Chinese practice of benchmarking and algorithm evaluation

Shan Xu, CAICT

Importance of benchmarking

- Where is the starting line?
- Where is the finish line?
- Who can be the participants?
- What is the technique score?
- Who will be the champion?

Health & medical field: Particularity & Sensitivity

Medical device approval of FDA

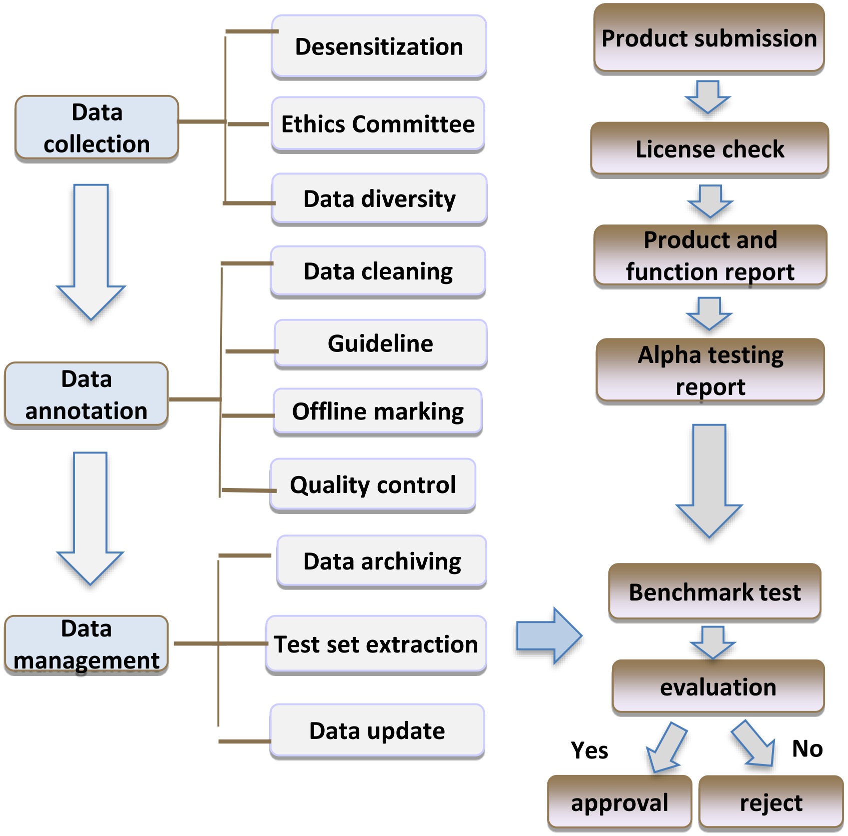
Products	Corporation	Expected Usage	FDA approval	FDA Test	
				Non-clinical Test	Clinical Test
Kardia Mobile/Band	AliveCor	ECG analysis, AF identification	510(k)	Complete all non-clinical tests, including ECG special requirements (using ECG standard library)	Have clinical trials, no details
CardioLogs ECG Analysis Platform	CardioLogs Technologies	Arrhythmia detection	510(k)	Have non-clinical tests (using ECG standard library)	No clinical tests
Wave Clinical Platform	Excel Medical	Physiological data analysis, risk-based decision support, disease warning	510(k)	Clearly conducted non-clinical performance tests	No clinical tests
Contact application	Viz.AI	Analysis of the risk of large blood vessel obstruction in the brain on CT images	De Novo	Calculated ROC on the standard performance test set	300 cases of retrospective clinical test
IDX-DR	IDX LLC	Detection of diabetic retinopathy	De Novo	No non-clinical test content	900 cases of prospective clinical test
Arterys Cardio DL	Arterys LLC	Analysis of blood flow and cardiac output on MRI images	510(k)	Non-clinical performance test using a test set of 1000 cases	
Arterys Oncology DL	Arterys LLC	Assist in confirming the presence of lesions and image segmentation on radiological images	510(k)	Clearly confirm and verify the non-clinical approach to the deep learning model	No clinical tests
iCAD PowerLook® Tomo	iCAD INC.	Breast CT/MRI image analysis	PMA	Non-clinical performance test using a test set of 240 lesions	603 cases of retrospective clinical tests
QVCAD System	Qview Medical Inc.	Ultrasound image analysis of breast	PMA	Non-clinical performance test using a test set of 398 cases	185 cases of retrospective clinical tests

Explore different paths in China

- **FDA approval paths:** not responsible for clinical test, companies provide data and testing report, and the review board monitor the progress of the process at any time.
- **China's exploration path :** trying to establish benchmark dataset and evaluation standards based on the advantages of massive data accumulation in hospitals.

	Time span	Quantities	storage
Data size	＞15 years	40 billion	
Medical records	＞7 years	1.8 million	
Examinations	＞10 years	＞250 million	
Imaging reports	＞6 years		180T
Lab results	＞10 years	＞400 million	
Medical orders	＞15 years	20 billion	
Drug information	＞15 years	30 billion	
……			

Massive data accumulation in one hospital in Changsha



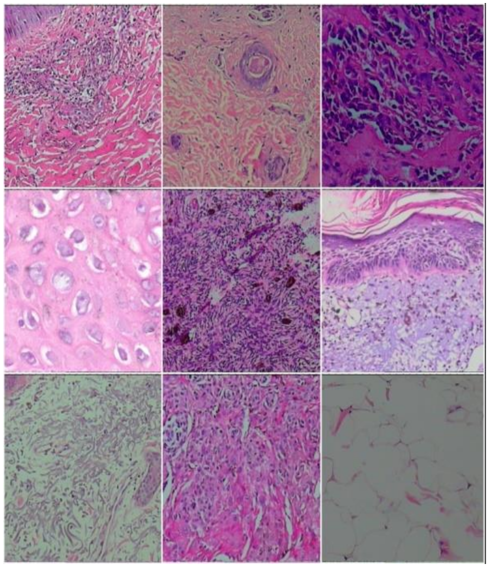
China’s benchmark dataset

Attributes	Dermatology data set	Ophthalmic data set	Ophthalmic data set	Lung data set	ECG data set
ownership	Xiangya Hospital	Union Hospital	NIFDC (CFDA)	NIFDC (CFDA)	Public
Data source	＞200 hospitals in 30 province	5 hospitals headed by Peking Union Medical College Hospital	11 hospitals in 10 province	22 hospitals in 9 province	11 hospitals
Data capacity	＞50,000 patients;＞0.2 million clinical images	**A** 3290 cases + **B** 1000 cases	6327 cases	623 cases , 4436 nodules	6877 cases
Data set use	**A** To meet the development of AI products for the diagnosis of skin disease; **B** To construct structured electronic medical records	**A** To meet AI product evaluation for diabetic retinopathy ； **B** To meet AI product evaluation for screening of fundus diseases	To meet AI product evaluation for diabetic retinopathy	To meet AI product evaluation for lung nodules	For the physiological signal analysis of one normal type and eight abnormal types of 12-lead ECG
Scope of Application	**A** Skin disease picture recognition, auxiliary diagnosis, skin pathological diagnosis; **B** High efficiency support of medical joint platform	**A** Lesions, staging, referral/non-referral, quality discrimination of diabetic retinopathy ； **B** Judgment of the presence/absence of 18 common fundus diseases	Staging, referral/non-referral, quality discrimination of diabetic retinopathy ；	Detection, classification, boundary segmentation, size measurement of lung nodules	Automatic identification of the rhythm/morphology abnormalities in clinical ECGs
Database classification	Clinical skin picture (skin tumor, erythema scaly skin disease, bullous skin disease, etc.), dermascopic picture, skin pathology picture	**A** Lesions, staging, referral/non-referral, quality of diabetic retinopathy； **B** The presence/ absence of fundus diseases	Diabetic retinopathy 0～4; Other fundus diseases; Unrecognizable images	**A** Intrapulmonary solid / partially solid / pure ground-glass / calcified nodules, **B** Pleural solid / calcified nodules	Normal, Atrial fibrillation (AF), first-degree atrioventricular block (I-AVB), left bundle brunch block (LBBB), right bundle brunch block (RBBB), premature atrial contraction (PAC), premature ventricular contraction (PVC), ST-segment depression (STD), ST-segment elevated (STE)

Skin Disease Picture Library

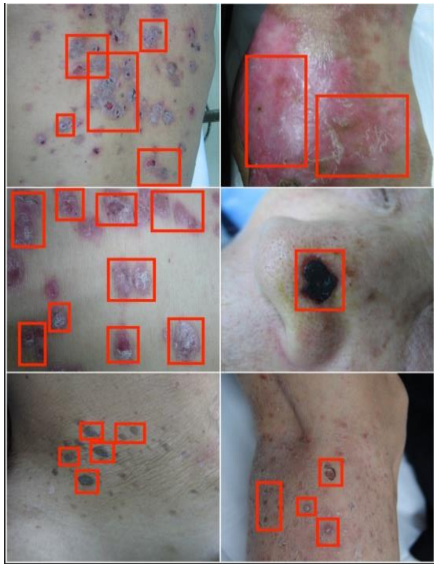
Skin disease big data acquisition platform

- Collected data of over 50,000 dermatology patients from 200 different hospitals
- Expanded to over 100 hospitals
- Acquired 5 software copyrights



Dermatopathology picture library
（1 million pics）

Standardized skin disease picture library
（0.4 million pics）



Tagged picture library
（20,000 pics）

Fundus Image Database

Data capacity

- **6327** cases
- **7 types**:
  Diabetic retinopathy 0～4;
  Other fundus diseases;
  Unrecognizable images

Data diversity

Data source: **11** hospitals in 10 provinces

- **Imaging equipment**:
  ①**≥13** fundus cameras;
  ②field angle:**27°-45°**;
  ③Meet the fundus camera line;
  ④**1 ~10 million** pixels

Reference standard

- **Labeling doctors**: Attending physician and above with **≥5 years** of experience
- **Selection test**:
  ①**2** rounds of exams, **120** images
  - Diabetic retinopathy 57%, other lesions 38%, unrecognizable 5%
  ②Results: **15 out of 47** doctors
  -Accuracy >80%, stability >85%, consistency > 0.75 (Fleiss Kappa)

Figure 1: Data types of the 6327 cases

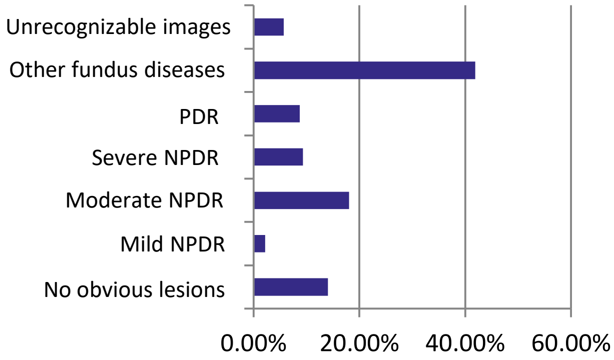


Figure 2: Distribution of the image pixels

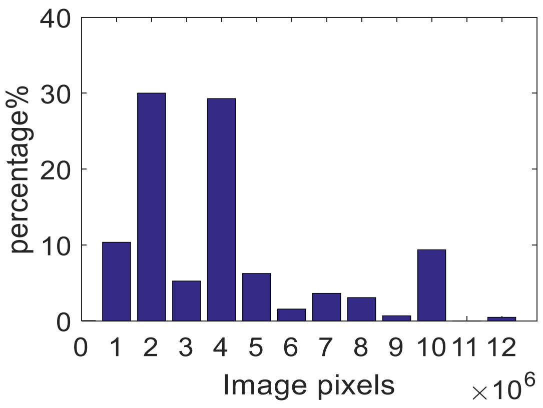
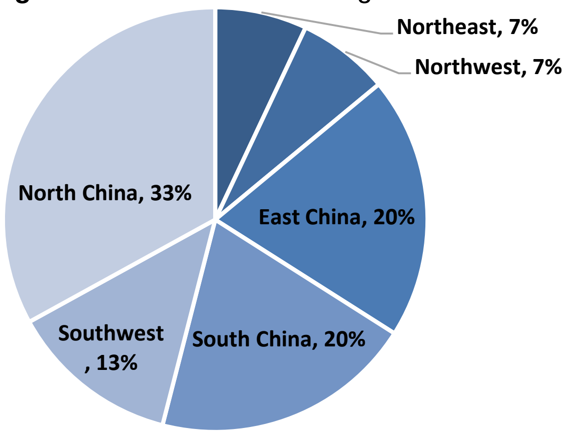


Figure 3 : Area of the labeling doctors



Lung Image Database

Data capacity

- **623** cases;**4436** nodules
- **6 types**: **Intrapulmonary** solid / partially solid / pure ground-glass / calcified nodules, **pleural** solid / calcified nodules

Data diversity

- **Data source**: **22** hospitals in 9 provinces
- **Imaging equipment**: ① **≥15** CT types； ② Routine + enhanced CT **66%**, low dose screening 34%;③ ≥3mm nodules **50.95%**, <3mm nodules 49.05%

Reference standard

- **Labeling doctors**: **24** labeling doctors + **15** arbitration experts from 220 doctors
  ① Accuracy >80%, classification accuracy >80%, average cross ratio >80%
  ② the lowest is the **intermediate level**, the subordinate senior level **56.4%**.
  ③ Average working years **> 13 years**
  ④ From **25 top three hospitals** from 13 provinces and autonomous regions

Figure 1: Area distribution of data source

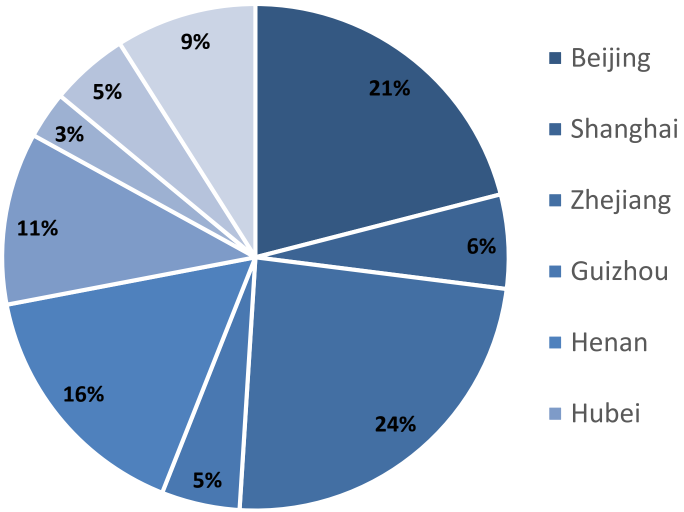
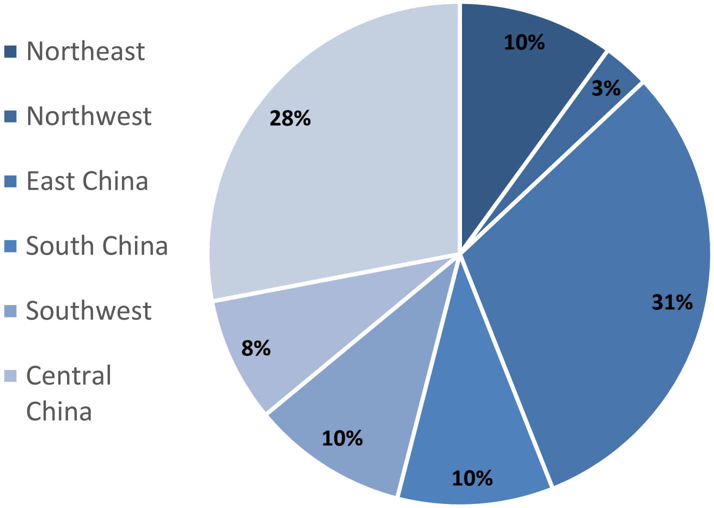


Figure 2: Area of the labeling doctors



ECG Benchmark Dataset

Introduction

- A platform for the open-source **data and algorithms** for cardiovascular disease (CAD) early detection in China.

Target

- To encourage the development of algorithms to identify the **rhythm/ morphology abnormalities** from 12-lead ECGs.

Content

- The **training set** contains **6,877** (female: 3178; male: 3699) 12 leads ECG recordings lasting from 6 s to just 60 s;
- The **test set** contains **2,954** ECG recordings with the similar lengths.

Table : Data profile for the training set according to the 'Frist label' annotations.

Type	#recording	Time length (s)				
		Mean	SD	Min	Median	Max
Normal	918	15.43	7.61	10.00	13.00	60.00
Atrial fibrillation (AF)	1098	15.01	8.39	9.00	11.00	60.00
First-degree atrioventricular block (I-AVB)	704	14.32	7.21	10.00	11.27	60.00
Left bundle branch block (LBBB)	207	14.92	8.09	9.00	12.00	60.00
Right bundle branch block (RBBB)	1695	14.42	7.60	10.00	11.19	60.00
Premature atrial contraction (PAC)	556	19.46	12.36	9.00	14.00	60.00
Premature ventricular contraction (PVC)	672	20.21	12.85	6.00	15.00	60.00
ST-segment depression (STD)	825	15.13	6.82	8.00	12.78	60.00
ST-segment elevated (STE)	202	17.15	10.72	10.00	11.89	60.00
Total	6877	15.79	9.04	6.00	12.00	60.00

Algorithm Evaluation in AI4H

Benchmark dataset

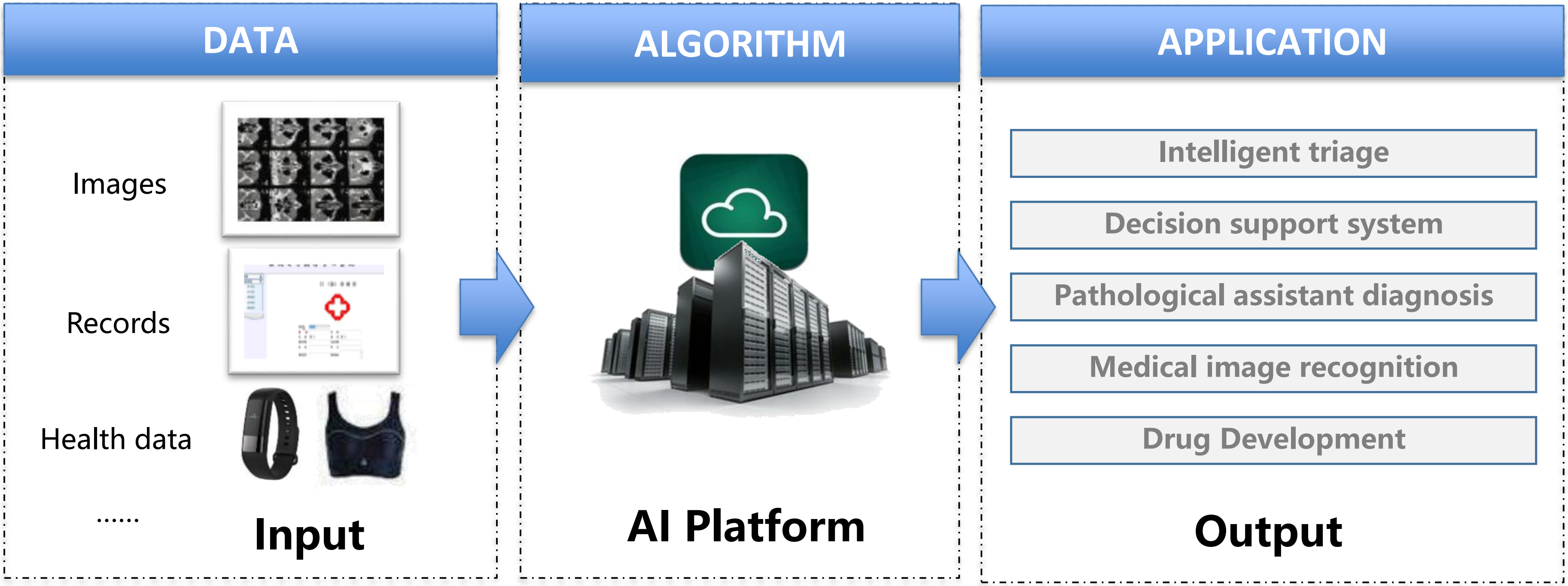
- Data diversity
- Data capacity
- Reference standard

Algorithm performance

- Algorithm accuracy
- Time complexity
- Space complexity

Service quality evaluation

- Service accuracy
- Customer satisfaction
- Response timeliness



DATA

Images

Records

Health data

……

Input

ALGORITHM

AI Platform

APPLICATION

Intelligent triage

Decision support system

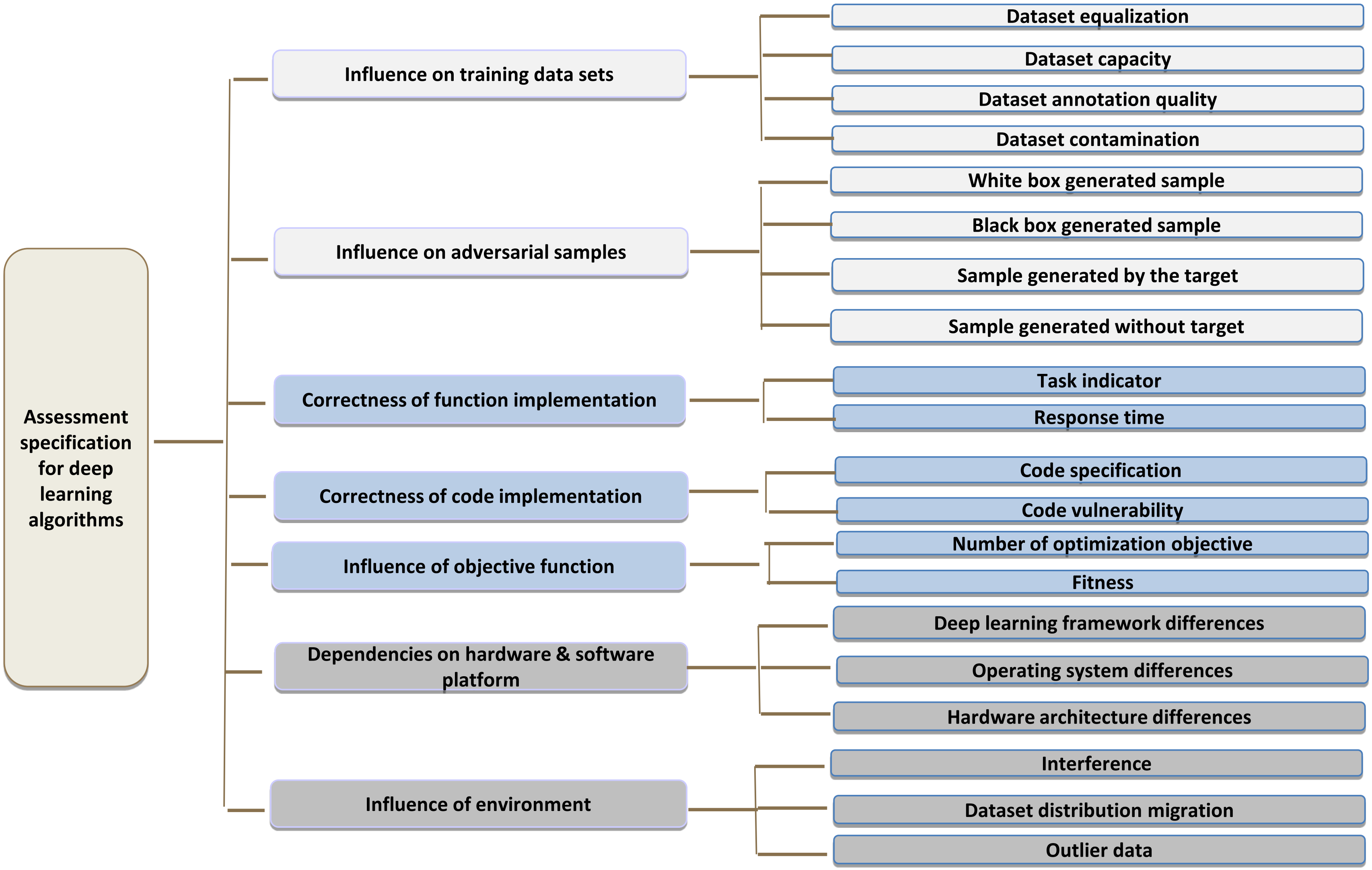
Pathological assistant diagnosis

Medical image recognition

Drug Development

Output

Evaluation criteria framework



- Assessment specification for deep learning algorithms
  - Influence on training data sets
    - Dataset equalization
    - Dataset capacity
    - Dataset annotation quality
    - Dataset contamination
  - Influence on adversarial samples
    - White box generated sample
    - Black box generated sample
    - Sample generated by the target
    - Sample generated without target
  - Correctness of function implementation
    - Task indicator
    - Response time
  - Correctness of code implementation
    - Code specification
    - Code vulnerability
  - Influence of objective function
    - Number of optimization objective
    - Fitness
  - Dependencies on hardware & software platform
    - Deep learning framework differences
    - Operating system differences
    - Hardware architecture differences
  - Influence of environment
    - Interference
    - Dataset distribution migration
    - Outlier data

* Artificial intelligence—Assessment specification for deep learning algorithms [AIOSS—01--2018]

Evaluation index in different stage

stage			Demand stage				Design stage			
Reliability goal			A	B	C	D	A	B	C	D
Evaluation index	Influence on training dataset	Data et equalization	--	--	--	--	●	●	●	●
		Dataset size	--	--	--	--	●	●	●	○
		Dataset annotation quality	--	--	--	--	●	●	○	○
		Dataset contamination	--	--	--	--	●	○	○	○
	Influence on adversarial samples	White box generated sample	--	--	--	--	--	--	--	--
		Black box generated sample	--	--	--	--	--	--	--	--
		Sample generated by the target	--	--	--	--	--	--	--	--
		Sample generated without target	--	--	--	--	--	--	--	--
	Correctness of algorithm function	Task indicator	●	●	●	●	●	●	●	●
		Response time	●	●	○	○	●	●	○	○
	Correctness of code implementation	Code specification	--	--	--	--	--	--	--	--
		Code vulnerability	--	--	--	--	--	--	--	--
	Influence of the objective function	Number of optimization objective	--	--	--	--	●	●	○	○
		Fitness	●	●	○	○	--	--	--	--
	Dependencies on hardware & software platform	Deep learning framework differences	●	●	●	○	--	--	--	--
		Operating system differences	●	●	○	○	--	--	--	--
		Hardware architecture differences	○	○	○	○	--	--	--	--
	Influence of environment	Interference	--	--	--	--	--	--	--	--
		Dataset distribution migration	--	--	--	--	--	--	--	--
		Outlier data	--	--	--	--	--	--	--	--

* Artificial intelligence—Assessment specification for deep learning algorithms [AIOSS—01--2018]

Evaluation index in different stage

stage			Implementation stage				Operational stage			
Reliability goal			A	B	C	D	A	B	C	D
	Influence on training dataset	Data et equalization	●	●	○	○	--	--	--	--
		Dataset size	--	--	--	--	--	--	--	--
		Dataset annotation quality	--	--	--	--	--	--	--	--
		Dataset contamination	--	--	--	--	--	--	--	--
	Influence on adversarial samples	White box generated sample	●	●	●	○	--	--	--	--
		Black box generated sample	●	●	○	○	--	--	--	--
		Sample generated by the target	●	●	○	○	--	--	--	--
		Sample generated without target	●	●	○	○	--	--	--	--
	Correctness of algorithm function	Task indicator	●	●	●	●	●	●	●	●
		Response time	●	●	○	○	●	●	○	○
	Correctness of code implementation	Code specification	●	●	●	○	--	--	--	--
		Code vulnerability	●	●	○	○	--	--	--	--
	Influence of the objective function	Number of optimization objective	--	--	--	--	--	--	--	--
		Fitness	--	--	--	--	--	--	--	--
	Dependencies on hardware & software platform	Deep learning framework differences	--	--	--	--	●	●	●	○
		Operating system differences	--	--	--	--	●	●	○	○
		Hardware architecture differences	--	--	--	--	●	○	○	○
	Influence of environment	Interference	--	--	--	--	●	●	●	○
		Dataset distribution migration	--	--	--	--	●	●	○	○
		Outlier data	--	--	--	--	●	○	○	○

* Artificial intelligence—Assessment specification for deep learning algorithms [AIOSS—01--2018]

Trial application of the Evaluation

Medical AI Evaluation Contest

Directed by

Ministry of Science and Technology of PRC

Organized by

Artificial Intelligence Industry Technology Innovation Strategic Alliance (AITISA)

Co-organized by

China Academy of Information and Communications Technology(CAICT)

National Medical Center of National Health Commission pf China

SHENZHEN Cyberspace laboratory

Tencent MIAIS

National Engineering Laboratory for Internet Medical Systems and Applications

National Engineering Laboratory for Medical Big Data Application Technology

Mobile Health Ministry of Education China Mobile Joint Laboratory

Contestants' scope

Companies

Universities

Hospitals

Individuals

2018.04 Collaboration

2018.05 Preparation

2018.6.21 Startup meeting

2018.09 Starting evaluation

2018.11 Primary Election

2019. 01 Semi-finals

Finals and awards in April 2019

Thank you !