ICPR 2020

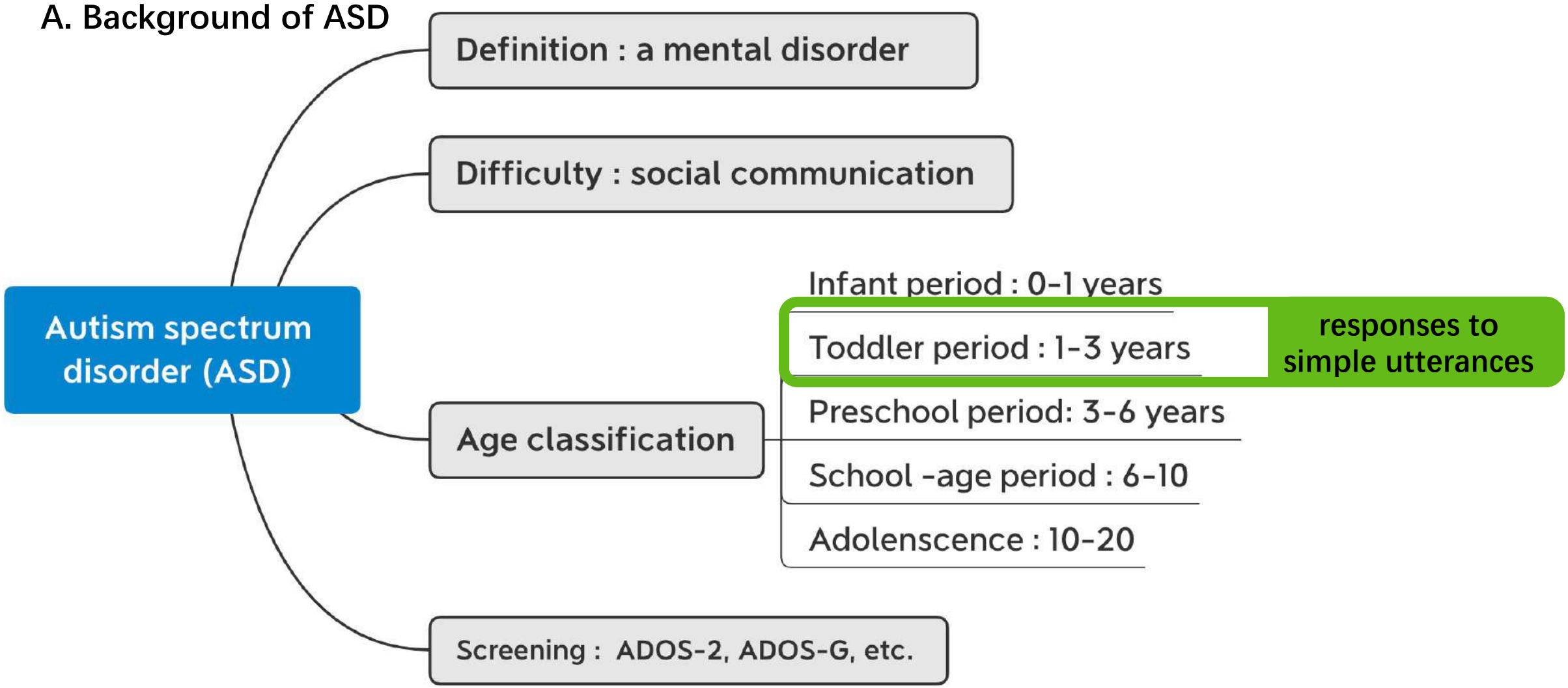
Responsive Social Smile: A Machine Learning based Multimodal Behavior Assessment Framework towards Early Stage Autism Screening

Yueran Pan[1], Kunjing Cai[2], Ming Cheng[1], Xiaobing Zou[3], Ming Li[1]
[1] Data Science Research Center, Duke Kunshan University, Kunshan, China
[2] Sun Yat-sen Universit Guangzhou, China
[3] The Third Affiliated Hospital, Sun Yat-sen University Guangzhou, China

OUTLINE

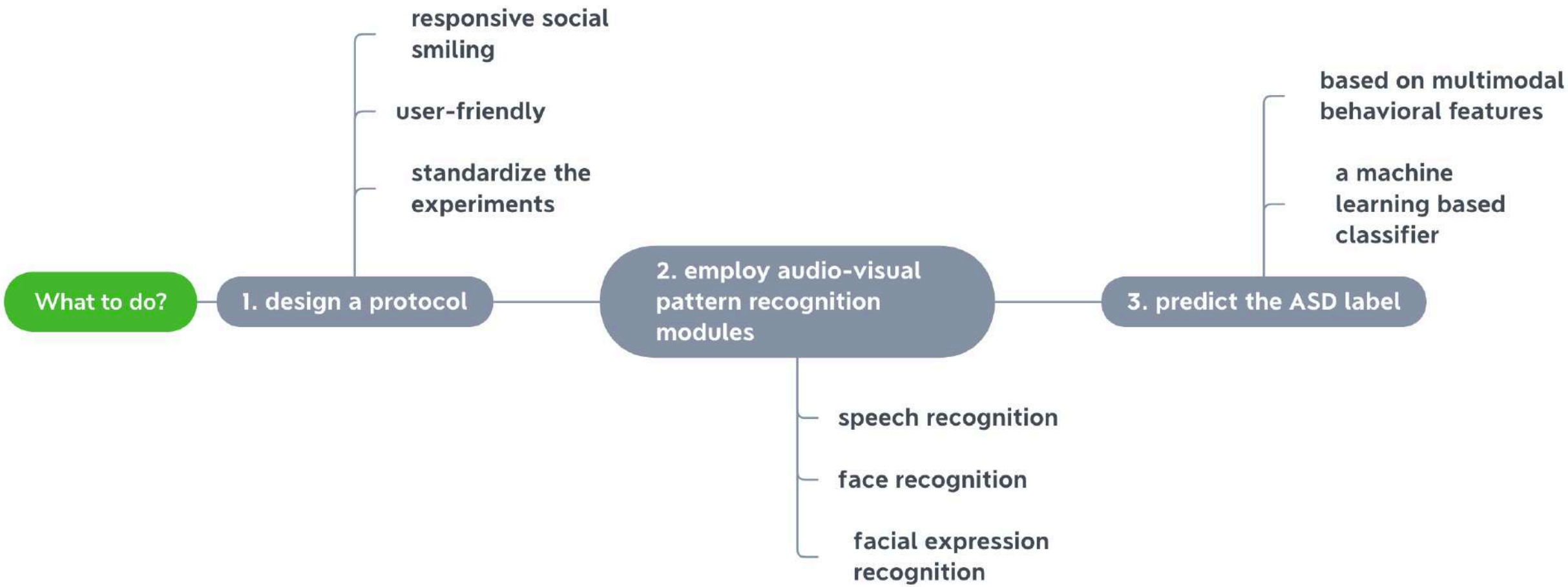
1. INTRODUCTION

A. Background of ASD



1. INTRODUCTION

B. Our proposed method



2. RELATED WORK

Technology towards ASD

TABLE I
COMPARISONS OF TYPICAL METHODS

Authors	Method	Algorithm	Accuracy	Sensitivity	Specificity	Data Scale (ASD/Non-ASD)	Age (Years)
Liu et al. [4]	Eye movement	K-means + SVM	88.51%	93.10%	86.21%	29/58	4-11
Li et al. [5]	Hand imitation tasks	Linear SVM	86.70%	85.70%	87.50%	16/14	2-4
Nakai et al. [6]	Abnormal prosody	SVM	76.00%	81.00%	73.00%	31/51	3-10
Heinsfeld et al. [7]	Neuroimaging	Neural Networks	70.00%	74.00%	63.00%	505/535	7-64
Ours	Responsive social smile	CNN + Decision Tree	80.49%	85.00%	77.27%	20/21	1-3

too expensive

Could be better

Not young

3. PROTOCOL AND DATABASE

A. Procedure of the responsive social smile protocol

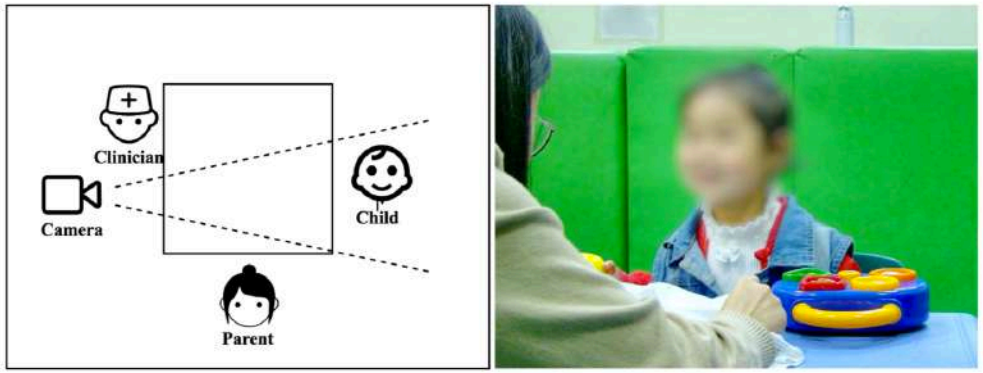


Fig. 1. The layout of the experimental environment and video recording example.

TABLE II
STIMULI AND KEY WORDS IN A PROTOCOL.

	Stimulus	Key Words	Voice Source
1	Greeting smile	"Hello!" + Children's names	Clinician
2	Praise words	"You are so cute/cool!"	Clinician
3	Hide and seek	"Let's play hide and seek'.'	Clinician
4	Hints of tickling	"I am going to tickling you!"	Clinician
5	Tickling	"Real tickling now!"	Clinician
6	Greeting smile	"Hello!" + Children's names	Parent

- Friendly environment :
  - green walls, colorful chairs and toys
- Audio-video recording
- Three paticipants

3. PROTOCOL AND DATABASE

B. Clinical Database

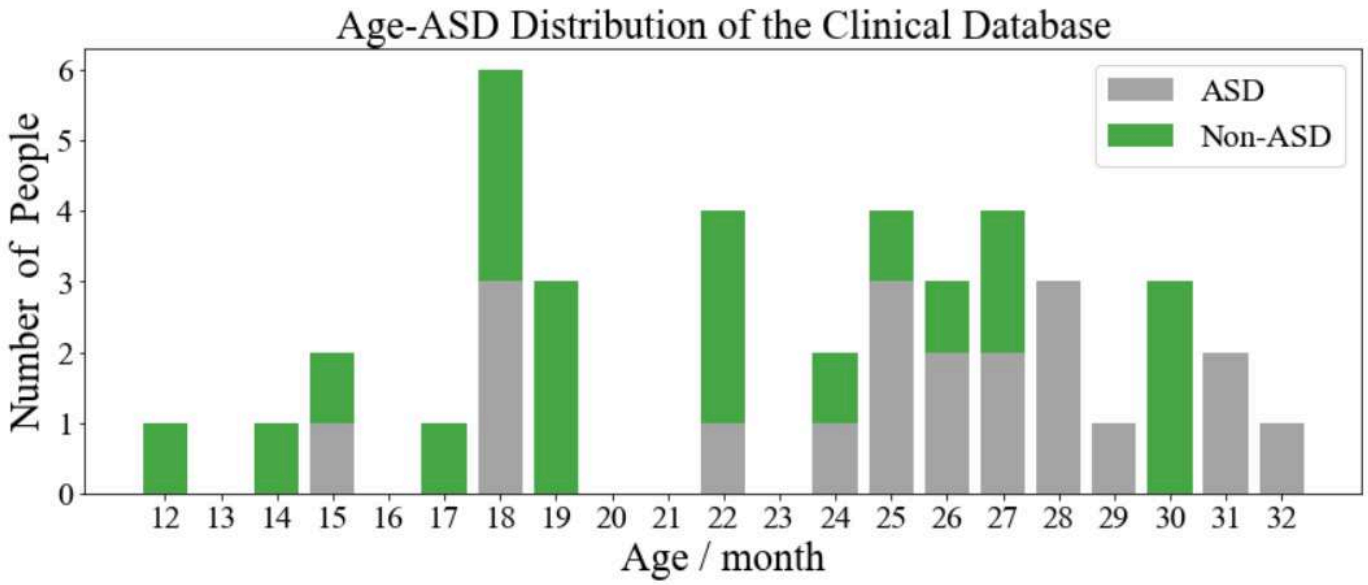
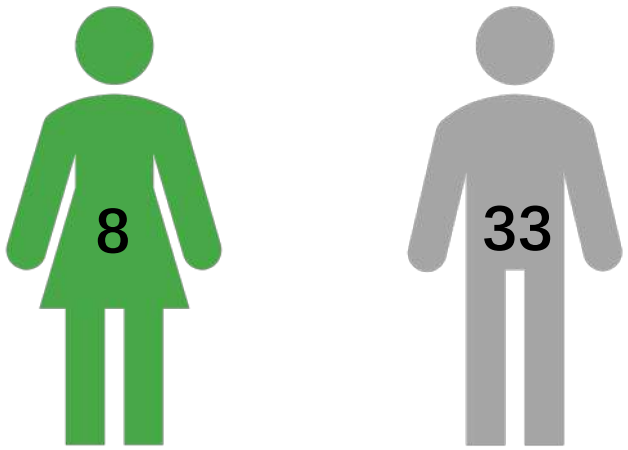
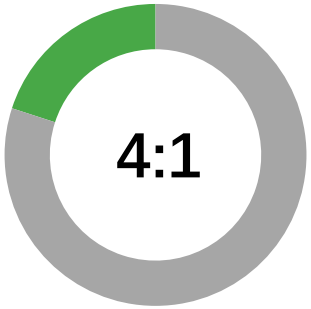


Fig. 2. Age-ASD distribution in the clinicial database.



Gender ratio of the clinicial database



Common gender ratio of ASD

4. MULTIMODAL ASSESSMENT FRAMEWORK

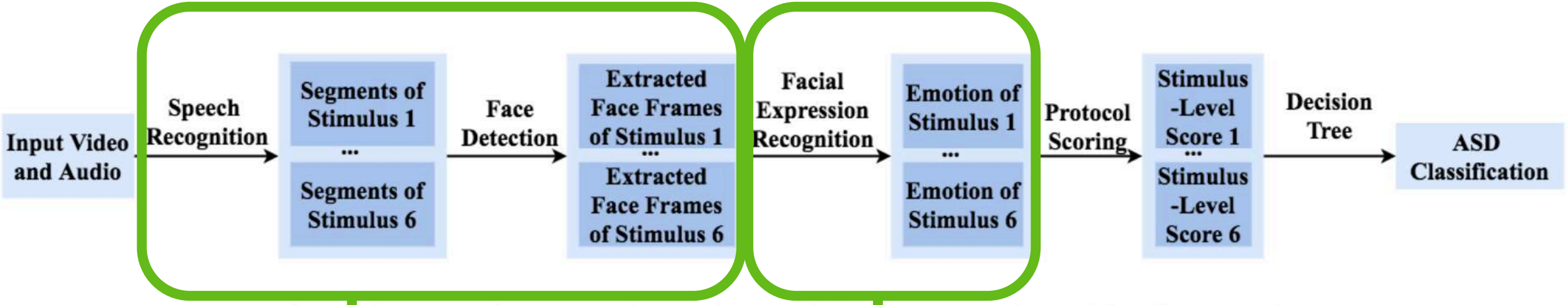


Fig. 3. The proposed assessment framework for analyzing the "Responsive Social Smile" protocol.

Refer others' work

Design our own model

4. MULTIMODAL ASSESSMENT FRAMEWORK

A. Temporal Stimulus Localization

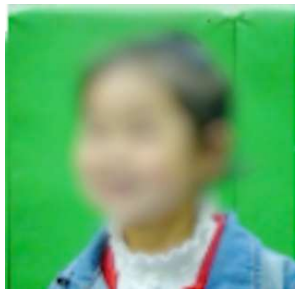
Kaldi + AISHELL-2 database → Our ASR system

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[24] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

B. Face Detection

OpenCV-DNN →



[25] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008.

4. MULTIMODAL ASSESSMENT FRAMEWORK

C. Facial Expression Recognition

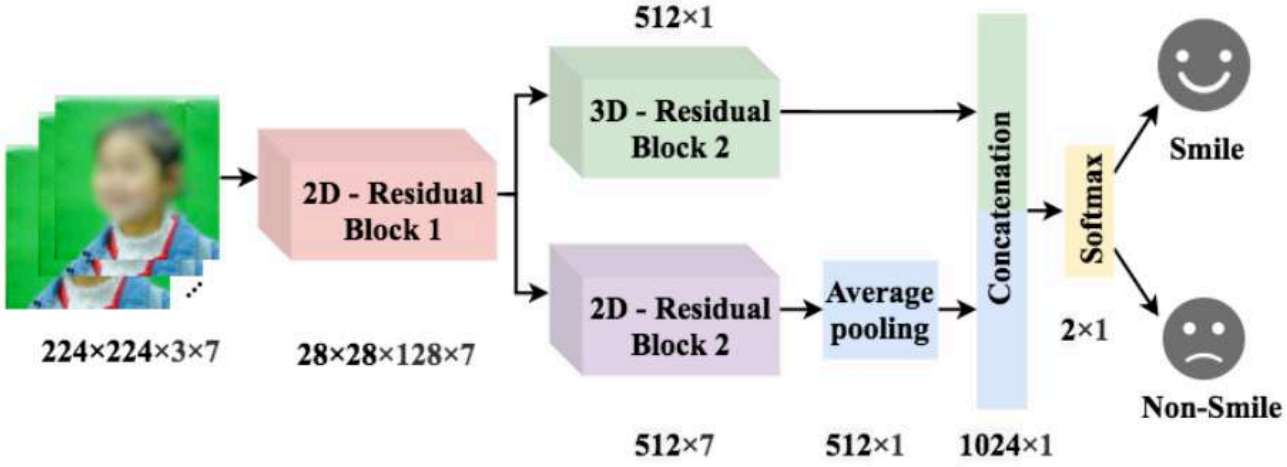


Fig. 4. Structure of the facial expression recognition neural network.

TABLE III
ARCHITECTURE OF THE FACIAL EXPRESSION RECOGNITION MODEL

Layer Name	2D CNN Branch	3D CNN Branch
conv1	7×7, 64, stride 2	
conv2	3×3 max pool, stride 2	
	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	
conv3	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$
conv4	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 2$
conv5	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 2$
pooling	average pooling	None
merge	concatenation, Softmax	

Spatial + Temporal

4. MULTIMODAL ASSESSMENT FRAMEWORK

C. Facial Expression Recognition

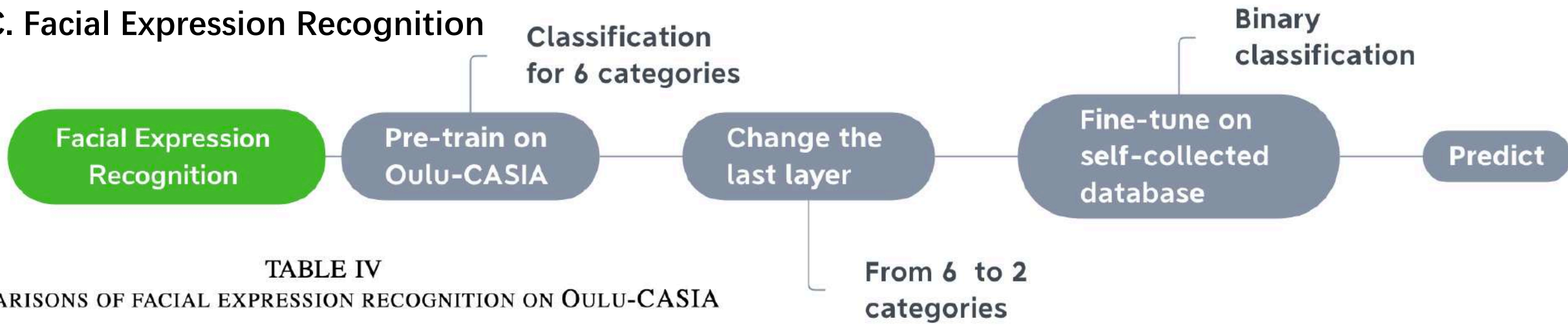


TABLE IV
COMPARISONS OF FACIAL EXPRESSION RECOGNITION ON OULU-CASIA

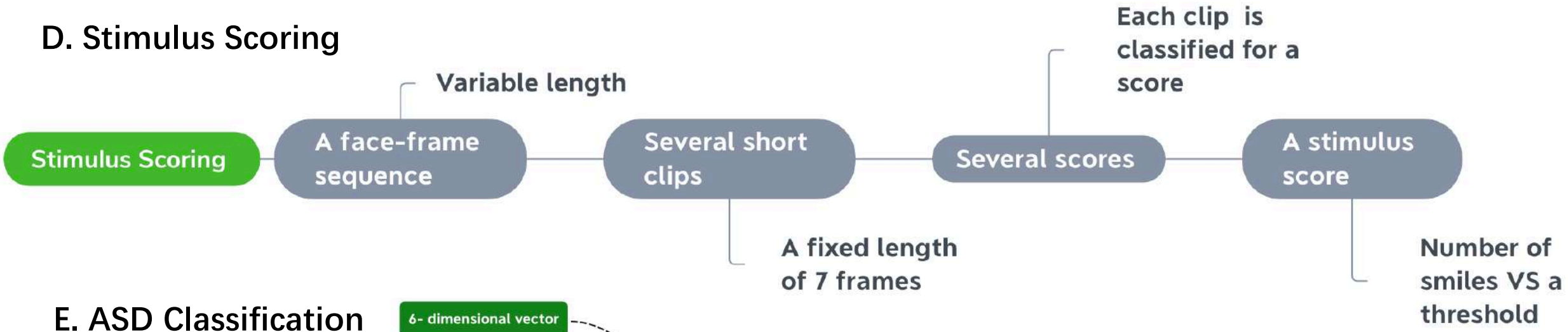
Method	Descriptor	Accuracy
Yu et al. [31]	DCPN	86.23%
Jung et al. [32]	CNN-DNN	81.46%
Zhang et al. [33]	PHRNN-MSCNN	86.25%
Kuo et al. [34]	CNN	91.67%
Ours	2D-3D CNNs	89.10%

Oulu-CASIA database :
480 videos (80 subjects by six expressions)

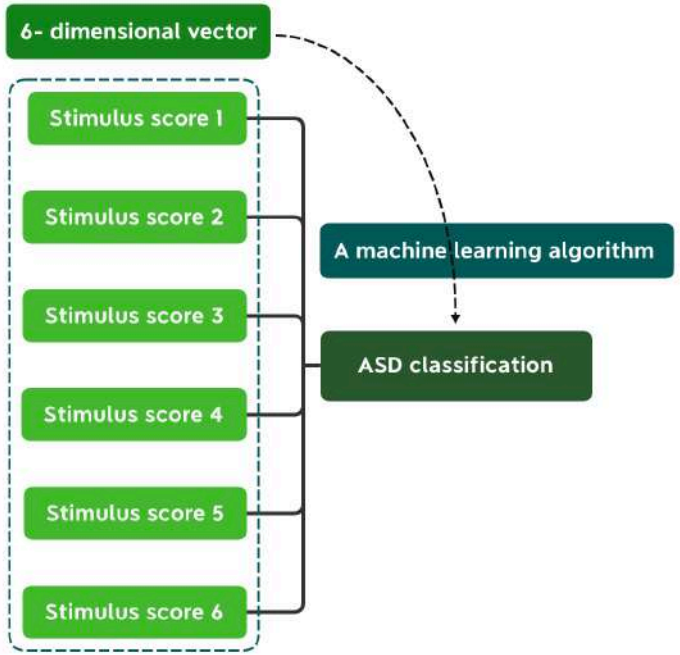
[30] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietika¨Inen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

4. MULTIMODAL ASSESSMENT FRAMEWORK

D. Stimulus Scoring



E. ASD Classification



$$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \longrightarrow \text{Non-ASD}$$

5. EXPERIMENTS

A. Experiment Settings

A stimulus-level video:

- 20 seconds
- under the condition of 24 FPS
- approximately 480 frames

Face images:

- resized to the shape of 224×224
- 7-frame clips

5. EXPERIMENTS

B. Fine-tuning FER Model

Two major problems：

- The output of the pre-trained model has six categories, which does not match with our binary classification.
  - **Change the last layer.**

- Most databases for facial expression recognition are collected from adults, which may not work well on young children.
  - **Fine-tune on a self-collected database containing 15,000 videos.**
  - **Achieve the accuracy of 92.60% for smile classification on the self-collected database.**

Self-collected database

5. EXPERIMENTS

C. Results of Stimulus Scoring

- **Children** : 41
- **Stimulus scores** : 196
- **Threshold** : 0.9
  which means the child must give a clear enough response to count as smiling.
- **Evaluation label** : majority voting from three clinicians' individual results.
- **Validation**: leave-one-out cross validation
- **Accuracy** : 85.20%

TABLE V
CONFUSION MATRIX OF STIMULUS SCORING ON THE COLLECTED CLINICAL DATABASE

		Predicted	
		Smile	Non-Smile
Actual	Smile	62	13
	Non-Smile	16	105

5. EXPERIMENTS

D. Results of ASD Classification

TABLE VI
CONFUSION MATRIX OF ASD CLASSIFICATION BASED ON PREDICTED ASD CLASSIFICATION BASED ON PREDICTED STIMULUS SCORES
STIMULUS SCORES

		Predicted	
		ASD	Non-ASD
Actual	ASD	17	3
	Non-ASD	5	16

- **Children** : 41
- **Input**: 6-dimensional feature vector consisting of all stimulus scores
- **Missing data** : mean of the other stimulus scores from the same child
- **Validation**: leave-one-out cross validation
- **Accuracy** : 80.49%
- **Evaluation** : predict with clinicians'' stimulus scores directly

TABLE VII

Algorithm	Accuracy	Sensitivity	Specificity
Logistic Regression	63.41%	66.67%	63.64%
Naive Bayes	68.29%	65.00%	68.42%
SVM	70.73%	70.00%	70.00%
Decision Tree	80.49%	85.00%	77.27%

TABLE VIII
ASD CLASSIFICATION BASED ON CLINICIAN'S STIMULUS SCORES

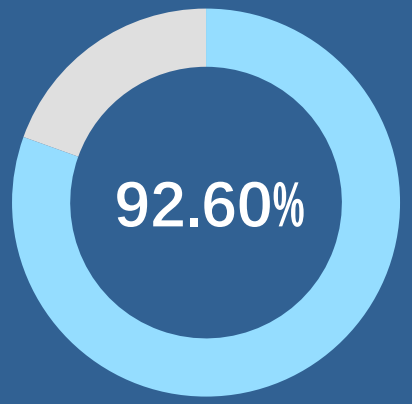
Algorithm	Accuracy	Sensitivity	Specificity
Logistic Regression	70.73%	70.00%	70.00%
Naive Bayes	73.17%	75.00%	71.43%
SVM	75.61%	70.00%	77.78%
Decision Tree	82.93%	80.00%	84.21%

5. EXPERIMENTS

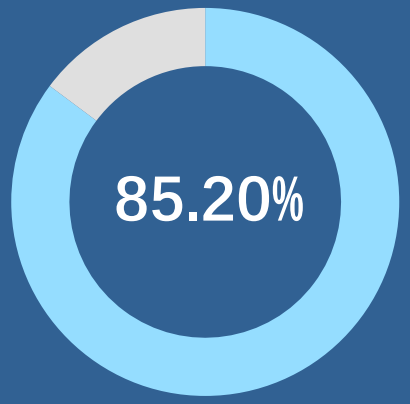
E. Failure Case Study

Fig. 5. Examples of failure cases.

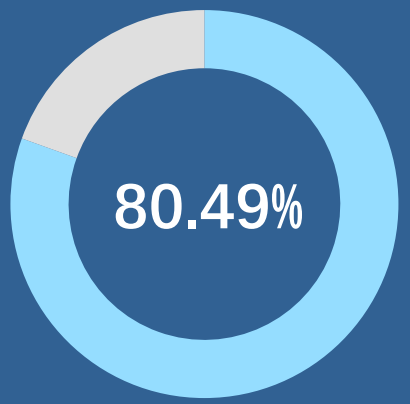
6. CONCLUSION



Self-collected emotion classification



Stimulus scoring



ASD classification

Future Research

- Judge whether children' s reactions are not due to designed stimulus.
- Fuse data from multiple complementary protocols of a child to further enhance the screening performance.

7. OURTEAM

AFFILIATIONS:

SMIIP LAB,
DATA SCIENCE RESEARCH CENTER,
DUKE KUNSHAN UNIVERSITY

CHILD DEVELOPMENT BEHAVIOR CENTER,
THE THIRD AFFILIATED HOSPITAL OF SUN YAT-SEN UNIVERSITY·LINGNAN HOSPITAL





AUTHORS: YUERAN PAN KUNJING CAI MING CHENG XIAOBING ZOU MING LI

Responsive Social Smile: A Machine Learning based Multimodal Behavior Assessment Framework towards Early Stage Autism Screening

昆山杜克大学
DUKE KUNSHAN UNIVERSITY