Disorder-Free Data are All You Need: Inverse Supervised Learning for Broad-Spectrum Head Disorder Detection

Yuwei He[1,3+], Yuchen Guo[1+], Jinhao Lyu[2+], Liangdi Ma[1,3+], Haotian Tan[1,3], Wei Zhang[5], Guiguang Ding[1,3], Hengrui Liang[6], Jianxing He[6], Feng Xu[1,3*], Xin Lou[2*], and Qionghai Dai[1,4*]

[1]Institute for Brain and Cognitive Sciences, BNRist, Tsinghua University, Beijing, China, 100084

[2]Department of Radiology, Chinese PLA General Hospital, Beijing, China, 100853

[3]School of Software, Tsinghua University, Beijing, China, 100084

[4]Department of Automation, Tsinghua University, Beijing, China, 100084

[5]Department of Radiology, Brain Hospital of Hunan Province, Hunan, China, 410013

[6]China State Key Laboratory of Respiratory Disease & National Clinical Research Center for Respiratory Disease, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China, 510120

[+]These authors contributed equally

[*]Corresponding authors, xufeng2003@gmail.com; louxin@301hospital.com.cn; daiqh@tsinghua.edu.cn

Collecting and annotating sufficient data containing disorders is crucial for the development of artificial intelligence (AI)-based medical systems. However, preparing data with complete disorder types and adequate annotations is challenging, which limits the ability of existing AI-based medical systems to diagnose specific disorders. In this paper, we introduce a novel AI-based system that achieves accurate and generalizable broad-spectrum disorder detection without requiring any data containing disorders. Specifically, we obtained a training dataset of $21,429$ **disorder-free head computed tomography (CT) scans. We then proposed a**

learning algorithm called Inverse Supervised Learning (ISL), which learns and understands disorder-free samples instead of disorder-contained ones. This approach enables the identification of all types of disorders. The system achieved AUC values of 0.883, 0.868, and 0.866 on retrospective (127 disorder types, $9,967$ scans), prospective (117 disorder types, $3,054$ scans), and cross-center (46 disorder types, 554 scans) datasets, respectively. These results demonstrate that the system can detect far more disorder types than previous AI-based systems. Additionally, the system provides visually understandable clues, and we developed a diagnosis and visualization software for clinical usage based on these advantages. Furthermore, the ISL-based systems achieved AUC values of 0.893 and 0.895 on pulmonary CT and retinal optical coherence tomography (OCT), respectively, demonstrating that ISL can generalize well to non-head and non-CT images.

Over the past decade, artificial intelligence (AI) has made significant strides and has been applied in various fields. In the medical field, the accumulation of medical image data has enabled many AI diagnostic techniques to achieve radiologist-level performance in recognizing, classifying, and quantifying specific diseases. For example, AI has been used for cerebral hemorrhage recognition[1] and COVID-19 recognition[2] from CT images. These breakthroughs have led us to envision that AI diagnostic techniques can assist in clinical decision-making from medical images and alleviate the severe shortage of expert radiologists in many areas and hospitals.

Despite the significant progress made in AI techniques, there is still a gap between these techniques and real clinical decision-making. Current AI techniques primarily focus on recogniz-

ing specific types of disorders from input medical data. However, for a clinical decision-making workflow, the most basic and essential task is to identify all possible disorder types that could be diagnosed from the medical image. For instance, in the case of brain CT, more than one hundred types of disorders could be diagnosed from it. Therefore, a decision-making diagnostic system for brain CT must be capable of detecting a broad spectrum of disorders, as missing the detection of any disorder type is unacceptable. Existing medical AI techniques are developed based on widely-used AI paradigms, which involve deciding the disorder types to be handled, collecting sufficient disorder-contained samples, and constructing recognition/localization/segmentation models for the disorders. This paradigm works well when the disorder types are limited, and the samples are easily accessible. However, developing a broad-spectrum disorder detection system using this paradigm requires collecting data and constructing models for all types of disorders, which is extremely difficult and inefficient, especially for unusual diseases. Therefore, it is impractical to use the widely-used AI paradigms to achieve real clinical decision-making.

To address the challenges mentioned above, we propose a novel AI solution called Inverse Supervised Learning (ISL). Instead of using disorder-contained data, which requires hundreds of disorder types and a large number of samples for each type, we use disorder-free medical images for supervision. In theory, we use the opposite problem (detecting no-disorder samples) to replace the original problem (detecting all types of disorders). Therefore, instead of training hundreds of models to recognize all possible types of disorders, we train just one model to understand the concept of disorder-free fully. Consequently, all disorders can be identified as they differ from the disorder-free samples used in training. With our paradigm, the challenges mentioned above are

fully resolved as there is no need for samples of all possible disorder types.

To achieve ISL, we utilize the traditional computer vision task of image inpainting in a novel framework. Image inpainting aims to restore the content of a partially missing image based on the context of non-missing information. Specifically, in this case, an image inpainting network is trained to replenish masked regions in a medical image, where the replenished content always reflects healthy tissue because the training dataset contains only disorder-free medical images. If any disorder exists in the image and the disorder region is masked off, the reconstructed disorder-free image would be significantly different from the original one. Conversely, for an image without disorders, no matter which region is masked, the reconstructed image should always be similar to the original one as they are both healthy and consistent with the rest of the healthy images. By masking, inpainting, and comparing all the image regions, ISL can detect various types of disorders and locate the disorder regions. Our proposed solution, ISL, (1) does not require the deliberate collection of data for any disorder type; (2) ensures that the data used to develop systems are easily accessible; (3) does not require experts to manually annotate the data; (4) enables the developed system to recognize broad-spectrum disorders rather than specific ones; and (5) provides experts with clinical clues, such as disorder locations.

In this study, we utilized ISL to construct a system for broad-spectrum disorder detection on unenhanced brain computed tomography (CT) scans.[3,4] CT is a first-line diagnostic modality for assessing brain abnormalities due to its quick acquisition and non-invasive nature. The ISL-based system was developed using only disorder-free head CT images. It achieved expert-level accuracies

on a retrospective dataset with 127 disorder types and a prospective dataset with 116 disorder types, surpassing the number of detectable disorder types in previous works. We also applied ISL to build two additional systems: one for pulmonary disorder detection in CT images and another for retinal disorder detection in optical coherence tomography (OCT) images. The results demonstrate that ISL can generalize well to non-brain and non-CT-based disorder detection.

Results

Building an ISL based system for clinically applicable broad-spectrum head disorder detection. Our proposed ISL-based disorder detection system for brain CT comprises a de-disorder network (DeDN), a disorder recognition network (DRN), and a disorder locating module. Firstly, a CT image is processed with specific window width and window locations and then fed into the DeDN to generate its corresponding de-disorder image. Next, we obtain a difference image by subtracting the original and generated images. Finally, the difference image is inputted into the DRN to determine whether any disorder exists in the image. Additionally, the disorder locating module can be used to locate the disorder.

To develop the system, we collected CT scans from the Chinese PLA General Hospital (PLAGH), a leading national hospital that serves patients throughout China. We constructed a training dataset of 21,429 healthy brain CT scans (March 2012 - July 2019) retrieved from the picture archiving and communication systems (PACS). The retrieval process involved matching the fixed description (”No abnormality is observed”) of healthy CT scans with historical diagnosis reports, resulting in a training dataset that was efficiently obtained without requiring expert effort

or disorder annotation.

Performance on the broad-spectrum head disorder detection. To evaluate the system, we obtained a retrospective test dataset from the PLAGH (9,967 scans, 88.23% with 127 types of disorders, March 2012 - July 2019) and a prospective test dataset from the PLAGH (3,054 scans, 88.70% with 116 types of disorders, July 2019 - August 2021). To demonstrate the clinical applicability of our system, we counted all types of disorders described in clinical reports from the PLAGH using a rule-based NLP algorithm and manual selection by radiologists. We sorted out 127 and 116 types of disorders for testing, respectively. To our knowledge, these test datasets have the broadest coverage of head disorder types. The number of scans for each disorder is shown in Table 1 and Table 2.

We employed a disorder-contained/free classification testing strategy for each type of disorder, with testing carried out at the scan-level. This means that the system predicted whether the entire scan contained any disorder or not. Scan-level classification is practical for clinical use as it enables radiologists to quickly identify the presence of disorders, which is particularly useful in emergency treatment.[5] The label of each scan was determined using disorder-related keywords in its associated report and then confirmed by radiologists based on the report and CT images.

Table 1 and Table 2 show the area under the receiver operating characteristic curve (AUC) for the two datasets. On the retrospective dataset, the average AUC for all disorder types was 0.883 (95% confidence interval [CI] 0.882 - 0.885), while on the prospective dataset, the average AUC was 0.868 (95% CI 0.866 - 0.869). Additionally, Table 1 and Table 2 also present the sensitivity

and specificity with 95% CI for the disorders. On the retrospective dataset, the system achieved an AUC > 0.95 for 43 disorders and an AUC > 0.90 for 74 disorders. On the prospective dataset, the system achieved an AUC > 0.95 for 30 disorders and an AUC > 0.90 for 50 disorders. These results demonstrate that our system is capable of detecting a broad spectrum of disorders in brain CT.

Analysis of system explainability. In clinical practice, a computer-aided diagnosis (CAD) system should provide understandable clues to support prediction results. Our model can quickly and intuitively locate the disorder region based on the generated de-disorder image, as shown in Fig.1a. Additionally, we developed a CAD software for clinical use, as shown in Fig.1b. The software takes a CT scan as input and outputs possible disorder regions, improving the diagnosis performance of radiologists. Our system not only detects the disorder location in a slice but also provides the disorder distribution in a scan. Fig.2b shows two example scans with different disorder distributions. Based on the distribution curves supplied by our system, we can observe that the disorders in the two scans have centralized and dispersive distributions, respectively. Fig.2c shows the average disorder distributions for some typical disorders in the retrospective dataset. The average distributions are close to the occurrence frequency at different brain tissues in reality, demonstrating the effectiveness of the system's explainability.

Performance contributions from different modules. This section elaborates on the reasons for adopting each module and demonstrates their contributions to the final performance. The results are presented in Fig. 3.

Table 1 | Performance on the retrospective dataset for head disorder detection.

retrospective				
	Num	**AUC**	**Sensitivity**	**Specificity**
Effusion	487	0.963 (0.952, 0.975)	0.920 (0.897, 0.945)	0.890 (0.872, 0.907)
Subarachnoid Hemorrhage	485	0.976 (0.967, 0.985)	0.928 (0.905, 0.953)	0.922 (0.906, 0.938)
Subdural Hematoma	482	0.958 (0.946, 0.971)	0.898 (0.873, 0.925)	0.890 (0.872, 0.909)
Pneumocephaly	474	0.967 (0.956, 0.979)	0.922 (0.899, 0.947)	0.915 (0.899, 0.933)
Parenchymal Hemorrhage	474	0.955 (0.943, 0.969)	0.901 (0.876, 0.928)	0.890 (0.873, 0.908)
Multiple Cerebral Infarctions	465	0.865 (0.844, 0.887)	0.738 (0.699, 0.776)	0.866 (0.847, 0.886)
Corona Radiata Cerebral Infarction	459	0.667 (0.636, 0.698)	0.560 (0.514, 0.606)	0.688 (0.662, 0.716)
Lacunar Infarction	456	0.687 (0.657, 0.719)	0.667 (0.623, 0.713)	0.596 (0.568, 0.624)
Basal Ganglia Ischemia	454	0.861 (0.839, 0.885)	0.731 (0.692, 0.773)	0.865 (0.847, 0.885)
Basal Ganglia Cerebral Infarction	453	0.716 (0.687, 0.747)	0.561 (0.512, 0.609)	0.778 (0.754, 0.802)
Ischemia	444	0.928 (0.914, 0.945)	0.867 (0.838, 0.899)	0.837 (0.816, 0.858)
Calcification	442	0.825 (0.801, 0.852)	0.692 (0.647, 0.738)	0.836 (0.814, 0.858)
Malacia Foci	436	0.917 (0.900, 0.935)	0.853 (0.821, 0.890)	0.833 (0.811, 0.856)
Contusion	424	0.954 (0.940, 0.969)	0.906 (0.880, 0.934)	0.897 (0.879, 0.914)
Cerebral Infarction Of Caudate Nucleus Head	422	0.898 (0.878, 0.916)	0.820 (0.784, 0.855)	0.813 (0.790, 0.836)
Periventricular Cerebral Infarction	397	0.803 (0.777, 0.829)	0.713 (0.668, 0.758)	0.733 (0.707, 0.759)
Tubercle	362	0.819 (0.792, 0.846)	0.671 (0.624, 0.721)	0.830 (0.810, 0.852)
Intraventricular Hemorrhage	323	0.986 (0.978, 0.995)	0.944 (0.923, 0.969)	0.942 (0.929, 0.957)
Brainswelling	315	0.952 (0.938, 0.969)	0.898 (0.867, 0.933)	0.880 (0.862, 0.900)
Sclerosis	272	0.840 (0.810, 0.872)	0.746 (0.695, 0.801)	0.823 (0.799, 0.844)
Space Occupying Lesions	261	0.925 (0.904, 0.946)	0.862 (0.820, 0.904)	0.847 (0.827, 0.869)
Epidural Hematoma	252	0.947 (0.927, 0.967)	0.893 (0.857, 0.929)	0.892 (0.875, 0.912)
Cerebral Edema	216	0.965 (0.952, 0.981)	0.903 (0.866, 0.944)	0.900 (0.882, 0.918)
Minor Hemorrhage	212	0.966 (0.953, 0.982)	0.910 (0.873, 0.953)	0.892 (0.874, 0.911)
Thalamic Cerebral Infarction	205	0.713 (0.671, 0.754)	0.610 (0.546, 0.673)	0.690 (0.662, 0.717)
Soft Tissue Swelling	203	0.937 (0.916, 0.962)	0.877 (0.833, 0.926)	0.852 (0.831, 0.873)
Arteriosclerosis	196	0.810 (0.771, 0.849)	0.668 (0.602, 0.735)	0.831 (0.809, 0.853)
Parenchymal Hematoma	176	0.982 (0.971, 0.995)	0.949 (0.920, 0.983)	0.938 (0.924, 0.952)
Centrum Semiovale Cerebral Infarction	151	0.733 (0.690, 0.781)	0.702 (0.636, 0.775)	0.663 (0.635, 0.693)
Parietal Lobe Cerebral Infarction	148	0.788 (0.742, 0.836)	0.622 (0.541, 0.703)	0.874 (0.856, 0.894)
Frontal Lobe Cerebral Infarction	123	0.729 (0.674, 0.787)	0.537 (0.447, 0.626)	0.880 (0.862, 0.899)
Arachnoid Cyst	119	0.844 (0.806, 0.886)	0.714 (0.639, 0.798)	0.831 (0.809, 0.852)
Hydrocephalus	108	0.999 (0.998, 1.000)	0.991 (0.981, 1.000)	0.969 (0.961, 0.980)
Cerebral White Matter Degeneration	107	0.831 (0.787, 0.878)	0.682 (0.598, 0.776)	0.851 (0.830, 0.871)
Paraventricular Ischemia	104	0.875 (0.834, 0.921)	0.702 (0.615, 0.798)	0.938 (0.925, 0.953)
Cavum Septum Pellucidum	102	0.842 (0.801, 0.887)	0.814 (0.745, 0.892)	0.689 (0.662, 0.716)
Subcutaneous Hematoma	102	0.896 (0.857, 0.936)	0.843 (0.775, 0.912)	0.816 (0.793, 0.837)
Temporal Bone Fracture	101	0.915 (0.875, 0.959)	0.871 (0.812, 0.941)	0.831 (0.809, 0.854)
Frontal Lobe Ischemia	92	0.842 (0.795, 0.893)	0.815 (0.739, 0.891)	0.753 (0.729, 0.778)
Subdural Hemorrhage	89	0.985 (0.971, 1.000)	0.955 (0.921, 1.000)	0.943 (0.930, 0.957)
Ventriculomegaly	83	0.992 (0.985, 1.000)	0.976 (0.952, 1.000)	0.817 (0.795, 0.839)
Parietal Fracture	83	0.909 (0.864, 0.960)	0.880 (0.807, 0.952)	0.831 (0.809, 0.854)
Occipital Fracture	82	0.922 (0.881, 0.971)	0.878 (0.817, 0.951)	0.867 (0.847, 0.886)
Occipital Lobe Cerebral Infarction	73	0.918 (0.877, 0.962)	0.849 (0.767, 0.932)	0.841 (0.819, 0.862)
Frontal Bone Fracture	71	0.875 (0.821, 0.936)	0.817 (0.732, 0.915)	0.760 (0.736, 0.784)
Osteoma	70	0.844 (0.790, 0.906)	0.743 (0.643, 0.843)	0.865 (0.845, 0.886)

retrospective				
	Num	**AUC**	**Sensitivity**	**Specificity**
Temporal Lobe Cerebral Infarction	66	0.915 (0.880, 0.958)	0.848 (0.773, 0.939)	0.829 (0.807, 0.851)
Old Cerebral Infarction	66	0.822 (0.759, 0.890)	0.636 (0.515, 0.758)	0.906 (0.889, 0.923)
Lipoma	63	0.887 (0.834, 0.943)	0.841 (0.762, 0.937)	0.753 (0.728, 0.779)
Scalp Hematoma	63	0.907 (0.855, 0.968)	0.873 (0.794, 0.968)	0.837 (0.816, 0.858)
Artifacts Shadow	62	0.969 (0.944, 1.000)	0.935 (0.887, 1.000)	0.866 (0.846, 0.886)
Brain Atrophy	59	0.953 (0.925, 0.989)	0.915 (0.847, 1.000)	0.805 (0.781, 0.829)
Cisterna Magna	59	0.870 (0.824, 0.921)	0.847 (0.763, 0.949)	0.719 (0.693, 0.746)
Herniation	55	0.999 (0.999, 1.000)	1.000 (1.000, 1.000)	0.969 (0.960, 0.980)
Skull Fracture	54	0.937 (0.894, 0.990)	0.889 (0.815, 0.981)	0.865 (0.846, 0.885)
Absence Of Bone	50	0.976 (0.960, 0.996)	0.940 (0.880, 1.000)	0.890 (0.872, 0.909)
Hemispheric Infarction	50	0.980 (0.962, 1.000)	0.940 (0.880, 1.000)	0.922 (0.907, 0.938)
Brainstem Infarction	48	0.890 (0.844, 0.939)	0.875 (0.792, 0.979)	0.760 (0.735, 0.786)
Multiple Hemorrhages	47	0.945 (0.903, 0.997)	0.915 (0.851, 1.000)	0.892 (0.875, 0.911)
Metal Shadow	40	0.948 (0.912, 0.991)	0.875 (0.775, 0.975)	0.844 (0.823, 0.865)
Metastases	40	0.936 (0.902, 0.974)	0.850 (0.750, 0.975)	0.801 (0.777, 0.825)
Centrum Semiovale Ischemia	39	0.845 (0.771, 0.928)	0.718 (0.590, 0.872)	0.846 (0.825, 0.867)
Atherosclerosis	38	0.948 (0.909, 0.995)	0.895 (0.816, 1.000)	0.819 (0.797, 0.842)
Midline Shift	37	0.941 (0.894, 1.000)	0.892 (0.811, 1.000)	0.880 (0.862, 0.900)
Facial Bone Fracture	31	0.990 (0.981, 1.000)	0.968 (0.935, 1.000)	0.893 (0.876, 0.912)
Corona Radiata Ischemia	31	0.895 (0.844, 0.956)	0.871 (0.774, 1.000)	0.774 (0.749, 0.800)
Leukoaraiosis	26	0.893 (0.831, 0.973)	0.846 (0.731, 1.000)	0.788 (0.764, 0.811)
Insular Infarction	26	0.965 (0.943, 0.993)	0.923 (0.846, 1.000)	0.876 (0.858, 0.895)
Parietal Lobe Ischemia	22	0.886 (0.811, 0.985)	0.909 (0.818, 1.000)	0.744 (0.719, 0.769)
Cerebellar Hemisphere Infarction	21	0.894 (0.830, 0.968)	0.857 (0.714, 1.000)	0.771 (0.747, 0.795)
Minor Haematoma	21	0.918 (0.849, 1.000)	0.905 (0.810, 1.000)	0.884 (0.865, 0.902)
Osseous Defects	20	0.947 (0.904, 1.000)	0.950 (0.900, 1.000)	0.893 (0.876, 0.913)
Lacunar Ischemia	19	0.935 (0.878, 1.000)	0.895 (0.789, 1.000)	0.894 (0.876, 0.912)
Zygomatic Fracture	17	0.889 (0.797, 1.000)	0.824 (0.647, 1.000)	0.757 (0.733, 0.781)
Widened Cisterna Magna	16	0.940 (0.907, 0.975)	1.000 (1.000, 1.000)	0.831 (0.809, 0.853)
Meningioma	16	0.943 (0.900, 1.000)	0.938 (0.875, 1.000)	0.754 (0.730, 0.781)
Tumour	16	0.970 (0.947, 0.999)	1.000 (1.000, 1.000)	0.825 (0.803, 0.846)
Widened Subarachnoid Space	15	0.907 (0.825, 1.000)	0.933 (0.867, 1.000)	0.841 (0.819, 0.862)
Sphenoid Fracture	14	0.931 (0.869, 0.999)	0.929 (0.857, 1.000)	0.725 (0.698, 0.750)
Malformation	13	0.992 (0.985, 1.000)	1.000 (1.000, 1.000)	0.902 (0.885, 0.920)
Cystic Foci	13	0.835 (0.699, 1.000)	0.846 (0.692, 1.000)	0.637 (0.609, 0.667)
Sella Turcica Enlargement	12	0.923 (0.875, 0.978)	1.000 (1.000, 1.000)	0.773 (0.748, 0.796)
Choroid Cyst	12	0.759 (0.616, 0.917)	0.833 (0.667, 1.000)	0.586 (0.556, 0.615)
Septum Pellucidum Cyst	12	0.743 (0.597, 0.917)	0.750 (0.500, 1.000)	0.620 (0.592, 0.649)
Wallerian Degeneration	11	0.973 (0.946, 1.000)	1.000 (1.000, 1.000)	0.760 (0.737, 0.786)
Senile Encephalopathy	11	0.926 (0.851, 1.000)	1.000 (1.000, 1.000)	0.265 (0.239, 0.289)
Enhancement Foci	11	0.950 (0.901, 1.000)	0.909 (0.818, 1.000)	0.878 (0.858, 0.898)
Skull Base Fracture	11	0.981 (0.962, 1.000)	1.000 (1.000, 1.000)	0.837 (0.815, 0.858)

retrospective				
	Num	**AUC**	**Sensitivity**	**Specificity**
Zygomatic Arch Fracture	11	0.870 (0.741, 1.000)	0.818 (0.636, 1.000)	0.757 (0.732, 0.781)
Epidural Hemorrhage	10	0.936 (0.873, 1.000)	1.000 (1.000, 1.000)	0.455 (0.427, 0.483)
Fat Deposition	9	0.969 (0.943, 1.000)	1.000 (1.000, 1.000)	0.878 (0.860, 0.898)
Fat Density	9	0.858 (0.745, 0.992)	0.778 (0.556, 1.000)	0.750 (0.726, 0.775)
Brainstem Ischemia	9	0.817 (0.642, 1.000)	1.000 (1.000, 1.000)	0.222 (0.198, 0.246)
Empty Sella	8	0.869 (0.743, 1.000)	0.875 (0.750, 1.000)	0.585 (0.556, 0.613)
Artery Tortuosity	8	0.786 (0.604, 1.000)	0.875 (0.750, 1.000)	0.333 (0.306, 0.361)
Aneurysm	8	0.822 (0.691, 0.979)	0.625 (0.375, 1.000)	1.000 (1.000, 1.000)
Ethmoid Fracture	7	0.901 (0.815, 1.000)	1.000 (1.000, 1.000)	0.537 (0.507, 0.566)
Tuberous Sclerosis Complex	6	0.806 (0.613, 1.000)	0.833 (0.667, 1.000)	0.699 (0.673, 0.726)
Gliomas	5	0.999 (0.999, 1.000)	1.000 (1.000, 1.000)	0.997 (0.995, 1.000)
Subacute Cerebral Infarction	5	0.872 (0.759, 1.000)	1.000 (1.000, 1.000)	0.525 (0.497, 0.553)
Insula Ischemia	5	0.828 (0.684, 1.000)	0.800 (0.600, 1.000)	0.785 (0.761, 0.811)
Cerebellar Lesion	4	0.887 (0.776, 1.000)	1.000 (1.000, 1.000)	0.624 (0.595, 0.651)
Ethmoid Paperboard Fracture	4	0.815 (0.630, 1.000)	0.750 (0.500, 1.000)	0.743 (0.718, 0.767)
Pons Ischemia	4	0.975 (0.950, 1.000)	1.000 (1.000, 1.000)	0.948 (0.935, 0.961)
Cerebral Fissure	3	0.987 (0.973, 1.000)	1.000 (1.000, 1.000)	0.960 (0.948, 0.972)
Demyelination	3	0.876 (0.752, 1.000)	1.000 (1.000, 1.000)	0.628 (0.601, 0.656)
Widened Ventricles	3	0.975 (0.950, 1.000)	1.000 (1.000, 1.000)	0.926 (0.910, 0.941)
Widened Subdural Space	3	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Cisterna Magna Cyst	3	0.695 (0.389, 1.000)	1.000 (1.000, 1.000)	0.282 (0.256, 0.309)
Cerebral Hemisphere Compression	2	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Bony Process	2	0.625 (0.249, 1.000)	1.000 (1.000, 1.000)	0.249 (0.224, 0.274)
Arterial Thickening	2	0.400 (0.229, 0.570)	1.000 (1.000, 1.000)	0.247 (0.223, 0.271)
Pineal Gland Enlargement	2	0.951 (0.902, 1.000)	1.000 (1.000, 1.000)	0.902 (0.885, 0.920)
Dysplasia	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Cerebellar Compression	1	0.739 (0.712, 0.764)	1.000 (1.000, 1.000)	0.739 (0.712, 0.764)
Negative Occupancy Effect	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Pellet Press	1	0.866 (0.847, 0.886)	1.000 (1.000, 1.000)	0.866 (0.847, 0.886)
Widened Tentorium	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Widened Lambdoid Sutures	1	0.932 (0.918, 0.948)	1.000 (1.000, 1.000)	0.932 (0.918, 0.948)
Ventricular Reduction	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Adenoid Hypertrophy	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Cystic Pineal Gland	1	0.937 (0.924, 0.951)	1.000 (1.000, 1.000)	0.937 (0.924, 0.951)
Frontal Lobe Cyst	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Neuroma	1	0.997 (0.995, 1.000)	1.000 (1.000, 1.000)	0.997 (0.995, 1.000)
Thyroid Cartilage Fracture	1	0.932 (0.917, 0.948)	1.000 (1.000, 1.000)	0.932 (0.917, 0.948)
Capsular Infarction	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Massive Hemorrhage	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
avg	127/127	0.883 (0.858, 0.911)	0.810 (0.765, 0.858)	0.835 (0.815, 0.856)

Table 2 | Performance on the prospective dataset for head disorder detection.

	prospective			
	Num	**AUC**	**Sensitivity**	**Specificity**
Ischemia	473	0.903 (0.884, 0.923)	0.833 (0.801, 0.867)	0.804 (0.761, 0.847)
Malacia Foci	460	0.931 (0.916, 0.948)	0.848 (0.817, 0.880)	0.847 (0.807, 0.887)
Parenchymal Hemorrhage	220	0.956 (0.941, 0.973)	0.900 (0.864, 0.941)	0.844 (0.801, 0.887)
Pneumocephaly	213	0.931 (0.906, 0.958)	0.878 (0.836, 0.925)	0.847 (0.807, 0.887)
Subarachnoid Hemorrhage	213	0.949 (0.928, 0.972)	0.897 (0.859, 0.939)	0.877 (0.844, 0.914)
Brainswelling	193	0.830 (0.793, 0.868)	0.777 (0.725, 0.839)	0.715 (0.669, 0.761)
Effusion	168	0.940 (0.918, 0.965)	0.875 (0.827, 0.929)	0.847 (0.807, 0.887)
Soft Tissue Swelling	142	0.825 (0.782, 0.871)	0.761 (0.690, 0.831)	0.748 (0.702, 0.798)
Subdural Hematoma	123	0.939 (0.911, 0.970)	0.886 (0.837, 0.943)	0.840 (0.801, 0.880)
Calcification	115	0.792 (0.736, 0.850)	0.757 (0.678, 0.835)	0.718 (0.672, 0.767)
Lacunar Infarction	109	0.662 (0.599, 0.729)	0.596 (0.505, 0.688)	0.684 (0.635, 0.733)
Contusion	107	0.931 (0.902, 0.965)	0.850 (0.785, 0.925)	0.844 (0.807, 0.883)
Basal Ganglia Cerebral Infarction	92	0.646 (0.577, 0.718)	0.543 (0.446, 0.641)	0.724 (0.675, 0.773)
Space Occupying Lesions	90	0.880 (0.836, 0.927)	0.800 (0.722, 0.878)	0.748 (0.699, 0.798)
Intraventricular Hemorrhage	83	0.956 (0.929, 0.990)	0.904 (0.843, 0.964)	0.847 (0.807, 0.883)
Sclerosis	81	0.854 (0.799, 0.910)	0.741 (0.642, 0.827)	0.850 (0.810, 0.890)
Tubercle	78	0.797 (0.735, 0.862)	0.718 (0.628, 0.821)	0.776 (0.730, 0.822)
Subcutaneous Hematoma	78	0.829 (0.771, 0.892)	0.782 (0.692, 0.885)	0.755 (0.712, 0.804)
Midline Shift	74	0.602 (0.519, 0.688)	0.500 (0.392, 0.608)	0.678 (0.629, 0.727)
Cavum Septum Pellucidum	72	0.747 (0.680, 0.817)	0.625 (0.514, 0.736)	0.788 (0.745, 0.831)
Cerebral Infarction Of Caudate Nucleus Head	72	0.889 (0.841, 0.943)	0.847 (0.764, 0.931)	0.782 (0.736, 0.828)
Arachnoid Cyst	67	0.658 (0.586, 0.738)	0.642 (0.537, 0.761)	0.604 (0.552, 0.656)
Cisterna Magna	66	0.785 (0.720, 0.860)	0.697 (0.591, 0.818)	0.755 (0.712, 0.804)
Artifacts Shadow	65	0.954 (0.924, 0.990)	0.908 (0.846, 0.985)	0.847 (0.810, 0.887)
Brain Atrophy	63	0.963 (0.939, 0.996)	0.921 (0.857, 1.000)	0.893 (0.859, 0.926)
Epidural Hematoma	62	0.958 (0.932, 0.990)	0.903 (0.839, 0.984)	0.868 (0.831, 0.905)
Basal Ganglia Ischemia	61	0.821 (0.758, 0.887)	0.738 (0.639, 0.852)	0.752 (0.706, 0.798)
Minor Hemorrhage	60	0.962 (0.937, 0.993)	0.917 (0.850, 0.983)	0.896 (0.862, 0.929)
Multiple Cerebral Infarctions	59	0.836 (0.770, 0.910)	0.763 (0.661, 0.881)	0.807 (0.767, 0.850)
Cerebral Edema	58	0.924 (0.881, 0.978)	0.914 (0.845, 1.000)	0.807 (0.764, 0.850)
Ventriculomegaly	55	0.945 (0.912, 0.986)	0.945 (0.891, 1.000)	0.755 (0.709, 0.801)
Paraventricular Ischemia	53	0.823 (0.763, 0.888)	0.660 (0.547, 0.792)	0.807 (0.767, 0.853)
Atherosclerosis	48	0.930 (0.893, 0.973)	0.875 (0.792, 0.979)	0.847 (0.810, 0.887)
Scalp Hematoma	48	0.797 (0.718, 0.878)	0.667 (0.542, 0.792)	0.837 (0.801, 0.877)
Periventricular Cerebral Infarction	47	0.666 (0.577, 0.760)	0.574 (0.447, 0.723)	0.693 (0.644, 0.742)
Subdural Hemorrhage	47	0.901 (0.856, 0.954)	0.894 (0.809, 1.000)	0.730 (0.684, 0.779)
Senile Encephalopathy	44	0.978 (0.961, 1.000)	0.977 (0.955, 1.000)	0.896 (0.865, 0.929)
Hydrocephalus	44	0.954 (0.916, 1.000)	0.932 (0.864, 1.000)	0.847 (0.810, 0.887)
Occipital Fracture	42	0.823 (0.742, 0.910)	0.810 (0.690, 0.929)	0.736 (0.690, 0.785)
Osteoma	41	0.789 (0.698, 0.887)	0.732 (0.610, 0.854)	0.755 (0.709, 0.804)
Parietal Fracture	35	0.836 (0.756, 0.924)	0.743 (0.600, 0.886)	0.773 (0.727, 0.819)
Frontal Lobe Ischemia	35	0.750 (0.662, 0.843)	0.629 (0.457, 0.800)	0.758 (0.712, 0.804)

prospective				
	Num	**AUC**	**Sensitivity**	**Specificity**
Parenchymal Hematoma	35	0.961 (0.934, 1.000)	0.914 (0.829, 1.000)	0.847 (0.810, 0.887)
Herniation	32	0.990 (0.979, 1.000)	1.000 (1.000, 1.000)	0.672 (0.620, 0.724)
Lipoma	30	0.856 (0.788, 0.932)	0.800 (0.667, 0.967)	0.785 (0.742, 0.828)
Arteriosclerosis	23	0.730 (0.601, 0.868)	0.565 (0.348, 0.783)	0.899 (0.868, 0.933)
Frontal Bone Fracture	23	0.731 (0.613, 0.862)	0.696 (0.522, 0.870)	0.666 (0.617, 0.718)
Hemispheric Infarction	23	0.991 (0.983, 1.000)	0.957 (0.913, 1.000)	0.951 (0.929, 0.975)
Absence Of Bone	22	0.823 (0.729, 0.927)	0.818 (0.682, 1.000)	0.656 (0.604, 0.709)
Temporal Lobe Cerebral Infarction	21	0.822 (0.709, 0.956)	0.810 (0.667, 1.000)	0.752 (0.706, 0.801)
Frontal Lobe Cerebral Infarction	21	0.712 (0.551, 0.882)	0.619 (0.429, 0.857)	0.868 (0.831, 0.908)
Corona Radiata Cerebral Infarction	18	0.614 (0.444, 0.794)	0.556 (0.333, 0.778)	0.752 (0.706, 0.801)
Parietal Lobe Ischemia	17	0.827 (0.721, 0.953)	0.824 (0.647, 1.000)	0.669 (0.617, 0.718)
Occipital Lobe Cerebral Infarction	16	0.943 (0.894, 1.000)	0.875 (0.750, 1.000)	0.804 (0.761, 0.850)
Zygomatic Fracture	15	0.904 (0.830, 0.993)	0.867 (0.733, 1.000)	0.675 (0.623, 0.724)
Temporal Bone Fracture	15	0.693 (0.511, 0.886)	0.600 (0.333, 0.867)	0.951 (0.929, 0.975)
Parietal Lobe Cerebral Infarction	15	0.854 (0.736, 1.000)	0.867 (0.733, 1.000)	0.721 (0.675, 0.770)
Malformation	14	0.800 (0.698, 0.910)	0.786 (0.571, 1.000)	0.727 (0.678, 0.776)
Widened Cisterna Magna	14	0.872 (0.789, 0.965)	0.857 (0.714, 1.000)	0.758 (0.712, 0.804)
Cystic Foci	14	0.675 (0.530, 0.827)	0.571 (0.286, 0.857)	0.718 (0.669, 0.767)
Wallerian Degeneration	13	0.997 (0.994, 1.000)	1.000 (1.000, 1.000)	0.979 (0.963, 0.994)
Empty Sella	13	0.931 (0.881, 1.000)	0.923 (0.846, 1.000)	0.816 (0.776, 0.856)
Septum Pellucidum Cyst	13	0.515 (0.350, 0.696)	0.462 (0.231, 0.769)	0.764 (0.718, 0.810)
Skull Fracture	13	0.900 (0.807, 1.000)	0.923 (0.846, 1.000)	0.733 (0.687, 0.779)
Metal Shadow	12	0.923 (0.856, 1.000)	0.917 (0.833, 1.000)	0.899 (0.868, 0.933)
Zygomatic Arch Fracture	12	0.895 (0.820, 0.993)	0.833 (0.667, 1.000)	0.788 (0.745, 0.834)
Centrum Semiovale Ischemia	12	0.671 (0.492, 0.865)	0.667 (0.417, 0.917)	0.727 (0.678, 0.773)
Brainstem Ischemia	12	0.841 (0.684, 1.000)	0.833 (0.667, 1.000)	0.604 (0.552, 0.660)
Osseous Defects	11	0.886 (0.799, 1.000)	0.909 (0.818, 1.000)	0.623 (0.571, 0.672)
Widened Ventricles	10	0.876 (0.752, 1.000)	0.900 (0.800, 1.000)	0.604 (0.549, 0.660)
Corona Radiata Ischemia	10	0.728 (0.549, 0.921)	0.700 (0.400, 1.000)	0.629 (0.577, 0.678)
Cerebral Fissure	8	0.729 (0.524, 0.987)	0.625 (0.375, 1.000)	0.963 (0.945, 0.985)
Fat Density	8	0.777 (0.598, 1.000)	0.875 (0.750, 1.000)	0.531 (0.475, 0.586)
Insular Infarction	8	0.855 (0.710, 1.000)	0.875 (0.750, 1.000)	0.537 (0.485, 0.595)
Cerebral White Matter Degeneration	7	0.604 (0.331, 0.906)	0.571 (0.286, 1.000)	0.764 (0.718, 0.813)
Metastases	7	0.970 (0.940, 1.000)	1.000 (1.000, 1.000)	0.868 (0.834, 0.905)
Thalamic Cerebral Infarction	7	0.536 (0.332, 0.747)	0.714 (0.429, 1.000)	0.457 (0.402, 0.509)
Old Cerebral Infarction	7	0.692 (0.457, 0.952)	0.857 (0.714, 1.000)	0.273 (0.224, 0.322)
Widened Subdural Space	6	0.860 (0.728, 1.000)	0.833 (0.667, 1.000)	0.794 (0.752, 0.840)
Meningioma	6	0.785 (0.601, 0.998)	1.000 (1.000, 1.000)	0.448 (0.393, 0.500)
Centrum Semiovale Cerebral Infarction	6	0.522 (0.267, 0.787)	0.667 (0.333, 1.000)	0.503 (0.451, 0.555)
Bony Process	5	0.434 (0.185, 0.675)	0.400 (0.000, 0.800)	0.733 (0.687, 0.779)
Fat Deposition	5	0.958 (0.915, 1.000)	1.000 (1.000, 1.000)	0.788 (0.748, 0.831)
Tumour	5	0.780 (0.577, 1.000)	0.800 (0.600, 1.000)	0.779 (0.733, 0.825)

prospective				
	Num	**AUC**	**Sensitivity**	**Specificity**
Multiple Hemorrhages	5	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Enhancement Foci	4	0.939 (0.879, 1.000)	1.000 (1.000, 1.000)	0.834 (0.794, 0.874)
Widened Tentorium	4	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Cisterna Magna Cyst	4	0.929 (0.859, 1.000)	1.000 (1.000, 1.000)	0.736 (0.690, 0.785)
Cerebellar Hemisphere Infarction	4	0.781 (0.561, 1.000)	1.000 (1.000, 1.000)	0.334 (0.282, 0.383)
Arterial Thickening	3	0.976 (0.953, 1.000)	1.000 (1.000, 1.000)	0.929 (0.905, 0.960)
Artery Tortuosity	3	0.856 (0.764, 0.985)	1.000 (1.000, 1.000)	0.721 (0.675, 0.770)
Widened Arterial	3	0.889 (0.777, 1.000)	1.000 (1.000, 1.000)	0.693 (0.638, 0.745)
Sphenoid Fracture	3	0.942 (0.883, 1.000)	1.000 (1.000, 1.000)	0.825 (0.785, 0.868)
Brainstem Infarction	3	0.940 (0.879, 1.000)	1.000 (1.000, 1.000)	0.868 (0.831, 0.908)
Subacute Cerebral Infarction	3	0.951 (0.902, 1.000)	1.000 (1.000, 1.000)	0.902 (0.871, 0.933)
Epidural Hemorrhage	3	0.984 (0.967, 1.000)	1.000 (1.000, 1.000)	0.951 (0.929, 0.975)
Brainstem Swelling	2	0.992 (0.985, 1.000)	1.000 (1.000, 1.000)	0.985 (0.972, 1.000)
Arterial Cilia	2	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Widened Subarachnoid Space	2	0.739 (0.632, 0.850)	1.000 (1.000, 1.000)	0.660 (0.610, 0.712)
Myeloma	2	0.502 (0.003, 1.000)	1.000 (1.000, 1.000)	0.003 (0.000, 0.003)
Dysplasia	1	0.816 (0.776, 0.859)	1.000 (1.000, 1.000)	0.816 (0.776, 0.859)
Cerebral Hemisphere Compression	1	0.991 (0.982, 1.000)	1.000 (1.000, 1.000)	0.991 (0.982, 1.000)
Pellet Press	1	0.985 (0.972, 1.000)	1.000 (1.000, 1.000)	0.985 (0.972, 1.000)
Leukoaraiosis	1	0.816 (0.773, 0.859)	1.000 (1.000, 1.000)	0.816 (0.773, 0.859)
Cerebral Cortical Atrophy	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Ventricular Reduction	1	0.991 (0.982, 1.000)	1.000 (1.000, 1.000)	0.991 (0.982, 1.000)
Pineal Gland Enlargement	1	0.991 (0.982, 1.000)	1.000 (1.000, 1.000)	0.991 (0.982, 1.000)
Sella Turcica Enlargement	1	0.979 (0.963, 0.997)	1.000 (1.000, 1.000)	0.979 (0.963, 0.997)
Choroid Cyst	1	0.629 (0.577, 0.681)	1.000 (1.000, 1.000)	0.629 (0.577, 0.681)
Frontal Lobe Cyst	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Facial Bone Fracture	1	0.258 (0.209, 0.301)	1.000 (1.000, 1.000)	0.258 (0.209, 0.301)
Pons Ischemia	1	0.908 (0.877, 0.939)	1.000 (1.000, 1.000)	0.908 (0.877, 0.939)
Insula Ischemia	1	0.736 (0.690, 0.785)	1.000 (1.000, 1.000)	0.736 (0.690, 0.785)
Minor Haematoma	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Massive Hemorrhage	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Cerebellar Vermis Hemorrhage	1	0.828 (0.788, 0.868)	1.000 (1.000, 1.000)	0.828 (0.788, 0.868)
avg	116/116	0.868 (0.821, 0.918)	0.809 (0.737, 0.888)	0.795 (0.753, 0.838)

Table 3 | Performance on the cross-center dataset for head disorder detection.

	cross-center			
	Num	**AUC**	**Sensitivity**	**Specificity**
Lacunar Infarction	100	0.862 (0.861, 0.864)	0.839 (0.837, 0.842)	0.781 (0.779, 0.784)
Senile Encephalopathy	99	0.879 (0.877, 0.880)	0.848 (0.845, 0.852)	0.788 (0.784, 0.791)
Subarachnoid Hemorrhage	96	0.903 (0.901, 0.904)	0.818 (0.815, 0.821)	0.869 (0.866, 0.872)
Parenchymal Hemorrhage	88	0.843 (0.840, 0.844)	0.778 (0.775, 0.781)	0.796 (0.793, 0.798)
Contusion	64	0.858 (0.854, 0.857)	0.731 (0.728, 0.734)	0.877 (0.874, 0.880)
Malacia Foci	56	0.843 (0.842, 0.846)	0.784 (0.781, 0.788)	0.809 (0.806, 0.812)
Subdural Hematoma	52	0.877 (0.875, 0.879)	0.834 (0.831, 0.837)	0.857 (0.854, 0.861)
Parenchymal Hematoma	43	0.798 (0.796, 0.801)	0.677 (0.674, 0.680)	0.854 (0.851, 0.857)
Intraventricular Hemorrhage	41	0.923 (0.922, 0.925)	0.855 (0.853, 0.857)	0.968 (0.966, 0.969)
Soft Tissue Injury	39	0.857 (0.855, 0.858)	0.719 (0.716, 0.722)	0.966 (0.965, 0.968)
Epidural Hematoma	22	0.872 (0.869, 0.873)	0.772 (0.769, 0.775)	1.000 (1.000, 1.000)
Leukoaraiosis	21	0.870 (0.867, 0.870)	0.821 (0.816, 0.826)	0.773 (0.769, 0.778)
Brain Atrophy	19	0.884 (0.881, 0.884)	0.789 (0.786, 0.791)	0.932 (0.930, 0.934)
Pneumocephaly	19	0.941 (0.941, 0.943)	0.895 (0.893, 897)	0.942 (0.941, 0.944)
Effusion	17	0.914 (0.913, 0.916)	0.904 (0.901, 0.906)	0.910 (0.907, 0.912)
Scalp Hematoma	14	0.971 (0.970, 0.972)	0.928 (0.926, 0.929)	1.000 (1.000, 1.000)
Brainswelling	12	0.924 (0.921, 0.923)	0.831 (0.829, 0.833)	1.000 (1.000, 1.000)
Herniation	9	0.833 (0.832, 0.835)	0.758 (0.754, 0.763)	0.797 (0.792, 0.802)
Soft Tissue Swelling	7	0.952 (0.952, 0.954)	0.859 (0.857, 0.861)	1.000 (0.999, 1.000)
Occipital Lobe Cerebral Infarction	6	0.944 (0.943, 0.945)	0.833 (0.831, 0.836)	1.000 (1.000, 1.000)
Basal Ganglia Cerebral Infarction	5	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Arachnoid Cyst	4	0.625 (0.622, 0.628)	0.530 (0.525, 0.535)	0.956 (0.948, 0.965)
Frontal Lobe Cerebral Infarction	4	0.625 (0.624, 0.629)	0.634 (0.627, 0.642)	0.640 (0.632, 0.647)
Calcification	3	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Parietal Lobe Cerebral Infarction	3	0.667 (0.661, 0.667)	1.000 (1.000, 1.000)	0.664 (0.661, 0.667)
Cerebellar Hemisphere Infarction	3	0.778 (0.774, 0.779)	0.665 (0.662, 0.668)	1.000 (1.000, 1.000)
Temporal Lobe Cerebral Infarction	3	0.778 (0.776, 0.780)	0.666 (0.663, 0.669)	1.000 (1.000, 1.000)
Wallerian Degeneration	3	0.889 (0.887, 0.890)	0.841 (0.831, 0.851)	0.853 (0.843, 0.863)
Insular Infarction	3	0.778 (0.775, 0.779)	1.000 (1.000, 1.000)	0.666 (0.663, 0.669)
Hemispheric Infarction	2	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Cerebral Edema	2	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Subdural Hemorrhage	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Multiple Hemorrhages	1	0.000 (0.000, 0.000)	0.000(0.000, 00)	0.000(0.000, 0.000)
Subcutaneous Hematoma	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Minor Hemorrhage	1	0.000 (0.000, 0.000)	0.000(0.000, 00)	0.000(0.000, 0.000)
Brainstem Infarction	1	0.000 (0.000, 0.000)	0.000(0.000, 00)	0.000(0.000, 0.000)
Temporal Bone Fracture	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Fat Density	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Midline Shift	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Thalamic Cerebral Infarction	1	0.000 (0.000, 0.000)	0.000(0.000, 00)	0.000(0.000, 0.000)
Cerebellar Vermis Hemorrhage	1	0.000 (0.000, 0.000)	0.000(0.000, 00)	0.000(0.000, 0.000)
Subacute Cerebral Infarction	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Malformation	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Hydrocephalus	1	0.000 (0.000, 0.000)	0.000(0.000, 00)	0.000(0.000, 0.000)
Centrum Semiovale Cerebral Infarction	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Periventricular Cerebral Infarction	1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
avg	46/46	0.866 (0.865, 0.868)	0.803 (0.800, 0.806)	0.852 (0.850, 0.855)

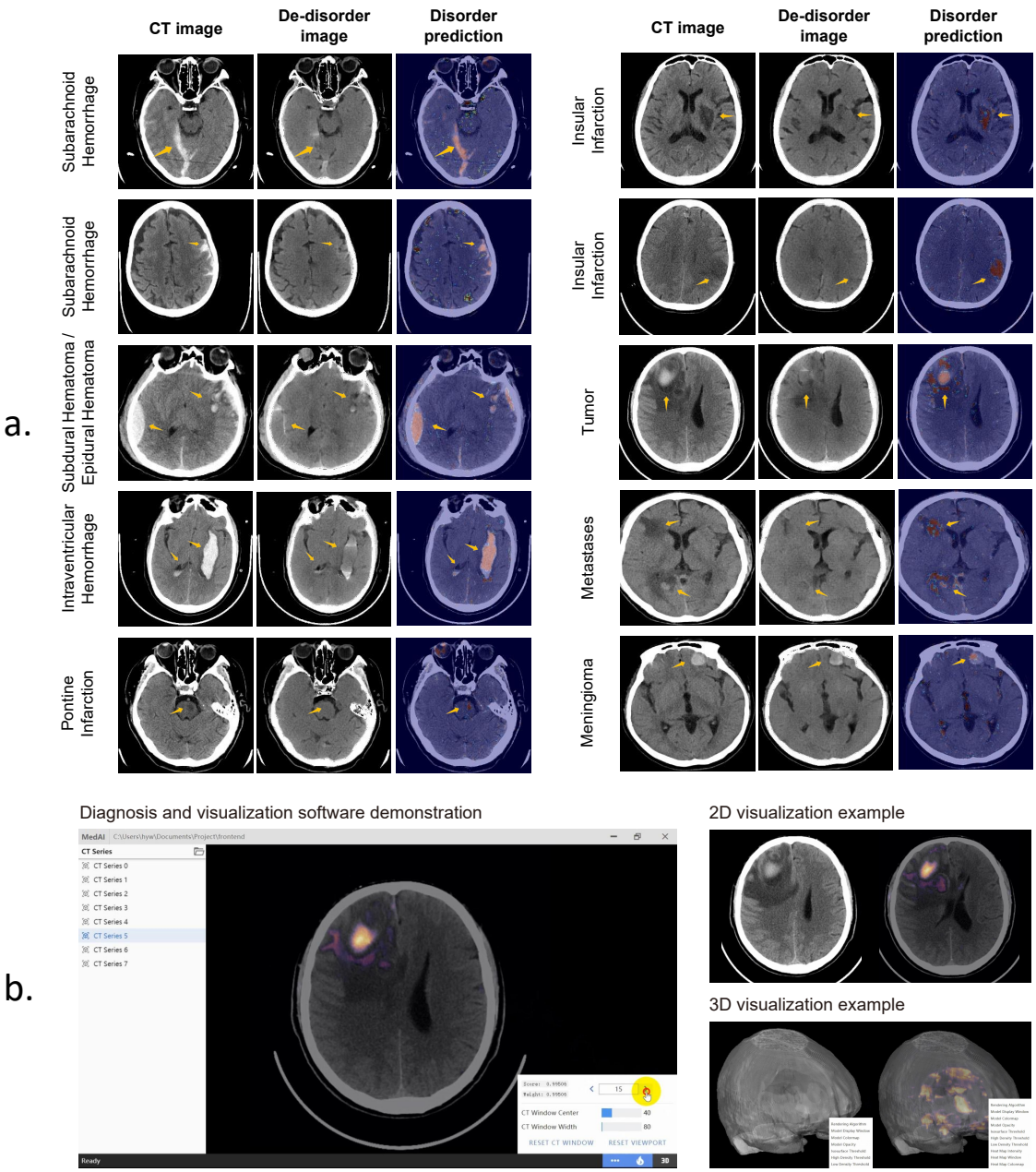

Fig. 1 | Visualization examples for head disorder detection. a) Typical examples of our system's performance are shown, including the original CT image, the corresponding de-disorder image generated by our system, and the heatmap indicating the probability of containing disorders. Warmer colors in the heatmap indicate a higher probability of disorders. The heatmaps provide visual clues to the system's decision. **b)** We also developed a diagnosis and visualization software that takes a CT scan as input and outputs possible disorder locations in the form of a heatmap. The heatmap can be displayed on a 2D slice or on a 3D reconstruction scan.

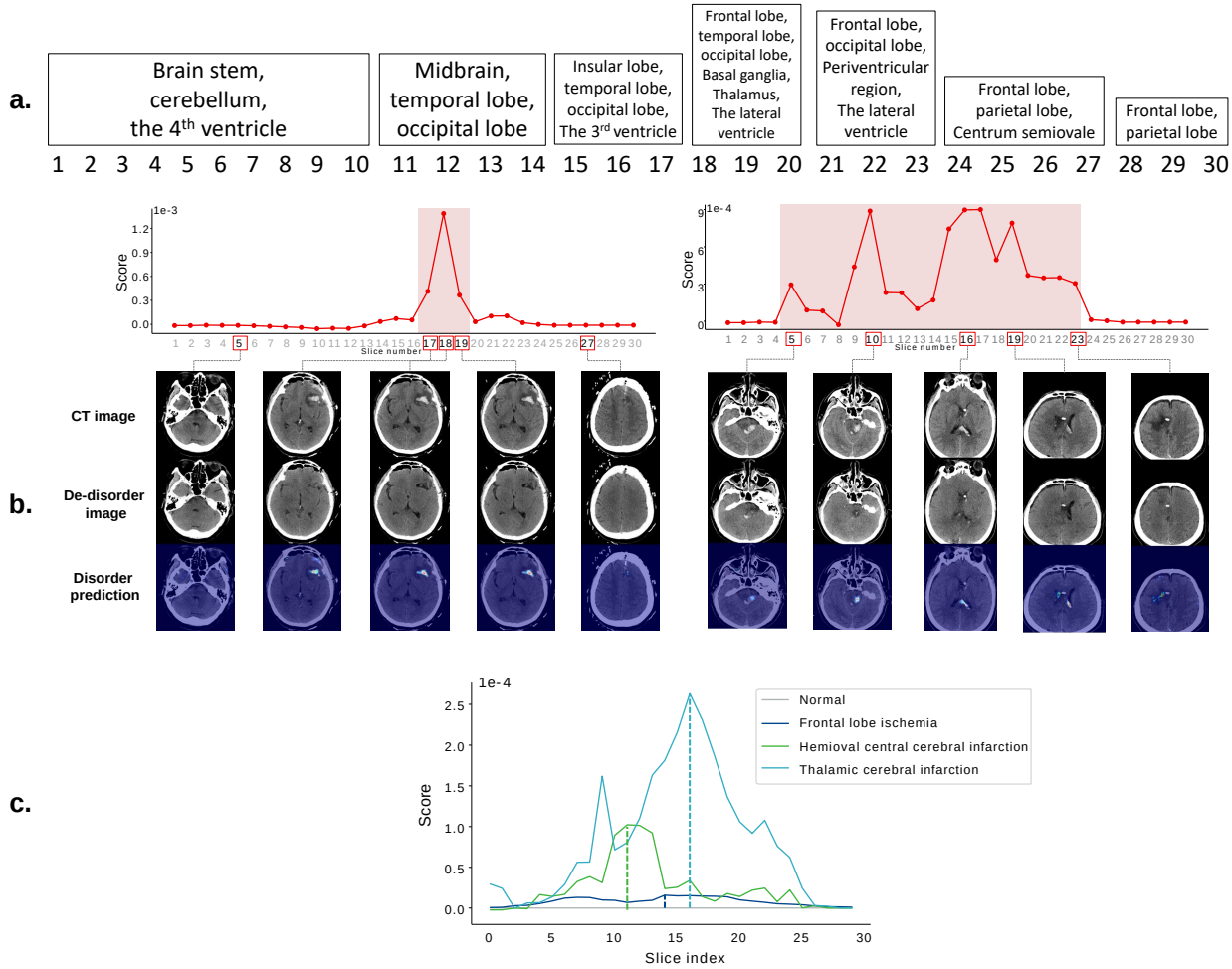

Fig. 2 | Examples of disorder distributions.The x-axis of the graph represents the slice index of a CT scan, while the y-axis represents a slice's abnormal score. A higher score indicates a greater likelihood of lesions in the slice. **a)** The correspondence between slice indexes and brain tissues is shown. **b)** Two example scans with dispersive and centralized distributions are presented. **c)** The average disorder distributions of some typical disorders in the retrospective dataset are displayed.

Performance contribution from de-disorder network. In ISL, the evaluation of the probability of disorder containment depends on the difference image $\mathbf{x}_{dif} = ||\mathbf{x} - \mathbf{x}_{dd}||$, where $\mathbf{x}_{dd}$ is a de-disorder image generated by a de-disorder network (DeDN). Original medical images are too complex for a system to learn disorder-related information solely based on them. Therefore, we do not directly apply original images for evaluation. Using difference images for evaluation is more intuitive, as

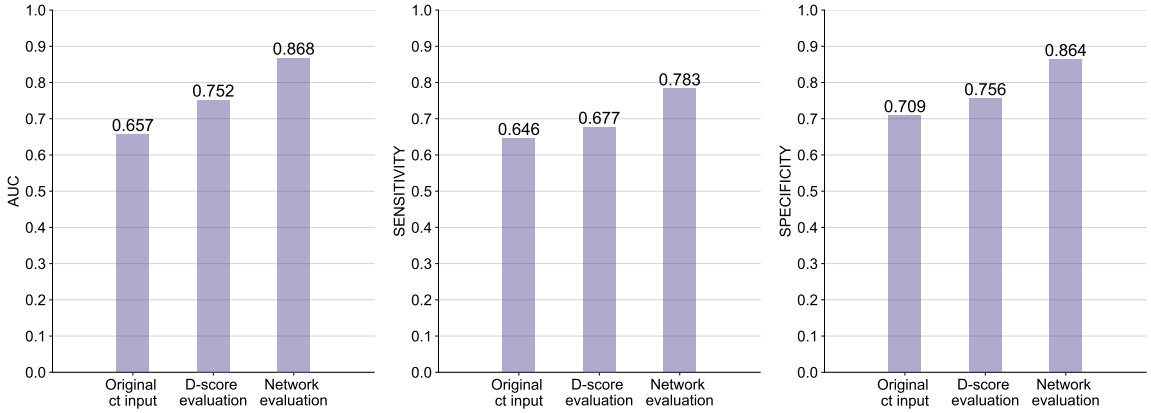

Fig. 3 | Iterative performance improvement by de-disorder network (DeDN) and disorder recognition network (DRN). Original ct input: original CT slices + DRN; **D-score evaluation**: DeDN + pixel value sum; **Network evaluation**: DeDN + DRN. The comparison was performed on the prospective test dataset.

the greater the difference between the original and de-disorder images, the higher the probability of disorder containment.

To numerically demonstrate the effectiveness of the difference image generated by the DeDN, we also applied the original-image-based method for evaluation. As shown in Fig. 3, on the prospective test dataset, the average AUC with 95% confidence interval is 0.657 and 0.752, respectively, where the result from the original-image-based method (**Original CT input**) is significantly lower than that from the difference-image-based method (**D-score evaluation**). The improved result highlights the value of the DeDN.

Performance contribution from disorder recognition network. After obtaining the difference image $\mathbf{x}_{dif}$ with a DeDN, we used a disorder recognition network (DRN) to evaluate the probability of disorder containment. Although we could determine the probability directly based on the pixel value sum of the difference image, we did not adopt this strategy. This is because a DeDN cannot

produce perfectly healthy tissue, meaning that even for a healthy area, the pixel value sum of that area may still be positive. As a result, the accumulated pixel value sum of all healthy areas would negatively influence the probability evaluation.

Instead, we used the pixel sum-based evaluation (**D-score evaluation**) and DRN-based evaluation (**Network evaluation**) based on the difference image, as shown in Fig. 3. The average AUC of the two methods on the prospective dataset were 0.752 and 0.868, respectively, demonstrating the superiority of the DRN.

Evaluation of system generalizability. To be practical, an AI-based system should be able to generalize to new data from different centers and hospitals. In order to evaluate the generalizability of the ISL-based system, we constructed a cross-center test dataset from the Brain Hospital of Hunan Province (BHHP), which served as an independent test cohort from PLAGH. This dataset consisted of 554 scans, of which 59.01% had 46 different types of disorders. The AUCs with 95% confidence interval (CI) for the 46 types of disorders are presented in Table 3, along with the sensitivity and specificity of the system for the cross-center test. The average AUC was 0.866, which was only 0.017 lower than that of the retrospective intra-center test. These results demonstrate the generalizability of the system across different centers.

Evaluation of inverse-supervised learning generalizability. To assess the generalizability of ISL across different body parts and medical image types, we employed it to develop two additional systems. The first system is designed for detecting pulmonary disorders in CT images, while the second system is designed for detecting retinal disorders in optical coherence tomography (OCT)

Table 4 | Performance on the pulmonary CT test dataset for pulmonary disorder detection.

pulmonary CT				
	Num	**AUC**	**Sensitivity**	**Specificity**
pneumothorax	82	0.992 (0.992, 0.992)	0.996 (0.995, 0.996)	0.937 (0.935, 0.939)
pneumonia	86	0.911 (0.909, 0.912)	0.874 (0.869, 0.878)	0.830 (0.828, 0.831)
bronchiectasis	96	0.811 (0.807, 0.811)	0.697 (0.689, 0.704)	0.816 (0.815, 0.817)
bullae	88	0.786 (0.782, 0.787)	0.589 (0.582, 0.595)	0.827 (0.826, 0.828)
atelectasis	82	0.952 (0.951, 0.952)	0.984 (0.982, 0.986)	0.853 (0.852, 0.855)
effusion	46	0.958 (0.956, 0.958)	0.999 (0.999, 1.000)	0.883 (0.880, 0.885)
avg	6/6	0.893 (0.891, 0.894)	0.838 (0.834, 0.842)	0.854 (0.853, 0.855)

Table 5 | Performance on the retinal OCT test dataset for retinal disorder detection.

retinal OCT				
	Num	**AUC**	**Sensitivity**	**Specificity**
CNV	1,220	0.939 (0.938, 0.940)	0.847 (0.844, 0.851)	0.904 (0.900, 0.907)
DME	1,600	0.913 (0.912, 0.914)	0.838 (0.835, 0.841)	0.890 (0.888, 0.893)
DRUSEN	1,220	0.827 (0.826, 0.829)	0.760 (0.757, 0.764)	0.762 (0.758, 0.765)
avg	3/3	0.895 (0.894, 0.896)	0.817 (0.814, 0.820)	0.855 (0.852, 0.858)

images.

Performance of pulmonary disorder detection. In addition to brain CT, we developed an ISL-based system for detecting disorders in pulmonary CT scans. The data used for system development were collected from the First Affiliate Hospital of Guangzhou Medical University (FAHGMU), another leading national hospital that serves patients from across China. We constructed a training dataset consisting of 3,410 healthy pulmonary CT scans and a test dataset that included 6 types of pulmonary disorders (82 pneumothorax, 86 pneumonia, 96 bronchiectasis, 88 bullae, 82 atelectasis, and 46 effusion), as well as 600 healthy scans. The AUCs and detection examples for each type of disorder are presented in Table 4 and Fig. 4a. The average AUC was 0.893, indicating that ISL can generalize well across different body parts.

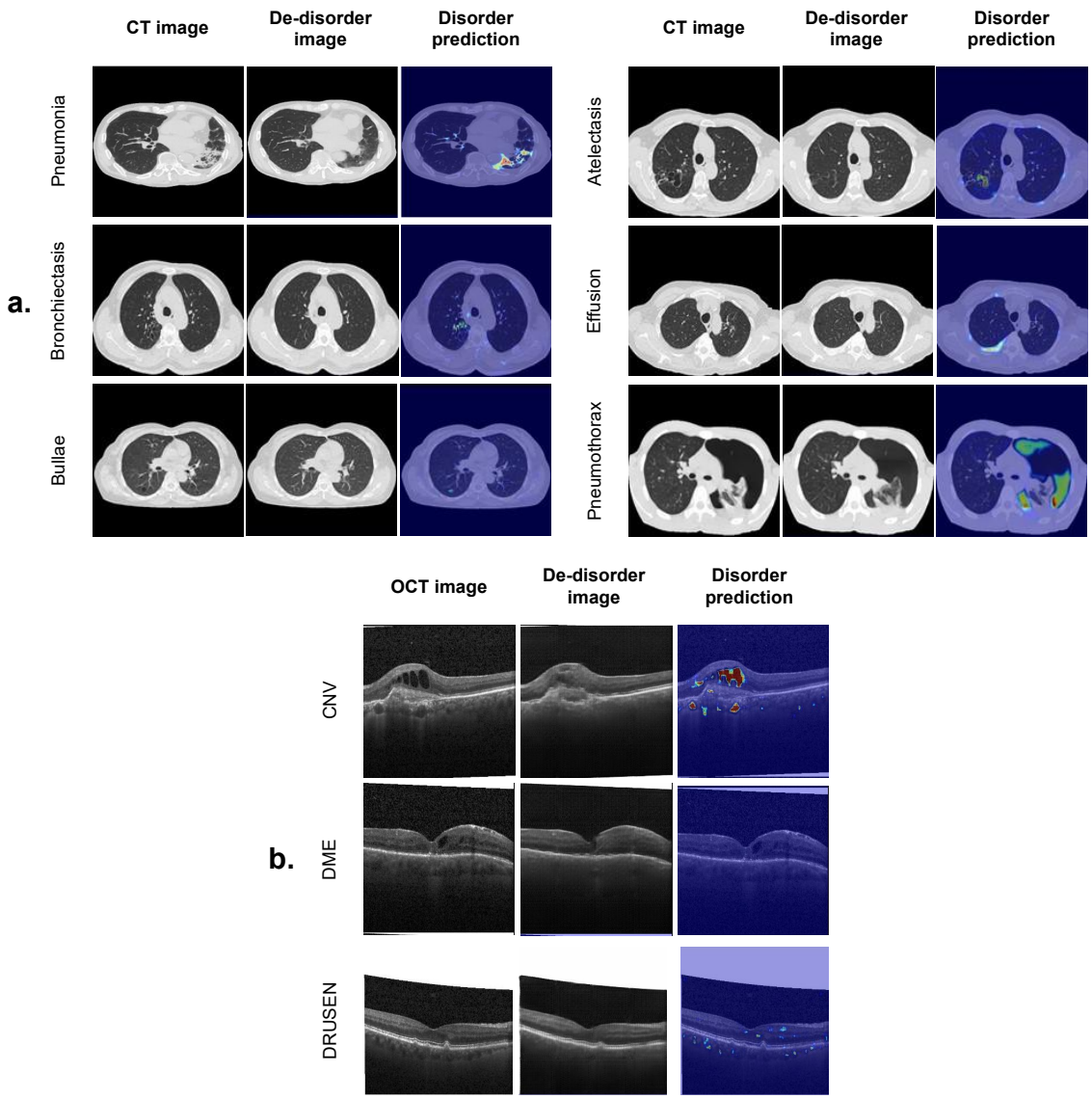

Fig. 4 | Examples of our system on the CT-based pulmonary disorder detection and the OCT-based retinal disorder detection.

Performance of retinal disorder detection. To demonstrate the ability of ISL to generalize across different medical image types, we developed a retinal disorder detection system based on optical coherence tomography (OCT) images. We used the dataset collected by Kermany et al.,[6] which includes 108,312 images (37,206 with choroidal neovascularization, 11,349 with diabetic macular edema, 8,617 with drusen, and 51,140 normal). Following the development procedure of ISL,

we used only the normal OCT images as the training dataset. The model was tested with 1,000 images (250 from each category) from 633 patients, as in Kermany et al.[6] The AUCs with 95% confidence interval (CI) on the scan-level are summarized in Table 5. The AUCs for choroidal neovascularization (CNV), diabetic macular edema (DME), and drusen were 0.939, 0.913, and 0.827, respectively. Despite being developed using only normal OCT images, our system achieved clinically acceptable results, indicating that ISL is applicable to different medical image types. Detection examples for each type of disorder are shown in Fig. 4b.

Discussion

We introduced a learning strategy called inverse supervised learning (ISL) and utilized it to develop a head disorder detection system that requires no disorder data or annotation during the development process. The system's detectable disorder coverage is comparable to that of a human expert. Additionally, the system's excellent generalizability and explainability enhance its clinical applicability.

Annotated and disorder-contained data. Most existing deep-learning medical systems rely on supervised learning, which requires a substantial amount of annotated data to achieve generalizability, accuracy, and recognition gratuity. However, obtaining sufficient annotated data in medical image research is challenging due to the time-consuming and expert knowledge-intensive nature of the notating process. For instance, even for an experienced expert, it may take several minutes to annotate a medical image at the region-level, which provides strong supervision for disorder detection by indicating the exact lesion region. Consequently, research works that rely on region-level

annotation, such as Nikolov et al.[7] and Monteiro et al.,[8] are limited by the amount of annotated data, which hinders the generalizability and accuracy of the system.

To reduce the dependence on annotated data, researchers have explored alternative learning strategies for medical image research. For instance, weakly-supervised learning[9,10,11] allows each training sample to lack a label or have an incorrect label, significantly reducing the annotation cost for experts. Unsupervised learning, on the other hand, uses unannotated training data to enhance the feature representation capacity of a deep learning network, thereby reducing the number of required annotated samples. Self-supervised learning is a recent representative unsupervised learning method[12,13] that annotates each sample by itself instead of relying on human experts. However, these learning strategies require a substantial amount of disorder-contained data to ensure accuracy. Collecting enough disorder-contained data is challenging for general researchers due to ethical and legal considerations, limiting related research to large medical institutions. For example, Chilamkurthy et al.[5] collected over 300,000 brain CT scans from more than 20 medical centers, which is beyond the reach of most researchers.

Compared to previously adopted learning strategies in medical image research, the proposed ISL tackles a challenging task where no annotated or disorder-contained data is available. The only available data is disorder-free data, which can be easily obtained by any medical institution capable of medical imaging scans.

Disorder coverage. The clinical application of medical image research is an important goal. However, most existing works focus on only one or two common disorder types,[5,14] even for systems

developed by institutions with abundant medical resources. For instance, the system[5] is derived from over 300,000 scans, yet it can only recognize four types of disorders. This challenge arises from two aspects. Firstly, it is impractical for researchers to construct models for each disorder type due to the difficulty of collecting and annotating medical images of many disorder types. Secondly, developing models for rare disorders with previous learning strategies is challenging when only a few samples are available. With ISL, researchers do not need to collect data or construct models for specific disorders, enabling the built system to achieve broad-spectrum disorder detection.

Anomaly Detection. Distinguishing disorder-contained images from disorder-free ones can be viewed as an anomaly detection problem, which is a popular research field in machine learning. An intuitive assumption is that anomalies lie outside the distribution of normal samples. Therefore, it is natural to train a classifier to differentiate abnormal samples from normal ones.[15,16]

Recent works have utilized generation networks for anomaly detection, employing two primary approaches: (1) utilizing the latent feature;[17,18,19] and (2) utilizing the reconstructed image.[20,21,22] In the first approach, a generation network produces a latent feature and a reconstructed image when an image is inputted. The latent features can be used to determine whether the image is abnormal. In the second approach, a generation network produces a corresponding normal image for a given image. If the original image contains abnormal characteristics, it can be recognized based on the difference between the original and generated images. In the field of medical image analysis, two types of methods have achieved certain results in specific diseases.[23,24,25] For instance, Yao et al.[23] used the second approach to generate healthy pulmonary CT images, which

were used to determine whether the lungs contained COVID-19.

However, both approaches have limitations when applied to broad-spectrum disorder detection in medical images. In the first approach, medical images are complex, which results in complex latent features. Therefore, recognizing disorder-contained images based solely on latent features is challenging. To demonstrate this, we compared our method with a baseline method that directly fed original medical images into the disorder recognition network. The baseline method achieved an average AUC of 0.653 on the prospective dataset, which is inferior to the results (AUC 0.868) obtained by our method. In the second approach, existing generation-based methods reconstruct the original disorder tissues of a medical image due to the strong feature representation capability of generative networks, which fails to achieve abnormal recognition. With the generation strategy in ISL, only context images and global structure information are provided, allowing the generation network to eliminate the interference of the original disorder tissue and conceive healthy tissues like a radiologist. To showcase the efficacy of our system in comparison to existing techniques, we conducted a comparative analysis with other representative reconstruction-based anomaly detection methods, specifically Auto-Encoder,[26] AnoGAN,[17] GANomaly,[27] pix2pix,[28] and Cycle-GAN.[29] The experiment was carried out on the task of detecting pulmonary disorders. The results of the analysis are presented in Supplementary Table 5. Our system outperformed the baselines, with the highest AUC of 0.846 achieved by GANomaly, which is 0.047 lower than our method. This significant improvement underscores the ability of our ISL-based system to successfully accomplish disorder recognition tasks.

Methods

CT scan collection. Initially, we retrieved 954,508 scans from the PACS of the PLAGH between March 2012 and July 2019. These scans contained CT images stored in DICOM (digital imaging and communications in medicine) format, and all DICOMs were de-identified before data analysis. We then screened the scans by excluding reconstructed scans (processed with algorithms in CT machines), non-axial-section scans (coronal section and sagittal section scans), non-head scans (scans of breast and full-body, etc.), and non-origin scans (scans of CTA and CTP, etc.). The inclusion and exclusion criteria of the screening process are detailed in Fig. 5. After screening, a total of 62,239 CT scans with an average slice number of 28 were selected.

Disorder types statistics. Each retrieved scan includes a clinical report written by an interpreting radiologist during the examination. To determine the disorder types to be interpreted by our system, we first applied a rule-based NLP algorithm to the clinical reports. This algorithm counted the occurrence frequencies of different word phrases. We then invited three radiologists to analyze the frequency statistics results and select the disorder types to be evaluated. Ultimately, 127 types of disorders were selected.

Training dataset and test dataset. The construction of the development and test datasets relied on the clinical reports, which were considered the gold standard. The training dataset only included disorder-free CT scans, and their reports uniformly described them as ”No abnormality is observed.” Therefore, we could efficiently obtain disorder-free CT scans. Ultimately, we selected 22,602 disorder-free scans, which were divided into two parts: the training dataset (21,429 scans)

and the negative samples in the retrospective test dataset (1,173 scans).

Regarding the positive samples (disorder-contained scans) in the test dataset, we initially retrieved 18,514 scans using stated disorder-related keywords. We then invited 30 board-certified radiologists with 6 to 15 years of experience to label each scan based on the images and its associated report. The radiologists assigned a binary label (i.e., 0, 1) to each scan, where 1 indicated that the scan contained the expected disorder type. Ultimately, 8,794 scans were labeled as 1 and selected as the positive samples in the retrospective test dataset. The prospective and cross-center test datasets were constructed similarly. However, due to the smaller data amounts in these two datasets compared to the retrospective test dataset, they also contained a smaller number of disorder types, specifically 116 and 46, respectively. Please refer to Supplementary Table 1-3 for the detailed data statistics.

Developing a system with inversed-supervised learning. ISL allows for the training of a deep learning network without accessing disorder-contained samples, enabling researchers with only general and healthy images to build a broad-spectrum disorder detection system. ISL is built on two technologies: missing information completion and data distribution estimation. Missing information completion enables a system to reconstruct the healthy tissue of masked parts of a medical image using a de-disorder network (DeDN) derived from general and healthy images. If a medical image contains any disorders, the reconstructed image will differ significantly from the original. Data distribution estimation requires the estimation of the distribution of healthy difference images, which are calculated using healthy images and their reconstructed images generated by a

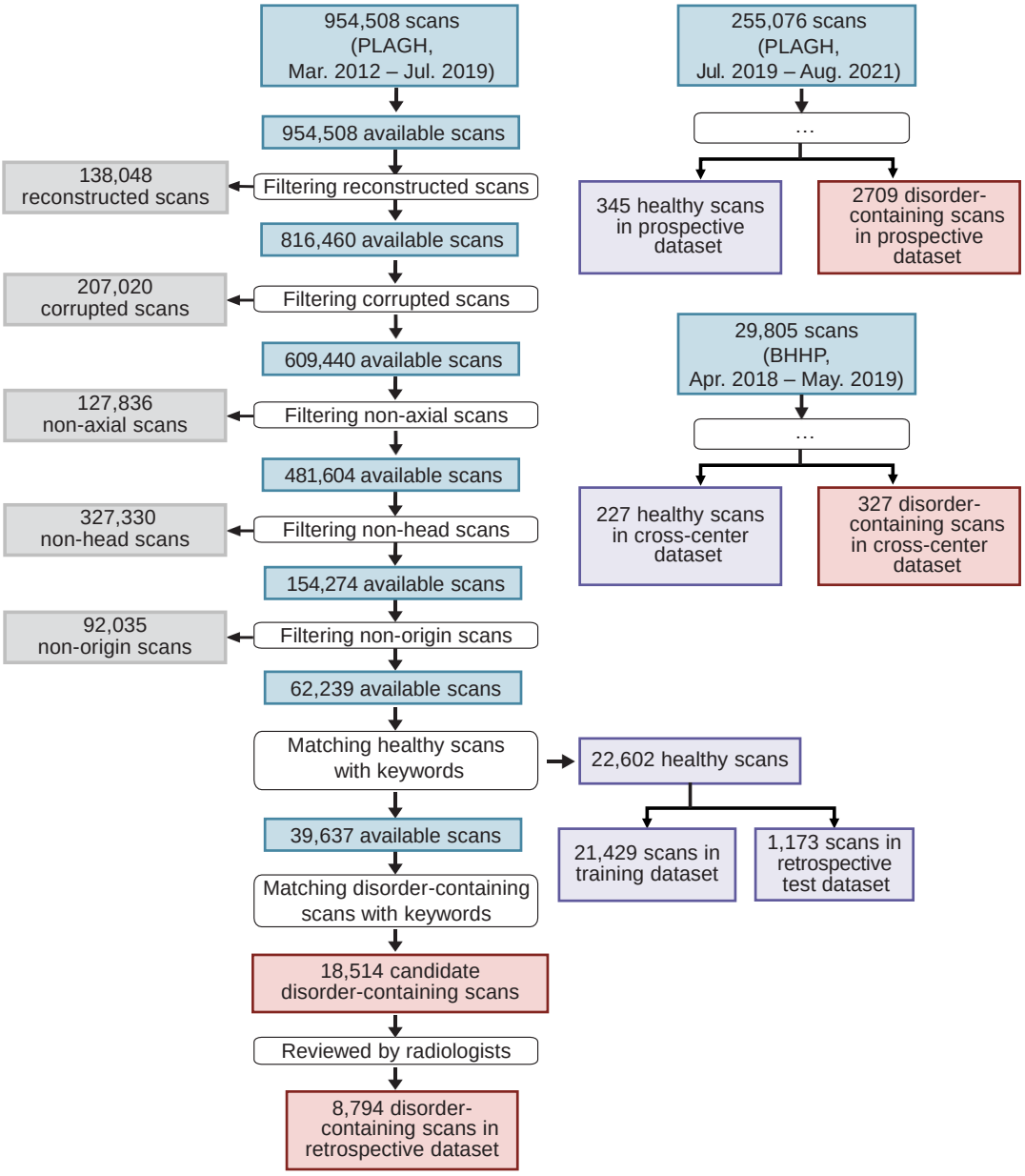

Fig. 5 | The process of constructing the datasets. *Left*, 954,508 scans (March 2012 - July 2019) were collected from the PLAGH by retrieving head CT-related keywords. After a series of filtering steps, a training dataset and a retrospective dataset were constructed. The training dataset consisted of 21,429 healthy scans, and the retrospective dataset consisted of 1,173 healthy scans and 8,794 disorder-containing scans. *Right*, we collected 255,076 scans (July 2019 - August 2021) from the PLAGH and 29,805 scans (April 2018 - May 2019) from the BHHP. We used these data to construct the prospective and cross-center test datasets using the same process.

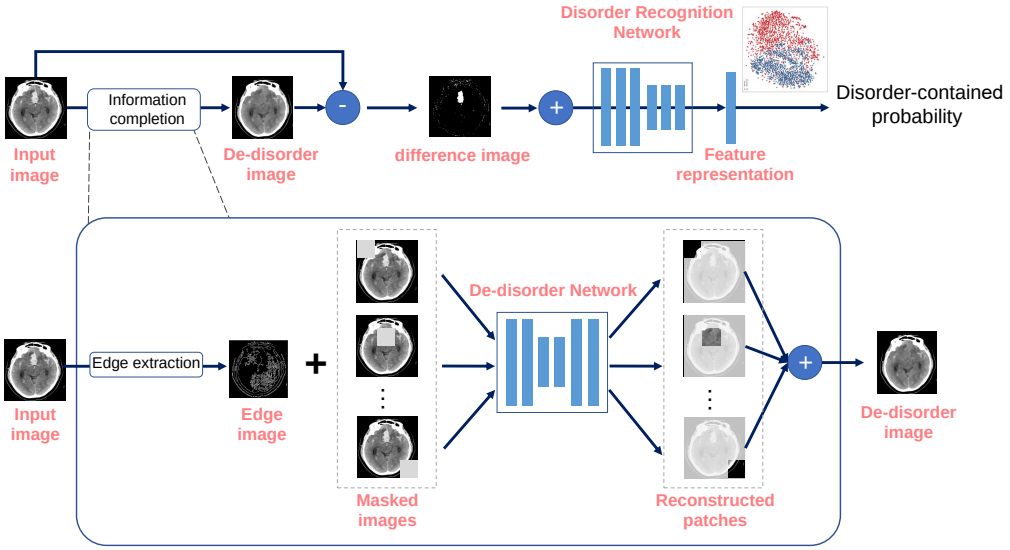

Fig. 6 | The overview of ISL, a learning algorithm for developing broad-spectrum disorder detection systems. The training dataset consists only of healthy scans, and a de-disorder network is learned to generate de-disorder images. A disorder recognition network is then employed to predict the probability of disorder containment based on the difference image obtained by subtracting the input and de-disorder image. This approach enables the developed model to achieve broad-spectrum disorder detection even without any disorder-contained data.

DeDN. If an image contains any disorders, its difference image will fall outside the distribution and be detected. Notably, unlike many existing disorder detection algorithms, ISL can detect a significantly increased number of disorder types.

CT slice conversion In our dataset, the pixel values in a CT scan are represented by 14-bit numbers, which exceed the range of human perception. To address this, we converted each CT slice into a 3-channel 8-bit image, conforming to the standard image format and suitable for display. Radiologists typically use specific window locations (WL) and window widths (WW) to observe various types of disorders. Building on this, we applied specific WL and WW settings for the image conversion, with the specific settings outlined in Supplementary Table 7.

Problem Formulation. ISL is designed to address binary-classification problems by predicting the probability of the presence of any disorder type in a medical image. For example, in brain CT scans, the input of an ISL-based system is a slice $\mathbf{x}_i$ from a brain CT scan for a CT scan $\{\mathbf{x}_i\}_{i=1}^{N}$, where N is the slice number of this scan. The output p_i from our system indicates the probability of any disorder type in the slice $\mathbf{x}_i$. During deployment, the slice-level outputs $\{p_i\}_{i=1}^{N}$ are aggregated to a scan-level p' output by averaging the probabilities of all the slices in the scan, where $p' = \frac{1}{N}\sum_{i=1}^{N} p_i$. As the ISL-based system processes individual slices, we have omitted the subscript number of slices in a scan for the sake of conciseness in the following method introduction.

An ISL-based system comprises two networks: a de-disorder network (DeDN) and a disorder recognition network (DRN). Given a medical image $\mathbf{x}$, we first use a DeDN to generate a de-disorder image $\mathbf{x}_{dd}$ of $\mathbf{x}$. If $\mathbf{x}$ contains a disorder, the disorder tissues in the area are converted into healthy ones. Then, the difference image $\mathbf{x}_{dif} = \|\mathbf{x} - \mathbf{x}_{dd}\|$ is input into the DRN network, which predicts the probability p of disorder containment in the image.

De-disorder network. Given a masked image $\tilde{\mathbf{x}} = \mathbf{x} \odot \mathbf{m}$, where $\mathbf{m}$ is an image mask of $\mathbf{x}$, a deep encoder-decoder network (DeDN) can predict the masked region and generate a reconstructed image $\hat{\mathbf{x}}$. In this study, we utilized the DeDN to generate de-disordered medical images. Specifically, we divided a medical image $\mathbf{x}$ into $K \times K$ grids of uniform size. For each grid with coordinates (i, j), where $1 \leq i \leq K$ and $1 \leq j \leq K$, we applied a mask $\mathbf{m}^{(i,j)}$ to erase it and obtain a masked image $\tilde{\mathbf{x}}^{(i,j)}$. The DeDN was then used to reconstruct the masked image and generate the reconstructed image $\hat{\mathbf{x}}^{(i,j)}$. Finally, we combined the $K \times K$ generated images into a reconstructed

de-disordered medical image using the following equation:

$$\mathbf{x}_{dd} = \sum_{i=1}^{K} \sum_{j=1}^{K} \hat{\mathbf{x}}^{(i,j)} \odot (1 - \mathbf{m}^{(i,j)}). \tag{1}$$

We will use $\mathbf{x}_{dd}$ for comparative analysis with the original image $\mathbf{x}$. A deep encoder-decoder network (DeDN) takes as input a masked image $\tilde{\mathbf{x}}^{(i,j)}$ and multiple image edge maps $\{\mathbf{e}_k\}_{k=1}^{n_e}$ of $\mathbf{x}$. Edge maps retain structural information of the masked region, which can improve the quality of the reconstructed image. Edge maps can be constructed using mature image processing schemes, such as the Canny Edge Detector.[30] The DeDN G generates the reconstructed image $\hat{\mathbf{x}}^{(i,j)}$ using the following equation:

$$\hat{\mathbf{x}}^{(i,j)} = G\left(\tilde{\mathbf{x}}^{(i,j)}, \{\mathbf{e}_k\}_{k=1}^{n_e}\right). \tag{2}$$

We trained the network using a joint loss:

$$\mathcal{L}_{de} = \lambda_{\ell_1}\mathcal{L}_{\ell 1} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_p\mathcal{L}_{perc} + \lambda_s\mathcal{L}_{style}, \tag{3}$$

which includes an $\ell 1$ loss, adversarial loss, perceptual loss, and style loss. The ℓ_1 loss $|\hat{\mathbf{x}}^{(i,j)} - \mathbf{x}|$ minimizes the reconstruction error between $\hat{\mathbf{x}}^{(i,j)}$ and $\mathbf{x}$. The adversarial loss $\mathcal{L}_{adv}$ ensures the reality of the generated image and is defined as:

$$\begin{aligned}\mathcal{L}_{adv} =& \mathbf{x}\mathbb{E}_{(\mathbf{x})}\left[\log D\left(\mathbf{x}\right)\right] \\ &+ \mathbb{E}_{(\mathbf{x},\mathbf{m}^{(i,j)})} \log\left[1 - D\left(G(\tilde{\mathbf{x}}^{(i,j)}, \{\mathbf{e}_k\}_{k=1}^{n_e})\right)\right],\end{aligned} \tag{4}$$

where D is the discriminator network. We also included perceptual loss $\mathcal{L}_{perc}$ and style loss $\mathcal{L}_{style}$, following Nazeri et al.[31] The perceptual loss $\mathcal{L}_{perc}$ penalizes reconstructed images that are not perceptually similar to the original ones and is defined as a distance measure between activation maps of a pretrained network:

$$\mathcal{L}_{perc} = \mathbb{E}_{(\mathbf{x})}\left[\sum_i \frac{1}{n_a} \left\|\phi_i\left(\mathbf{x}\right) - \phi_i\left(\hat{\mathbf{x}}\right)\right\|_1\right], \tag{5}$$

where ϕ_i is the activation map of the ith layer of the pretrained VGG-19 network, and n_a is the number of layers. We chose five layers[31] of the VGG-19 network pretrained on the ImageNet dataset.[32] The style loss is calculated based on these activation maps and is an effective tool to alleviate the ”checkerboard” artifacts caused by transpose convolution layers. The loss measures the differences between covariances of the activation maps and is defined as:

$$\mathcal{L}_{style} = \mathbb{E}_{(\mathbf{x})}\left[\sum_i \frac{1}{n_a} \left\|G_i^{\phi_i}(\mathbf{x}) - G_i^{\phi_i}(\hat{\mathbf{x}})\right\|_1\right], \tag{6}$$

where $G_i^{\phi_i}$ is a Gram matrix constructed from ϕ_i.

Disorder recognition network. To minimize the impact of reconstructed noise on disorder detection, we developed a disorder recognition network that takes the difference image $\mathbf{x}_{dif}$ as input and extracts an embedded representation in the latent space. We designated the embedding distribution of difference images from disorder-free data as the reference distribution. The disorder recognition network should ensure that the embeddings of disorder-free data are centralized and compact, while the embeddings of disorder-contained data are random and as far as possible from

the reference distribution. In this case, the distance between an embedded representation and the center of the reference distribution can effectively indicate the possibility of disorder presence.

Inspired by the support vector data description (SVDD) algorithm[33] and the contrastive learning method,[34] we developed the DRN based on augmentation views. In addition to a given medical image $\mathbf{x}$, the DRN uses two augmented views, $\mathbf{x}^-$ and $\mathbf{x}^+$, generated from $\mathbf{x}$ for network training. $\mathbf{x}^-$ and $\mathbf{x}$ are considered the same class (negative class), while $\mathbf{x}^+$ is of a different class (positive class). With augmentations, $\mathbf{x}^-$ maintains a similar appearance to the healthy image, while $\mathbf{x}^+$ disrupts the inherent distribution of the healthy image. In this study, we randomly applied rotation and flipping transformations to generate $\mathbf{x}^-$ and used cutout transformation to generate $\mathbf{x}^+$. It should be noted that for medical images, cutout can damage healthy tissue, and the view generated based on it should be of a different class. Therefore, the embeddings of $\mathbf{x}^-$ and $\mathbf{x}$ should be similar, while that of $\mathbf{x}^+$ should be distinct.

After training the DeDN, the input of the DRN consists of three parts: the reference difference image $\mathbf{x}_{dif}$, the negative difference image $\mathbf{x}^-_{dif} = \left\|\mathbf{x}^- - \hat{\mathbf{x}}^-_{dd}\right\|$, and the positive difference image $\mathbf{x}^+_{dif} = \left\|\mathbf{x}^+ - \hat{\mathbf{x}}^+_{dd}\right\|$, where $\hat{\mathbf{x}}^-$ and $\hat{\mathbf{x}}^+$ denote the reconstructed images of $\mathbf{x}^-$ and $\mathbf{x}^+$, respectively. The DRN extracts embedded representations of the difference images, denoted as $\mathbf{h}$, $\mathbf{h}^-$, and $\mathbf{h}^+$. The DRN learns reasonable embedded representations of disorder-free input by maximizing the similarity between $\mathbf{h}^-$ and $\mathbf{h}$ while distinguishing $\mathbf{h}^+$ from $\mathbf{h}$. In this study, we used the Euclidean distance as the metric to measure the similarity of the embeddings. We pre-trained the network using MoCo[12] and averaged the embeddings of the wide-ranging training dataset. The

averaged embedding $\mathbf{c}$ is considered the center of the reference distribution. Setting the center as an anchor, we designed a compactness loss to maximize the similarity between the negative embeddings:

$$\mathcal{L}_{com} = \mathbb{E}_{(\mathbf{h})}\left[\left\|\mathbf{h} - \mathbf{c}\right\|_2\right] + \mathbb{E}_{(\mathbf{h}^-)}\left[\left\|\mathbf{h}^- - \mathbf{c}\right\|_2\right], \tag{7}$$

where $\mathcal{L}_{com}$ minimizes the distances between the embeddings $\mathbf{h}, \mathbf{h}^-$ and the reference center, which ensures that the DRN can extract consistent features for disorder-free difference images. To further improve the discriminative ability of the network, we used a discrimination loss $\mathcal{L}_{dis}$, which forces the network to maximize the distance between the reference center and the embedded representation of $\mathbf{x}^+$:

$$\mathcal{L}_{dis} = -\mathbb{E}_{(\mathbf{h}^+)}\left[\left\|\mathbf{h}^+ - \mathbf{c}\right\|_2\right]. \tag{8}$$

The overall loss function utilized to train the DRN is defined as:

$$\mathcal{L} = \lambda_c \mathcal{L}_{com} + \lambda_d \mathcal{L}_{dis} \tag{9}$$

where λ_c and λ_d are the weights of the loss functions $\mathcal{L}_{com}$ and $\mathcal{L}_{dis}$, respectively.

Disorder visualization. The ISL-based system is capable of identifying the locations of disorders, which is particularly crucial for clinical applications.[35,36,1] Higher pixel values in the image regions of $\mathbf{x}_{dif}$ indicate a higher likelihood of the presence of a disorder. To enhance the visual appeal of the results, we conducted several post-processing steps on $\mathbf{x}_{dif}$: (1) Eliminating the bias caused by the normal range reconstruction. Pixels with values below a certain threshold t were set to zero. (2)

Reducing reconstruction noise. After normalizing the pixel values to the range of [0, 1], we added the values of the $s \times s$ region surrounding each pixel to itself. This smoothing technique reduced the noise in the image. (3) Enhancing the disorder area. We utilized an exponential function to manipulate the pixel values, resulting in an amplification of the differences in values among pixels. This process serves to accentuate the presence of disorder within the region of interest.

Model selection and statistical analysis. Since we were unable to access data containing disorders for model evaluation during training, we selected the model when the training loss did not decrease for 5 consecutive epochs. The primary parameters of our system, including the network architectures, hyperparameter values, and optimization strategies, are presented in Supplementary Tables 6 and 7. To ensure statistical significance, we applied 95% confidence interval (CI). Specifically, for each iteration, we randomly sampled 30% of CT scans from the test dataset for evaluation. We repeated this procedure 1,000 times and calculated the 95% CI of the evaluation metrics for the model.

Code and software availability. The system was developed using standard libraries and scripts available in PyTorch. The developing code is at *https://gitlab.com/heyuwei403/islcode*. The code will be made publicly available after the acceptance. A demo video of our diagnosis and visualization software is at *https://gitlab.com/heyuwei403/isl-system-demo*.

Data availability. The datasets to develop our head and pulmonary disorder detection system are not publicly available due to the privacy requirement of the PLAGH and FAHGMU. The cross-center dataset (554 scans and experts' annotations) from BHHP is allowed to be distributed for

research purposes from the corresponding author upon reasonable request. The development and evaluation dataset for retinal OCT disorder detection can be downloaded from *https://data.mendeley.com/datasets/rscbjbr9sj/2*.

References

1. Lundberg, S. M. *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* **2**, 749–760 (2018).

2. Liang, H. *et al.* Artificial intelligence for stepwise diagnosis and monitoring of covid-19. *European Radiology* **32**, 2235–2245 (2022).

3. Motwani, M. *et al.* Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *European heart journal* **38**, 500–507 (2017).

4. Hadamitzky, M. *et al.* Optimized prognostic score for coronary computed tomographic angiography: results from the confirm registry (coronary ct angiography evaluation for clinical outcomes: An international multicenter registry). *Journal of the American College of Cardiology* **62**, 468–476 (2013).

5. Chilamkurthy, S. *et al.* Deep learning algorithms for detection of critical findings in head CT scans: A retrospective study. *The Lancet* **392**, 2388–2396 (2018).

6. Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018).

7. Nikolov, S. *et al.* Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430* (2018).

8. Monteiro, M. *et al.* Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head ct using deep learning: an algorithm development and multicentre validation study. *The Lancet Digital Health* (2020).

9. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine* **25**, 1301–1309 (2019).

10. Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering* **5**, 555–570 (2021).

11. Guo, Y. *et al.* Deep learning with weak annotation from diagnosis reports for detection of multiple head disorders: a prospective, multicentre study. *The Lancet Digital Health* **4**, e584–e593 (2022).

12. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735 (2020).

13. He, K. *et al.* Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009 (2022).

14. Lee, H. *et al.* An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature Biomedical Engineering* **3**, 173–182 (2019).

15. Ruff, L. *et al.* Deep one-class classification. In *International conference on machine learning*, 4393–4402 (PMLR, 2018).

16. Schölkopf, B. *et al.* Support vector method for novelty detection. In *NIPS*, vol. 12, 582–588 (Citeseer, 1999).

17. Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U. & Langs, G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, 146–157 (Springer, 2017).

18. Akcay, S., Atapour-Abarghouei, A. & Breckon, T. P. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, 622–637 (Springer, 2018).

19. Abati, D., Porrello, A., Calderara, S. & Cucchiara, R. Latent space autoregression for novelty detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 481–490 (2019).

20. Tian, Y. *et al.* Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4975–4986 (2021).

21. Liznerski, P. *et al.* Explainable deep one-class classification. *arXiv preprint arXiv:2007.01760* (2020).

22. Fan, Y. *et al.* Video anomaly detection and localization via gaussian mixture fully convo-

lutional variational autoencoder. *Computer Vision and Image Understanding* **195**, 102920 (2020).

23. Yao, Q., Xiao, L., Liu, P. & Zhou, S. K. Label-free segmentation of covid-19 lesions in lung ct. *IEEE transactions on medical imaging* **40**, 2808–2819 (2021).

24. Baur, C., Graf, R., Wiestler, B., Albarqouni, S. & Navab, N. Steganomaly: Inhibiting cyclegan steganography for unsupervised anomaly detection in brain mri. In *International conference on medical image computing and computer-assisted intervention*, 718–727 (Springer, 2020).

25. Stepec, D. & Skocaj, D. Unsupervised detection of cancerous regions in histology imagery using image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3785–3792 (2021).

26. Baur, C., Wiestler, B., Albarqouni, S. & Navab, N. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, 161–169 (Springer, 2019).

27. Akcay, S., Atapour-Abarghouei, A. & Breckon, T. P. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, 622–637 (Springer, 2019).

28. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional

adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134 (2017).

29. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232 (2017).

30. Ding, L. & Goshtasby, A. On the canny edge detector. *Pattern Recognition* **34**, 721–725 (2001).

31. Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z. & Ebrahimi, M. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212* (2019).

32. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. In *Proc. of the 22nd IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (IEEE, 2009).

33. Tax, D. M. & Duin, R. P. Support vector data description. *Machine learning* **54**, 45–66 (2004).

34. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722* (2019).

35. Kux, L. *Clinical and Patient Decision Support Software; Draft Guidance for Industry and Food and Drug Administration Staff* (US FDA, 2017).

36. Muelly, M. C. & Peng, L. Spotting brain bleeding after sparse training. *Nature Biomedical Engineering* **3**, 161 (2019).

Acknowledgements

We thank the radiologists from the PLAGH for their efforts in labeling the test data.

Author's contributions

F.X., X.L., Q.D., Y.H. and Y.G. designed the research; X.L., J.L., Y.H., W.Z. and H.L. collected the data; Y.H., Y.G., L.M., J.L., F.X. and X.L. verified the raw underlying data, which had been accessed by all authors; Y.H., Y.G., L.M. and H.T. developed the system; Y.H., Y.G., L.M., G.D., J.L., H.L. and J.H. analyzed the results; Y.H., L.M. and J.L. co-wrote the manuscript; F.X., S.L., H.Q. and Y.G. critically revised the manuscript; and all the authors discussed the results and provided feedback regarding the manuscript.

Competing interests

The authors declare no competing interests.