort study. Nevertheless, further studies using the gold standard are warranted to determine the accurate performance of the model. Fourth, this study did not reflect the complexity of clinical practice. Multiple factors can influence the results of a clinical simulation test, including the case difficulty, queue size of the CT scan, readers' expertise level, image-processing time, patient acuity, and interruption by other examinations. Therefore, our results may vary with these factors. To address this issue, multicentered and prospective validation studies are warranted.

In conclusion, we developed an ADA with a deep generative network trained only on normal brain CT images from healthy individuals. Our model achieved moderate but consistent performance in detecting emergency brain CT scans using internal and external ED screening cohorts. In the clinical simulation test, our study also highlighted the feasibility of the ADA as a triage system to reprioritize radiology worklists and accelerate the diagnosis of various emergency conditions.

Table 4. Comparison of outcomes pre-ADA and post-ADA triage

			Pre-ADA	Post-ADA	Difference between pre- and post-ADA	*p* value[a]	*p* value[b]
Emergency (n = 197)	WT	Median (IQR)	422.5 (299.0)	70.5 (168.0)	-294.0 (352.0)	<0.001	<0.001
		Mean (±SD)	436.6 (±192.2)	147.4 (±184.0)			
		Min-Max	1–997	1–803			
	RT	Median (IQR)	29.0 (12.5)	30.0 (11.0)	0.0 (13.0)	0.38	0.006
		Mean (±SD)	29.7 (±9.2)	30.3 (±7.7)			
		Min-Max	9–76	7–79			
	TAT	Median (IQR)	445.0 (298.0)	88.5 (179.0)	-297.5 (347.0)	<0.001	<0.001
		Mean (±SD)	457.9 (±195.4)	168.7 (±183.2)			
		Min-Max	63–1017	6–847			
Control (n = 1,598)	WT	Median (IQR)	327.0 (357.0)	364.8 (307.4)	79.3 (197.9)	<0.001	
		Mean (±SD)	335.1 (±217.1)	366.0 (±192.9)			
		Min-Max	1–1053	1–1000			
	RT	Median (IQR)	31.00 (11.5)	28.00 (11.5)	-1.5 (14.0)	<0.001	
		Mean (±SD)	31.2 (±8.9)	29.7 (±9.2)			
		Min-Max	9–79	8–79			
	TAT	Median (IQR)	357.0 (352.0)	393.0 (303.4)	72.8 (202.3)	<0.001	
		Mean (SD)	364.3 (218.0)	393.2 (192.1)			
		Min-Max	12–1095	9–1045			

Data are expressed as the mean (SD, standard deviation) or median [interquartile range, IQR] (seconds). All statistical tests were two-sided, and statistical significance was set at $p = 0.05$. [a]The Wilcoxon signed-rank test was used for comparison between pre- and post-ADA triage. [b]The Wilcoxon rank-sum test was used for comparison between emergency and non-emergency cases.

Table 5. Comparison of outcomes pre- and post-ADA triage among false negatives and false positives

			Pre-ADA	Post-ADA	Difference between pre- and post-ADA	*p* value[a]
False Negatives (n = 44)	WT	Median (IQR)	358.0 (291.5)	449.8 (199.3)	71.0 (145.0)	0.009
		Mean (±SD)	400.5 (±192.2)	445.0 (±150.4)		
		min–max	72–922	146–803		
	RT	Median (IQR)	28.8 (10.0)	28.8 (9.4)	-0.3 (9.9)	0.68
		Mean (±SD)	29.4 (±8.6)	29.6 (±7.8)		
		min–max	9–76	12–62		
	TAT	Median (IQR)	384.3 (300.9)	471.0 (205.0)	70.3 (143.6)	0.02
		Mean (±SD)	421.7 (±196.2)	464.5 (±150.9)		
		min–max	82–951	155–847		
False Positives (n = 305)	WT	Median (IQR)	357.0 (366.0)	101.0 (104.0)	-220.5 (360.5)	<0.001
		Mean (±SD)	342.9 (±220.4)	111.1 (±76.8)		
		min–max	1–957	1–449		
	RT	Median (IQR)	32.5 (11.5)	35.5 (13.0)	2.5 (14.5)	<0.001
		Mean (±SD)	33.6 (±9.1)	36.6 (±9.6)		
		min–max	9–79	9–79		
	TAT	Median (IQR)	378.0 (357.5)	134.50 (111.5)	-223.5 (361.5)	<0.001
		Mean (SD)	374.0 (±220.8)	143.7 (±78.1)		
		min–max	12–1007	9–508		

Data are expressed as the mean (SD, standard deviation) or median [interquartile range, IQR] (seconds). All statistical tests were two-sided, and statistical significance was set at $p = 0.05$. [a]The Wilcoxon signed-rank test was used for comparison between pre- and post-ADA triage

참고문헌

1 Goyal, M. *et al.* Randomized assessment of rapid endovascular treatment of ischemic stroke. *N Engl J Med.* **372**, 1019-1030, https://doi.org/10.1056/NEJMoa1414905 (2015).

2 Jahan, R. *et al.* Association Between Time to Treatment With Endovascular Reperfusion Therapy and Outcomes in Patients With Acute Ischemic Stroke Treated in Clinical Practice. *JAMA* **322**, 252-263, https://doi.org/10.1001/jama.2019.8286 (2019).

3 Sheth, S. A. *et al.* Time to endovascular reperfusion and degree of disability in acute stroke. *Ann Neurol.* **78**, 584-593, https://doi.org/10.1002/ana.24474 (2015).

4 Seyam, M. *et al.* Utilization of artificial intelligence–based intracranial hemorrhage detection on emergent noncontrast CT images in clinical workflow. *Radiol Artif Intell.* 4, e210168, https://doi.org/10.1148/ryai.210168 (2022).

5 Morey, J. R. *et al.* Real-World Experience with Artificial Intelligence-Based Triage in Transferred Large Vessel Occlusion Stroke Patients. *Cerebrovasc Dis.* **50**, 450-455, https://doi.org/10.1159/000515320 (2021).

6 O'Neill, T. J. *et al.* Active Reprioritization of the Reading Worklist Using Artificial Intelligence Has a Beneficial Effect on the Turnaround Time for Interpretation of Head CTs with Intracranial Hemorrhage. *Radiol Artif Intell.* 3, e200024, https://doi.org/10.1148/ryai.2020200024 (2020).

7 Arbabshirani, M. R. *et al.* Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *NPJ Digit Med.* **1**, 9, https://doi.org/10.1038/s41746-017-0015-z (2018).

8 Chen, X. & Konukoglu, E. Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. Preprint at http://arxiv.org/abs/*1806.04972* (2018).

9 Choi, H. *et al.* Deep learning only by normal brain PET identify unheralded brain anomalies. *EBioMedicine.* **43**, 447-453, https://doi.org/10.1016/j.ebiom.2019.04.022 (2019).

10 Fujioka, T. *et al.* Efficient anomaly detection with generative adversarial network for breast ultrasound imaging. *Diagnostics (Basel)* **10**, 456, https://doi.org/10.3390/diagnostics10070456 (2020).

11 Seah, J. C., Tang, J. S., Kitchen, A., Gaillard, F. & Dixon, A. F. Chest radiographs in congestive heart failure: visualizing neural network learning. *Radiology* **290**, 514-522, https://doi.org/ 10.1148/radiol.2018180887 (2019).

12 Baur, C. *et al.* Modeling Healthy Anatomy with Artificial Intelligence for Unsupervised Anomaly

Detection in Brain MRI. *Radiol Artif Intel.l* **3**, e190169, https://doi.org/10.1148/ryai.2021190169 (2021).

13 Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G. & Schmidt-Erfurth, U. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Med Image Anal.* **54**, 30-44, https://doi.org/10.1016/j.media.2019.01.010 (2019).

14 Abdal, R., Qin, Y. & Wonka, P. Image2stylegan++: How to edit the embedded images? Preprint at http://arxiv.org/abs/1911.11544 (2019)

15 Chapman, B. E., Lee, S., Kang, H. P. & Chapman, W. W. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. Journal of biomedical informatics 44, 728-737 (2011).

16 Brix, M. K. *et al.* The Evans' Index revisited: new cut-off levels for use in radiological assessment of ventricular enlargement in the elderly. *Eur J Radiol* **95**, 28-32, https://doi.org/ 10.1016/j.ejrad.2017.07.013 (2017).

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

18 Prevedello, L. M. *et al.* Automated critical test findings identification and online notification system using artificial intelligence in imaging. *Radiology* **285**, 923-931, https://doi.org/ 10.1148/radiol.2017162664 (2017).

19 Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at http://arxiv.org/abs/1409.1556 (2014).

20 Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. *In Proceedings of the IEEE conference on computer vision and pattern recognition.* 586-595 (2018)

21 Zhu, J., Shen, Y., Zhao, D. & Zhou, B. In-domain gan inversion for real image editing. In *European Conference on Computer Vision.* 592-608 (2020).

22 Goodfellow, I. J. *et al.* Generative adversarial networks. Preprint at http://arxiv.org/abs/*1406.2661* (2014).

23 Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4401-4410 (2019)

24 Zhong, Z., Zheng, L., Kang, G., Li, S. & Yang, Y. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 13001-13008 (2020)

25 Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. Preprint at http://arxiv.org/abs/*1912.01703* (2019).

26 Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Preprint at http://arxiv.org/abs/*1412.6980* (2014).

27 Wulff, J. & Torralba, A. Improving inversion and generation diversity in stylegan using a gaussianized latent space. Preprint at http://arxiv.org/abs/*2009.06529* (2020).

28 Zhu, P., Abdal, R., Qin, Y., Femiani, J. & Wonka, P. Improved StyleGAN Embedding: Where are the Good Latents? Preprint at http://arxiv.org/abs/*2012.09036* (2020).

29 Bartz, C., Bethge, J., Yang, H. & Meinel, C. One Model to Reconstruct Them All: A Novel Way to Use the Stochastic Noise in StyleGAN. Preprint at http://arxiv.org/abs/*2010.11113* (2020).

30 Akkus, Z., Kostandy, P. M., Philbrick, K. A. & Erickson, B. J. Extraction of brain tissue from CT head images using fully convolutional neural networks. In Medical imaging 2018: Image processing. SPIE, 10574, 514-520 (2018)

31 Rathnayake, S., Nautsch, F., Goodman, T. R., Forman, H. P. & Gunabushanam, G. Effect of Radiology Study Flow on Report Turnaround Time. *AJR Am J Roentgenol* **209**, 1308-1311, https://doi.org/10.2214/ajr.17.18282 (2017).

32 Boland, G. W., Guimaraes, A. S. & Mueller, P. R. Radiology report turnaround: expectations and solutions. *Eur Radiol* **18**, 1326-1328, https://doi.org/10.1007/s00330-008-0905-1 (2008).

33 Sung, J. *et al.* Added Value of Deep Learning–based Detection System for Multiple Major Findings on Chest Radiographs: A Randomized Crossover Study. *Radiology* **299**, 450-459, https://doi.org/10.1148/radiol.2021202818 (2021).

34 Han, C. *et al.* MADGAN: Unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction. *BMC Bioinformatics* **22**, 1-20, https://doi.org/10.1186/s12859-020-03936-1 (2021).

35 Titano, J. J. *et al.* Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med* **24**, 1337-1341, https://doi.org/10.1038/s41591-018-0147-y (2018).

36 Wood, D. A. *et al.* Deep learning models for triaging hospital head MRI examinations. *Med Image Anal.* **78**, 102391, https://doi.org/10.1016/j.media.2022.102391 (2022).

영문요약

영문요약

Triage is essential for the early diagnosis and reporting of neurologic emergencies. Herein, we report the development of an anomaly detection algorithm (ADA) with a deep generative model trained on brain computed tomography (CT) images of healthy individuals that reprioritizes radiology worklists and provides lesion attention maps for brain CT images with critical findings. In the internal and external validation datasets, the ADA achieved area under the curve values (95% confidence interval) of 0.85 (0.81–0.89) and 0.87 (0.85–0.89), respectively, for detecting emergency cases. In a clinical simulation test of emergency cohorts, the median wait time was significantly shorter post-ADA triage than pre-ADA triage by 294 s (422.5 s [interquartile range, IQR 299] to 70.5 s [IQR 168]), and the median radiology report turnaround time was significantly faster post-ADA triage than pre-ADA triage by 297.5 s (445.0 s [IQR 298] to 88.5 s [IQR 179]) (all $p < 0.001$).