



HAL
open science

Diagnosis after Zooming in: A Multi-label Classification Model by Imitating Doctor Reading Habits to Diagnose Brain Diseases

Ruiqian Wang, Guanghui Fu, Jianqiang Li, Yan Pei

► **To cite this version:**

Ruiqian Wang, Guanghui Fu, Jianqiang Li, Yan Pei. Diagnosis after Zooming in: A Multi-label Classification Model by Imitating Doctor Reading Habits to Diagnose Brain Diseases. Medical Physics, 2022, 10.1002/mp.15871 . hal-03760525

HAL Id: hal-03760525

<https://inria.hal.science/hal-03760525v1>

Submitted on 27 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Diagnosis after Zooming in: A Multi-label Classification Model by Imitating Doctor Reading Habits to Diagnose Brain Diseases

Ruiqian Wang¹, Guanghui Fu^{2,3,4,5,6,7}, Jianqiang Li¹, Yan Pei⁸

1. Beijing University of Technology, Beijing 100124, China

2. Sorbonne Université, Paris, France

3. Institut du Cerveau - Paris Brain Institute - ICM, Paris, France

4. Inserm, Paris, France

5. CNRS, Paris, France

6. AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France

7. Inria, Aramis project-team, Paris, France

8. University of Aizu, Aizuwakamatsu 965-8580, Japan

Corresponding Author: Jianqiang Li, email: lijianqiang@bjut.edu.cn

Abstract

Purpose: Computed tomography (CT) has the advantages of being low cost and noninvasive and is a primary diagnostic method for brain diseases. However, it is a challenge for junior radiologists to diagnose CT images accurately and comprehensively. It is necessary to build a system that can help doctors diagnose and provide an explanation of the predictions. Despite the success of deep learning algorithms in the field of medical image analysis, the task of brain disease classification still faces challenges: researchers lack attention to complex manual labeling requirements and the incompleteness of prediction explanations. More importantly, most studies only measure the performance of the algorithm, but do not measure the effectiveness of the algorithm in the actual diagnosis of doctors.

Methods: In this paper, we propose a model called DrCT2 that can detect brain diseases without using image-level labels and provide a more comprehensive explanation at both the slice and sequence levels. This model achieves reliable performance by imitating human expert reading habits: targeted scaling of primary images from the full slice scans and observation of suspicious lesions for diagnosis. We evaluated our model on two open-access datasets: CQ500 and the RSNA Intracranial Hemorrhage Detection Challenge. In addition, we defined three tasks to comprehensively evaluate model interpretability by measuring whether the algorithm can select key images with lesions. To verify the algorithm from the perspective of practical application, three junior radiologists were invited to participate in the experiments, comparing the effects before and after human-computer cooperation in different aspects.

Results: The method achieved F1-scores of 0.9370 on CQ500 and 0.8700 on the RSNA dataset. The results show that our model has good interpretability under the premise

39 of good performance. Human radiologist evaluation experiments have proven that our
40 model can effectively improve the accuracy of the diagnosis and improve efficiency.

41 **Conclusions:** We proposed a model that can simultaneously detect multiple brain
42 diseases. The report generated by the model can assist doctors in avoiding missed
43 diagnoses, and it has good clinical application value.
44

45 **Keywords:** Medical image classification, Interpretability, Attention mechanism,
46 Human-AI Interaction.

47 Contents

48	I. Introduction	1
49	II. Related Works	4
50	II.A. Sequence-level Image Classification	4
51	II.B. Interpretive Medical Image Analysis Model	5
52	II.C. Human-computer Interaction (HCI) in Medical Image Analysis Domain	6
53	III. Task Analysis	8
54	IV. Sequence-level Interpretive Evaluation Tasks	9
55	V. Methods	12
56	V.A. Primary Network	12
57	V.A.1. Feature Attention	14
58	V.A.2. Dependencies Learning	14
59	V.A.3. Slice Attention	14
60	V.B. Attention Proposal Slices Network	15
61	V.C. Knowledge Fusion	16
62	V.D. Loss Functions	16
63	VI. Evaluation	16
64	VI.A. Algorithm Performance Evaluation	17
65	VI.A.1. Dataset and Training Details	17
66	VI.A.2. Supervised with Different Loss Functions	18
67	VI.A.3. Effect of APS Module	19
68	VI.A.4. Performance on Interpretive Evaluation Tasks	22
69	VI.A.5. Comparison with Image-level Algorithm	23
70	VI.B. Human Radiologists Evaluation	24
71	VII. Discussion	26
72	VIII. Conclusion	30

73 **IX. Acknowledgments**

31

74 **References**

31

1. Introduction

Brain disease is one of the ailments that threaten health and damages the life of humans¹. Due to the increase in patients, professional doctors are insufficient². Training an experienced doctor usually takes more than seven years. Therefore, it is necessary to build a computer-aided medical diagnosis system, which can help doctors diagnose effectively and accurately. In recent years, deep learning algorithms have been applied in brain disease classification tasks, e.g.^{3,4,5,6}. Although these algorithms have achieved good performance, this domain still faces challenges. Most studies label sequence CT images at the image level, which is time-consuming. More important, separately labeling images or diagnosing by algorithm violates common medical knowledge. Diagnosing sequenced medical images requires the observation of adjacent images. Doctors usually browse adjacent slices and capture the changes between the slices to make a judgment on the disease. There are subtle differences between adjacent brain CT images, but they play a decisive role in disease judgment. The key images and points almost determine all the judgments of the disease. How to focus on them to obtain more information about the disease is essential.

By considering all the challenges we analyzed before, we proposed a model called DrCT2. The method is inspired by the diagnostic habits of radiologists. As shown in Figure 1, in clinical diagnosis, radiologists browse a complete set of brain CT scans and focus on the key images that may reflect diseases, observing closer to get more information about the lesion. Our model imitates the doctor's reading habit by selecting key images and zooming in to obtain detailed information at different scales. It consists of three parts: primary network, attention proposal slices network (APS), and knowledge fusion network. The primary network includes feature attention, dependencies learning, and slice attention module. These two attention modules focus on key features and images that may reflect the lesion, respectively. Dependencies learning module can learn the relationship between sequence slices. The APS zooms in on key images proposed by the slice attention module and obtains more details from different scales. The knowledge fusion network merges the knowledge learned from two neural networks to make a final judgment. Our model inputs sequence CT, and only requires the label at the sequence level. It is not necessary to label each CT image as in most studies, which reduces the workload of data labeling and is more reasonable.

106 The medical computer vision (CV) model is different from the normal CV model. The
107 normal CV model focuses on the improvement of algorithm performance, while the medical
108 CV model needs to focus more on whether the model can truly assist doctors in diagnosis
109 under the premise of ensuring performance. In this paper, we put forward a new perspective:
110 for medical models, it is not only to maximize the prediction accuracy of the algorithm itself
111 but to measure the benefits that practitioners obtained after interaction with the algorithm.
112 We defined 3 tasks to evaluate whether the model can select key images. The evaluation
113 of key slices selection proves the reliability and interpretability of our model. Selecting
114 slices accurately is also very important for prompting radiologists to avoid misdiagnosis.
115 We further invited junior radiologists to simulate real diagnosis scenarios for evaluation.
116 The results show that our model can effectively improve diagnostic efficiency and accuracy.
117 Compared with previous work⁷, we improve the performance of the algorithm and propose
118 interpretive evaluation tasks. In the evaluation experiments, the effect of the model to
119 assist doctors in decision-making is verified. We not only evaluated the test set like most
120 studies but also focused on doctors to evaluate the potential of our model to assist clinical
121 decision-making.

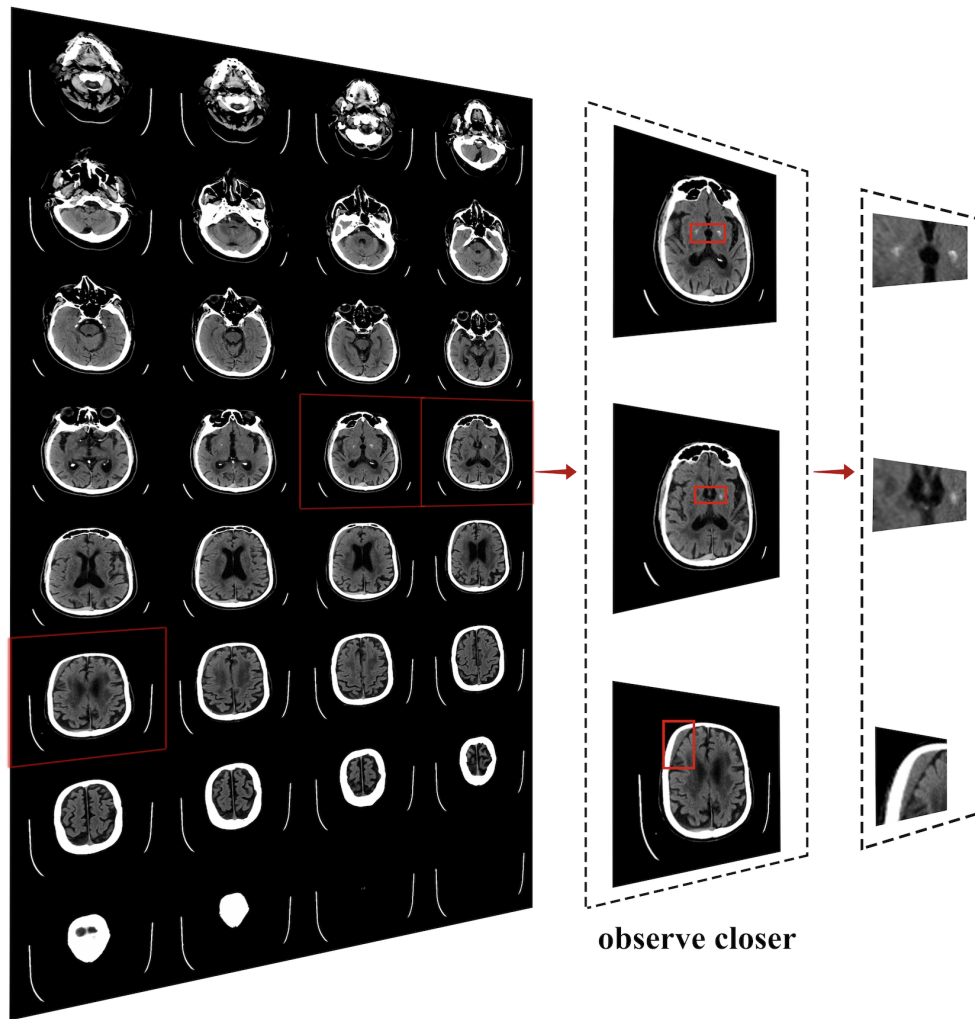


Figure 1: Some diseases are straightforward to misdiagnose, which requires observing and considering adjacent images. In this CT scan, we need to observe adjacent images to distinguish between a bleeding point or calcification. It is easy to ignore subdural hemorrhage without zooming in on the image.

122 II. Related Works

123 In this paper, we propose a model inspired by radiologists' diagnostic habits by zooming in
124 on key slices, analyzing at the sequence level, and giving multi-label classification results.
125 Our model can provide the interpretability of the model, and we evaluate this through a
126 sequence-level interpretative evaluation task. In Section II.A., we first analyze some sequence-
127 level image classification algorithms from two types of research directions, 2D and 3D. Our
128 research pays special attention to interpretability, so we also analyze some interpretable
129 medical image analysis algorithms in Section II.B..

130 II.A. Sequence-level Image Classification

131 In reference^{8,9,10,11}, the temporal attention mechanism which focuses on important frames
132 has been widely used. The algorithm proposed by Yang et al.¹² can adaptively capture the
133 regions of interest in each frame and learn the key features based on these areas. Yu et al.¹³
134 proposed a method for generating sentences to describe a video and Tu et al.¹⁴ proposed
135 the spatial-temporal attention (STAT) method for the video description task. These two
136 studies^{13,14} takes into account both the spatial and temporal information in a video.

137 Although these studies have been successful, they cannot solve the problem of sequence-
138 level brain disease classification. Many studies^{15,16,17} have tried to use 3D convolutional
139 neural networks to solve the problem of sequence-level brain disease classification. Nie et
140 al.¹⁸ proposed a novel 3D convolutional neural network architecture for learning supervised
141 features. Gao et al.¹⁹ integrated 2D and 3D CNN networks to classify brain diseases. Nawali
142 et al.¹⁶ and Ker et al.²⁰ proved the performance of 3D CNN in the classification task of a
143 cerebral hemorrhage.

144 These studies prove the effectiveness of 3D convolutional neural networks in medical
145 image processing tasks. However, the 3D architecture requires a large amount of calculation,
146 low-resolution images are selected as input to preserve the complete brain structure during
147 training. If the resolution of the image is reduced, the need for judging subtle bleeding
148 points and lesions cannot be met. Therefore, if the model can fuse the input multi-resolution
149 images, such as using high-resolution images for key images, it can provide more sufficient
150 information from different aspects which may achieve better performance. Multi-resolution

151 image fusion analysis has many advantages²¹. For sequence-level medical images, key images
152 and areas play a decisive role in disease judgment. Fu et al.⁷ proved its reliability and
153 reasonableness. Jiang et al.²² proposed a model called MFI-Net that can avoid the loss of
154 coarse-grained feature information in the shallow layer by extracting local and global feature
155 information at different resolutions. Liu et al.²³ proposed a multi-resolution medical image
156 fusion network with iterative back-projection (IBPNet). Experimental results show that it
157 has better performance in visual perception and objective evaluation. Li et al.²⁴ proposed a
158 model for lung nodule detection that employed patch-based multi-resolution CNNs to extract
159 the features and employed four different fusion methods for classification. In this work, we
160 focus on key slices and areas at the same time and try to obtain information about lesions
161 on slices from multiple resolutions.

162 Dependencies between slices should be considered, and some research has recently
163 learned slice dependencies through 3D networks. Zhuang et al.²⁵ proposed a new self-
164 supervised learning model that contains a Rubik’s cube recovery task. It can pre-train
165 3D neural networks from raw 3D medical data. Compared with the training strategy from
166 scratch, it can achieve better performance on various tasks. Zhu et al.²⁶ further devel-
167 oped this method, enriched the pre-training tasks, and achieved better performance in the
168 downstream tasks. Zhu et al.²⁷ proposed a novel SSL approach for 3D medical image clas-
169 sification. It embeds task knowledge into training 3D neural networks. The experimental
170 results demonstrate the effectiveness of embedding lesion-related prior knowledge into neu-
171 ral networks for 3D medical image classification. The above research proves that modeling
172 and learning dependencies between slices are effective. In our model, we learn about the
173 dependencies between slices through the cooperation of the primary and auxiliary (APS)
174 networks.

175 II.B. Interpretive Medical Image Analysis Model

176 In the medical image analysis domain, the interpretability of the model is significant. Only
177 giving the prediction result is not credible; the basis for the model’s decision also needs to
178 be explained. Pranav et al.²⁸ proposed a convolutional neural network, CheXNet, that can
179 output the probability of disease and use the class activation mapping method to indicate
180 the lesion areas of the lung disease. Zoom-in-Net²⁹ can generate four bounding boxes based

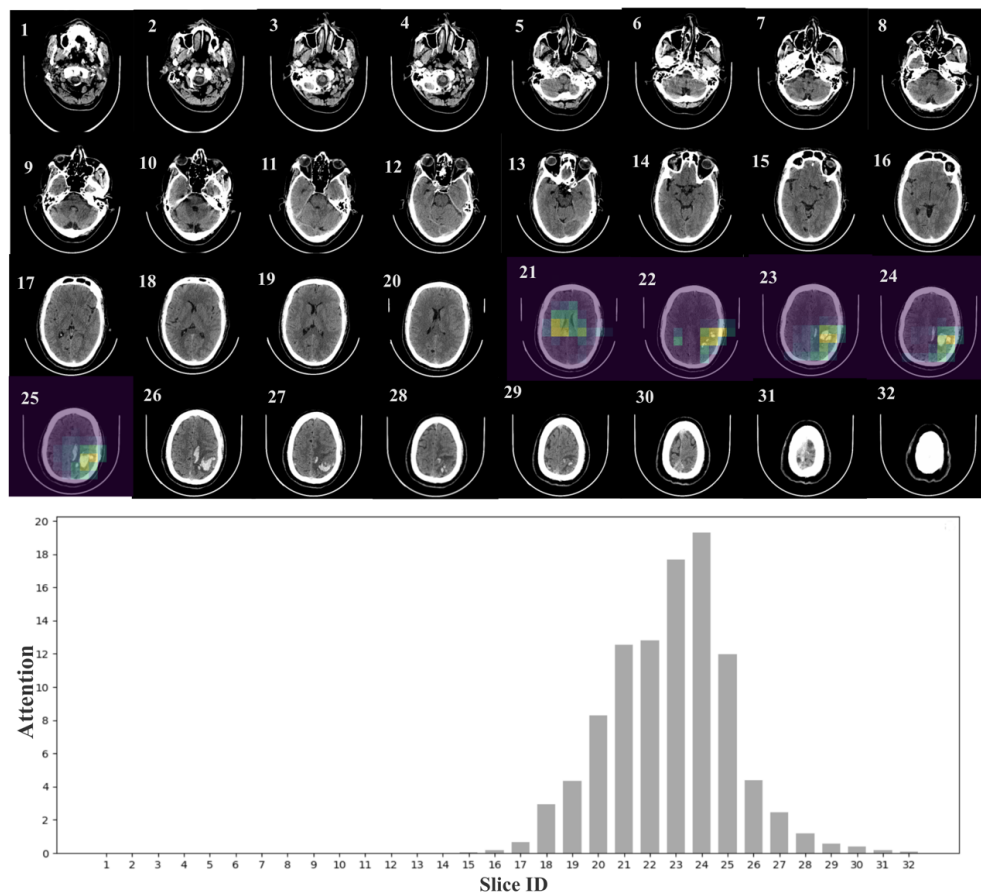
181 on the attention maps of the highlighted suspicious area, and four bounding boxes can
182 cover 80% of the lesions. Zhang et al.³⁰ proposed a solution for pathological diagnosis and
183 interpretation. The model proposed by this team can generate an interpretive report for the
184 pathologist’s reference. The report not only displays the selected regions of interest but also
185 generates explanatory text for different regions. Wang et al.³¹ proposed an algorithm for the
186 classification of intracranial hemorrhage diseases, which can increase the interpretability of
187 the model by outputting a prediction basis and image-level attention maps.

188 These studies all indicated suspicious areas on a single image. However, in our task, the
189 doctor is concerned not only about the suspicious part of the image but also about which
190 slice in the sequence CT can reflect the lesion. It is significant to automatically select key
191 images, which can improve the efficiency of doctors’ reading and provide an explanation for
192 the prediction. In addition, the key images selected by the model can be used as a reference
193 for the doctor’s diagnosis, prompting the doctor to avoid missed diagnosis or misdiagnosis.
194 As shown in Figure 2, our model can generate reports to assist doctors in diagnosis.

195 II.C. Human-computer Interaction (HCI) in Medical Image Anal- 196 ysis Domain

197 The focus of the medical image analysis model is to assist doctors in their work, so it should
198 be human-centered³². The combination of AI-based frameworks with complementary human
199 intervention could result in synergistic effects inpatient management, interpretation, and di-
200 agnosis³³. Sayres et al.³⁴ invited 10 ophthalmologists to diagnose the diabetic retinopathy
201 disease in three conditions (unassisted, grades only, or grades plus heatmap) based on retinal
202 fundus images to evaluate their model. The results found that algorithm-assisted diagnosis
203 improved the diagnostic accuracy and confidence of ophthalmologists, and the diagnosis time
204 was reduced after providing model explanations (the third condition). Zhao et al.³⁵ devel-
205 oped a deep learning model to help doctors diagnose musculoskeletal tumors. The expert
206 evaluation experiments showed that their models improved the sensitivities of six of seven
207 doctors and accuracy in three of seven doctors, while they did not significantly reduce the
208 specificities of any. Choi et al.³⁶ compared the diagnostic performance of radiologists on two
209 datasets (breast ultrasound images alone or with computer-aided diagnosis) and showed that
210 the computer-aided diagnosis method could improve radiologists’ diagnostic performance by

211 increasing specificity, accuracy, and positive predictive value. Ding et al.³⁷ invited 20 gas-
 212 troenterologists to conduct experiments, and the results showed that an algorithm-assisted
 213 method can identify abnormalities more sensitively and rapidly than conventional diagno-
 214 sis by gastroenterologists. Esmaeili et al.³⁸ trained some explainable deep learning models
 215 based on the Grad-CAM mechanism and evaluated whether the trained model could localize
 216 tumor regions. The results show that deep learning models may classify some tumor brains
 217 based on other non-relevant features. We believe that to evaluate model performance should
 218 not only consider under experimental data but more importantly, compare the effects of
 219 assisting doctors.



This patient may have the following diseases: ICH, IPH, Mass Effect

Figure 2: An example of the diagnostic report generated by the DrCT2 model. The model can identify the 5 slices most likely to reflect the lesion and highlight the suspicious area automatically. The histogram is the model's attention to each image. It can be seen from the figure that our algorithm accurately selects the key images and key areas.

220 III. Task Analysis

221 **Difficulty in labeling data** Labeling medical data is a costly and time-consuming task.
222 To avoid the waste of expert resources, the algorithm proposed in this paper is based on
223 the sequence level: can be trained without labeling on a single image. Labeling data at
224 the sequence level is easier to implement, and expert diagnosis based on full slice CT obeys
225 medical commonsense.

226 **Adjacent slice dependence** In clinical practice, a patient’s condition cannot be inferred
227 from a single slice, and doctors usually browse adjacent slices and catch the changes between
228 the slices to diagnose. As shown in Figure 3, observing changes in continuous CT images
229 is conducive to the diagnosis of the disease and reduces the occurrence of misdiagnosis and
230 missed diagnosis. The sequence-level model we proposed is more in line with the doctor’s
231 diagnostic habits.

232 **Dependency between diseases** There are dependency relationships between brain dis-
233 eases. For example, skull fractures are often accompanied by epidural hemorrhage. Hy-
234 pertensive cerebral parenchymal hemorrhage may break into the ventricle and cause intra-
235 ventricular hemorrhage and form a mass effect. For our model, the dependencies learning
236 module learns the sequence images from positive and negative directions and considers the
237 relationship between different diseases at the same time.

238 **Easy to miss diagnosis** Some diseases are easily missed in the judgment of brain diseases
239 for the doctor, such as subtle bleeding points and subarachnoid hemorrhage³⁹. As shown
240 in Figure 1, if the doctor does not observe the brain CT carefully, it is difficult to find the
241 bleeding point, because some small bleeding points are not significantly different from cal-
242 cification. An algorithm with interpretability can provide references for doctors’ diagnoses.
243 For our task, the algorithm should be able to pick out the key images and highlight the
244 suspicious areas in the images.

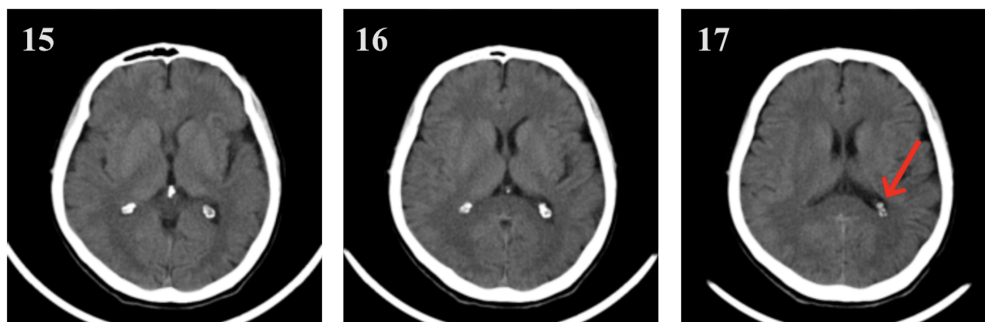


Figure 3: The three slices in this figure are adjacent. There is a small white dot where the arrow points. It is difficult to judge whether there is intraventricular hemorrhage only by observing the slice with ID 17. However, by observing the changes between the three slices, we can see that the white spot is calcification, not intraventricular hemorrhage.

245 IV. Sequence-level Interpretive Evaluation Tasks

246 Whether the key images can be selected correctly is crucial to assist in diagnosis. It is also an
247 important indicator for evaluating model performance. To evaluate the performance under
248 different disease conditions, we define the following tasks from different levels.

249 **Easy condition task** We define in a set of CT (28 images/set), having **five or more**
250 images that can reflect the disease, which is the simple condition for our algorithm. We
251 evaluated how many of the top five images with high attention selected by the algorithm
252 reflected the lesion.

253 **Difficult condition task** We defined in a set of CT images (28 images/set) as having **five**
254 **or fewer** images that can reflect the disease as the difficult condition for our algorithm. We
255 evaluated whether the top five images with high attention selected by the algorithm could
256 reflect lesions. This task can evaluate whether the algorithm can achieve good performance in
257 the case of an easily missed diagnosis. If the algorithm can achieve good performance under
258 this task, it proves that it has the potential to assist doctors in discovering difficult-to-detect
259 diseases.

260 **Comprehensive condition task** There are h images in a set of CT that can reflect the
261 disease, to evaluate whether the top h images selected by our algorithm with high attention

262 can cover. This task can comprehensively evaluate the explanatory performance at the
263 sequence level.

264 The pseudo-code for our three tasks can be seen in algorithm 1. U is a dataset containing
265 n_{scan} scans, and the composition of the dataset is different under different disease conditions.
266 S_i represents the i -th set of scans. The function $Model$ is the trained model, which inputs
267 a set of slices and outputs the attention value of each slice. The function $Sort$ sorts the
268 attention value in descending order. The function $Select$ select the slices (T_i) corresponding
269 to the top m_{select} values of the sort result. s_j represents the slice in T_i . For easy task and
270 difficult task, m_{select} is equal to 5; for the comprehensive task, m_{select} is the number of slices
271 that can reflect the lesion in the set of slices. D_i is the number of images selected correctly.
272 $Score_i$ is the evaluation result of the model on CT scan S_i . Sum is a variable used to record
273 scan evaluation results. We use the average accuracy for evaluation.

274 Please note that the easy and difficult tasks defined here are only defined from the
275 number of diseased slices, not all from the prediction mechanism of the algorithm or from
276 the perspective of radiologists.

Algorithm 1: Evaluation algorithm for key image picking

Input: $U = \{S_i\}, i \in [1, n_{scan}]$ **Output:** *Accuracy*: evaluation results

```

1 Initialize:  $Accuracy \leftarrow 0; Sum \leftarrow 0;$ 
2 for each  $S_i \in U$  do
3    $D_i \leftarrow 0$ 
4    $Score_i \leftarrow 0$ 
5    $A_i \leftarrow Model(S_i)$ 
6    $V_i \leftarrow Sort(A_i)$ 
7    $T_i \leftarrow Select(V_i, S_i)$ 
8   if easy condition task then
9     for each  $s_j \in T_i$  do
10      if slice  $s_j$  is the key slice then
11         $D_i = D_i + 1$ 
12      end
13    end
14     $Score_i = D_i/5$ 
15     $Sum = Sum + Score_i$ 
16  end
17  if difficult condition task then
18    if  $T_i$  contains key slices then
19       $Sum = Sum + 1$ 
20    end
21  end
22  if comprehensive condition task then
23    for each  $s_j \in T_i$  do
24      if slice  $s_j$  is the key slice then
25         $D_i = D_i + 1$ 
26      end
27    end
28     $Score_i = D_i/m_{select}$   $Sum = Sum + Score_i$ 
29  end
30 end
31  $Accuracy = (Sum/n_{scan}) * 100\%$ 

```

277 V. Methods

278 As shown in Figure 4, our proposed model DrCT2 contains three modules: primary network,
 279 attention proposal slices network (APS), and knowledge fusion network. The primary net-
 280 work learns the dependencies between slices and diseases. Two-step attention mechanisms
 281 keep the algorithm focused on key images and suspicious parts. The APS network zooms
 282 in key images proposed by the primary network and learns more details on different scales.
 283 The knowledge fusion network merges all the information learned from two modules and
 284 gives a final prediction. The cooperation of the three modules makes the model have good
 285 performance and explanatory.

286 V.A. Primary Network

287 The primary network is an encoder-decoder network, which we call DrCT1⁷. It inputs a
 288 set of brain CT images that after sampling (we sampled fix-length images from a set of CT
 289 images with different numbers of slices) to the encoder and outputs a feature matrix (each
 290 image is represented as a feature vector, and we combine them into a feature matrix). A set
 291 of brain CT images after sampling contains m slices.

$$292 \quad S = (s_1, s_2, \dots, s_m) \quad (1)$$

293 We used the VGG16⁴⁰ model pre-trained on ImageNet⁴¹ to extract features from each
 294 slice. We froze the convolutional layers and discarded the fully-connected layers during
 295 feature extraction. The adaptive max-pooling method is used to compress the feature maps
 296 (512 channels, 7×7 size) into 512-dimensional vectors (512 channel, 1×1 size feature
 297 map). Each slice is represented by a 512-dimensional feature vector:

$$298 \quad s_t = (z_1^{<t>}, z_2^{<t>}, \dots, z_{512}^{<t>}), t \in [1, m] \quad (2)$$

299 The decoder consists of three parts: the feature attention part, the dependencies learning
 300 part, and the slice attention part. The feature vectors passed through these three modules
 301 to learn the probability of each disease.

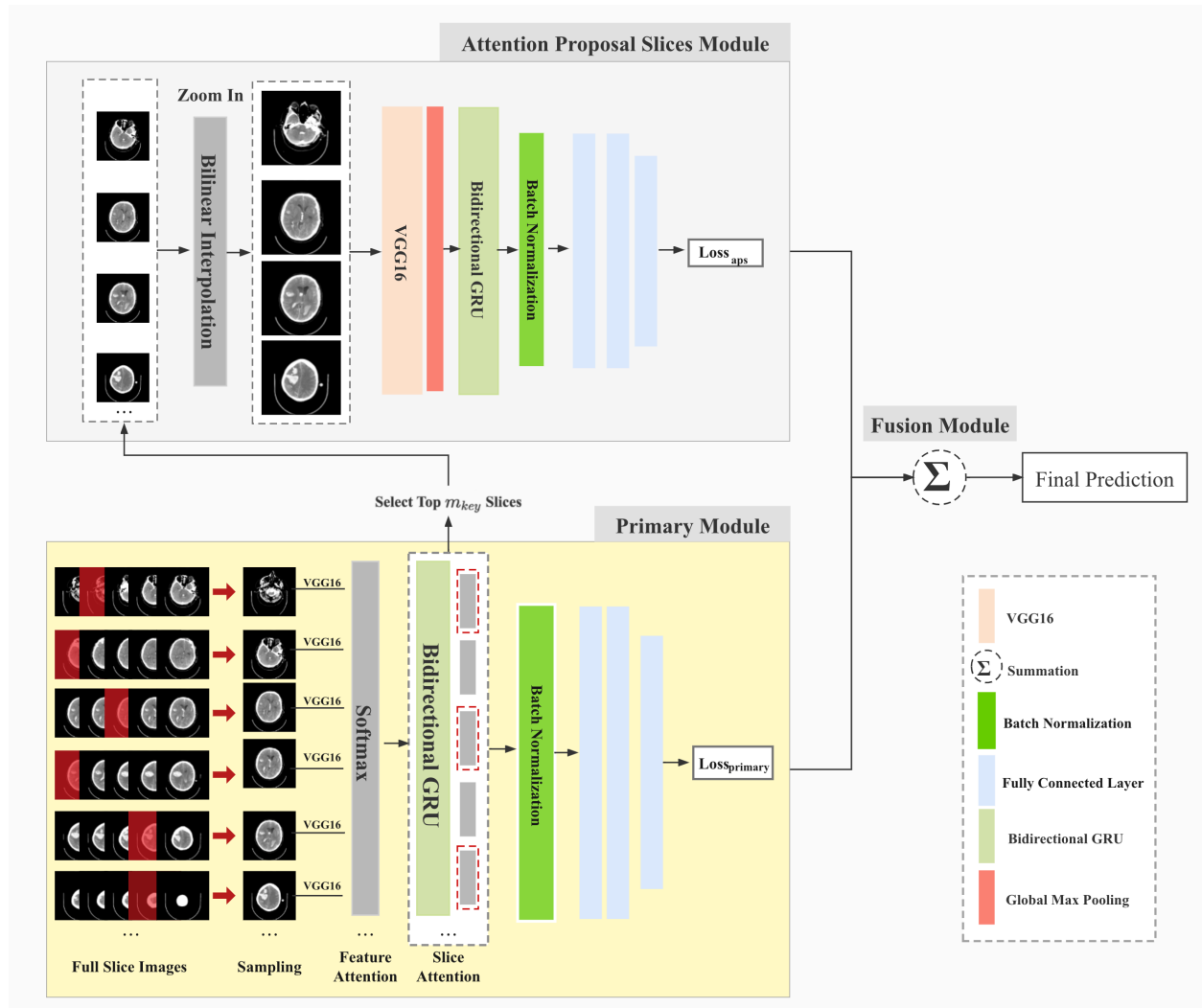


Figure 4: An overview of our proposed model consists of three parts: i) The primary network is encoder-decoder-based and assigns more weight to areas or slices that may contain lesions through the attention mechanism. ii) The APS network selects the m_{key} slices most likely reflecting the lesions and zooms in them as the input of the APS network to learn more detailed information. iii) The fusion network considers the knowledge learned by the two networks to make judgments about the disease.

302 V.A.1. Feature Attention

303 The feature vector learned by CNN is the high-dimensional feature representation of each im-
 304 age. Feature attention can assign different weights to each node. Nodes with high attention
 305 can be visualized to explain model prediction. It can be defined as:

$$306 \quad w_i^{<t>} = \frac{\exp(z_i^{<t>})}{\sum_{j=1}^{512} \exp(z_j^{<t>})} \quad (3)$$

307 where $z_i^{<t>}$ is the value of the i -th dimension vector in a 512-dimensional vector, and $w_i^{<t>}$
 308 is the weight assigned to the i -th dimension vector. The feature vectors $(z_1^{<t>}, \dots, z_{512}^{<t>})$ of
 309 the original slice (s_t) merge with the assigned weights $(w_1^{<t>}, \dots, w_{512}^{<t>})$ to obtain a new slice
 310 feature representation x_t .

311 V.A.2. Dependencies Learning

312 This module can learn the dependencies between slices through a bidirectional recurrent
 313 neural network (BRNN)⁴². It can capture information from both positive and negative
 314 directions. Gated recurrent unit (GRU⁴³) as the base unit of BRNN. GRU solves the long-
 315 term dependency problem in RNN networks. The new feature representation $(x_1, \dots, x_t, \dots, x_m)$
 316 is put into the BGRU for dependencies learning. The input of base units at each time step is
 317 the new feature representation weighted by the feature attention module. We use element-
 318 wise multiplication to merge these two vectors in this step. The output formula of each time
 319 step of BGRU is:

$$320 \quad \hat{y}^{<t>} = g(W_y[\vec{a}^{<t>}, \overleftarrow{a}^{<t>}] + b_y) \quad (4)$$

321 where $\hat{y}^{<t>}$ is the output at the t -th time step, $\vec{a}^{<t>}$ and $\overleftarrow{a}^{<t>}$ represent the information
 322 in the positive and the negative directions respectively, W_y and b_y are weight and bias
 323 respectively, and function g represents the activation function.

324 This module learns the dependencies between sequence-weighted features learned by
 325 feature attention.

326 V.A.3. Slice Attention

327 This module can assign weights to each vector as shown below:

$$a_t = Mapping(\hat{y}^{<t>}) \quad (5)$$

$$w_t = \frac{exp(a_t)}{\sum_{j=1}^m exp(a_j)} \quad (6)$$

where $\hat{y}^{<t>}$ is the output vector of at the t -th time step in the dependencies learning part. The *Mapping* function represents a fully connected layer, which maps the output vector to a value of length 1. m represents the number of slices after sampling. w_t represents the importance of the t -th slice in disease judgment. $(\hat{y}^{<1>}, \dots, \hat{y}^{<m>})$ and (w_1, \dots, w_m) are fused (using the element-wise multiplication method) to obtain the result after slice attention.

Highly attended feature vectors are proposed for the APS network. Weighted feature vectors pass through three fully connected layers and output the probability of each disease, and use the dropout⁴⁴ layer to prevent over-fitting, the dropout rate is 0.5. The activation function of the first two fully-connected layers is ReLU, and for the output layer, we used sigmoid function.

V.B. Attention Proposal Slices Network

The APS network can learn more detailed information from key images. The model can improve performance with the help of this network. The APS network zooms in on m_{key} key images proposed by the slice attention module and learns more details from different scales.

$$key\ images = (k_1, k_2, \dots, k_{m_{key}}) \quad (7)$$

We used bilinear interpolation to zoom-in on m_{key} key images from 224×224 to 512×512 . In large-resolution images, key details are easier to obtain. The formula is as follows:

$$K_i = BilinearInterpolation(k_i) \quad (8)$$

$$F_i = P(K_i) \quad (9)$$

$$F_l = merge(F_1, \dots, F_{m_{key}}) \quad (10)$$

where the function *BilinearInterpolation* represents the bilinear interpolation algorithm; k_i is the i -th slice in the m_{key} key slices, and K_i is the i -th slice after zoom-in operation. The function P is the VGG16 model pre-trained on ImageNet. F_i is the extracted feature vector, and F_l is the result of combining m_{key} feature vectors. This feature matrix passed through

357 a fully connected layer after using batch normalization⁴⁵. Then, the dependencies learning
 358 module and three fully connected layers were passed through to obtain the probabilities of
 359 diseases.

360 V.C. Knowledge Fusion

361 In clinical diagnosis, doctors make the final diagnosis by considering overall and details. The
 362 knowledge fusion module merges the knowledge from the primary network and the APS
 363 network. We summed the two feature matrices and after three fully connected layers, we
 364 obtained the final prediction.

365 V.D. Loss Functions

366 We optimized three losses during the overall training:

$$367 \quad \text{Loss} = \text{Loss}_{\text{primary}} + \text{Loss}_{\text{aps}} + \text{Loss}_{\text{fusion}} \quad (11)$$

368 $\text{Loss}_{\text{primary}}$ and Loss_{aps} is the mean square error (MSE):

$$369 \quad \text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2 \quad (12)$$

370 $\text{Loss}_{\text{fusion}}$ is the binary cross entropy error (BCE):

$$371 \quad \text{BCE} = -\frac{1}{n} \sum_{i=1}^n (y_i * \log(y'_i) + (1 - y_i) * \log(1 - y'_i)) \quad (13)$$

372 y' is the prediction of our model and y is the true label. n is the number of samples.
 373 DrCT2 contains three modules: the primary network and the APS network use the MSE
 374 loss function for supervision and the BSE loss for the fusion module.

375 VI. Evaluation

376 We evaluated the performance of our model from two perspectives. In the first part (Section
 377 VI.A.), we evaluate from the perspective of algorithm performance, conduct ablation experi-
 378 ments, and compare experiments. However, the medical task-oriented model is different from

379 the open domain computer vision model. The evaluation of computer vision models is more
380 inclined to improve the performance of the algorithm, and the medical task is more focused
381 on measuring whether it could assist doctors with diagnosis in the actual applications. More
382 emphasis is placed on the cooperation between humans and machines. In the second part
383 (Section VI.B.), from the perspective of human-computer interaction, we invited three junior
384 radiologists to measure the impact of utilizing our model on the human diagnosis.

385 VI.A. Algorithm Performance Evaluation

386 VI.A.1. Dataset and Training Details

387 Our experiment was evaluated using two datasets: CQ500⁴⁶ and the dataset provided by
388 the RSNA Intracranial Hemorrhage Challenge (represented as RSNA)⁴⁷. Table 1 shows
389 the data distribution of these two datasets. We use a fixed-length sampling method to
390 sample scans with different numbers of slices to the same number (32 for CQ500, 28 for
391 RSNA). The dataset after sampling is called the sampled dataset. For example, in the
392 0.625 mm scan data, 256 brain CTs will be generated, which will increase the computational
393 complexity of the algorithm. The unsampled data are used to generate a new CT set for data
394 enhancement. We use k -fold cross-validation (k equal to 10 for the CQ500 dataset, and 5
395 for the RSNA dataset). We keep the same data distribution as the first place solution in the
396 RSNA challenge for comparison reasons. We use PyTorch⁴⁸ as a framework for implementing
397 our deep learning models. The experiments are performed on a workstation equipped with
398 a 3.90 GHz CPU and 2070 GPU with 8 GB memory. The batch size is 32 for the CQ500
399 dataset and 256 for the RSNA dataset. We trained for 100 epochs and used the Adam⁴⁹
400 optimizer for optimization. In DrCT2, we used different learning rates for the three modules:
401 0.0003 for APS and 0.0005 for the primary network and fusion network. In the following
402 experiments, we used the Macro-F1 score as our evaluation metric. Macro-F1 can treat each
403 category equally, and it will be more sensitive about rare categories⁵⁰. In this task, we hope
404 that the model can consider the impact of fewer diseases on the performance of the model;
405 while Micro-F1 is more easily affected by common categories. Therefore, we choose Macro
406 F1 as our evaluation metrics.

Table 1: Data distribution of CQ500 and RSNA.

Dataset	CQ500	RSNA
Patients	490	18923
Scans	1181	21540
Diseases	9	6
Number of scans	32-396	21-60
After Sampling	32	28
Disease label statistics (at sequence level)		
Any(ICH)	701	7855
Intraparenchymal(IPH)	356	4688
Intraventricular(IVH)	72	3236
Subdural(SDH)	102	3388
Epidural(EDH)	20	309
Subarachnoid(SAH)	128	3468
Calvarial fracture	181	
Mass Effect	248	
Midline shift	173	

407 VI.A.2. Supervised with Different Loss Functions

408 We chose the MSE loss as the loss function of the primary network and the auxiliary network
 409 (APS). However, in principle, both modules can use the BCE as the loss function. We
 410 construct experiments on two datasets (CQ500 and RSNA) to compare the performance of
 411 models supervised by different loss functions. The result can be seen in Table 2. It can be
 412 seen from the experimental results that better performance can be obtained by using the
 413 MSE as the loss function. Therefore in the following experiments, we use the MSE as the
 414 loss function for these two modules.

Table 2: Comparative experiment of using different loss functions for $Loss_{Primary}$ and $Loss_{Aps}$.

Model	Dataset	Loss Function	Precision	Recall	F1 Score
DrCT2	CQ500	BCE	94.23%	93.94%	0.9360
		MSE	94.29%	94.10%	0.9370
	RSNA	BCE	87.93%	88.10%	0.8682
		MSE	88.05%	88.23%	0.8700

415 VI.A.3. Effect of APS Module

416 The APS module zooms in on m_{key} slices and puts them into the network. The selection of
 417 m_{key} slices depends on the slice attention mechanism, and they are the key images selected
 418 by the algorithm. We construct the following experiments to verify the role of this module
 419 from 3 aspects.

- 420 • With or without APS Network;
- 421 • Under a different number of slices, select through attention mechanism or random;
- 422 • Input under different resolutions.

423 **With or without APS Network Under Different Datasets** To evaluate the effect of
 424 the APS network, we evaluated the performance of the DrCT1 (without APS module) and
 425 DrCT2 (with APS module) model on two datasets are shown in Table 3.

Table 3: Experiment between the DrCT1 and DrCT2 models on the two datasets.

Dataset	Model	Precision	Recall	F1 Score
CQ500	DrCT1	93.10%	93.14%	0.9262
	DrCT2	94.29%	94.10%	0.9370
RSNA	DrCT1	87.51%	87.83%	0.8650
	DrCT2	88.05%	88.23%	0.8700

426 The precision-recall curve of DrCT2 in the training process is drawn in Figure 5.

427 From the results, we can see that with the help of the APS network, the DrCT2 model
 428 has better performance (+1.08% F1 for CQ500; +0.5% F1 for RSNA) than DrCT1. This
 429 proves the importance of the targeted fusion of multi-scale information.

430 **Key Slices Selection** We design experiments to compare the performance of random
 431 selection or selection through attention mechanisms. The architecture of our network is as
 432 described above, and the resolution of the image input to the APS network is 512×512 .
 433 The two methods select m_{key} slices from m brain CT images (m is the number of slices after
 434 sampling). We selected 4 to 28 slices in the following experiments, and the results are shown
 435 in Figure 6. We can see that it is effective to input after selection through the attention

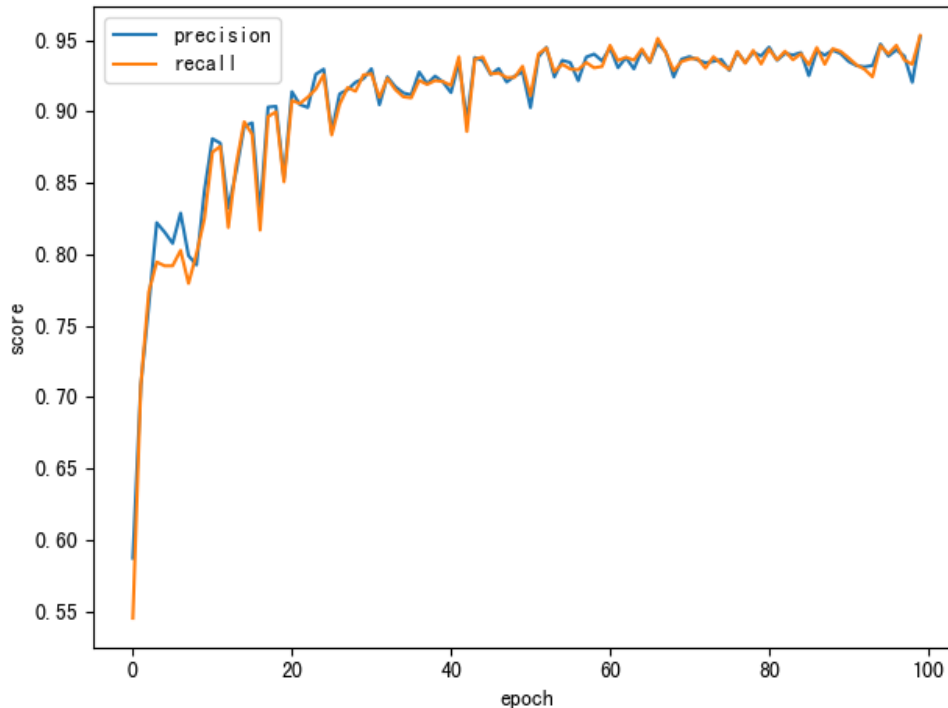


Figure 5: Recall and precision curve when training the DrCT2 model on dataset CQ500.

436 mechanism under a different number of slices. In the DrCT2 model, the best results were
 437 achieved when 16 images were selected. Therefore, in the following experiments, we set
 438 $m_{key} = 16$ as the default setting.

439 To further explore the potential rules of model selection of slices, we statistically an-
 440 alyzed the 5 slices in which the model gave the highest attention weight. On the CQ500
 441 dataset, there are 744 sets of CT scans that reflect the disease. The statistical graph is
 442 shown in Figure 7. We found that the key slices are often concentrated from the 16th to the
 443 22nd slices.

444 **Under Different Resolutions** To evaluate the effect of the zoom-in operation, we design
 445 the following 5 experiments. For the DrCT2 model, we input images of different resolutions
 446 (224×224 vs. 512×512) into the APS network to evaluate its performance. For the DrCT1
 447 model, we construct experiments to evaluate the performance under original resolution (224
 448 $\times 224$) or high resolution (512×512). After statistical analysis, as shown in Figure 7, we
 449 found that the key slices were often concentrated from the 16th to the 22nd slices. So we

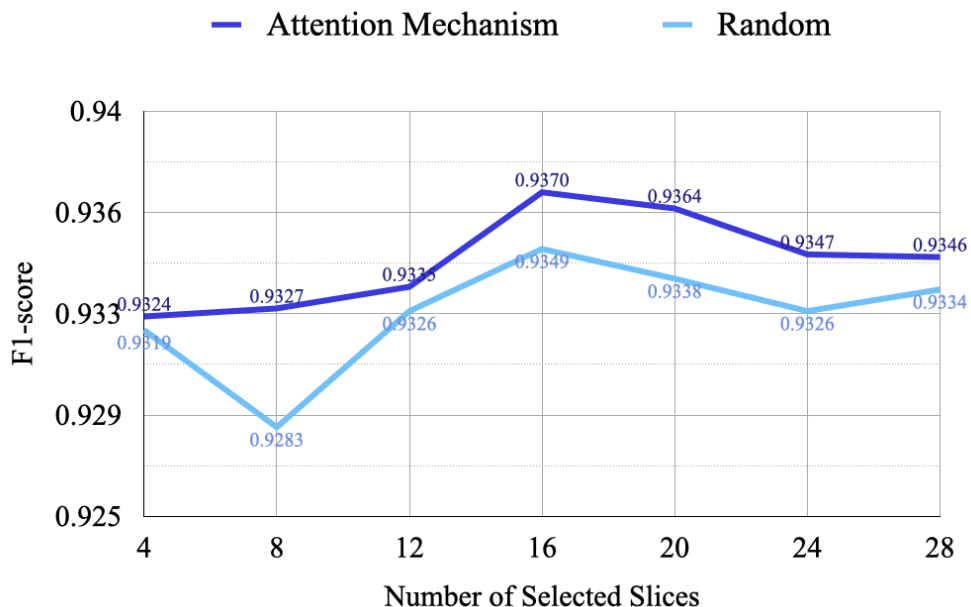


Figure 6: Evaluation results when m_{key} slices are input to the APS network by random selection and selection by our attention mechanism.

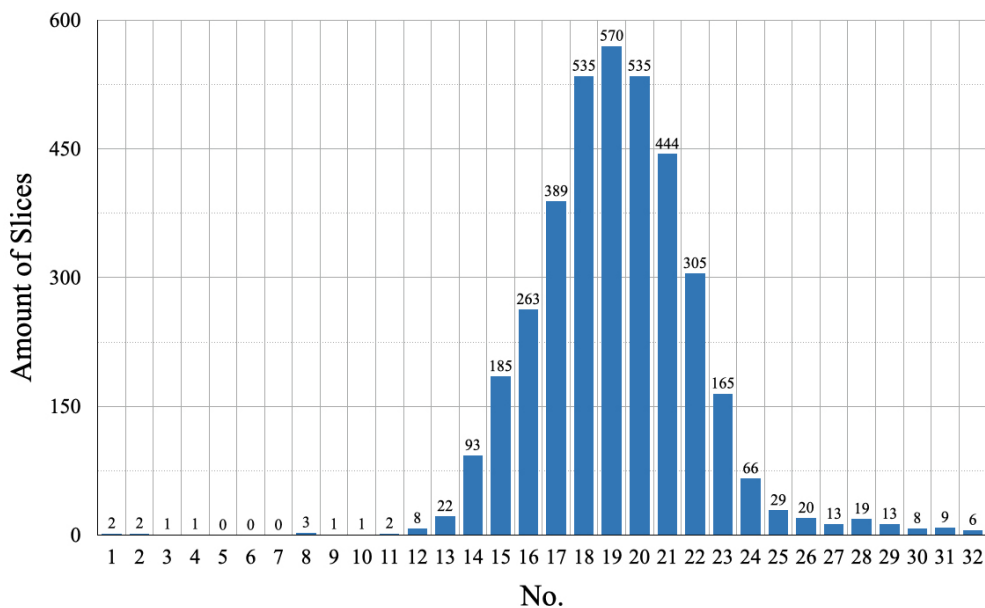


Figure 7: Statistics of the top five slice positions with high attention selected by our model in 744 sets of diseased brain CT (CQ500’s total diseased CTs).

450 simulate the mechanism of DrCT2 and replace these slices (from 16th to 22nd) with high-
 451 resolution images, while the rest still maintain low-resolution images. We call this strategy

452 the model fusion of DrCT1. Note that all the experiments are evaluated based on the CQ500
 453 dataset. The experiment result can be seen in Table 4.

Table 4: Comparative experiment on different strategies for DrCT1 and DrCT2.

Model	Dataset	Strategy	Input image resolution	Precision	Recall	F1 Score
DrCT1	CQ500	Original	Primary: 224×224	93.10%	93.14%	0.9262
		High-resolution Input	Primary: 512×512	91.70%	91.39%	0.9105
		Model Fusion	Primary 1: 224×224 Primary 2: 512×512	92.57%	92.27%	0.9179
DrCT2		With APS	Primary: 224×224; APS: 224×224	93.96%	93.62%	0.9327
		With APS	Primary: 224×224; APS: 512×512	94.29%	94.10%	0.9370

454 VI.A.4. Performance on Interpretive Evaluation Tasks

455 Model performance is certainly important in computer-assisted diagnosis algorithms, but our
 456 research not only focuses on performance but also model interpretability. This model can
 457 jointly select the key slices that the model focuses on through the attention mechanism and
 458 the auxiliary network (APS network). We elaborated in detail in Section IV.. The experiment
 459 is based on the dataset RSNA because it has annotations for each slice. We determined the
 460 number of slices for the three tasks we defined, and evaluated the performance of DrCT1
 461 and DrCT2 on these three tasks, as shown in Table 5.

Table 5: The results of the two models in the interpretive evaluation task.

Task	Scans Number	DrCT1	DrCT2
Easy Task	6,529	81.59%	83.92%
Difficult Task	1,759	58.76%	62.67%
Comprehensive Task	7,855	65.72%	67.15%

462 To evaluate the performance of the key images selection algorithm on different diseases,
 463 we calculated the accuracy for each disease: for all slices that reflect a certain disease, how
 464 many slices our algorithm can correctly pick out. The results are shown in Table 6. The
 465 results show that our model can be selected more accurately than DrCT1.

Table 6: The accuracy of the key slice selection for each disease.

Disease	Amount	DrCT1 Accuracy	DrCT2 Accuracy
EDH	2376	77.90%	79.46%
IPH	28270	76.70%	79.11%
IVH	20473	80.71%	81.96%
SAH	27784	80.74%	80.96%
SDH	36672	84.82%	84.84%
Any	83923	77.41%	78.23%

466 VI.A.5. Comparison with Image-level Algorithm

467 The first place in the RSNA Intracranial Hemorrhage Challenge⁵¹ is an image-level algorithm
 468 (represented as RSNA-1). The RSNA-1 utilizes total labels for each slice in a scan, and we
 469 only use one label in a scan. In the case of the RSNA dataset, for a set of sequences with
 470 28 slices, RSNA-1 has 28 slice-level labels, while we only have 1 sequence-level label. From
 471 this perspective, it is unfair to compare our model trained with weak labels with a model
 472 trained with full labels such as RSNA-1. However, to compare the differences between the
 473 two models, we did the following experiment.

474 We maintained the same distribution as them based on the results they submitted⁵¹.
 475 We convert the image-level prediction of the RSNA-1 algorithm to the sequence-level (if a
 476 slice reflects a certain disease, the whole CT set will also reflect it), and compare it with the
 477 results of DrCT1 and DrCT2. The precision, recall, and F1 score were used to evaluate the
 478 performance of the three models, and the results as shown in Table 7.

Table 7: Comparative experiment between sequence-level (DrCT1 and DrCT2) and image-level models (RSNA-1) on RSNA dataset.

Dataset	Model	Precision	Recall	F1 Score
RSNA	RSNA-1	93.59%	93.03%	0.9265
	DrCT1	87.51%	87.83%	0.8650
	DrCT2	88.05%	88.23%	0.8700

479 The results show that our proposed algorithm has great potential. Compared to image-
 480 level algorithms, we can obtain relatively good performance (-5.65% F1 score) using only
 481 weakly supervised labels (without using the image-level label). To further analyze the per-
 482 formance in predicting each disease, we calculated the accuracy for the three models. The
 483 prediction results are shown in Figure 8. We can see that our model performs better than the

484 DrCT1 model. For EDH, the performance of our model is even better than that of RSNA-1
 485 (+0.04% accuracy).

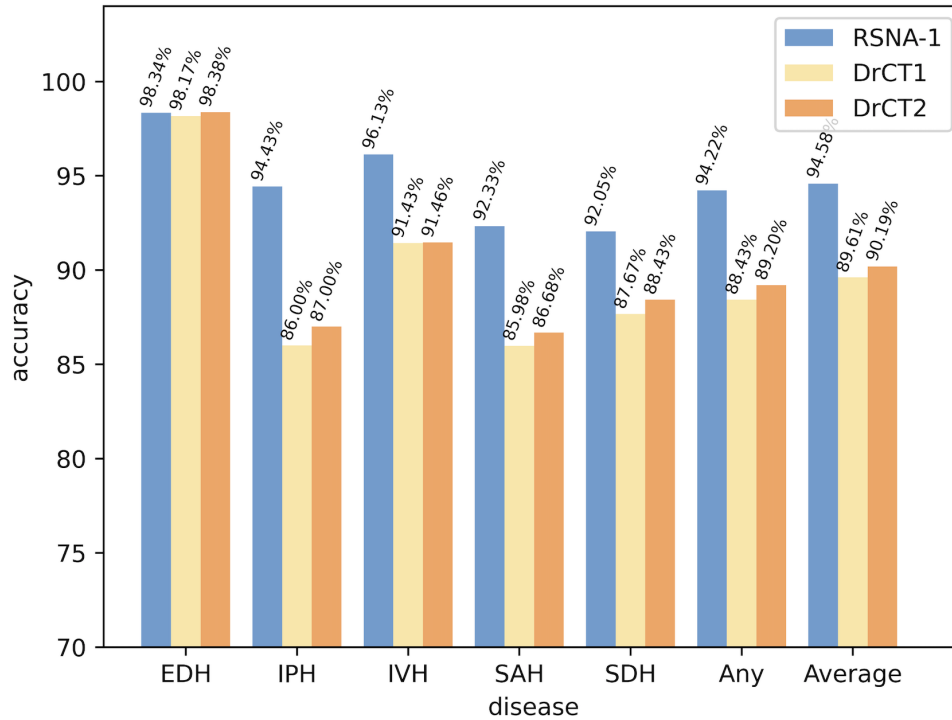


Figure 8: The accuracy of the three models on each disease.

486 VI.B. Human Radiologists Evaluation

487 Although these evaluation metrics in the above experiments can reflect the performance of
 488 the algorithm, how much help the artificial intelligence algorithm can bring to the radiologists
 489 in the real scene needs to be evaluated. For this reason, we designed experiments to simulate
 490 clinical scenarios to verify whether junior radiologists can be helped with our algorithms.
 491 We invited three junior radiologists (junior radiologists A, B, and C) with more than 2 years
 492 of experience to perform these experiments. Note that, we did not notify junior radiologists
 493 of the performance of the model in advance to prevent humans from having some prior
 494 judgments on the results of our model. We measured it from three aspects: diagnostic
 495 accuracy, diagnostic time, and diagnostic confidence. We use precision, recall, and F1-score
 496 to evaluate diagnostic accuracy. The diagnostic confidence is divided into four levels from 0
 497 to 3: 0 represents a completely uncertain diagnosis; 1 represents a slightly certain diagnosis;

2 represents a high probability diagnosis, and 3 represents full confidence in the diagnosis. The experimental data are 148 sets of brain CT scans in the RSNA dataset. And they are randomly selected from the test set to ensure that the model has never seen the data during the training process. The labels given by the RSNA dataset are used as the gold standard. The resolution of the CT slice is 224×224 . The details of the experiments can be seen below:

- Experiment 1 (radiologist only): The radiologists diagnose the selected brain CT scan, and record the radiologist’s diagnosis time and diagnosis conclusion for each set of data. The three radiologists need to score their diagnosis with four diagnostic confidence levels (the scoring time was not included in the diagnosis time) and record the degree of confidence.
- Experiment 2 (radiologist + AI): After an interval of one week from the start of experiment 1, maintain the radiologists are in the same state and randomly shuffle the data sequence to ensure that they forget the impression of the data in experiment 1. The radiologists combine the report generated by the DrCT2 algorithm to diagnose the selected brain CT scan. The experimental report contains three parts, as shown in Figure2, including all 28 CT slices, their attention weight histograms; the top 5 images with high attention, and their highlighted areas generated by slice attention and feature attention respectively; and the model prediction results.

Table 8: Human Radiologists evaluation task for Experiment 1 (radiologist only) and Experiment 2 (radiologist + AI model).

	Experiment 1			Experiment 2			Experiment
	JR A	JR B	JR C	JR A	JR B	JR C	DrCT2
Date (2021)	11.03	11.04	11.06	11.11	11.10	11.13	×
Precision	0.7411	0.6154	0.7127	0.7558	0.7522	0.8081	0.8551
Recall	0.7364	0.5715	0.7045	0.7474	0.7457	0.8048	0.8454
F1-score	0.7295	0.5780	0.6992	0.7423	0.7348	0.7977	0.8337
Confidence	2.48	2.05	2.10	2.41	2.65	2.48	×
Total time	3h9m27s	2h52m2s	2h39m44s	2h52m28s	2h25m43s	2h22m4s	2m1s
Time/scan(s)	76.80s	68.88s	64.76s	69.92s	62.29s	57.59s	0.82s

The results of our experiment can be seen in Table 8. From the experimental results, we can find that our model can be more accurate than junior radiologists. With the help of

519 our model, radiologists have effectively improved the diagnostic accuracy (+0.16 F1-score),
520 efficiency (-10s average diagnosis time), and more confidence. For radiologist B, we found
521 that even with the help of our model, it still cannot surpass the diagnostic accuracy of our
522 model. The reason for this problem is insufficient trust in our model, although the diagnostic
523 confidence has been improved. We analyzed this issue in detail in the Discussion section.

524 VII. Discussion

525 **From the perspective of algorithm performance:** Our model combines normal res-
526 olution and high-resolution input by the primary network and the APS network. It can
527 achieve better performance evaluated by the above experiments. Under the same primary
528 network, DrCT2 effectively improves the performance compared with DrCT1. To compare
529 the performance of DrCT2 and DrCT1 on the CQ500 and RSNA datasets, we reported the
530 precision, recall, and F1 score, and determined statistical significance by using a t-test with
531 a threshold of 0.05 ($P < 0.05$). The P-values for precision, recall, and F1 scores are 0.01940,
532 0.04971, and 0.02634, respectively. This improvement is because of the APS network. The
533 cooperation of the APS network and the primary network can help the model to improve
534 performance and also help guide the attention mechanism to better select the key slices. As
535 shown in Tables 5 and 6, DrCT2 achieved better performance than DrCT1 in the interpre-
536 tive evaluation task. Selecting slices more accurately not only helps improve performance
537 but also enables the model to generate more accurate reports to better assist doctors. It is
538 worth noting that improving interpretability is more important than improving performance
539 because interpretability is a prerequisite for the clinical application of medical models.

540 We did not use some mechanism to force the model to focus on adjacent slices, but
541 the model automatically focuses on sequence slices from 15 to 21 under the supervision
542 of the APS network. This proves that our auxiliary (APS) network is more inclined to
543 make judgments in conjunction with adjacent slices rather than making decisions based on
544 scattered slices. This is consistent with the doctor’s diagnosis habit, which is to diagnose
545 by adjacent slices. We explained its necessity in Section III. These slices are indeed more
546 likely to reflect cerebral hemorrhage from a medical point of view. Both the performance
547 improvement and the medical point of view have proven that the main and auxiliary network
548 (APS network) mechanism we proposed is effective.

549 From Table 4, we found that it is meaningless to learn only on high-resolution slices.
550 For example, DrCT1 achieved the lowest performance under 512×512 resolution (-1.57%
551 F1 score than original image resolution). Even under the model fusion strategy, it is less
552 effective than the original 224×224 resolution (-0.83% F1 score). Our strategy of learning
553 the key slices through the auxiliary network (APS network) again can achieve better results,
554 regardless of whether the original resolution or high-resolution images are input in the aux-
555 iliary network (APS network). If we input high-resolution slices in the APS network, the
556 best performance will be achieved. This is also in line with the characteristics of doctors'
557 diagnosis, that is, to pay more attention to the key slices. In addition, key slices were zoomed
558 in on to find detailed clues for diagnosis.

559 In Table 7, we can see that although we still have a gap with the image-level supervision
560 algorithm. However, we have achieved better performance (+0.04% accuracy) in the EDH
561 category with the least data, even if we only use relatively weak labels. This reflects the
562 potential of our model on small datasets.

563 There are some limitations to the deep learning method applied in the medical domain.
564 The main factors could include: enough high-quality annotated data, temporality (the dis-
565 eases are always progressing and changing over time in a non-deterministic way), domain
566 complexity (e.g., different imaging protocols, different types of data) and interpretability⁵².
567 In our research, we attempt to address data limitations with pre-training on ImageNet, but
568 natural images are different from medical images. We provide explanations at the sequence
569 and slice level, but in clinical applications, doctors expect the AI model to provide more
570 explanation for its predictions. How to learn more information from unlabeled medical data
571 through self-supervised learning, and provide more explanatory is our future direction.

572 **From the perspective of human-computer interaction:** In medical applications,
573 more attention should be given to how the algorithm assists the doctor in improving the
574 diagnostic accuracy (rather than maximizing the prediction accuracy of the algorithm it-
575 self). In addition, one may also focus on how an algorithm could minimize clinician time or
576 maximize confidence in the diagnosis, for example.

577 From this point of view, we evaluated the algorithm using two experiments. The first
578 one focused on explainability (whether the model pays attention to suspicious slices) and the

579 second on the prediction gain from the human-computer interaction. Through the experi-
580 mental results as shown in Table 5 and Table 6 we have proven that our model can better
581 select the key slices than DrCT1. We measure the performance changes after AI model assis-
582 tance from three perspectives: diagnostic accuracy, diagnostic time, diagnostic confidence.
583 From Table 8, we can summary that:

- 584 • Our model has better performance than three junior radiologists;
- 585 • Radiologists have improved the diagnostic efficiency and accuracy with the help of AI;
- 586 • Diagnosed with the aid of AI will be more confident usually, but there are also situations
587 that confidence has declined, such as junior radiologist A. Decline in confidence causes
588 low accuracy of improvement.

589 Both DrCT1⁷ and this study considers the important role of doctors in medical AI. The
590 differences are as follows:

- 591 • Experiment participants: in DrCT1, the participant in the evaluation is a medical
592 expert; in this paper, the participants are three junior radiologists. Doctors at two
593 professional levels can evaluate and experience from different perspectives.
- 594 • Evaluation angle and the role of AI: In DrCT1, the reliability of the model was eval-
595 uated by whether an expert can make a diagnosis based on the selected key images.
596 There was no AI involved in decision-making. In this study, the algorithm generates
597 reports to assist doctors in diagnosis, and experiments can demonstrate the impact of
598 our AI model on doctors' diagnostic decisions.

599 Our innovation not only includes the improvement of algorithm performance, but also
600 takes into account the role of doctors, and is committed to building models that are more
601 likely to be applied to the real world.

602 In the clinic, doctors not only give a diagnosis through medical imaging but also need to
603 comprehend the clinical case history and body examination. Therefore, we will comprehen-
604 sively consider multi-modal data for more comprehensive prediction in our future research.

605 We found that our model has the potential to learn abnormal brain structures. As
606 shown in Figure 9, our model predicts correctly: there is no intracranial hemorrhage in

607 this set of CT scans. However, if the doctor pays attention to the proposed slices with high
 608 attention, the focus of encephalomalacia will be found. This may help doctors to avoid
 609 missed diagnoses.

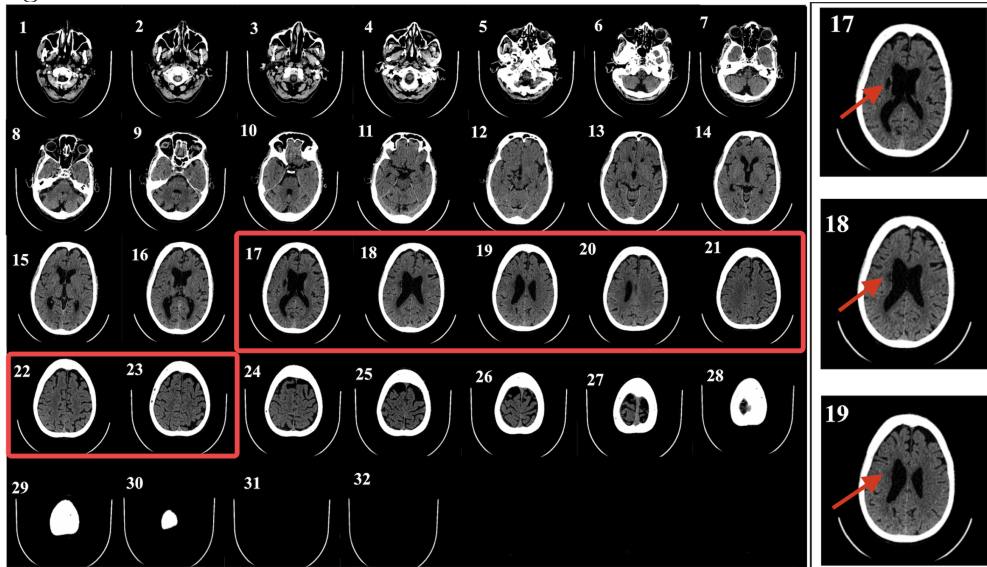


Figure 9: The red box circles the slices selected by the slice attention mechanism. The slices with ID 17, 18, and 19 contain cerebral infarction lesions.

610 **Insufficient trust in the AI model & how to gain the trust of humans:** We observed
 611 (from Table 8 that the use of artificial intelligence models could improve the diagnostic ac-
 612 curacy of junior radiologists. However, the combined (human-computer) accuracy remained
 613 lower than that achieved using AI only. One reason for this may be the lack of trust in the
 614 AI model, as seen for JR A in Table 8. This emphasizes the need for explainability of the
 615 AI systems, which can convince the practitioners of their pertinence. From our interviews
 616 with radiologists, they emphasized that prompting suspicious slices and highlighting key
 617 areas will help AI assist practitioners to clarify their diagnosis. Future directions include
 618 improving human-computer interaction performance, assisting practitioners, and improving
 619 patient care.

620 VIII. Conclusion

621 Our contributions are as follows: first, we proposed a model DrCT2 that can simultaneously
622 detect multiple brain diseases, which imitates the reading habits of human experts: observing
623 closer at key images from a set of slice scans and observing suspicious lesions for diagnosis.
624 The performance of the model was evaluated on two open-access datasets, with the F1 scores
625 of 0.9370 and 0.8700. Second, we proposed three tasks to evaluate the performance of the
626 algorithm for selecting key images. The accuracy of our model on these three tasks was
627 81.59%, 58.76%, and 65.72%, and it achieved better performance than DrCT1. This proves
628 that our primary and auxiliary network coordination mechanism can achieve better results.
629 The three tasks are of great significance for evaluating the interpretability of the model.
630 The key slices selected by the algorithm can provide a reference for the doctor's diagnosis
631 and reduce the occurrence of misdiagnosis and missed diagnosis. From the perspective of
632 human-computer interaction, we invited three junior radiologists to verify the diagnostic
633 effect after using the model. Experimental results prove that our model can effectively
634 improve the accuracy of diagnosis and help doctors improve efficiency. The algorithm avoids
635 complex annotations, is easy to implement and explanatory, and has good application value.
636 In future work, we will explore the potential of the algorithm in small sample data and
637 continue to increase the application potential of the algorithms.

638 IX. Acknowledgments

639 This study is supported by the National Key R&D Program of China with project no.
640 2020YFB2104402. Guanghui Fu is supported by the Chinese Government Scholarship pro-
641 vided by China Scholarship Council (CSC). The authors are grateful to Olivier Colliot and
642 Baptiste Couvy-Duchesne for their comments on the manuscript.

644 References

- 645 ¹ M. Naghavi et al., Global, regional, and national age-sex specific mortality for 264 causes
646 of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016,
647 *The Lancet* **390**, 1151–1210 (2017).
- 648 ² R. J. McDonald, K. M. Schwartz, L. J. Eckel, F. E. Diehn, C. H. Hunt, B. J. Bartholmai,
649 B. J. Erickson, and D. F. Kallmes, The effects of changes in utilization and technological
650 advancements of cross-sectional imaging on radiologist workload, *Academic radiology*
651 **22**, 1191–1198 (2015).
- 652 ³ G. Yang, Y. Zhang, J. Yang, G. Ji, Z. Dong, S. Wang, C. Feng, and Q. Wang, Automated
653 classification of brain images using wavelet-energy and biogeography-based optimization,
654 *Multimedia Tools and Applications* **75**, 15601–15617 (2016).
- 655 ⁴ X. W. Gao, R. Hui, and Z. Tian, Classification of CT brain images based on deep
656 learning networks, *Computer methods and programs in biomedicine* **138**, 49–56 (2017).
- 657 ⁵ R. Jain, N. Jain, A. Aggarwal, and D. J. Hemanth, Convolutional neural network
658 based Alzheimer’s disease classification from magnetic resonance brain images, *Cognitive*
659 *Systems Research* **57**, 147–159 (2019).
- 660 ⁶ J. Islam and Y. Zhang, A novel deep learning based multi-class classification method
661 for Alzheimer’s disease detection using brain MRI data, in *International Conference on*
662 *Brain Informatics*, pages 213–222, Springer, 2017.
- 663 ⁷ G. Fu, J. Li, R. Wang, Y. Ma, and Y. Chen, Attention-based full slice brain CT image
664 diagnosis with explanations, *Neurocomputing* **452**, 263–274 (2021).

- 665 ⁸ L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, Describ-
666 ing videos by exploiting temporal structure, in Proceedings of the IEEE international
667 conference on computer vision, pages 4507–4515, 2015.
- 668 ⁹ S. Sharma, R. Kiros, and R. Salakhutdinov, Action recognition using visual attention,
669 arXiv preprint arXiv:1511.04119 (2015).
- 670 ¹⁰ P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, Hierarchical recurrent neural encoder
671 for video representation with application to captioning, in Proceedings of the IEEE
672 Conference on Computer Vision and Pattern Recognition, pages 1029–1038, 2016.
- 673 ¹¹ S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko,
674 Sequence to sequence-video to text, in Proceedings of the IEEE international conference
675 on computer vision, pages 4534–4542, 2015.
- 676 ¹² Z. Yang, Y. Han, and Z. Wang, Catching the temporal regions-of-interest for video
677 captioning, in Proceedings of the 25th ACM international conference on Multimedia,
678 pages 146–153, 2017.
- 679 ¹³ H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, Video paragraph captioning using hier-
680 archical recurrent neural networks, in Proceedings of the IEEE conference on computer
681 vision and pattern recognition, pages 4584–4593, 2016.
- 682 ¹⁴ Y. Tu, X. Zhang, B. Liu, and C. Yan, Video description with spatial-temporal attention,
683 in Proceedings of the 25th ACM international conference on Multimedia, pages 1014–
684 1022, 2017.
- 685 ¹⁵ C. Yang, A. Rangarajan, and S. Ranka, Visual explanations from deep 3D convolutional
686 neural networks for Alzheimer’s disease classification, in AMIA Annual Symposium
687 Proceedings, volume 2018, pages 1571–1580, American Medical Informatics Association,
688 2018.
- 689 ¹⁶ K. Jnawali, M. R. Arbabshirani, N. Rao, and A. A. Patel, Deep 3D convolution
690 neural network for CT brain hemorrhage classification, in Medical Imaging 2018:
691 Computer-Aided Diagnosis, volume 10575, page 105751C, International Society for Op-
692 tics and Photonics, 2018.
-

- 693 ¹⁷ K. Han, H. Pan, R. Gao, J. Yu, and B. Yang, Multimodal 3D Convolutional Neural
694 Networks for Classification of Brain Disease Using Structural MR and FDG-PET Images,
695 in International Conference of Pioneering Computer Scientists, Engineers and Educators,
696 pages 658–668, Springer, 2019.
- 697 ¹⁸ D. Nie, H. Zhang, E. Adeli, L. Liu, and D. Shen, 3D deep learning for multi-modal
698 imaging-guided survival time prediction of brain tumor patients, in International
699 conference on medical image computing and computer-assisted intervention, pages 212–
700 220, Springer, 2016.
- 701 ¹⁹ X. W. Gao, R. Hui, and Z. Tian, Classification of CT brain images based on deep
702 learning networks, *Computer methods and programs in biomedicine* **138**, 49–56 (2017).
- 703 ²⁰ J. Ker, S. P. Singh, Y. Bai, J. Rao, T. Lim, and L. Wang, Image thresholding improves 3-
704 dimensional convolutional neural network diagnosis of different acute brain hemorrhages
705 on computed tomography scans, *Sensors* **19**, 2167 (2019).
- 706 ²¹ O. S. Faragallah, H. El-Hoseny, W. El-Shafai, W. A. El-Rahman, H. S. El-Sayed, E.-
707 S. M. El-Rabaie, F. E. A. El-Samie, and G. G. N. Geweid, A Comprehensive Survey
708 Analysis for Present Solutions of Medical Image Fusion and Future Directions, *IEEE*
709 *Access* **9**, 11358–11371 (2021).
- 710 ²² Y. Jiang, C. Wu, G. Wang, H.-X. Yao, and W.-H. Liu, MFI-Net: A multi-resolution
711 fusion input network for retinal vessel segmentation, *Plos one* **16**, e0253056 (2021).
- 712 ²³ C. Liu and B. Yang, A Multi-resolution Medical Image Fusion Network with Iterative
713 Back-Projection, in Chinese Conference on Pattern Recognition and Computer Vision
714 (PRCV), pages 41–52, Springer, 2021.
- 715 ²⁴ X. Li, L. Shen, X. Xie, S. Huang, Z. Xie, X. Hong, and J. Yu, Multi-resolution con-
716 volutional networks for chest X-ray radiograph based lung nodule detection, *Artificial*
717 *intelligence in medicine* **103**, 101744 (2020).
- 718 ²⁵ X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng, Self-supervised feature learning
719 for 3d medical images by playing a rubik’s cube, in International Conference on Medical
720 Image Computing and Computer-Assisted Intervention, pages 420–428, Springer, 2019.

- 721 ²⁶ J. Zhu, Y. Li, Y. Hu, K. Ma, S. K. Zhou, and Y. Zheng, Rubik’s cube+: A self-supervised
722 feature learning framework for 3d medical image analysis, *Medical image analysis* **64**,
723 101746 (2020).
- 724 ²⁷ J. Zhu, Y. Li, Y. Hu, and S. K. Zhou, Embedding Task Knowledge into 3D Neural
725 Networks via Self-supervised Learning, arXiv preprint arXiv:2006.05798 (2020).
- 726 ²⁸ P. Rajpurkar et al., Chexnet: Radiologist-level pneumonia detection on chest x-rays
727 with deep learning, arXiv preprint arXiv:1711.05225 (2017).
- 728 ²⁹ Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, Zoom-in-net: Deep mining
729 lesions for diabetic retinopathy detection, in *International Conference on Medical Image*
730 *Computing and Computer-Assisted Intervention*, pages 267–275, Springer, 2017.
- 731 ³⁰ Z. Zhang et al., Pathologist-level interpretable whole-slide cancer diagnosis with deep
732 learning, *Nature Machine Intelligence* **1**, 236–245 (2019).
- 733 ³¹ H. Lee et al., An explainable deep-learning algorithm for the detection of acute in-
734 tracranial haemorrhage from small datasets, *Nature Biomedical Engineering* **3**, 173–182
735 (2019).
- 736 ³² W. Xu, Toward human-centered AI: a perspective from human-computer interaction,
737 *Interactions* **26**, 42–46 (2019).
- 738 ³³ H. Arabi and H. Zaidi, Applications of artificial intelligence and deep learning in molec-
739 ular imaging and radiotherapy, *European Journal of Hybrid Imaging* **4**, 1–23 (2020).
- 740 ³⁴ R. Sayres et al., Using a deep learning algorithm and integrated gradients explanation
741 to assist grading for diabetic retinopathy, *Ophthalmology* **126**, 552–564 (2019).
- 742 ³⁵ K. Zhao, M. Zhang, Z. Xie, X. Yan, S. Wu, P. Liao, H. Lu, W. Shen, C. Fu, H. Cui,
743 Q. Fang, and J. Mei, Deep Learning Assisted Diagnosis of Musculoskeletal Tumors Based
744 on Contrast-Enhanced Magnetic Resonance Imaging, *Journal of Magnetic Resonance*
745 *Imaging* **n/a**.
- 746 ³⁶ J. S. Choi, B.-K. Han, E. S. Ko, J. M. Bae, E. Y. Ko, S. H. Song, M.-r. Kwon, J. H. Shin,
747 and S. Y. Hahn, Effect of a deep learning framework-based computer-aided diagnosis
-

- 748 system on the diagnostic performance of radiologists in differentiating between malignant
749 and benign masses on breast ultrasonography, *Korean journal of radiology* **20**, 749–758
750 (2019).
- 751 ³⁷ Z. Ding et al., Gastroenterologist-level identification of small-bowel diseases and normal
752 variants by capsule endoscopy using a deep-learning model, *Gastroenterology* **157**, 1044–
753 1054 (2019).
- 754 ³⁸ M. Esmaeili, R. Vettukattil, H. Banitalebi, N. R. Krogh, and J. T. Geitung, Explain-
755 able Artificial Intelligence for Human-Machine Interaction in Brain Tumor Localization,
756 *Journal of Personalized Medicine* **11**, 1213 (2021).
- 757 ³⁹ M. J. Vermeulen and M. J. Schull, Missed diagnosis of subarachnoid hemorrhage in the
758 emergency department, *Stroke* **38**, 1216–1221 (2007).
- 759 ⁴⁰ K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image
760 recognition, *arXiv preprint arXiv:1409.1556* (2014).
- 761 ⁴¹ J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, Imagenet: A large-scale
762 hierarchical image database, in 2009 IEEE conference on computer vision and pattern
763 recognition, pages 248–255, Ieee, 2009.
- 764 ⁴² M. Schuster and K. K. Paliwal, Bidirectional recurrent neural networks, *IEEE Trans-*
765 *actions on Signal Processing* **45**, 2673–2681 (1997).
- 766 ⁴³ J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, Empirical evaluation of gated recurrent
767 neural networks on sequence modeling, *arXiv preprint arXiv:1412.3555* (2014).
- 768 ⁴⁴ N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: a
769 simple way to prevent neural networks from overfitting, *The journal of machine learning
770 research* **15**, 1929–1958 (2014).
- 771 ⁴⁵ S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by
772 reducing internal covariate shift, *arXiv preprint arXiv:1502.03167* (2015).
- 773 ⁴⁶ S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal,
774 V. Mahajan, P. Rao, and P. Warier, Deep learning algorithms for detection of critical
775 findings in head CT scans: a retrospective study, *The Lancet* **392**, 2388–2396 (2018).

- 776 ⁴⁷ A. E. Flanders et al., Construction of a machine learning dataset through collaboration:
777 the RSNA 2019 brain CT hemorrhage challenge, *Radiology: Artificial Intelligence* **2**,
778 e190211 (2020).
- 779 ⁴⁸ N. Ketkar, Introduction to pytorch, in *Deep learning with python*, pages 195–208,
780 Springer, 2017.
- 781 ⁴⁹ D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv preprint
782 arXiv:1412.6980 (2014).
- 783 ⁵⁰ R. H. Pinheiro, G. D. Cavalcanti, R. F. Correa, and T. I. Ren, A global-ranking local
784 feature selection method for text categorization, *Expert Systems with Applications* **39**,
785 12851–12857 (2012).
- 786 ⁵¹ SeuTao, Code for 1st Place Solution in Intracranial Hemorrhage Detec-
787 tion Challenge @ RSNA2019, [https://github.com/SeuTao/RSNA2019_](https://github.com/SeuTao/RSNA2019_Intracranial-Hemorrhage-Detection/)
788 [Intracranial-Hemorrhage-Detection/](https://github.com/SeuTao/RSNA2019_Intracranial-Hemorrhage-Detection/), Accessed March 1, 2020.
- 789 ⁵² R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, Deep learning for healthcare:
790 review, opportunities and challenges, *Briefings in bioinformatics* **19**, 1236–1246 (2018).
-