

Investigation of mental state estimation to support learning
based on behavior measurements

学習支援のための行動指標を利用した心的状態推定手法の検討

WANG GUAN YUN
王 冠 云

Graduate School of Information Sciences
Tohoku University

November, 2023

Abstract

In an e-learning environment, it is difficult for teachers to track students' engagement in learning or detect whether they need help. The thesis estimates two mental states of students when they are learning with an intelligent tutoring system(ITS): the engagement state and the help-seeking state. We recruited 13 participants from Japan and 22 participants from Taiwan and asked them to solve a problem used in the International Olympiad of Linguistics in 2018. When solving the problem, their facial video, records of clicking the hints button, and answer sheet are recorded. We extracted features, including Action Unit codes (AU, head pose, and gaze with open-source software, OpenFace. Those features consisted of nine types of feature sets: Basic AUs, Head Pose, Co-occurring AUs, Gaze feature set, etc. The classifications of mental states are trained by lightGBM and SVM models. We evaluate the models by AUC, F1 score, and accuracy. The results suggest that facial features effectively estimate the engagement and help-seeking states. Furthermore, the performance of lightGBM is better than SVM. We used SHAP (Shapley Additive exPlanations) value to calculate the importance of facial features. The important features to estimate engagement state focus on the upper face, but the features from the lower part of the face are more important to predict the help-seeking state. The results from Japanese and Taiwanese data are similar. The intra-person learning and inter-person learning were both conducted in this thesis. The application of this thesis is the potential to be implemented in ITS and improve students' learning performance.

Keywords: *Engagement, Help-seeking, LightGBM, OpenFace, SHAP*

Contents

Chapter 1: Introduction	1
1. Engagement and help-seeking behaviors	2
1.1 Engagement.....	2
1.2 Help-seeking Behavior.....	4
2. The potential measurements and computer vision.....	4
3. The structure of this thesis	6
Chapter 2: From an Intelligent Tutoring System to Human Judgements.....	9
1. Webpage design	9
2. Problem-solving task: Linguistic Olympiad problem.....	15
3. Users' Behavioral Data	16
4. Data Annotation of Engagement State	20
Chapter 3: Study 1: Predicting learners' engagement and help-seeking behaviors in an e-learning environment by using facial and head pose features	28
1. Introduction.....	28
2. Research Review.....	29
2.1 Estimating engagement levels using facial expressions	29
2.2 Help-seeking Behavior in Intelligent Tutoring System	30
2.3 Current Study	31
3. Methods.....	31
3.1 Participants.....	31
3.2 Materials	31
3.3 Procedures.....	33
3.4 Mental States Categorization	33
3.5 Facial Feature Extraction	34
3.6 Machine Learning Models	36
4. Results.....	37
4.1 Behavioral Results	37
4.2 Classification of the Engagement State	38
4.3 Classification of the Help-seeking State	42
5. Discussion.....	47
Chapter 4: Study 2: Cultural comparison on estimating learners' engagement and help-seeking behaviors by facial and head features.....	51
1. Introduction.....	51
2. Research Review.....	51
2.1 Mental states estimation by facial expressions	51
2.2 Cultural differences from face and their learning behavior	52

Contents

2.3	Current Study	53
3.	Methods.....	54
3.1	Participants.....	54
3.2	Materials	54
3.3	Procedures of the Experiment.....	55
3.4	Mental States Categorization	55
3.5	Feature Engineering	56
3.6	Machine Learning and SHAP analysis	57
4.	Estimating Mental State with Taiwan’s data.....	58
4.1	Behavioral results.....	59
4.2	Classification of the engagement state.....	59
4.3	Classification of the help-seeking state.....	66
5.	Comparison between Taiwan’s data and Japan’s data	73
5.1	Comparison of Behavioral Results	73
5.2	Comparison of the Classification Results	74
6.	Discussion.....	76
Chapter 5: Inter-person Learning models on Estimating the mental states		78
1.	Introduction.....	78
2.	Methods.....	78
2.1	Participants and materials	78
2.2	Feature Engineering	79
2.3	Machine Learning and SHAP analysis	80
3.	Results.....	81
3.1	Estimation on the engagement state.....	82
3.2	Estimation on the help-seeking state.....	84
3.3	Nationality Classification by Action Units	89
4.	Discussion.....	90
Chapter 6: Apply Machine Learning Methods to Questionnaire Datasets		92
1.	Introduction.....	92
2.	Methods.....	92
2.1	Participants.....	92
2.2	Independent Variables (Features).....	93
2.3	Dependent Variables (Output).....	93
2.4	Machine learning	95
3.	Results.....	96
3.1	Regression by LightGBM.....	96
3.2	SHAP Analysis.....	100
4.	Discussion.....	106

Contents

5. Conclusion	109
Chapter 7: General Discussions	110
1. Estimation of the mental states	110
2. Machine learning	112
3. Cultural differences and cross training and testing.....	113
4. Future Issues and Limitation.....	115
Chapter 8: Conclusions	117
1. Major Findings.....	117
1.1 Mental states classification by facial videos	117
1.2 Cultural differences between Japan and Taiwan	117
1.3 Machine learning on inter-person model	117
1.4 Machine learning applied on a questionnaire dataset	118
2. Concluding Remarks.....	118
Appendix.....	119
Appendix A. The Manual of annotation.....	119
Appendix B. The Results of SHAP analysis	122
B-1. Intra-person learning results of estimation of engagement states	122
B-2. Intra-person learning results of classifying help-seeking states	127
B-3. Inter-person learning results of estimation of engagement states	145
B-4. Inter-person learning results of classifying help-seeking states	158
References.....	175

Chapter 1: Introduction

Nowadays, the information and communication technology (ICT) becomes more and more important in educational fields. Artificial Intelligent plays an important role in many learning management systems, such as Moodles, to improve students' learning performance or to facilitate their learning motivation (Gasevic et al., 2016; Lerche & Kiel, 2018). The trend of implementing AI into educational systems is increasing beyond the COVID-19 generation due to the plenty of e-learning platforms and online classes established.

To make technology beneficial to human beings, the aspect from psychological is important. For example, students' mental states can be monitored by ICT in e-learning environments. Therefore, teachers can track students' learning process and improve teaching skill to motivate learners to learn and perform well even in a long-distance situation. The essential issue here is related to learning analysis and artificial intelligence. This thesis is standing on the boarder of these fields and tries to utilize the AI tools to make benefits on learning and teaching. The advantages of using AI on education are especially expected to improve students and teachers' skills by adopting to their needs (Gasevic et al., 2016; Viberg et al., 2018).

I have conducted experiments related to social robots that apply to educational fields(Wang, 2020). The results showed that, compared with high empathic robots, students preferred a neutral robot that can provides them with objective and straightforward support. The experiments recruited undergraduate students and high school students. Both groups of participants showed the similar results about their expectation of technology. Even though the design of the educational tools is important, the learners' experience and needs on it are more important than the system design. Therefore, it triggers me to explore more on the side of users, which is the side of learners, to estimate their mental states during learning

The reason to focus on the mental states is because of the blooming of online learning after COVID-19 pandemic. No matter teachers and students are e-learning adopters or not, all of them should learn how to use technologies to teach and learn, especially during the lockdown period. Unlike traditional face-to-face classes, in an e-learning environment, teachers have difficulty tracking students' mental states. Students are easily tired during online learning; for example, when taking video lectures, they will act mind wandering behaviors (Edyburn & Development, 2021; Risko et al., 2012), which is detrimental to their understanding of course content (Hong et al., 2022).

Furthermore, the risk perception, information behavior, and protection behavior during COVID-19 period influenced people's daily life. Many places, including schools, universities, shops, restaurants, and other public places were locked down because of

the pandemic. We also have conducted a survey to investigate how people handle the situation of the severe period.

Technically, this research mainly aims to estimate students' mental state so that the results can be applied on learning support systems and improve students' learning performance. The tool that first come up with our mind is web camera, which is commonly used in online meeting, and most of the personal computer and tablet are equipped with one. It is possible to utilize this tool to capture student' facial videos during learning and make the data analyzable for artificial intelligence. The techniques of computer vision application on education can be integrated into different ICTs. For example, a social robots can be equipped with a camera to analysis the interaction with users.

This thesis focuses on an AI support e-learning tool called "intelligent tutoring system (ITS)" is mainly used and discussed in educational research(Aleven & Koedinger, 2000; Tang et al., 2021). A common ITS is usually made for web users, and it can deliver adaptive guidance and instruction to learners, evaluate learners' performance, and combine with the learners' models to classify or cluster them (Mousavinasab et al., 2021). Besides, the subjects taught by ITS are various(Tang et al., 2021), such as computer engineering, science(Graesser et al., 2018), and languages(Graesser et al., 2018). Moreover, a typical ITS includes dialogue modules to provide knowledge contents(Aleven & Koedinger, 2002).

Beyond the learning with ITS, the current thesis uses non-verbal information from students' videos to identify their mental state like a real human teacher. From the aspect of a teacher, student engagement and their intention of seeking help are essential to know. Engagement is related to a student's involvement on learning and the help-seeking behavior is related to a student's obstacle on learning. Instead of monitoring students' engagement, the intention of seeking help to avoid frustration might be another critical role of e-learning. I believe that these two mental states represent the positive and negative mental states. To make an ITS detect these mental states automatically, this thesis is focusing on developing a method by using students' behaviors which videos can take.

1. Engagement and help-seeking behaviors

First of all, this part explains why this thesis focuses on engagement and help-seeking behaviors. The operational definitions are also provided in this session.

1.1 Engagement

Educational research defines engagement as behavioral, emotional, and cognitive

engagement (Fredricks et al., 2004). In their definition, behavioral engagement is related to students' conduct and on-desk behavior, which concerns students' involvement in learning, including behaviors such as effort, persistence, concentration, attention, asking questions, and contributing to class discussion. Emotional engagement refers to students' affective reactions, including interest, boredom, happiness, etc. Finally, cognitive engagement is also written in "self-regulation" by some research, and it refers to how students use their learning strategies to plan, monitor and evaluate their work when they are accomplishing their tasks.

In HCI (human-computer interaction) fields, engagement refers to user engagement, which focuses on the interactions of humans and computers, virtual agents, or mobile apps (Karimah & Hasegawa, 2022). Engagement is also commonly seen as an outcome, which indicates how well a computer system is accepted and used by users (Karimah & Hasegawa, 2022; O'Brien et al., 2022), or how often users are willing to access the system (Davenport Huyer et al., 2020). For the aspects of ITS research, most automatic engagement research focus on emotional engagement (Karimah & Hasegawa, 2022), and they are trying to classify the basic emotions based on Ekman's research, such as anger, surprise, disgust, enjoyment, fear, and sadness (Ekman et al., 1978; Kouahla et al., 2022). Engagement in some research is viewed as one of the emotional states and named flow (D'Mello et al., 2007; Mills et al., 2014), which indicates intense concentration. Research on authentic problem-solving tasks concerns cognitive engagement, and they estimated the level of students' mental investments in learning (Li et al., 2021). On the other hand, some ITS research focused on perceived engagement judged by observers (Monkaresi et al., 2017; Whitehill et al., 2014), and they were concerned about how the e-learning system can perceive their engagement.

This thesis conceptualized engagement as behavioral engagement judged by external observers such as human or automatic AI models. A machine learning method based on computer vision imitates a teacher who observes students' behaviors. However, it should be noted that it is difficult to divide the different types of engagement since their definitions are overlapped and are highly related to each other (Fredricks et al., 2004). For example, students who behave to pay attention to learning task is usually effortful and tend to be good at self-regulation, whose behaviors are seen as behaviorally engaged and cognitively engaged. Besides, previous research showed that engagement is not a stable predictor of student learning performance since many factors affect engagement, and students might use self-regulation strategies to regulate their effort to engage in learning (Li et al., 2021). This motivated this thesis to investigate other learning states by learners' interactive behavior with ITS when puzzling with the learning task.

1.2 Help-seeking Behavior

Research suggested that help-seeking behaviors are related to self-efficacy, gender difference, or social skills (Ryan et al., 1998; Ryan et al., 1997), and their academic performance can be enhanced since they ask for more help when they use online learning systems (Bartholomé et al., 2006; Broadbent, 2017; Roll et al., 2011).

Learning with an intelligent tutoring system usually happens when a student is alone and apart from a real human teacher in long-distance learning or taking video-based lectures. Keep asking questions when students face difficulties is beneficial for students learning. In addition, although research revealed that using metacognitive feedback as a hint can motivate students to seek help actively, they only provide a hint after students have an error (Roll et al., 2011). It might be a trade-off that if students are hesitant to seek help and take fewer actions since they are afraid of making mistakes, they will get fewer hints than other actively engaging students, eventually affecting their learning performance. Research showed that students with insufficient metacognitive skills might wait too long to seek help since they cannot monitor their learning self by themselves (Alevén & Koedinger, 2000).

However, the reasons preventing students from asking for help are that they are learning alone in e-learning and will not be able to or willing to verbally talk to ITS, which is probably not equipped with a natural language processing function. In those cases, an ITS containing computer vision functions should solve that problem because it can observe students learning through videos if learners trust and allow the system to take their facial video to help them monitor their learning. Besides, when students learn alone, systems log such as the behavior of clicking on buttons, typing on blanks, or the like can be recorded for learning analysis.

In this thesis, the help-seeking behavior is defined by the system. I designed a hint button that allows learners to inquire without a human teacher companion. During problem-solving, learners might have some difficulties. The hint buttons aim to fulfill learners' needs to ask for help, facilitate their learning motivation, and improve their learning performance. The ITS designed will be explained in the next chapter.

2. The potential measurements and computer vision

To make an ITS automatically provide learning support, an automatic function to detect learners' mental state is crucial. This thesis focuses on the two mental states, including the engagement and the help-seeking states. The measurements of the mental states include subjective and objective when the measurement is by self-report or equipment.

For example, for measuring engagement, there are several methods to conduct data

pre-processing. Although self-report and questionnaires are commonly used in measuring engagement, some objective techniques are proposed by research. The most widely used method for automatic detecting engagement is the method of machine or deep learning(Karimah & Hasegawa, 2022), which requires feature extraction before training models. For example, students' behaviors record, such as dialogues between students and ITS(D'Mello et al., 2007), reading pattern(Mills et al., 2014), and prosodic data from students' speech(Pellet-Rostaing et al., 2023), can be used as features to predict students' emotional states or engagement.

On top of that, facial features, including facial expression, gaze, and head pose, are revealed as applicable data(Dragon et al., 2008; Kato et al., 2022a, 2022b; Miao et al., 2022; Sato et al., 2022; Shioiri et al., 2021). Researchers also applied this approach to different learning tasks, such as taking video lectures(Kawamura & Murase, 2020; Miao et al., 2022; Son et al., 2020) or doing mental calculations (Kato et al., 2022a, 2022b). The facial features can be extracted as low-level features, including low-dimensional geometry and appearance descriptors such as head nodding or smile(Karimah & Hasegawa, 2022; Pellet-Rostaing et al., 2023). High-level features of facial features are extracted by aggregating low-level features, such as facial action units (FAUs), which are also the most used in automatic facial detection research. FAUs encode facial muscle movement by Action Unit (AU) codes(Zhi et al., 2020). Furthermore, OpenFace(Baltrusaitis et al., 2018), an open-source software, is conveniently used to analyze and extract facial features, including facial landmarks coded by AU, head pose, and gaze.

To the best of my knowledge, there is no research focusing on detecting help-seeking behavior by learners' behaviors, self-report, and facial expressions. In a real classroom, learners are able to ask questions if they want. Researcher is easy to track their inquiry behavior. However, in an e-learning environment, if the system does not provide a user-friendly environment to ask question, learners cannot ask question even if they want to do that. This thesis tried to challenge this situation. Based on the various results of estimating the engagement state, this thesis used facial video and extracted features from the face, head, and gaze to estimate learners' mental state.

The advantages of the computer vision methods allow us to develop a tool to apply in a large scale. The invasive method by using a web camera would not bother learners when they are learning. Although learners might intently behave well in front of the camera, they still finally ignore the camera when the time goes by. Especially in nowadays, many online classes are developed, and the video quality of web cameras is also upgrading. Therefore, it can be believed that the utilization of a web camera can be more benefit to education than ever before. This dissertation was conducting in the early-2020s, the generation of COVID-19. I believe that this research also has

contribution in the blooming of the e-learning after the pandemic. Furthermore, the features of the machine learning modeled not only from the facial video, we further investigate if the machine learning methods can be applied to a questionnaire data. Therefore, the application can have a variety range that contains not only about computer vision, but also about social psychological studies which are taken by questionnaire survey.

3. The structure of this thesis

This thesis focuses on the mental states estimation, and focuses on the invasive measurement of facial video taken by a web camera.

There are many results suggested that facial expression is useful to estimate learners' learning states as mentioned (Shioiri, et al, 2021; Sato, et al., 2022; Miao, et al, 2023; Kato, et al., 2022). I also have several results on using facial expression to estimate learners' mental state (Wang, et al, 2022; Wang et al, 2023).

First of all, in order to simulating an e-learning environment, I built a website as an intelligent tutoring system. The website integrated the real teaching experiences and educational theories into the system. Technically, the interactable functions allow learners to reflect during problem-solving, ask question if facing difficulties and complete the problem like a real competition. The learning task was a Linguistic Olympiad's problem since it doesn't require prior knowledge to solve the problem. Any bias of ability and skills was expected to be reduced in this research. On top of that, the function of web camera also integrated into the website and learners' behavior are recorded synchronously. The videos of learners' face are annotated by a team of labelers. The details are explained in the Chapter 2.

Secondly, the Chapter 3 is the study about the estimation of the mental states on Japanese participants. The website was confirmed that they can solve the problem smoothly with it. The machine learning models include two: Support Vector Machine (SVM) and Light Gradient Boosting Machine (LightGBM).

Furthermore, since we noticed that the facial expression has differences between different cultures, and present previous studies were focusing on single culture. This trigger me to explore the cultural differences. In Chapter 4, the study showed the results of estimating the mental states on Taiwanese participants, since the culture difference in individualism has a distance of one standard deviation between the two cultures. The study also discuss about the cultural differences and further tested on the identification of the two cultures to examine the differences on the facial expressions. In addition, the framework of the study 1 and study 2 is shown in Figure 1. The details are introduced in the following chapters.

However, the analysis in Chapter 3 and Chapter 4 were based on intra-person learning. That is, in the machine models, the data from the same person would be divided into the training and the testing data. In this thesis, we further explore the analysis on inter-person learning. In this approach, people with their data who are sampled to the training dataset would not be counted in the testing dataset. Chapter 5 conducted the inter-person learning in both Japanese and Taiwanese dataset, and their engagement state and help-seeking state were estimated.

In Chapter 6, we tried to apply the machine learning method to questionnaire datasets which collected by a previous study. The study investigate risk perception, information behaviors, and protection behaviors by questionnaire during severa COVID-19 period. The previous work used four-way ANOVA to predict the psychological variables, which cause to the limitation of limited independent variables. Therefore, in this chapter, we implemented the machine learning approach to expand more independent variables as features for machine learning.

Last but not least, the general discussion and the conclusion are given in the Chapter 7. The facial videos are useful for estimating the mental state, including the engagement state and the help-seeking state. The features extracted by facial videos, including Action Units, head pose, and gaze, are useful to build a machine learning model. However, the generalization issue still remains, for example, the people from western country and other mental states still needs to be explored in the future.

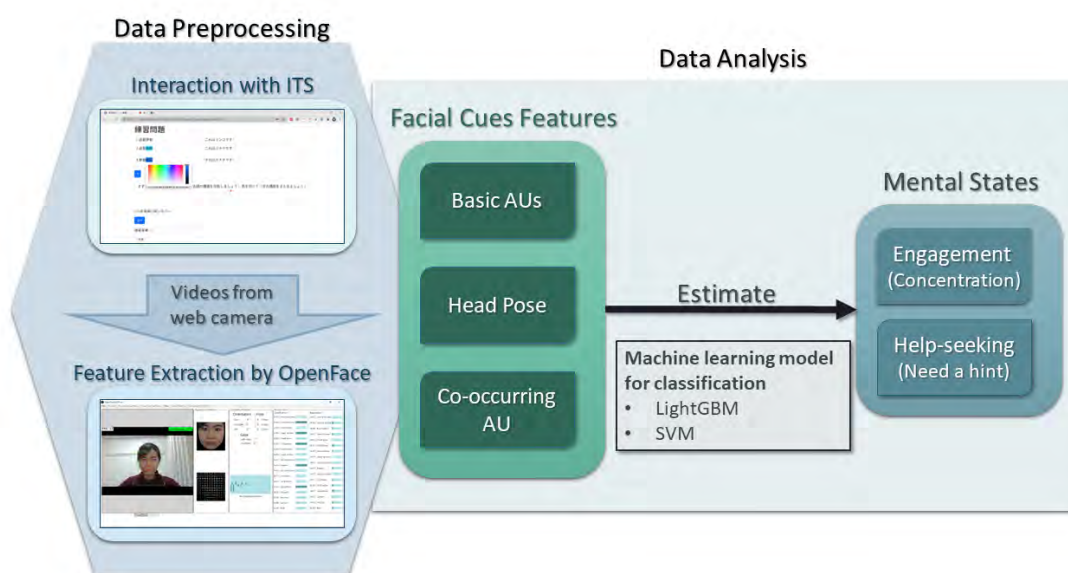


Figure 1 Framework of the current data analysis. While the participants interacted with our website, the web camera took their facial expressions, and the videos were saved on the local computer. After the experiment, we extract features by OpenFace 2.0, including Action Units (AUs) and head pose data. In data analysis, we set up three kinds of feature sets: basic AUs feature

Chapter 1

set, head pose feature set, and co-occurring AU feature set. The engagement state and help-seeking state are estimated by LightBGM and SVM models.

Chapter 2: From an Intelligent Tutoring System to Human

Judgements

1. Webpage design

The current research used a website to simulate an intelligent tutoring system as our experimental environment, which can be accessed on the Internet(Wang, Nagata, et al., 2023). On the website, several interactive functions help students complete the problem-solving task of linguistics. Participants can draw different colors on the words of sentences to categorize them, and they can use the blank table to summarize their analysis. Participants can utilize these functions to help them solve the problem smoothly.

In addition, the website also provides several hint buttons for the participants. Those hint buttons provide hints based on the principles.

In addition, it has been shown that hints provided by ITS are usually beneficial for students(Aleven et al., 2016b). In the current study, we proposed a website as a simulating ITS prototype, which can monitor learners' engagement and provide interactions with feedback hints for learners. In an e-learning environment, students might face questions about learning content, boredom with online classes, or other things that frustrate learning. Research also indicated that making students perceive effectiveness during online learning will reduce cognitive fatigue and mind-wandered behavior(Hong et al., 2022). Considering the problem, using digital technology with interaction should be a solution to let students engage more in learning(Ha & Im, 2020).

An intelligent tutoring system commonly provides feedback and hints to help students. Specifically, ITS can provide two categories of hints: bottom-out hints and principle-based hints. Bottom-out hints almost point out the final answer to learners, whereas principle-based ones only tell learners the principles related to the problem. Both hints won't tell learners the definitive answer. For analysis, to what extent the student understands the learning content relies on learners' "self-explanation," which indicates the process that learners can obtain the skills and knowledge during learning by explaining the principles of content in their mind(Aleven et al., 2016b). In addition, the interaction between students and the system creates an environment that enables students to clearly understand their learning process will help scaffold their mindset of the learning contents rather than passively waiting for the teacher's instruction(Jonassen et al., 1998).

Participants can click the hint buttons if they need help. After the button is clicked, the hints will be shown on the website unless the button is clicked again. Participants can solve the question considering hints. They are asked to finish ten questions on the website with ten blanks they need to fill in. After they finish the problem set, they can submit their answers via the website.

The website is mainly written in a JavaScript library called “jsPsych” (de Leeuw, 2015), which is user-friendly for psychological experiments on web pages. Their time completing the problem, logs of clicking the buttons, and answers are recorded. The website can be shown in Japanese or Traditional Chinese depending on the participant’s language. The website can be accessed on <https://oooo2552.github.io/linguisticpuzzle/>.

<h3>Instructions for Language Puzzle Experiment with Facial Expression Analysis</h3> <p>Hello. Thank you for participating in this experiment.</p> <p>My name is Ganyun Wang and I am from the Graduate School of Information Sciences at Tohoku University.</p> <p>Today's experiment involves solving a language puzzle according to the instructions on the website. This problem is the second problem of the 2018 International Linguistics Olympiad (IOL).</p> <p>This experiment is expected to take about an hour, but there is no time limit. Please use the website to solve the problem to <u>the best of your ability.</u></p> <p>One of the purposes of this experiment is to understand human facial expressions during learning, so <u>we will be using a webcam to take a picture of your face during the experiment. Please be aware of this. Also, during the experiment, please try not to move your head too much.</u></p> <p>If there are no questions, let's begin the experiment. Thank you.</p> <p><input type="button" value="English(Demo)"/></p> <p>Note: The English version is not complete. It is only for demonstration. (The English version was translated by ChatGPT)</p>	<h3>実験のインストラクション</h3> <p>こんにちは。ご参加ありがとうございます。</p> <p>私は東北大学情報科学研究科の王冠云（ワングアンユン）と申します。</p> <p>本日の実験はウェブサイトの指示に従って、言語パズルを解いてもらうものになっています。この問題は、2018年国際言語学オリンピック（略称IOL）の2番目の問題です。</p> <p>本実験は1時間ほどを想定していますが、特に時間制限はありません。全力でウェブサイトを利用して、問題を解いてください。</p> <p>本実験の目的の一つは、学習する人間の顔表情を理解することなので、<u>実験の過程でウェブカメラを用いて、あなたの顔を撮影します。</u>ご了承ください。また、実験中、なるべく<u>頭の位置を動かさないようご協力お願いします。</u></p> <p>問題がなければ、実験を始めましょう。よろしくお願いたします。</p> <p><input type="button" value="日本語"/></p>	<h3>實驗說明</h3> <p>您好，非常感謝您的參與。</p> <p>我是就讀東北大學資訊科學研究科的王冠云。</p> <p>這個實驗需要您根據網頁的指示，進行一個語言學的解謎。實驗所使用的謎題的出處來自於2018年國際語言學奧林匹亞競賽(簡稱IOL)個人賽的第二大題。</p> <p>雖然這個實驗預計將花費您1小時的時間，但我們並沒有設定時間限制。請盡全力的好好利用這個網站，<u>努力解決問題吧！</u></p> <p>此外，本實驗為了瞭解學習時學習者的臉部表情，<u>在實驗的過程中，將使用您的網路攝影機拍攝您的臉。</u>請您理解，我們不會將您的臉公開發表，僅作為研究分析用途。另外，再麻煩您於實驗中，<u>盡量不要大動作移動頭部的位子。</u></p> <p>如果沒有問題的話，就讓我們開始實驗吧！請多多指教！</p> <p><input type="button" value="繁體中文"/></p>
--	---	--

Fig. 1 The introduction of the experimental website. The Japanese and Chinese version are made for the experiment. The English version is not completed and only for demonstration. The participant will choose the language version to fit their mother tongue.

Welcome to Linguistic Puzzles

What is Linguistic Puzzles?

Linguistic puzzles are problems used in the "Linguistics Olympiad," an international science competition. These problems are designed to be enjoyable for those who like solving mysteries and are centered around the theme of language.

What are the characteristics of the problems?

The problems are akin to analyses conducted in actual linguistic research, where one uncovers hidden patterns from "**language data seen for the first time.**" Similar to solving mysteries or puzzles, they require analytical skills, information processing abilities, logical thinking, and the power to experiment and learn from mistakes. In this regard, these abilities are thought to be similar to the fundamental skills in mathematics and programming.

A noteworthy feature of the Linguistics Olympiad problems is their **self-contained nature. All the information necessary to write an answer is concealed within the problems.** Thus, "memorization" is not required. Unlike language exams or speech competitions, the proficiency to speak or write in English or other specific languages is not demanded. Instead, linguistic knowledge contributes to constructing the crucial "**relational worldview**" that is significant for language analysis.

Source: [Japan Linguistics Olympiad](#)

Start!

Fig. 2 The Introduction of Linguistic Puzzles. Before the experiment, the participants can read the introduction of the problem they are going to solve. The core information in this page wants to convey is that to solve the problem does not need any prior language and the answer can be analyzed by the problem itself.

Participants ID(eg.001)

Name(English)

Age

Gender(female/male/others)

Continue

Fig. 3 The participants will type their basic information on the website. The "ID" and the "Name" are only used for corresponding to the video files and the log files. In the back platform, their personal information would become random numbers and disconnected.

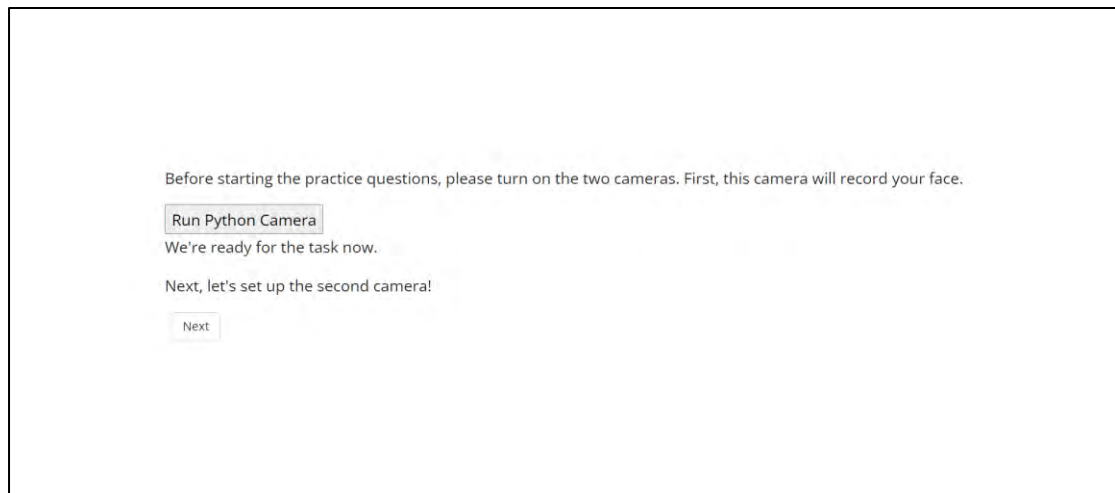


Fig. 4 The participants will turn on the web camera by themselves. This process make sure that all participants are knowing that their faces are recorded by the web camera. The camera can be activated by the website since I use Flask to build a Python API to connect to the web browser. It should be noted that, although the camera is activated by the website, the camera is run by the local computer and the video files are also saved in local.

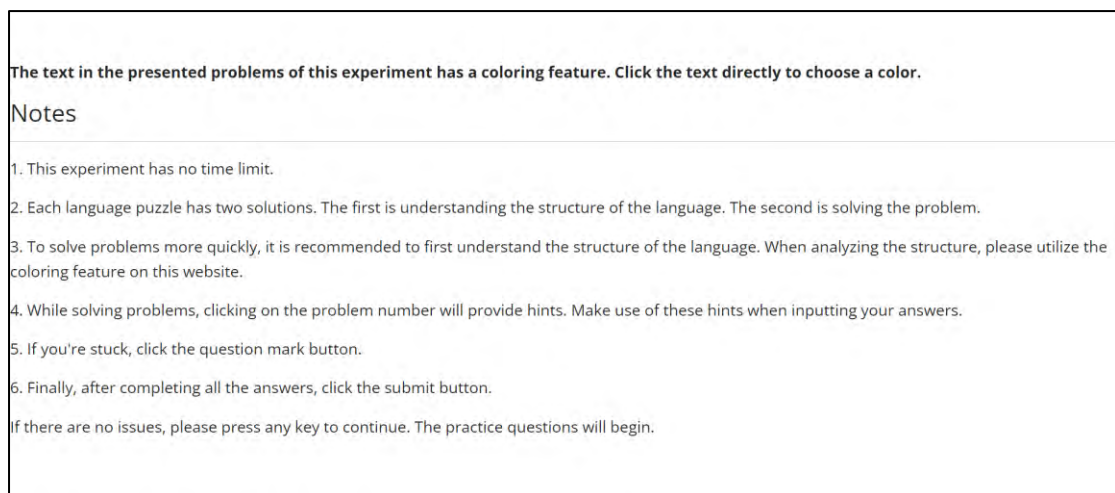


Fig. 5 The instruction of the experiment. Before the participants go to the problem-solving section, they will read the instructions first. This introduces about the rules and remind the participants that there is no time limit in the experiment. To make sure the participant read the instruction, this part

the button is removed, and the participants need to press the keyboard instead of clicking.

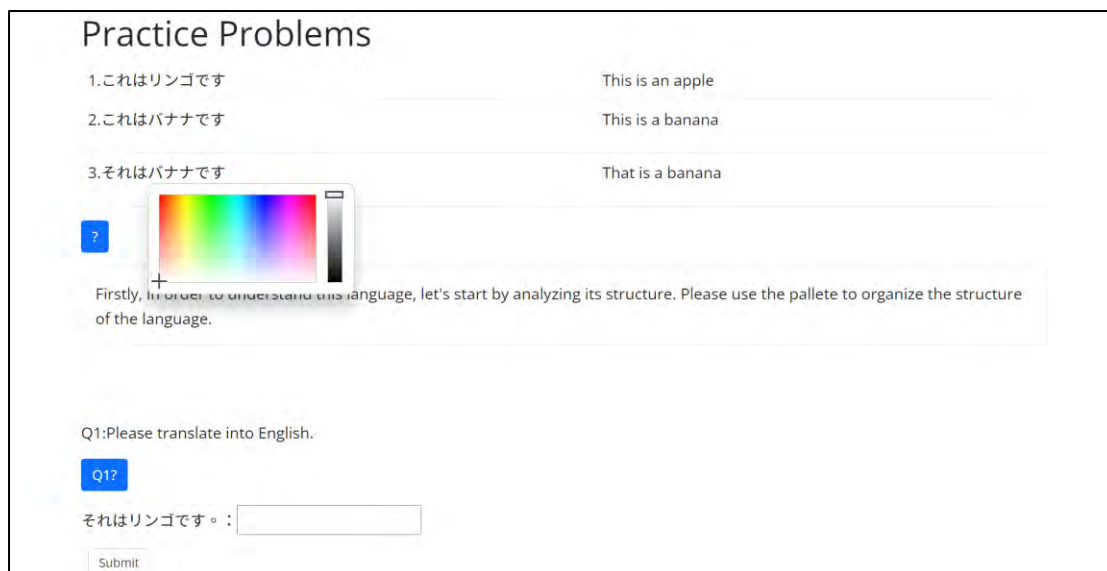


Fig. 6 The practice problem on the website. This is a simple problem to analysis the sentences “This is an apple.”, “This is a banana.”, and “That is a banana.” According to these three sentences and the translation, we can infer how to say, “That is an apple.” If the participant feel unsure or need help, they can click the blue buttons, which are hint buttons. They can also highlight the words to help them do the analysis. This practice problem is to let the participants get familiar with the linguistics puzzles and the interface of the website.

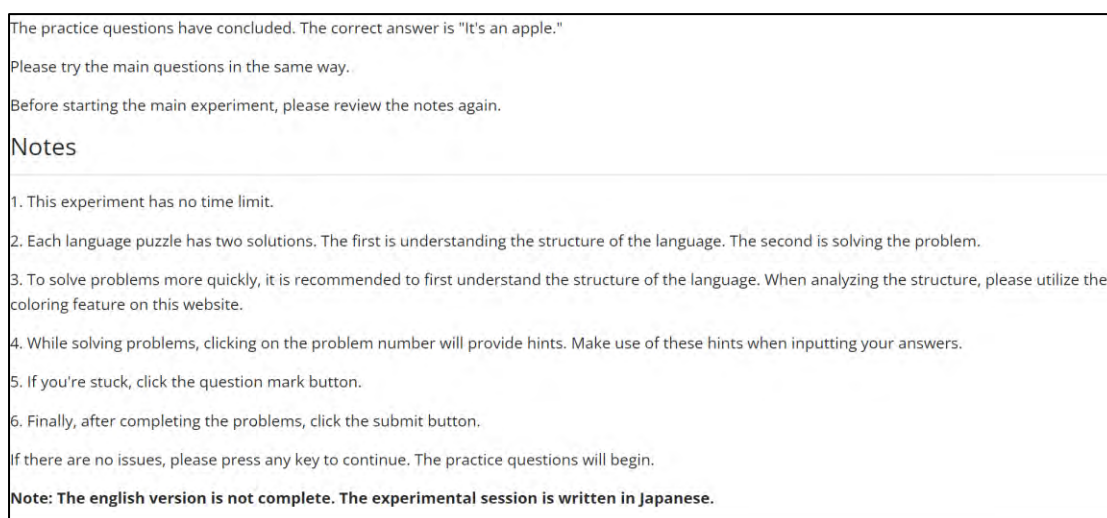


Fig. 7 After the participants finish the practice problem, they can continue to the experimental problem. The instruction will be showed again, they can take a rest if they need.

問題2.

以下にハクン語の文章とその和訳がある:

- 1. ɲa ka kɾ ne - 私は行くか?
- 2. nɾ ʒip tu? ne - あなたは眠ったか?
- 3. ɲabə ati lapkʰi tɾ? ne - 私は彼を見たか?
- 4. nirum ki ɾa? - 私達はあなた達を知っているか?
- 5. nɾba ɲa ɾa? - 彼らはあなた達を見るか?
- 6. tarum kəmə nɾ lan tʰu ne - 彼らはあなたを殴ったか?
- 7. nu?rum kəmə at lapkʰi kan ne - あなた達は彼を見るか?
- 8. nɾba ati cʰam tu? ne - あなたは彼を知っていたか?
- 9. tarum kəmə nirum lapkʰi ri ne - 彼らは私達を見るか?
- 10. ati kəmə ɲa lapkʰi tʰɾ ne - 彼は私を見たか?



ヒント1
2番と8番の文の中で、同じところを探して着色してください。

ヒント2
そして、5番を見て、2番と8番の文と比べて、同じところを着色してください。

ヒント3 **ヒント4** **ヒント5**
7番と9番の文も、他の文と比べて、同じ部分に同じ色を付けてください。

ヒント6 **ヒント7** **ヒント8** **ヒント9**

	私を	私達を	あなたを	あなた達を	彼を	彼らを
私は						
私達は						
あなたは						
あなた達は						
彼は						
彼らは						

この表は?

表の書き方のヒント

知っている主語と目的語は入力した後、表の他の枠も同じ規則で埋めてみてください。

質問:

(a)日本語に訳しなさい:

- 1. nɾ ʒip ku ne **Q1ヒント**
- 2. ati kəmə nirum lapkʰi tʰi ne **Q2ヒント**
- 3. tarum kəmə nu?rum cʰam ran ne **Q3ヒント**
- 4. nirum kəmə tarum lan ki ne **Q4ヒント**
- 5. nirum kəmə nɾ cʰam ti? ne **Q5ヒント**
- 6. nirum ka ti? ne **Q6ヒント**

(b)ハクン語に訳しなさい:

- 7. 私はあなたを殴ったか? **Q7ヒント**
- 8. 彼らは私を見たか? **Q8ヒント**

質問8の文も、上記の例文と同じ過去形ですが、主語と目的語については例文を利用して考えましょう。

- 9. 彼はあなたを知っているか? **Q9ヒント**
- 10. あなた達は眠るか? **Q10ヒント**

必ずデータを保存してください。

Submit

Fig. 8 In the experimental problem, the whole problem set will be shown, and the hint buttons can be clicked if the participants need. They can also highlight the word and utilize the table to facilitate their problem-solving process. The task will be further explained in the next section.

2. Problem-solving task: Linguistic Olympiad problem

The problem-solving task cited from the second problem in the International Olympiad of Linguistics in 2018. I used the actual problem set used in the competition. The problem set has several language versions. Therefore, the narratives of questions and the answers are not controversial but fair to solvers who speak different mother tongues. Besides, in a real competition, five problem sets will usually be given, and the challengers should finish all the problem sets within 6 hours. Considering the experimental settings, it can be assumed that our participants could complete one problem within 72 minutes.

I have 3-year experience training students to participate in the Linguistic Olympiad exam. She also has experience teaching a national representative team from Taiwan and Hong Kong. The problem of Linguistic Olympiad exam is designed for 13-18 high school students. Although solving this problem doesn't need prior knowledge, linguistic knowledge still helps solve the problems quicker. As for the experiment, an actual problem from the Linguistic Olympiad is used, and some hints to help our participants solve the problem based on our teaching experiences and linguistic knowledge are also provided on the website. In other words, our hints are based on principles expected to help students learn better. Besides, even if our participants are undergraduate students, they still might struggle to solve the problem since they are beginners.

第十六回国際言語学オリンピック (2018)。
個人戦 問題

2

問題2(20点)。以下にハクン語の文章とその和訳がある:

1. $\eta a ka kx ne$ — 私は行くか?
2. $nx \zeta ip tu? ne$ — あなたは眠ったか?
3. $\eta ab\bar{e} ati lapk^h i t\bar{y}? ne$ — 私は彼を見たか?
4. $nirum k\bar{e}m\bar{e} nu?rum c^h am ki ne$ — 私達はあなた達を知っているか?
5. $n\bar{y}b\bar{e} \eta a lapk^h i r\bar{y} ne$ — あなたは私を見るか?
6. $tarum k\bar{e}m\bar{e} nx lan t^h u ne$ — 彼らはあなたを殴ったか?
7. $nu?rum k\bar{e}m\bar{e} ati lapk^h i kan ne$ — あなた達は彼を見るか?
8. $n\bar{y}b\bar{e} ati c^h am tu? ne$ — あなたは彼を知っていたか?
9. $tarum k\bar{e}m\bar{e} nirum lapk^h i ri ne$ — 彼らは私達を見るか?
10. $ati k\bar{e}m\bar{e} \eta a lapk^h i t^h y ne$ — 彼は私を見たか?

(a) 日本語に訳しなさい:

1. $nx \zeta ip ku ne$
2. $ati k\bar{e}m\bar{e} nirum lapk^h i t^h i ne$
3. $tarum k\bar{e}m\bar{e} nu?rum c^h am ran ne$
4. $nirum k\bar{e}m\bar{e} tarum lan ki ne$
5. $nirum k\bar{e}m\bar{e} nx c^h am ti? ne$
6. $nirum ka ti? ne$

(b) ハクン語に訳しなさい:

7. 私はあなたを殴ったか?
8. 彼らは私を見たか?
9. 彼はあなたを知っているか?
10. あなた達は眠るか?

△ ハクン語はシナ・チベット語族のサル語派に属する。インドの東端とその近辺に位置するミャンマー国内のいくつかの地区で約10,000人が使用している。

\bar{e} と \bar{y} は母音である。 $c^h, k^h, \eta, t^h, \zeta, ?$ は子音である。

—Peter Arkadiev

Fig. 9 The original problem sheet of the second problem in the International Olympiad of Linguistics in 2018

3. Users' Behavioral Data

Due to technical problems, we deleted the data of 4 Japanese participants and 1 Taiwanese participant. In sum, the total numbers of participants, which allows us to

calculate their score and completion time, are 9 Japanese and 21 Taiwanese participants. We use a *t*-test to compare these two groups of participants and Pearson's correlation to examine the correlation between behaviors. The significant level is $\alpha=0.05$.

The average time to complete the problem was about 50 minutes (the longest: 87 minutes; the fastest: 18 minutes). The completion time of the linguistic problem task is not significantly different ($t(28)=0.80, p=0.40$) from Japan's samples ($M=46, SD=16.07$) and Taiwan's samples ($M=50$ min, $SD=18.51$).

The score of all correct would be 20 points consisting of 10 questions with 2 points. The average score is 14 points (the highest: 19; the lowest: 9). The score is not significantly different ($t(28)=0.89, p=0.38$) from Japan's samples ($M=14.88, SD=2.08$) and Taiwan's samples ($M=14.14, SD=2.05$).

The average clicks of the hint buttons were 35.53 (the most: 78 times; the least: 7 times). The score is not significantly different ($t(28)=0.12, p=0.89$) from Japan's samples ($M=34.77, SD=20.12$) and Taiwan's samples ($M=35.86, SD=20.84$).

No significant correlation was found between the score and completion time ($r=0.28, t(28)=1.51, p=0.41$), and no significant correlation was found between the score and clicking times ($r=0.30, t(28)=1.65, p=0.11$). The correlation plot is shown in Fig. 10 and Fig. 11. However, there is a significant correlation between the completion time and the clicks of the hint buttons ($r=0.53, t(28)=3.21, p < .05$). The correlation plot is shown in Fig. 12.

This indicates that the hints we designed can help students to some extent. The results suggest that the much time the learners spend on the website, the more times they will click on the hint buttons. Besides, the results still show a positive trend of the correlation the significance was not showed might due to the sample size of the participants.

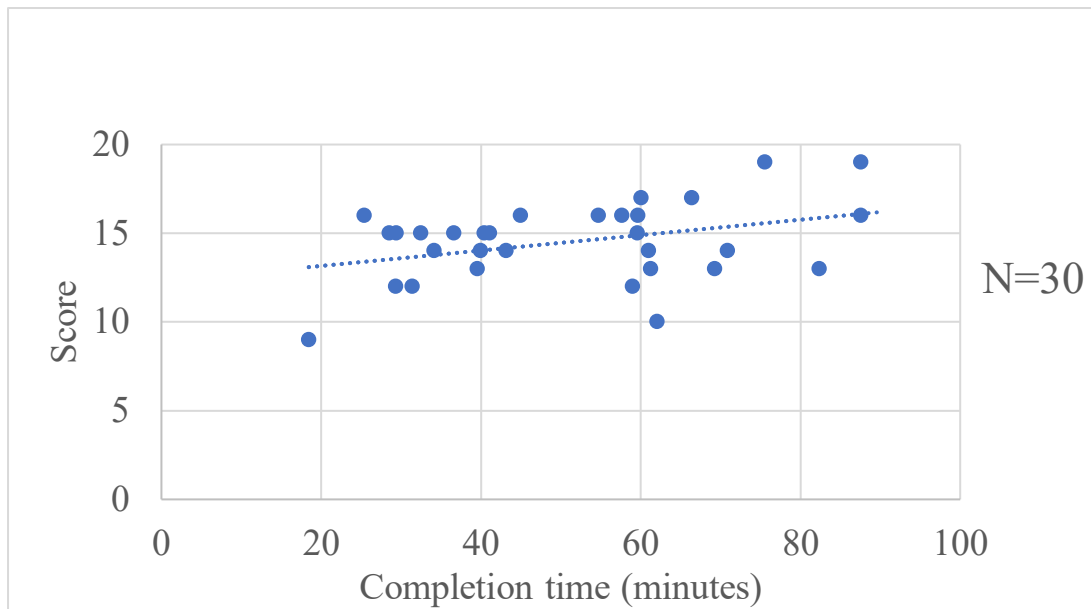


Fig. 10 Correlation between score and completion time.

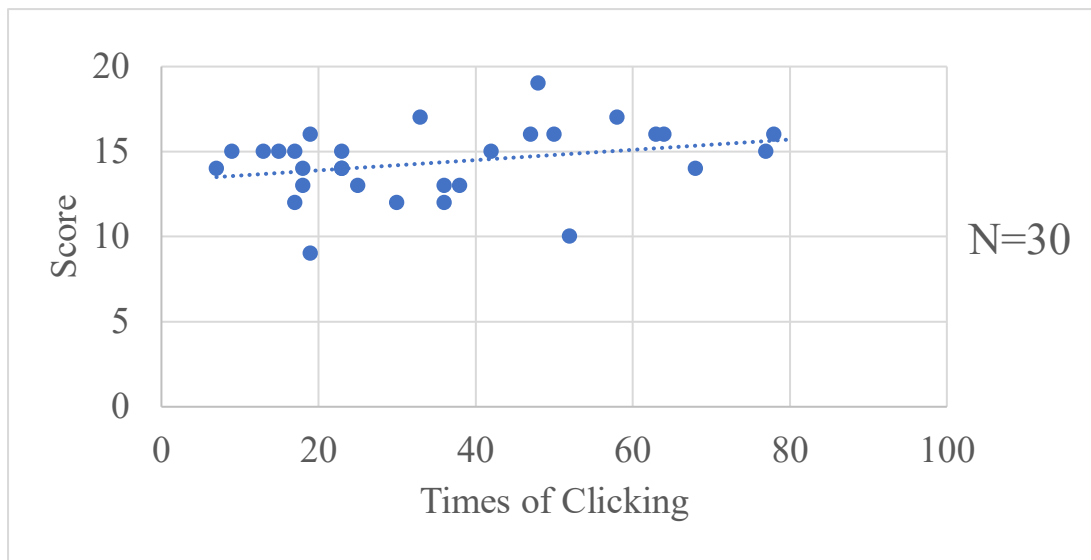


Fig. 11 Correlation between score and times of clicking the hint buttons

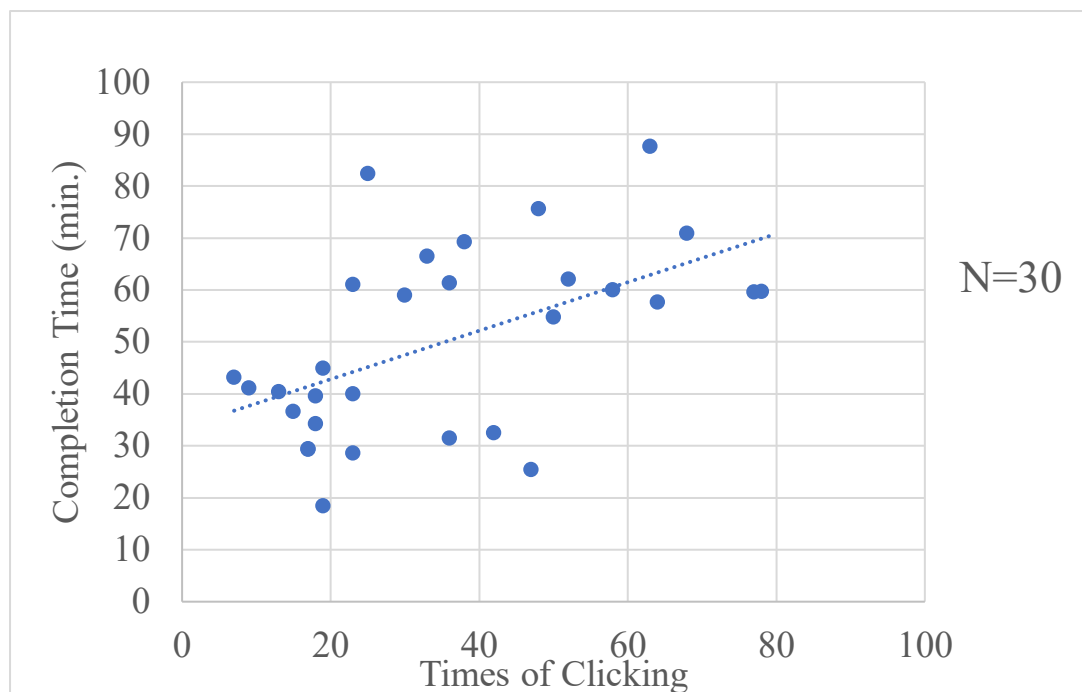


Fig. 12 Correlation between score and times of clicking the hint buttons. ($p < 0.05$)

Furthermore, the behavioral data between Taiwanese participants and Japanese participants did not have significant differences. The results showed that the website we used is no bias for different cultures' participants. The further comparison will be explained in the Chapter 4.

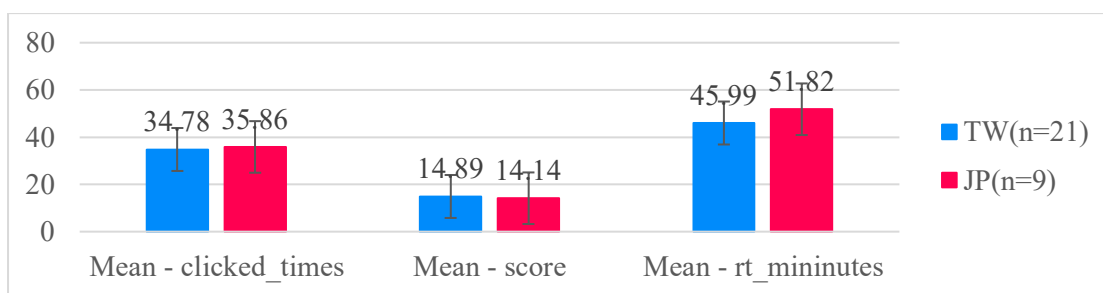


Fig. 13 The behavioral data of Taiwanese participants and Japanese participants

4. Data Annotation of Engagement State

Our rating standard is the same as the previous study, and the engagement levels are rated from 1 (Not engaged at all) to 4 (Very engaged) (Whitehill et al., 2014). The detail is as follows (cited from Whitehill et al. on page 89(Whitehill et al., 2014)):

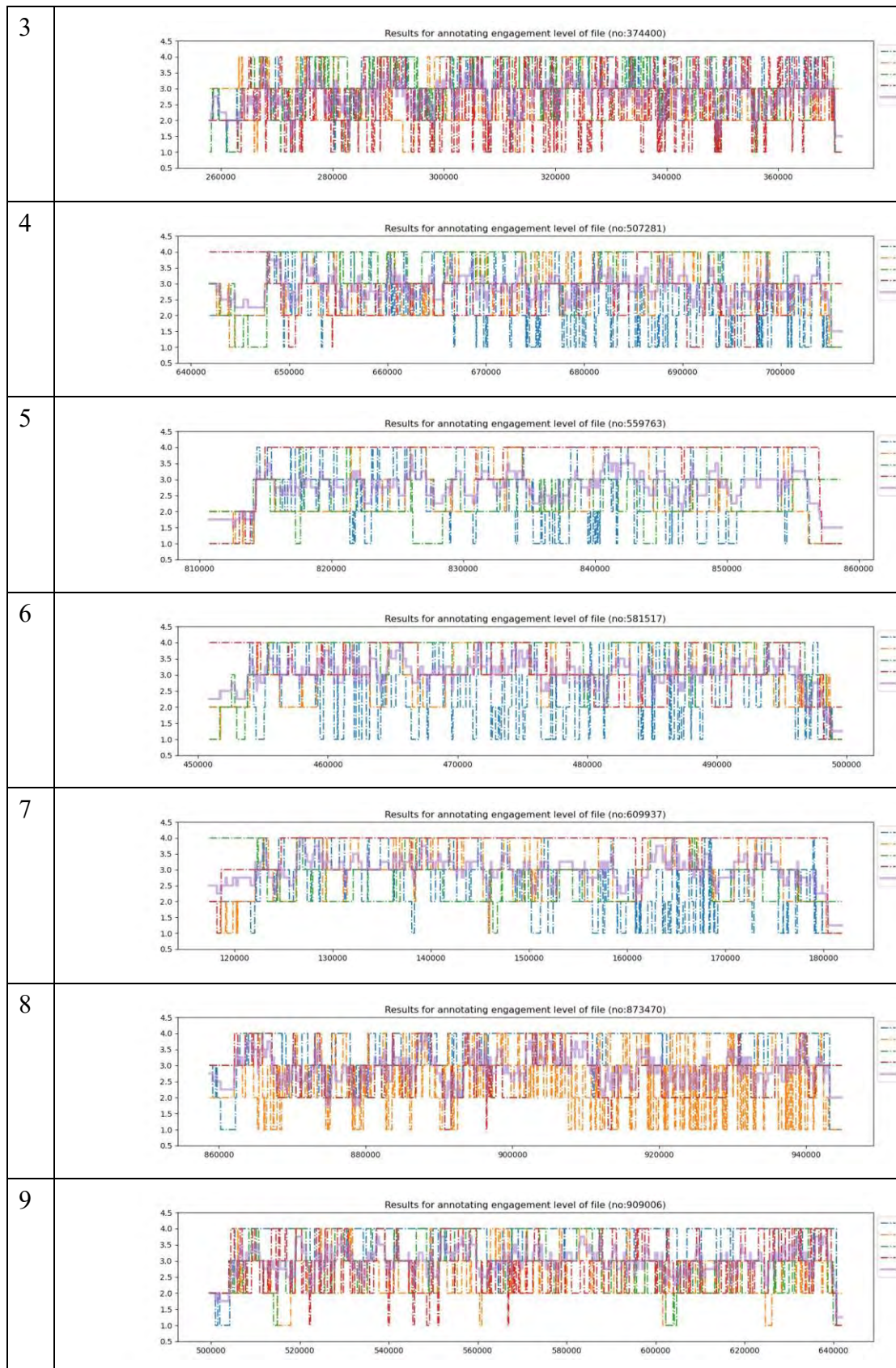
- 1: Not engaged at all – e.g., looking away from the computer and obviously not thinking about the task, eyes completely closed.
- 2: Nominally engaged – e.g., eyes barely open, clearly not “into” the task.
- 3: Engaged in the task – a student requires no admonition to “stay on task.”
- 4: Very engaged – a student could be “commended” for his/her level of engagement in the task.

Labelers were instructed to label the videos with “How engaged does the participant *appear to be*.” Our labelers have teaching experiences ranging from 0 to more than six years, including tutoring as a home teacher, online foreign language teacher, and in-person small class teacher. They labeled the participants as an external observer, and they all followed the manual (Appendix A.).

The details of all participant’s annotation results are showed as following. The participants of the experiment of this research are 13 Japan’s students and 21 Taiwan’s students. However, the 4 of Japan’s students did not complete all experiment task, and 1 of Japan’s student’s video was broken, so their data was not counted in the analysis. The details of participants will introduce in Chapter 3 and 4.

Table 1 The results of the Japan’s participants (n=12)

#	Graph
1	<p>Results for annotating engagement level of file (no:105449)</p>
2	<p>Results for annotating engagement level of file (no:176257)</p>



Chapter 2

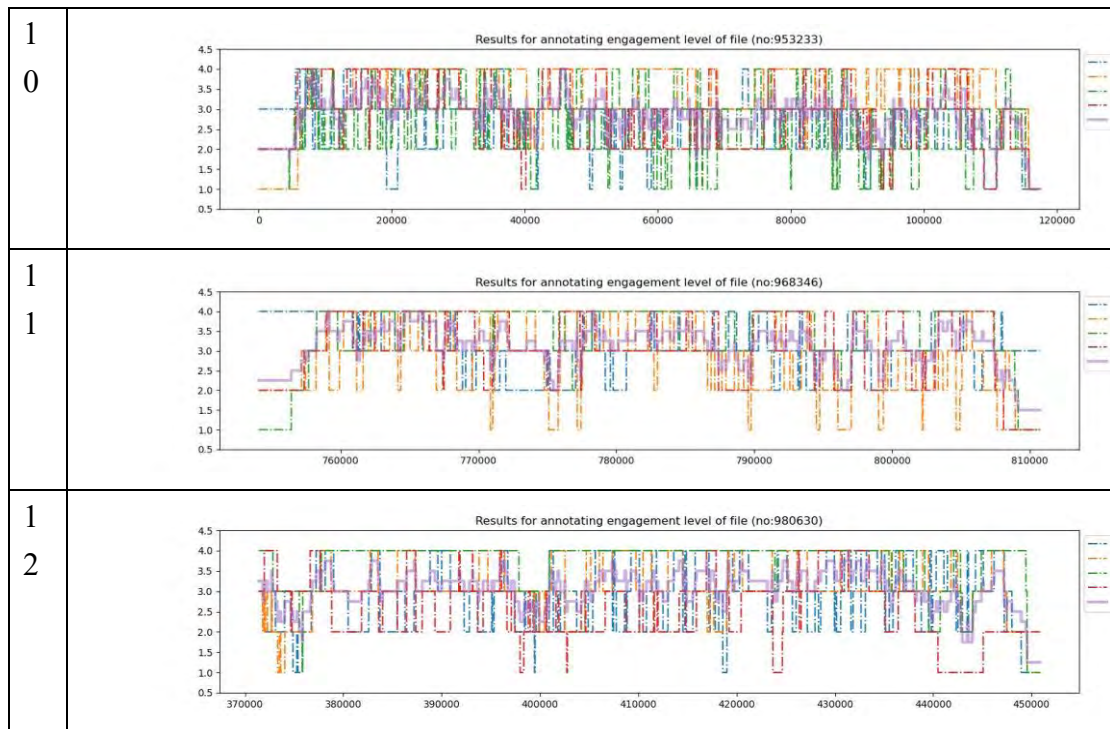
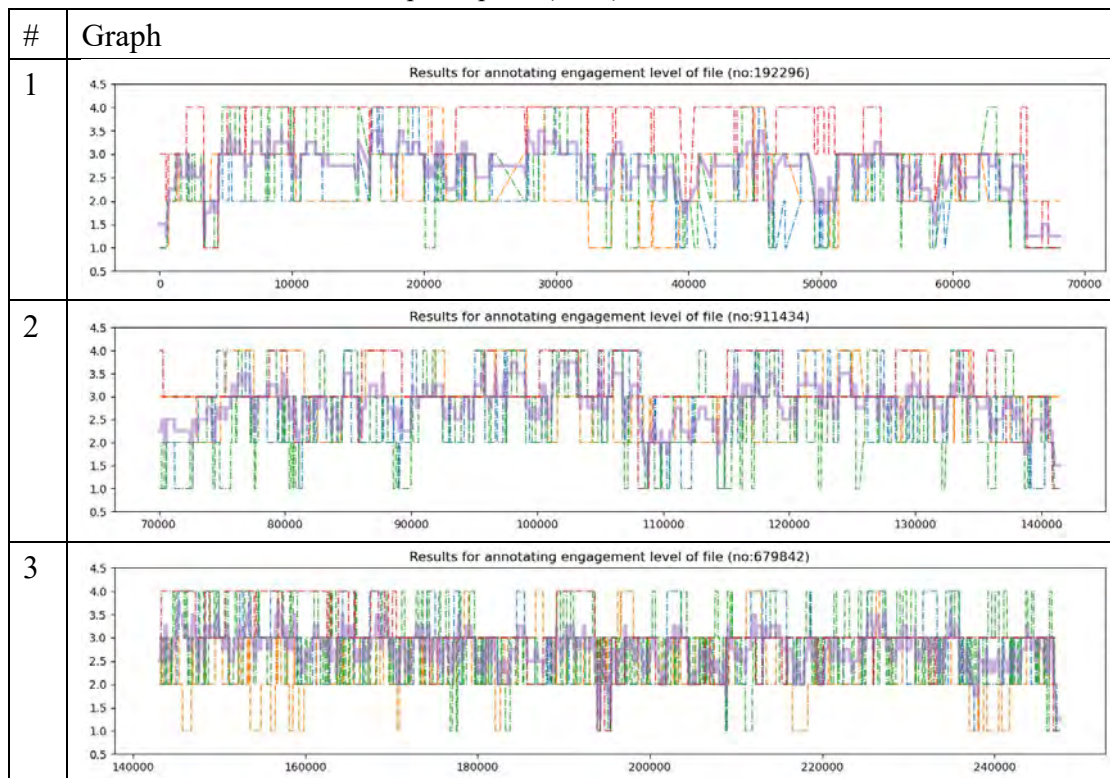
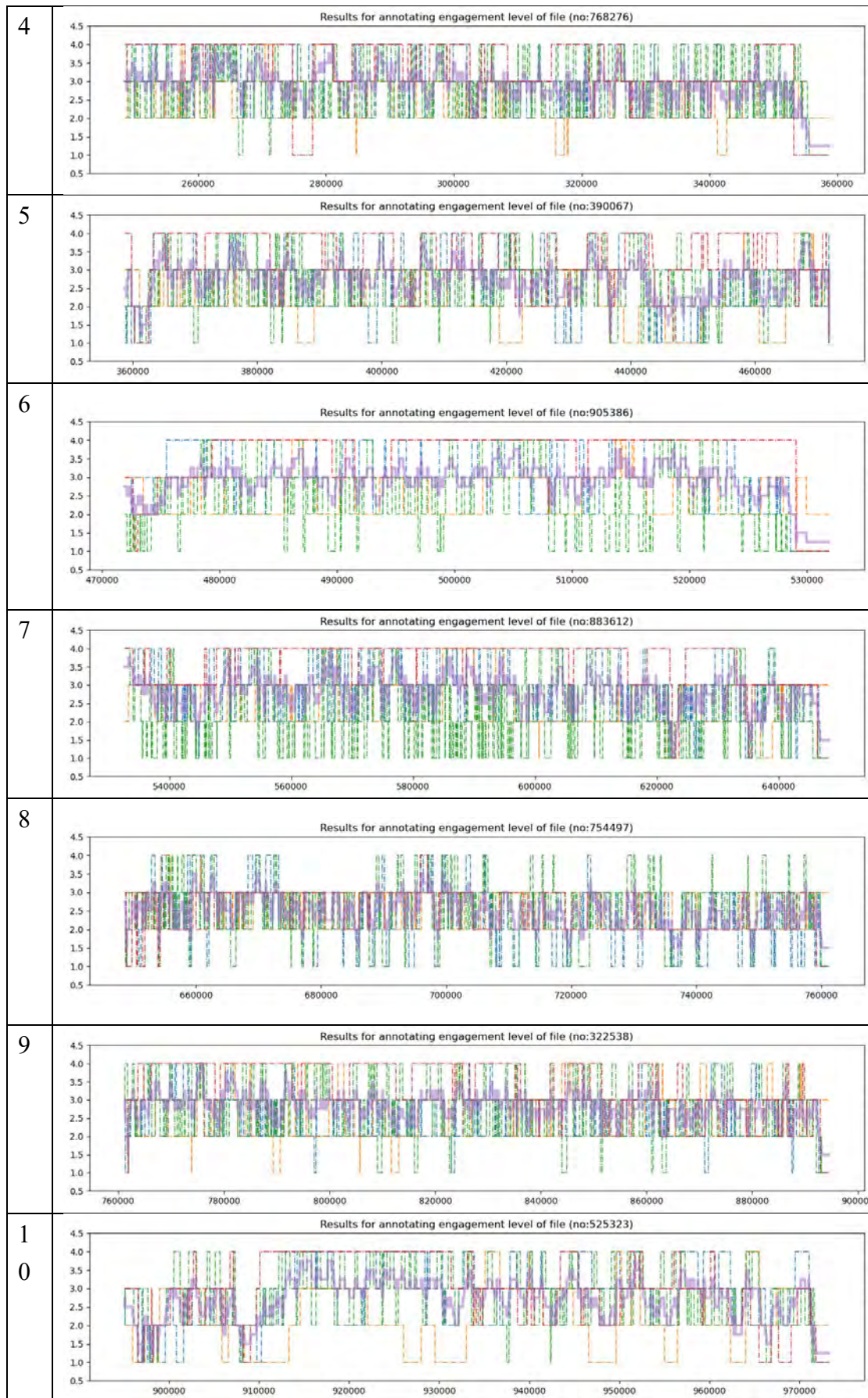
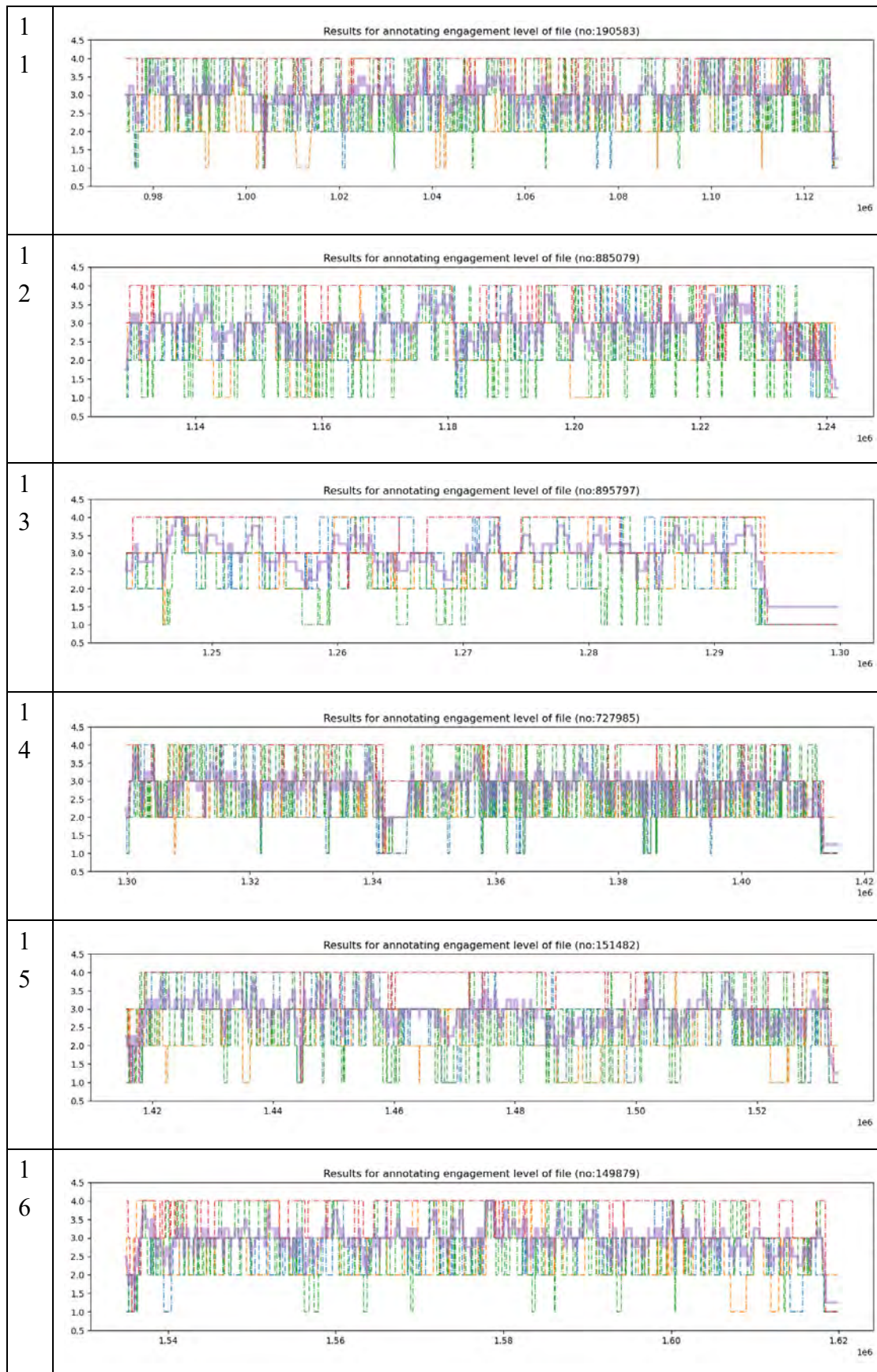
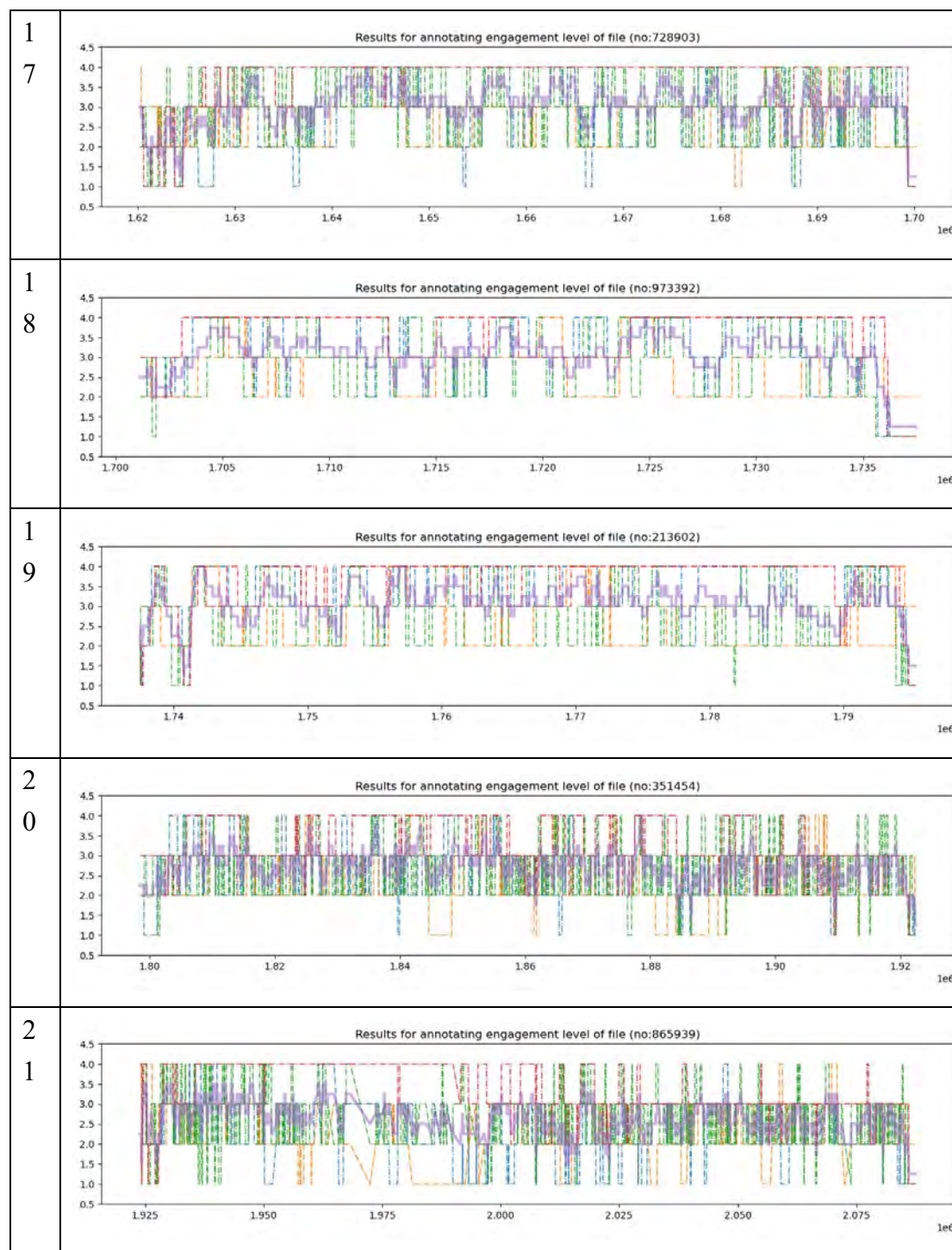


Table 2 The results of the Taiwan's participants (n=21)









The overall annotation results are shown in Fig. 14 and Fig. 15. Every participants' engagement are average by different labelers of annotation. Besides, the values of kappa were 0.21 for Japan's participants and 0.23 for Taiwan's participants, which both were fair agreement.

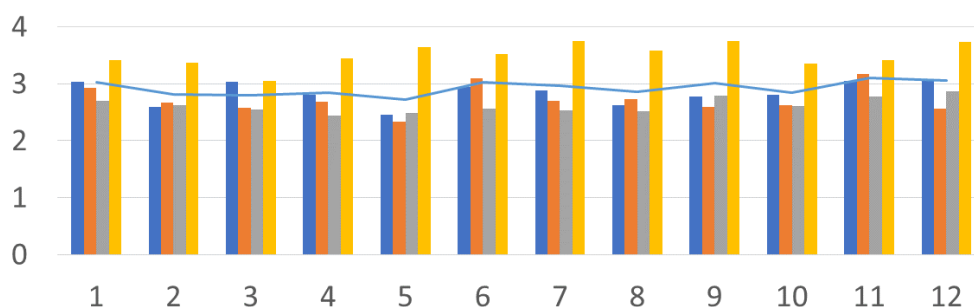


Fig. 14 The annotation results of Japanese participants. The horizontal axis showed the 12 participants, the vertical axis showed the levels of engagement, and the color of the bars showed different labelers.

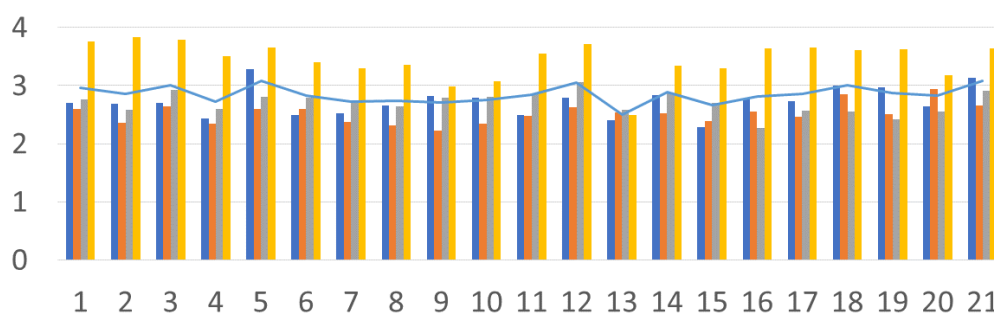


Fig. 15 The annotation results of Taiwanese participants. The horizontal axis showed the 21 participants, the vertical axis showed the levels of engagement, and the color of the bars showed different labelers.

Last but not least, the correlation of the engagement score and the problem-solving score has no significant relationship ($r=-0.1$, $t(28)=-0.58$, $p=0.71$). The results showed in This result suggest that the learning performance is not highly correlated to the engagement. One of the reason is that the engagement score are the mean of all videos of a certain participant. However, a person might adjust their engagement resources during learning to save the cognitive loading. That is, their intensity of engagement might be changed during time pass by. The result motivated to us that the contribution of this thesis is to find out the mental state of the student. As this thesis aims to estimate learners' engagement states and help-seeking states, the little scale of learners' learning behaviors but large scale of data training are expected to explore the issue that what is the crucial key to influence learners' learning performance. For the aspect of the supporting learning system, we would not average the data, but we can use narrower

Chapter 3: Study 1: Predicting learners' engagement and help-seeking behaviors in an e-learning environment by using facial and head pose features

Keywords: Machine learning, Facial expression, Hint processing, Action Units, Engagement, Help-seeking

1. Introduction

During the COVID-19 pandemic, many schools and universities developed online and on-demand lectures. Regardless of whether teachers and students are e-learning adopters, all of them by now have learned how to use technologies to teach and learn when most classes are online. Unlike traditional face-to-face classes, teachers have difficulty tracking students' mental states in e-learning environments. Students often have difficulty focusing throughout a lecture, and mind wandering likely occurs when they lose attention (Edyburn & Development, 2021; Risko et al., 2012), which is detrimental to learners' understanding of lecture contents (Hong et al., 2022). Besides, unlike face-to-face classes, online classes lack interactions with teachers. While learners might still face difficulty in classes, they cannot ask questions about content as easily, which will frustrate their learning. While monitoring learners' engagement is the target in many of the studies, the intention of seeking help is another important aspect of education betterment.

Digital technologies with interaction could help students to engage more in learning (Ha & Im, 2020). An AI-support e-learning tool called an intelligent tutoring system (ITS) has been investigated in educational research (Aleven & Koedinger, 2000; Aleven & Koedinger, 2002; Tang et al., 2021) with different subjects, such as science and languages (Graesser et al., 2018). According to a systematic review, a typical ITS is equipped with artificial intelligence techniques that deliver adaptive guidance and instruction, evaluate learners, define their models, and classify or cluster learners (Mousavinasab et al., 2021). One crucial advantage of ITS is providing hints, which is known to benefit learners (Aleven et al., 2016a).

The current study aims to investigate methods that automatically predict learners' mental states of engagement and help-seeking. In face-to-face classes, teachers and students can communicate through non-verbal information, such as eye contact, body

language, and verbal information. Teachers could detect learners who have trouble understanding the learning materials from non-verbal communication and take care of them in the classroom. However, teachers' skills of detection and support do not work in e-learning with ITS. Thus, the current study aims to extract non-verbal information from learners' facial images from video recordings. Furthermore, we focus on developing a method using machine learning models to make an ITS detect mental states automatically. For this purpose, we designed a website to simulate an ITS prototype, which provides interactions by giving hints to learners. We also used it to collect data on learners' behaviors and facial information.

2. Research Review

2.1 *Estimating engagement levels using facial expressions*

Engagement is a term used with different meanings in different contexts; for example, engagement in human-computer interaction research is considered as how users engage in using technology (O'Brien & Toms, 2010). On the other hand, educational research pointed out three aspects of engagement: emotional, behavioral, or cognitive (Fredricks et al., 2004). Emotional engagement describes positive or negative reactions to education environments, and behavioral engagement is involvement in learning, such as attention, concentration, asking questions, and contributing to class discussion. Cognitive engagement is a psychological investment in learning, such as planning, monitoring, and evaluating cognition. In the education studies focusing on emotional engagement (Karimah & Hasegawa, 2022), well-established emotion models such as Ekman's basic emotions (i.e., anger, surprise, disgust, enjoyment, fear, and sadness (Ekman et al., 1978; Kouahla et al., 2022) or positive or negative valences models are used to classify emotional features to estimate engagement (Elbawab & Henriques, 2023). Some other studies evaluated behavioral engagement (Bosch & D'Mello, 2021; Monkaresi et al., 2017; Whitehill et al., 2014) and investigated learners' engagement levels by appearance. In contrast, a study on problem-solving tasks concerned cognitive engagement and estimated engagement levels given by self-report from the participants (Li et al., 2021).

Our interest is in behavioral engagement to investigate the method to estimate learners' engagement states for learning. Some research used learners' log data of involvement and interaction with course material to predict their engagement by using machine learning or deep learning methods (Ayouni et al., 2021; Hussain et al., 2018; Sghir et al., 2023; Wang, 2019). On the other hand, it has also been shown that engagement can be predicted based on facial expressions, such as extracting AUs and head pose with open-source software, OpenFace (Baltrusaitis et al., 2018), is applicable

to estimate engagement level (Betto et al., 2023; Bosch & D'Mello, 2021; Kato et al., 2022b; Li et al., 2021; Miao et al., 2023; Sato et al., 2022; Shioiri, 2022). Log data of interaction activities from learners relies on a built system to track student behavior. However, the advantage of using an appearance-based method, such as video analysis, is that it requires no special equipment other than a camera. Since only a camera is required, a wide range of applications can be expected. Therefore, this current study also attempted to build a model that used facial images to predict engagement levels.

2.2 Help-seeking Behavior in Intelligent Tutoring System

Another highly-relevant mental state when learners require help (i.e., help-seeking state), such as when learners are stuck in a thought loop, is essential to increase the learning benefit. Asking questions when students face difficulties is beneficial for learning if answers or hints are given, even from a computer. We investigate the help-seeking state in addition to the engagement state as students' academic performance is generally enhanced by asking for help in online learning (Bartholomé et al., 2006; Broadbent, 2017; Roll et al., 2011). In a real classroom, some students may hesitate to ask questions because of their low academic self-efficacy or because they feel threats from peers or teachers (Ryan et al., 1998; Ryan et al., 1997). A potential benefit of using an intelligent tutoring system (ITS) is that students usually learn alone without the presence of human teachers or peers, creating a safe and comfortable environment.

There are several methods to provide help in ITSs. A typical method is for the system to offer a hint as help after learners have an error or when they inquire about a hint (Aleven & Koedinger, 2002; Roll et al., 2011). Another approach uses conversational agents to improve learning (Graesser et al., 2017), such as an ITS with a natural language process (NLP) to provide hints and feedback when a learner verbally asks for help (Graesser, 2016). Providing a hint automatically helps learners with insufficient metacognitive skills, particularly because they might wait too long to seek help due to their lack of ability to monitor their learning process (Aleven & Koedinger, 2000). There is also an approach to using a robot to help learners with step-by-step hints during problem-solving tasks by generating conversations with learners (Wang, 2020). All the methods mentioned above determine the timing for an ITS to provide help depending on verbal conversations between the systems and users.

To detect a need for help before a verbal request, an ITS needs a function to detect the learners' mental state of help-seeking with exclusive non-verbal information. The current study examines whether ITSs utilizing computer vision tools could achieve this goal with the learners' facial videos. To the best of our knowledge, there are only limited investigations of help-seeking behavior during learning, and none was able to provide hints before an error or a request detection, except for our preliminary report at a

conference (Wang, Nagata, et al., 2023). We attempted to predict the help-seeking state before an act of inquiry, and our machine learning model can predict that with 85% accuracy. Based on the previous success, the current study is an elaborated study with a comprehensive analysis, including the engagement estimation and more feature importance analysis than the previous one.

2.3 Current Study

The purpose of the current study is to examine whether features from facial videos can be indicators of the two mental states: engagement and help-seeking. For this purpose, we recorded the participants' facial expressions when they attended our learning experiment, a linguistic puzzle chosen from the Linguistic Olympiad's problems. The experiment did not require prior knowledge to solve the problem (Amaro, 2016; Hudson & Sheldon, 2013).

In summary, the research questions of the current study are: (RQ1) Can a machine learning method estimate learners' engagement from facial features and head poses obtained by a video camera? (RQ2): Can a machine learning method predict learners' struggles or help-seeking during learning? For both questions, the results showed that the face images taken by video are useful.

3. Methods

3.1 Participants

We recruited nine students (1 female) from Tohoku University as participants. Their average age was 22.78 ± 1.20 . All of them participated voluntarily in the experiment. They do not have any Linguistics Olympiad experience, and none of them majored in linguistics and other related fields.

3.2 Materials

3.2.1 Linguistic Olympiad problem

The task was to solve the second problem set in the Japanese version of the International Olympiad of Linguistics (IOL) in 2018, which consisted of 10 questions in total. Participants read ten translation pairs between an unknown language and Japanese, and learned to deduct the corresponding terms between the two languages. All the participants are new to IOL, and their majors are not in linguistics, literature, foreign languages, or other related fields. The task provided a fairground to ensure all

participants' starting points were identical. In addition, studies using similar problems reported that learners usually enjoyed solving the problem (Wang, 2020; Wang, Nagata, et al., 2023).

3.2.2 Webpage design

The current study used a website (Fig. 17), which we made for the experiment as an intelligent tutoring system written in a JavaScript library called “jsPsych” (de Leeuw, 2015). We made a video recorder using a Python program with a web camera, which set the resolution of the videos to 640×480 pixels with a sampling rate of 20 Hz. The website recorded the completion time of the problem, logs of clicking the buttons, and answers during the experiment.

The website has several interactive functions to help learners solve the problem. Participants can use colors to categorize words in different sentences and use a table to summarize their analysis. Most importantly, the website also provides several hint buttons for the participants. Participants can click buttons for a hint when they need help. After the button is clicked, the hints will appear and stay on the website till the button is clicked again to remove it. After answering the questions, participants can submit their answers via the website.

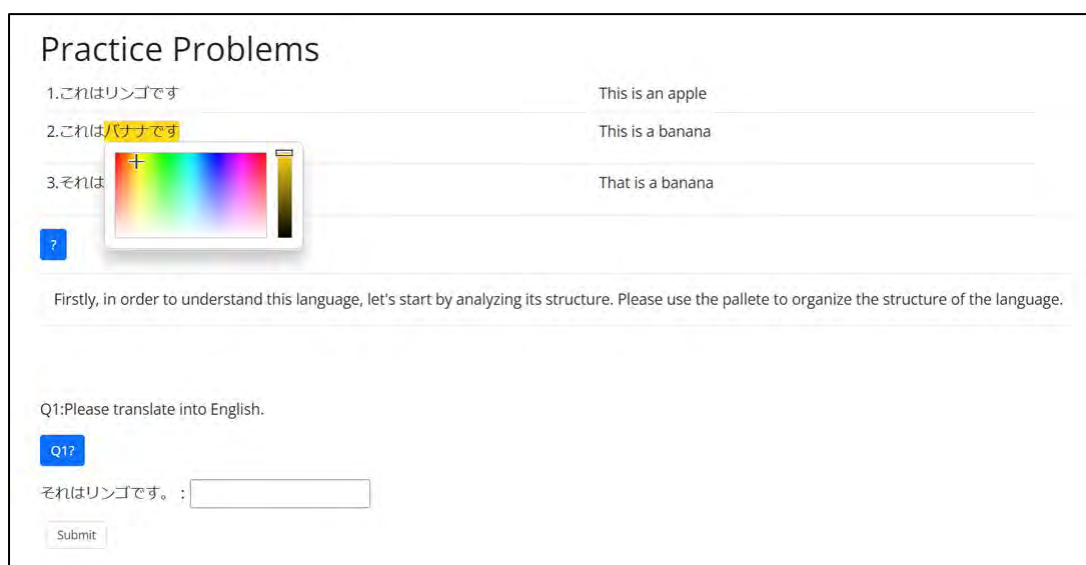


Fig. 17. The demonstration of the experimental webpage interface. Note: The screenshot shows a simple question to analyze Japanese; this demonstrates a practice session for participants to familiarize themselves with the website. The palette will show when users click on the words. The hint buttons are blue, which provide principle-based hints after clicking.

3.3 *Procedures*

The experiment started after participants clicked to select the language they used. Then, the webpage showed texts on the screen explaining the introduction of Linguistic Olyampaid, and this problem would not require previous knowledge of the target language, which means that all necessary information would be in the problem itself. Explanation of hints followed, saying participants can obtain hints by clicking hint buttons. After the introduction, the participants filled in basic information such as gender and age and clicked a button to turn on the web camera to start recording their faces throughout the experiment.

The participants were given a simple linguistic problem to solve in the practice session, aiming to familiarize participants to interact with this website and the task. Then, they started the experimental session after a short break if needed. There was no break during the experiment session, and the participants were allowed to solve the problem at their own pace. There was no time limit in the experiment session, and the experiment ended when the participants finished answering all the questions. They were instructed to avoid actions that would frame out from the face recording camera, such as unnecessary head and body movements while solving the problem.

The Ethics Committee of the Research Institute of Electrical Communication, Tohoku University, approved this study.

3.4 *Mental States Categorization*

To build models to estimate mental states, we must establish a ground truth first. The current study treats the two mental states independently. For example, the same frames of facial video can represent both a high engagement and a help-seeking state. For labeling the engagement states, we used subjective judgments from a team of four labelers from Tohoku University with varied nationalities: Japanese, Rumanian, and Chinese. We provided the labelers with an instruction manual beforehand, explaining how to label engagement based on the participants' facial appearances (see appendix).

The labelers used an annotation software, VGG Image Annotator (VIA) (Dutta & Zisserman, 2019), to rate the engagement by selecting one of four levels. They performed annotation when the videos continuously played, but they could pause the video when necessary. Each labeler annotated the nine participants' videos for the whole experiment session. A previous study rated each piece of a 10-second video clip by a score since they had a technical constraint (Whitehill et al., 2014). However, we can conquer the technical issue in this current study thanks to the annotation software. We did not opt to have a single number to an entire clip because the software we used here helped us precisely rate engagement states flexibly and simply.

Labelers were instructed to rate the face in videos: "How engaged does the

participant *appear to be.*” The engagement levels were rated from 1 (Not engaged at all) to 4 (Very engaged), following the previous study (Whitehill et al., 2014). The annotation scores were averaged over four labelers in every frame. The inter-rater reliability estimated by Fleiss Kappa was $\kappa = 0.22$, suggesting a fair agreement of judgments across labelers (Landis & Koch, 1977). To have two categories with a similar number of samples, we divided the data into high and low levels of engagement, thresholding at the score of 3 (high ≥ 3 and low < 3). The high and low engagement bins were 54407 (42.4%) and 40082 (57.6%) respectively.

To define help-seeking states, we used moments of button clicks recorded by the website. Since participants needed reaction time after the moment of a help-seeking decision, we defined the time interval between 4 to 1 second before the moment of clicking as the period of the help-seeking state. We discarded the time interval between 1 to 0 seconds right before the moment of clicking because the clicking behavior itself might influence the facial behavior. A 3-second interval was chosen because some pilot attempts yielded better predictions for help-seeking with the basic action units (AUs; the details are described in the next section) than 10s and 15s intervals. In contrast with the help-seeking state, we defined the rest of the time as a “working state,” indicating learners are working on the problem. We randomly chose 3-second intervals from the “working state” to analyze equivalent data for the working and the help-seeking states. There are 758 samples of the help-seeking state (48.8%) and 794 samples of the working state (51.2%). Although we sampled the working state with the same number as the number of clicks, some clicks were too close to separate. Since the frames were overlapped, the samples of the help-seeking state are less than the working state.

3.5 Facial Feature Extraction

We used OpenFace 2.2.0 to extract the participants’ facial features (Baltrusaitis et al., 2018) from each video frame. The positional changes of facial landmarks, such as the boundaries of eyes, eyebrows, and mouth, are used by OpenFace to evaluate the degree of facial muscle activity, i.e., Action Units (AUs) (Ekman et al., 1978). OpenFace uses two types of indexes to indicate the strength of AUs. One is the presence of features (0 and 1), used for all 18 AUs (Table 3), and the other is the intensity, expressed by continuous numbers between 0 to 5, used for 17 AUs, except for AU45. OpenFace also extracts head pose parameters, including the three coordinates indicating the head’s location and the head’s rotation angles: pitch, yaw, and roll.

As for eye gaze direction vector in world coordinates, the left eye and right eye are recorded in x, y, and z axes (6 types). The values were normalized by OpenFace. Besides, the eye gaze direction in radians in world coordinates averaged for both eye. An individual looking left-right and up-down (2 types) will result in changes of the

values. The Gaze features have 8 types.

Table 3. Description of 18 AU features that OpenFace can extract

AU	Description	AU	Description
1	Inner Brow Raiser	14	Dimpler
2	Outer Brow Raiser	15	Lip Corner Depressor
4	Brow Lowerer	17	Chin Raiser
5	Upper Lid Raiser	20	Lip stretcher
6	Cheek Raiser	23	Lip Tightener
7	Lid Tightener	25	Lips part
9	Nose Wrinkler	26	Jaw Drop
10	Upper Lip Raiser	28	Lip Suck
12	Lip Corner Puller	45	Blink

We used three feature sets following the previous studies (Bosch & D'Mello, 2021). The first one is the Basic AU feature set. It has 35 AUs with six descriptive statistics, including mean, median, standard deviation, minimum, maximum, and range (which is the difference between maximum and minimum) every ten frames (0.5s), resulting in 210 AU features (6 indexes \times 35 AUs). The second one is the Head Pose feature set. It has head pose features, including three coordinates and three rotation angles summarized with six descriptive statistics every ten frames (0.5s), which yielded 36 head pose features (6 indexes \times 6 head poses). The third one is the Co-occurring AU feature set. It has combinations of every intensity AU pair. We estimated AU co-occurrences using Bosch and D'Mello's method and equations (Bosch & D'Mello, 2021). The Co-occurring AU feature is based on the similarity between two AU's distribution within 0.5s using Jensen-Shannon divergence (JSD), which measures the symmetric relationship between AUs. To be specific, JSD Equation (1) is an extension of Kullback-Leibler divergence (KLD) Equation (2), which measures the information lost by using a prior distribution Q to approximate a posterior distribution P , given probability density functions p and q for P and Q respectively. In our case, p and q are given by two AU distributions, and we computed 136 JSD features, including all combinations of 17 AU intensities. The Gaze features has 8 gaze indexes and summarized in meand and standard deviation every ten frames (0.5s), restulsting in 16 Gaze features.

$$KLD(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \quad (1)$$

$$JSD(P||Q) = \frac{1}{2} KLD(P||\frac{1}{2}[P + Q]) + \frac{1}{2} KLD(Q||\frac{1}{2}[P + Q]) \quad (2)$$

3.6 Machine Learning Models

Fig. 18 shows the framework of the analysis with a Machine Learning technique. We used LightGBM (Light Gradient Boosting Machine) (Ke et al., 2017) to model how to predict mental states from facial features. LightGBM is built in a gradient-boosting framework that uses tree-based learning algorithms and is known as a fast method for training. We also used the Support Vector Machine (SVM) as a comparison since previous studies showed that SVM performs better than other models, including Naïve Bayes, k-NN, Random Forest, DNN, etc., for face image analysis (Bosch & D'Mello, 2021; Li et al., 2021). We trained these two models for three feature sets (Basic AUs, Head Pose, and Co-occurring AUs) to predict the engagement and help-seeking states separately. The models are evaluated by 5-fold cross-validation, where 80% of data were used for training and 20% for testing. Random selection divided data into five groups after pooling results from all participants.

We used three indexes for the model evaluation, following the previous studies (Bosch & D'Mello, 2021; Li et al., 2021): Area Under the curve (AUC) of Receiver Operating Characteristic (ROC), the F_1 score, and the rate of correct judgments. The ROC curve connects points with coordinates of false positive and true positive with variable classification threshold. AUC varies between 0.5 (random classification, straight line connecting (0,0) and (1,1)) and 1 (perfect classification, the line connecting (0,1) and (1,1)). The F_1 score is the harmonic mean of precision (the rate of true positive against positive data) and recall (the rate of true positive against positive responses). Accuracy is the proportion of frames classified in the correct label in all classified frames.

To explore the importance of each factor to the LightGBM prediction results, we used an analysis called SHapley Additive exPlanations (SHAP) (Lundberg et al., 2019), which was widely used (Ikeda et al., 2022; Miao et al., 2023) and known for its strength in estimating feature importance in prediction and identifying the most important feature (Belle & Papantonis, 2021). Among five validations, we calculated SHAP values for the Basic AUs and Head Pose feature sets with the highest predictive capacity (highest AUC).

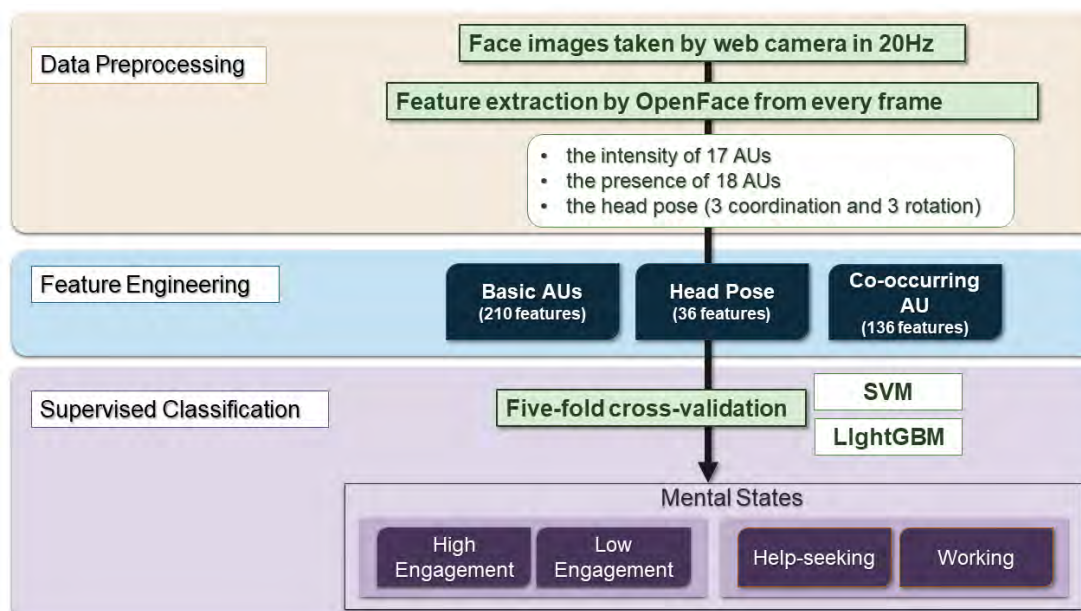


Fig. 18. Framework of the current study. While the participants interacted with our website, the web camera took their facial expressions, and the videos were saved on the local computer. After the experiment, we extract features by OpenFace 2.0, including Action Units (AUs) and head pose data. In data analysis, we set up three kinds of feature sets: Basic AUs feature set, Head Pose feature set, and Co-occurring AUs feature set. The engagement state and help-seeking state are estimated by LightBGM and SVM models.

4. Results

4.1 Behavioral Results

The average time to complete the problem was about 46.0 minutes (the longest was 75.6 minutes; the shortest was 31.4 minutes). The average number of hint button clicks was 34.8 (varied between 13 and 78 times). The perfect score was 10 points, with one point for each of the ten questions. If the answer included grammar errors, the score was deducted to 0.5. The average score is 7.44 points (the highest of 9.5 and the lowest of 6).

We compared the results of completion time, score, and the clicks of the hint buttons and found no statistical significance between them, while positive correlations were shown for all combinations. The Pearson's correlation coefficient and the results of the statistical test of significance are shown in Table 4. Since the samples were small, it was hard to find statistical differences. However, the results all show a positive correlation between each other. They suggested that the more learners involved in the problem and the more they use the hint buttons, the higher the score they can get.

Table 4. Pearson's correlation coefficient between behavioral variables and the results of a statistical test of no correlation.

	Completion time	Score	Hint button clicks
Completion time	-	-	-
Score	0.58 ($t(7)=1.90, p=0.09$)	-	-
Hint button clicks	0.51 ($t(7)=1.59, p=0.15$)	0.32 ($t(7)=0.89, p=0.40$)	-

4.2 Classification of the Engagement State

The model performance for predicting engagements using the LightGBM and SVM is summarized in Fig. 19 and Fig. 21. The results for the three feature sets (Basic AUs, Head Pose, and Co-occurring AUs) were evaluated by F_1 , AUC, and accuracy. These values are the average values of five repetitions from the five-fold cross-validation. The F_1 was between 0.72 and 0.83, the AUC was between 0.61 and 0.87, and the accuracy was between 0.61 and 0.79. The results showed that prediction was higher than chance performance (random classification, $AUC=0.5$), and the facial and head features are useful for estimating engagements. A one-sample t-test showed that all AUC scores are significantly above the chance level ($t(5)=5.41, p<0.05$). The overall performance of LightGBM was significantly better than SVM ($t(8)=2.64, p<0.05$). The ROC curves shown in Fig. 21 also indicate the prediction that facial and head features are useful for estimating engagements. The overall performance is shown in Fig. 19 and Fig. 21.

Furthermore, the complete comparison of the different feature sets also conducted. The details were shown in Fig. 20. In order to save training time, this part only trained the LightGBM models. The F_1 score ($F(8,36)=326.43, p<0.001$), accuracy ($F(8,36)=717.53, p<0.001$), and the AUC ($F(8,36)=1694.97, p<0.001$) have significant differences. The results suggested that the head pose features are beneficial to classify the engagement states, and the fusion feature set of Basic AUs, Head pose and Gaze features was well-performed.

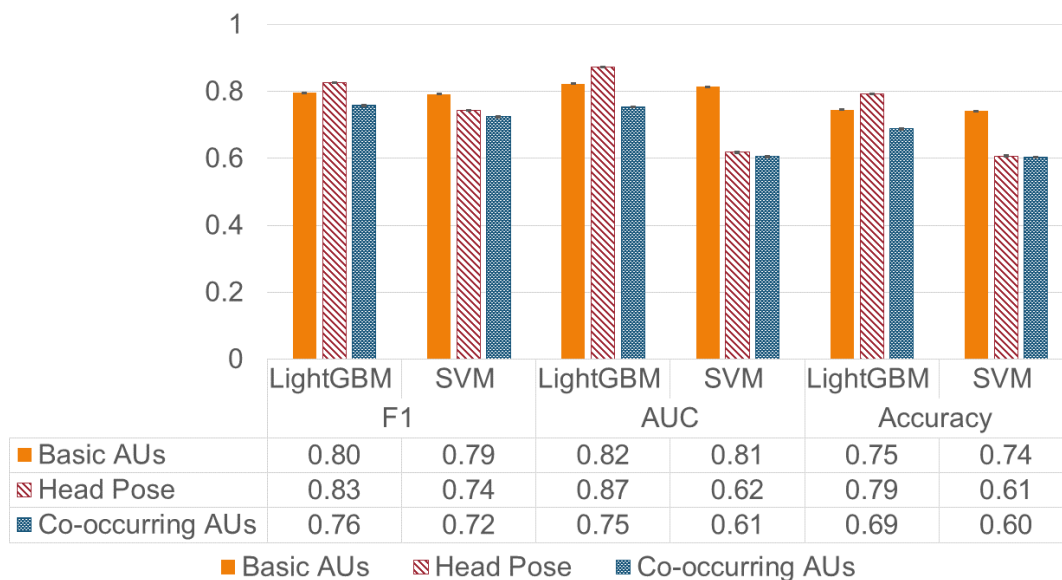


Fig. 19. Results of engagement classification. Note: the errors obtained from the standard deviations of the 5-fold cross-validation results, but the error bars are small and hard to see in this bar plot.

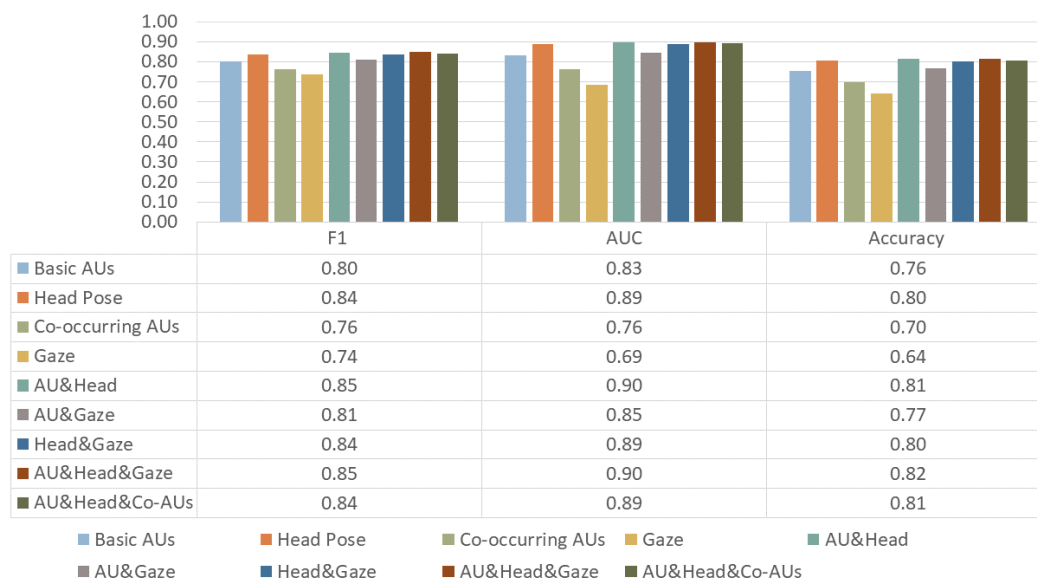


Fig. 20 Results of engagement classification for all features in LightGBM.

Table 5 The Results of multiple comparison by Tukey's HSD when estimate the engagement state

Metrics	Graph																				
F1	<p>Multiple Comparisons Between All Pairs (Tukey)</p> <table border="1"> <caption>Approximate Mean Values for F1 Score</caption> <thead> <tr> <th>Feature Set</th> <th>Mean Value</th> </tr> </thead> <tbody> <tr><td>head-pose</td><td>0.845</td></tr> <tr><td>gaze&head</td><td>0.840</td></tr> <tr><td>gaze</td><td>0.740</td></tr> <tr><td>co_occurAU</td><td>0.770</td></tr> <tr><td>basicAUs</td><td>0.805</td></tr> <tr><td>AU&head&gaze</td><td>0.850</td></tr> <tr><td>AU&gaze</td><td>0.815</td></tr> <tr><td>AU&Head&co_AU</td><td>0.840</td></tr> <tr><td>AU&Head</td><td>0.840</td></tr> </tbody> </table>	Feature Set	Mean Value	head-pose	0.845	gaze&head	0.840	gaze	0.740	co_occurAU	0.770	basicAUs	0.805	AU&head&gaze	0.850	AU&gaze	0.815	AU&Head&co_AU	0.840	AU&Head	0.840
Feature Set	Mean Value																				
head-pose	0.845																				
gaze&head	0.840																				
gaze	0.740																				
co_occurAU	0.770																				
basicAUs	0.805																				
AU&head&gaze	0.850																				
AU&gaze	0.815																				
AU&Head&co_AU	0.840																				
AU&Head	0.840																				
AUC	<p>Multiple Comparisons Between All Pairs (Tukey)</p> <table border="1"> <caption>Approximate Mean Values for AUC Score</caption> <thead> <tr> <th>Feature Set</th> <th>Mean Value</th> </tr> </thead> <tbody> <tr><td>head-pose</td><td>0.890</td></tr> <tr><td>gaze&head</td><td>0.885</td></tr> <tr><td>gaze</td><td>0.700</td></tr> <tr><td>co_occurAU</td><td>0.760</td></tr> <tr><td>basicAUs</td><td>0.830</td></tr> <tr><td>AU&head&gaze</td><td>0.895</td></tr> <tr><td>AU&gaze</td><td>0.850</td></tr> <tr><td>AU&Head&co_AU</td><td>0.885</td></tr> <tr><td>AU&Head</td><td>0.885</td></tr> </tbody> </table>	Feature Set	Mean Value	head-pose	0.890	gaze&head	0.885	gaze	0.700	co_occurAU	0.760	basicAUs	0.830	AU&head&gaze	0.895	AU&gaze	0.850	AU&Head&co_AU	0.885	AU&Head	0.885
Feature Set	Mean Value																				
head-pose	0.890																				
gaze&head	0.885																				
gaze	0.700																				
co_occurAU	0.760																				
basicAUs	0.830																				
AU&head&gaze	0.895																				
AU&gaze	0.850																				
AU&Head&co_AU	0.885																				
AU&Head	0.885																				
Accuracy	<p>Multiple Comparisons Between All Pairs (Tukey)</p> <table border="1"> <caption>Approximate Mean Values for Accuracy Score</caption> <thead> <tr> <th>Feature Set</th> <th>Mean Value</th> </tr> </thead> <tbody> <tr><td>head-pose</td><td>0.810</td></tr> <tr><td>gaze&head</td><td>0.805</td></tr> <tr><td>gaze</td><td>0.650</td></tr> <tr><td>co_occurAU</td><td>0.700</td></tr> <tr><td>basicAUs</td><td>0.760</td></tr> <tr><td>AU&head&gaze</td><td>0.815</td></tr> <tr><td>AU&gaze</td><td>0.770</td></tr> <tr><td>AU&Head&co_AU</td><td>0.805</td></tr> <tr><td>AU&Head</td><td>0.805</td></tr> </tbody> </table>	Feature Set	Mean Value	head-pose	0.810	gaze&head	0.805	gaze	0.650	co_occurAU	0.700	basicAUs	0.760	AU&head&gaze	0.815	AU&gaze	0.770	AU&Head&co_AU	0.805	AU&Head	0.805
Feature Set	Mean Value																				
head-pose	0.810																				
gaze&head	0.805																				
gaze	0.650																				
co_occurAU	0.700																				
basicAUs	0.760																				
AU&head&gaze	0.815																				
AU&gaze	0.770																				
AU&Head&co_AU	0.805																				
AU&Head	0.805																				

In order to understand the facial expression and the head pose. Firstly, we used SHAP to estimate the importance of the features in the LightGBM model with the Basic AUs feature set (Fig. 22). The left panel of Fig. 22 shows absolute SHAP values, which indicate the summarized effect from all sample points. The right panel of Fig. 22 shows the SHAP value (horizontal axis) and each index value (coded by color: red for higher value) of an estimation (data of a time bin of the test set). The SHAP result shows that “AU02_c_mean (outer brow raiser)”, “AU23_c_mean (lip tightener)”, and “AU04_r_mean (brow lowerer)” are the top three important AU indexes in the model and that the first two are correlated positively, and the third one is done negatively.

The importance of the features in the LightGBM model with the Head Pose feature set was also calculated by SHAP. The results showed that “pose_Tx_min,” “pose_Tz_min,” and “pose_Ty_max” are the top three important head pose features in the model. It suggested that the location of the head with respect to the camera was important.

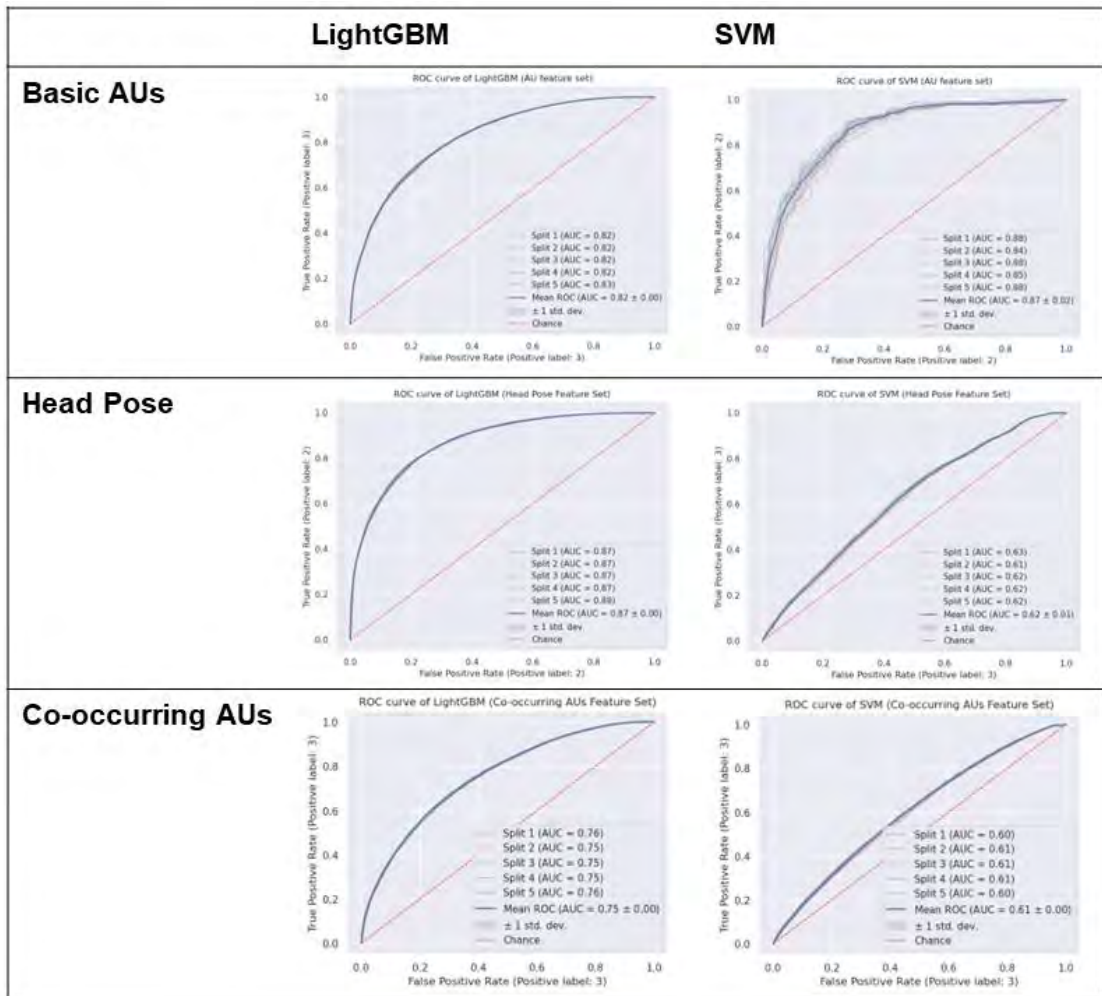


Fig. 21. ROC curve results of high/low engagement classifications by LightGBM and SVM model

on three feature sets (Basic AU, Head Pose, and Co-occurring AUs)

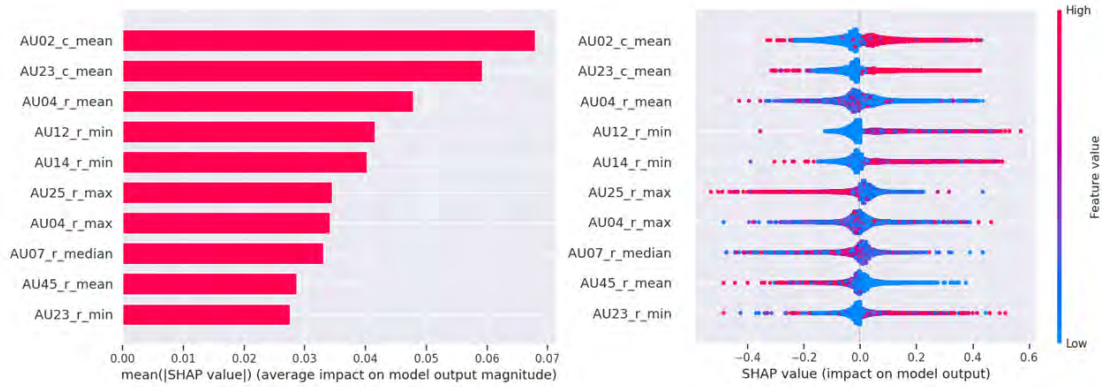


Fig. 22. SHAP summary plot of LightGBM classifying engagement state by Basic AU feature set. The suffix of the features indicate which statistics we calculated; “min” indicates minimum; “max” indicates maximum. Besides, “r” indicates the intensity of the AU; “c” indicates the presence of the AU.

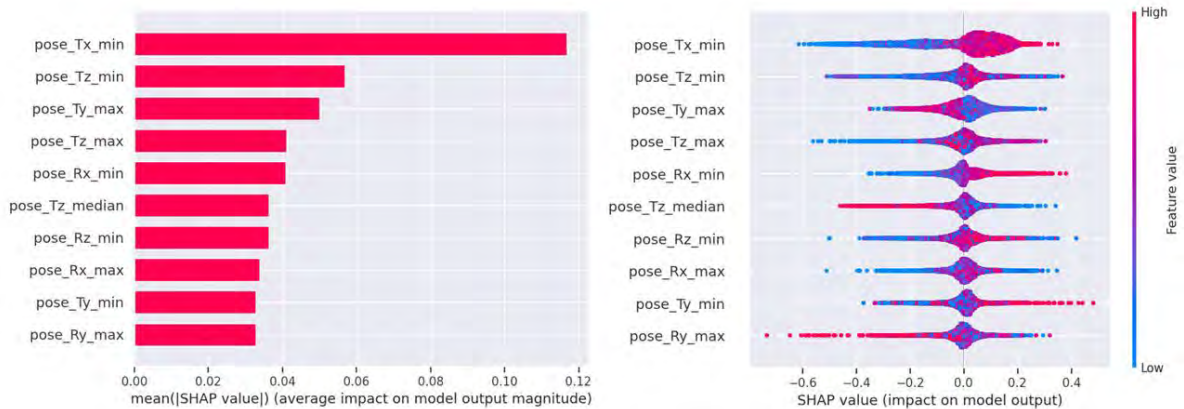


Fig. 23. SHAP summary plot of LightGBM classifying engagement state by Head Pose feature set. The suffix of the features indicate which statistics we calculated; “min” indicates minimum; “max” indicates maximum. Besides, “Tx/Ty/Tz” indicates the location of the head with respect to the camera in millimeters (positive Z is away from the camera); “Rx/Ry/Rz” indicates rotation that in radians around X, Y, Z axes, which can be seen as pitch, yaw, and roll separately (left-handed positive sign).

4.3 Classification of the Help-seeking State

We also used three feature sets, basic AUs, head pose, co-occurring AUs, and two machine learning methods, the LightGBM and SVM models. The F_1 score, AUC, and accuracy results were the average of the five-fold cross-validation and are shown in Fig. 24. The F_1 was between 0.52 and 0.92, the AUC was between 0.61 and 0.92, and the accuracy was between 0.57 to 0.92. A one-sample t -test showed that all scores of all

AUCs are higher than the chance level ($t(5)=5.06, p<0.05$). The overall performance of LightGBM was significantly better than that of SVM ($t(8)=4.22, p<0.05$). The ROC curves in Fig. 26 are above the chance level line except for the SVM of head pose sets.

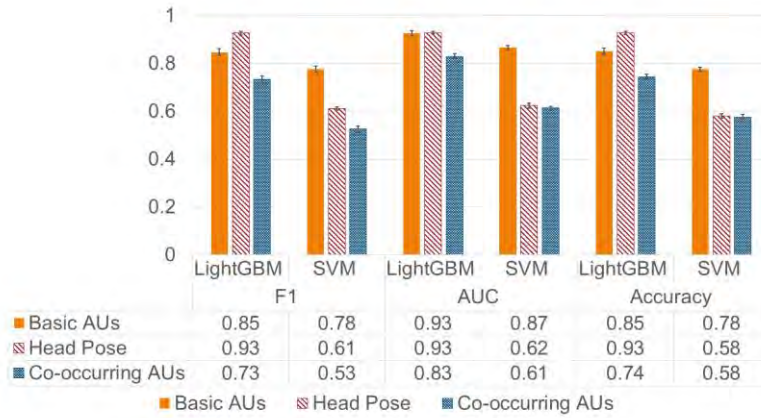


Fig. 24. Results of help-seeking state classification. Note: the errors obtained from the standard deviations of 5-fold cross-validation results.

Furthermore, focusing on the lightGBM, we also trained the models by different feature sets, including gaze, BasicAU&Head Pose, Basic AUs & Gaze, Head Pose & Gaze, BasicAUs & Head Pose & Gaze, and Basic AUs & Head Pose & Co-occurring AUs. The F1 score was between 0.69 to 0.92, the AUC was between 0.78 to 0.98, and the accuracy was between 0.70 to 0.93.

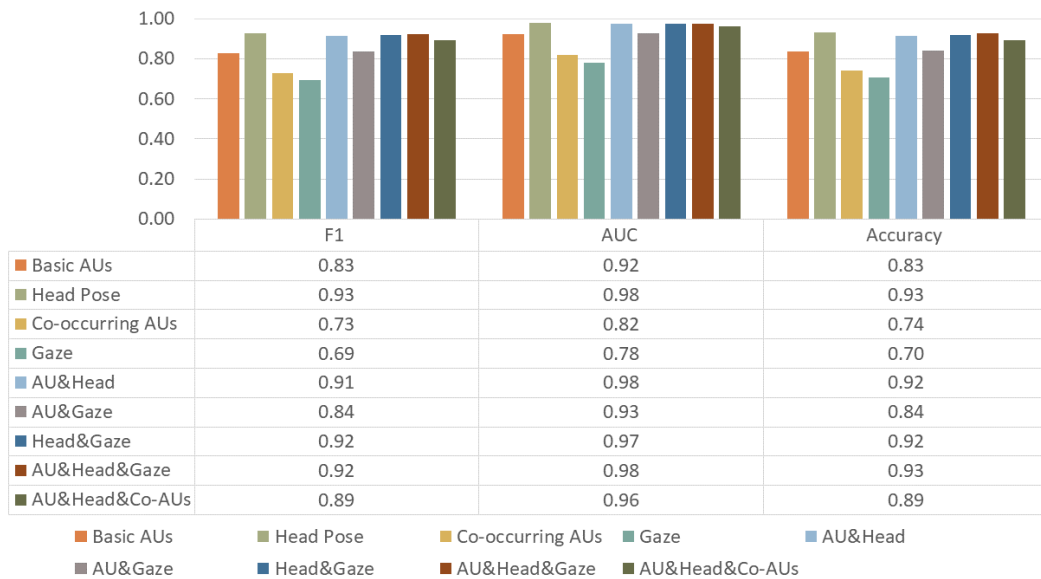


Fig. 25 Expanded Results of help-seeking state classification.

According to the result of one-way ANOVA analysis, there were differences

between F1 score ($F(8,36)=120.99, p<0.001$), AUC ($F(8,36)=144.24, p<0.001$), and Accuracy($F(8,36)=125.81, p<0.001$) across those 9 feature sets. The multiple comparison by Tukey’s HSD are shown in Table 6. The feature sets of gaze and co-occurring AUs performed worse than other feature sets. Besides, when the feature sets includes head pose features, the performance would become better than others.

Table 6 The Results of multiple comparison by Tukey’s HSD when estimate the help-seeking state

Metrics	Tukey’s HSD																																								
F1	<p style="text-align: center;">Multiple Comparisons Between All Pairs (Tukey)</p> <table border="1"> <caption>Approximate F1 Score Data from Tukey's HSD Plot</caption> <thead> <tr> <th>Feature Set</th> <th>Mean F1 Score</th> <th>Lower Bound</th> <th>Upper Bound</th> </tr> </thead> <tbody> <tr><td>head-pose</td><td>0.93</td><td>0.91</td><td>0.95</td></tr> <tr><td>gaze&head</td><td>0.92</td><td>0.90</td><td>0.94</td></tr> <tr><td>gaze</td><td>0.69</td><td>0.67</td><td>0.71</td></tr> <tr><td>co_occurAU</td><td>0.73</td><td>0.71</td><td>0.75</td></tr> <tr><td>basicAUs</td><td>0.83</td><td>0.81</td><td>0.85</td></tr> <tr><td>AU&head&gaze</td><td>0.92</td><td>0.90</td><td>0.94</td></tr> <tr><td>AU&gaze</td><td>0.84</td><td>0.82</td><td>0.86</td></tr> <tr><td>W&Head&co_AU</td><td>0.89</td><td>0.87</td><td>0.91</td></tr> <tr><td>AU&Head</td><td>0.92</td><td>0.90</td><td>0.94</td></tr> </tbody> </table>	Feature Set	Mean F1 Score	Lower Bound	Upper Bound	head-pose	0.93	0.91	0.95	gaze&head	0.92	0.90	0.94	gaze	0.69	0.67	0.71	co_occurAU	0.73	0.71	0.75	basicAUs	0.83	0.81	0.85	AU&head&gaze	0.92	0.90	0.94	AU&gaze	0.84	0.82	0.86	W&Head&co_AU	0.89	0.87	0.91	AU&Head	0.92	0.90	0.94
Feature Set	Mean F1 Score	Lower Bound	Upper Bound																																						
head-pose	0.93	0.91	0.95																																						
gaze&head	0.92	0.90	0.94																																						
gaze	0.69	0.67	0.71																																						
co_occurAU	0.73	0.71	0.75																																						
basicAUs	0.83	0.81	0.85																																						
AU&head&gaze	0.92	0.90	0.94																																						
AU&gaze	0.84	0.82	0.86																																						
W&Head&co_AU	0.89	0.87	0.91																																						
AU&Head	0.92	0.90	0.94																																						
AUC	<p style="text-align: center;">Multiple Comparisons Between All Pairs (Tukey)</p> <table border="1"> <caption>Approximate AUC Score Data from Tukey's HSD Plot</caption> <thead> <tr> <th>Feature Set</th> <th>Mean AUC Score</th> <th>Lower Bound</th> <th>Upper Bound</th> </tr> </thead> <tbody> <tr><td>head-pose</td><td>0.98</td><td>0.96</td><td>1.00</td></tr> <tr><td>gaze&head</td><td>0.97</td><td>0.95</td><td>0.99</td></tr> <tr><td>gaze</td><td>0.78</td><td>0.76</td><td>0.80</td></tr> <tr><td>co_occurAU</td><td>0.82</td><td>0.80</td><td>0.84</td></tr> <tr><td>basicAUs</td><td>0.92</td><td>0.90</td><td>0.94</td></tr> <tr><td>AU&head&gaze</td><td>0.97</td><td>0.95</td><td>0.99</td></tr> <tr><td>AU&gaze</td><td>0.93</td><td>0.91</td><td>0.95</td></tr> <tr><td>W&Head&co_AU</td><td>0.96</td><td>0.94</td><td>0.98</td></tr> <tr><td>AU&Head</td><td>0.97</td><td>0.95</td><td>0.99</td></tr> </tbody> </table>	Feature Set	Mean AUC Score	Lower Bound	Upper Bound	head-pose	0.98	0.96	1.00	gaze&head	0.97	0.95	0.99	gaze	0.78	0.76	0.80	co_occurAU	0.82	0.80	0.84	basicAUs	0.92	0.90	0.94	AU&head&gaze	0.97	0.95	0.99	AU&gaze	0.93	0.91	0.95	W&Head&co_AU	0.96	0.94	0.98	AU&Head	0.97	0.95	0.99
Feature Set	Mean AUC Score	Lower Bound	Upper Bound																																						
head-pose	0.98	0.96	1.00																																						
gaze&head	0.97	0.95	0.99																																						
gaze	0.78	0.76	0.80																																						
co_occurAU	0.82	0.80	0.84																																						
basicAUs	0.92	0.90	0.94																																						
AU&head&gaze	0.97	0.95	0.99																																						
AU&gaze	0.93	0.91	0.95																																						
W&Head&co_AU	0.96	0.94	0.98																																						
AU&Head	0.97	0.95	0.99																																						

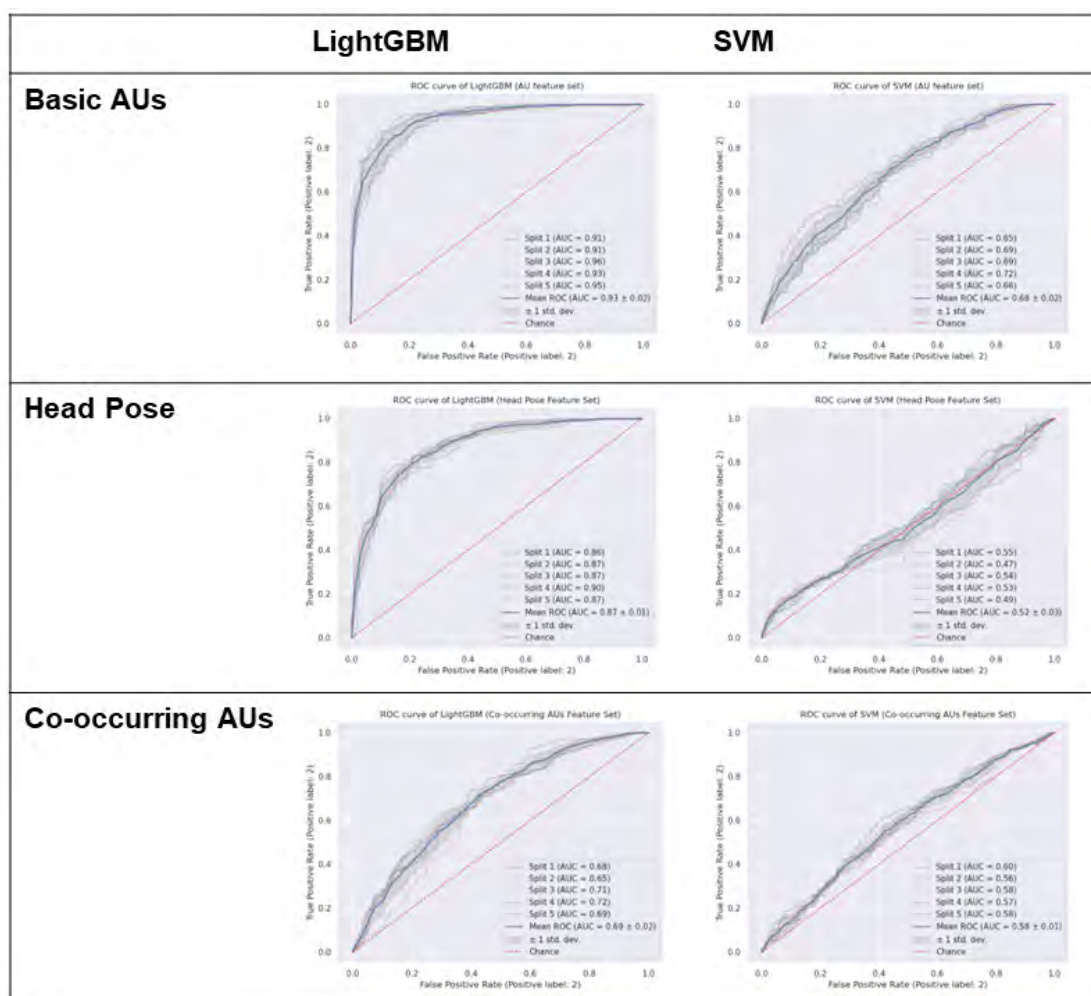
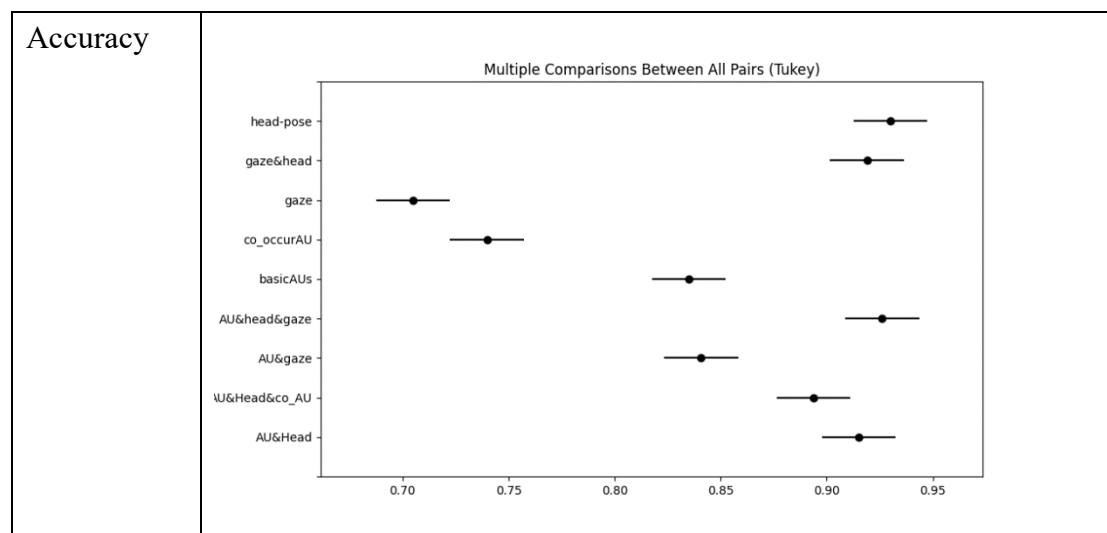


Fig. 26. The ROC curve results of help-seeking/working state classifications by LightGBM and SVM model on the three feature sets (Basic AU, Head Pose, and Co-occurring AUs)

We calculated the SHAP values of the LightGBM model trained with the basic AUs

and Head Pose feature sets for the help-seeking state. The summary plots of the SHAP value are shown in Fig. 27 and Fig. 28. The figures indicate the SHAP value of the highest AUC classification model. SHAP analysis showed that “AU04_r_median (brow lowerer)”, “AU23_r_mean (lip tightener)”, and “AU14 (dimpler)” are important features in the Basic AU feature set for estimating the help-seeking state, whereas “pose_Ty_min,” “pose_Tz_mean,” and “pose_Tx_min” are important features in the Head Pose feature set for estimating the help-seeking state. The results of the SHAP analysis suggest that (1) the lower part of the face is more important in general, even if the top one is related to eyebrow furrowing; (2) the location of the head with respect to the camera is more important than the rotation of the head.

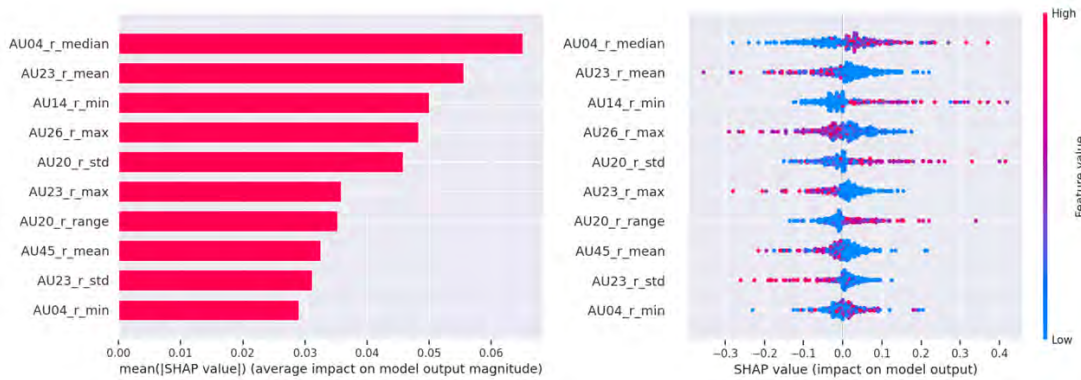


Fig. 27. SHAP summary plot of LightGBM classifying help-seeking state by Basic AU feature set. The suffix of the features indicates which statistics we calculated; “min” indicates minimum; “max” indicates maximum; “std” indicates standard deviation. Besides, “r” indicates the intensity of the AU.

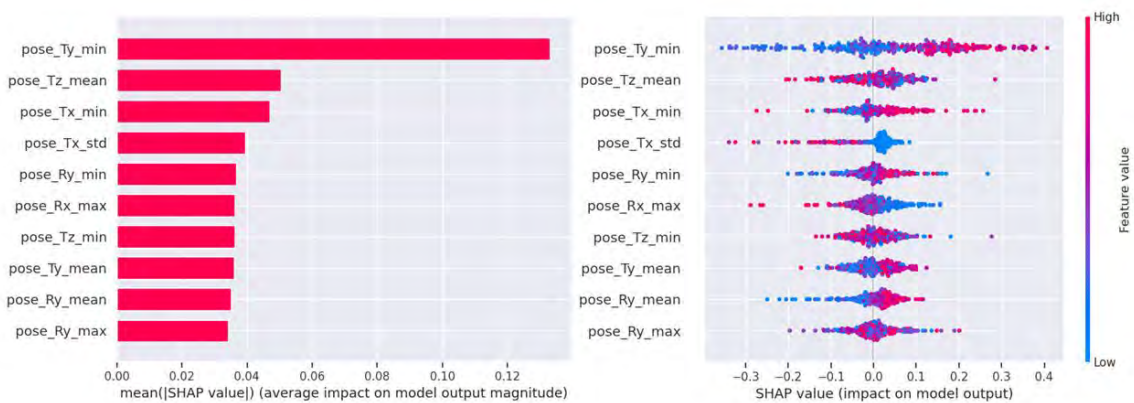


Fig. 28. SHAP summary plot of LightGBM classifying help-seeking state by Head Pose feature set. The suffix of the features indicates which statistics we calculated; “min” indicates minimum; “max” indicates maximum; “std” indicates standard deviation. Besides, “Tx/Ty/Tz” indicates the

location of the head with respect to the camera in millimeters (positive Z is away from the camera); “Rx/Ry/Rz” indicates rotation that in radians around X, Y, Z axes, which can be seen as pitch, yaw, and roll separately (left-handed positive sign).

5. Discussion

We developed an intelligent tutoring system in our experiments to solve a problem at the Linguistic Olympiad. Using the system, the current study examined whether a computer system can predict two mental states from learners’ face videos during learning. One mental state is the engagement state, which we classified as the high and low engagement states. The other is the help-seeking state, which we defined four to one seconds before learners requested a hint. With machine learning methods, we used facial features and head poses extracted from videos recorded during the experiment to predict the engagement and help-seeking states. Results revealed that facial features and head poses were effective indicators for both state classifications. The overall prediction used LightGBM of the accuracy is higher than 70% for both. These accuracy scores are higher than those of previous studies for engagement prediction (Bosch & D’Mello, 2021; Li et al., 2021), and we extended the method to predict mental states to the help-seeking state. The accuracy level reached a level for classroom environments (Sümer et al., 2021).

We compared the three feature sets and the two machine learning models. The prediction performance was better by LightGBM than SVM for all three feature sets. This study further compared the unimodal, bimodal, and multimodal models to build the feature sets. Among unimodal models, the current study suggested that the Head Pose feature set is the best, while the Basic AUs feature set shows similarly good predictions among the three feature sets. In contrast, the Co-occurring AUs and Gaze feature set results had lower accuracy than the other two. If facial expressions expressed by combinations of AUs are factors reflecting the mental states, the co-occurring AUs could predict mental states at least slightly better than Basic AU features. However, the worse prediction with the Co-occurring AUs feature set indicates that the AUs used as low-level features, instead of their combinations, are enough to contribute to estimating the mental state of engagements and help-seeking. In addition, only relying on Gaze features is not enough for predicting both engagement states and help-seeking states.

As for the bimodal and multimodal models, the results suggested that if the models contains Head Pose features, the performance would become better than other models. This results suggested that to determine the behavioral performance of the learners, the movement and location of the head play an important roles. A previous study suggest that using gaze and facial expression data was useful for estimating behavioral

engagement (Xiao et al., 2022), but in our case, we found head pose data was more useful than gaze data. In Xiao et al. (2022)'s study, they did not use head pose data to predict engagement, but in other studies (Li et al., 2021; Sümer et al., 2021), head pose features were commonly used.

The SHAP values are calculated to investigate further which features are important. The results showed that the brow and lower part of the face are the two most important features for help-seeking (Fig. 9). Several reports indicated that the brow and lower part of the face are related to negative feelings, even though those studies did not analyze AUs. For example, when people feel unsure about the problem, their lips move apart, and the shape of their brows changes (el Kaliouby & Robinson, 2005). Besides, people tend to depress their lips and furrow their eyebrows when watching videos related to banking, fuel, or pharmaceuticals rather than videos about pet care, entertainment, or baby care that let them smile more (McDuff & Kaliouby, 2017). In sum, when people are unsure about something or watching more serious videos, they smile less, lower their brows, and apart or depress their lips, which are expressions related to AU04 (brow lowerer), AU25 (lips part), AU26 (jaw drop) and other AUs around lips. The current results for help-seeking are on the same line as those from these studies. It could be the case that there is a typical mind process for serious and careful thinking, which is related to facial expressions with depressing lips and furrowing eyebrows.

Furthermore, the effectiveness of the Head Poses feature set estimating mental states is consistent with previous studies (Li et al., 2021; Sümer et al., 2021). SHAP analysis of head poses in the current study showed that the location of the head with respect to the camera is important to estimate both the engagement state and help-seeking state. The results suggested that the head positions can determine whether a learner behaves well or not, such as sitting up straight in front of the monitor. Therefore, the web camera installed on the top of the laptop's screen is useful to monitor learners in front of the screen. Besides, the important features to estimate the engagement state have more features related to head pitch. In contrast, the important features to estimate the help-seeking state have more features related to head yaw. A previous study showed that learners concentrate or think when tilting their heads (el Kaliouby & Robinson, 2005). To be more specific than the previous study, this study's results suggested that the features of the head's pitch are related to engagement since they might indicate sleepiness and concentration. On the other hand, when learners intend to ask questions, a head yaw allows them to receive more information on the screen, which might help them solve the problem until they finally click the hint buttons.

The current study attempted to estimate not only the engagement state, a major target of learning studies, but also the help-seeking state, which, we believe, is another important mental state during learning. As far as we know, no previous studies

compared or investigated the relationship between learners' engagement and help-seeking states. Our results showed differences in behavior indexes extracted from facial videos (facial expressions and head poses) described above. We confirmed the differences by applying the predicting model of engagement to help-seeking and vice versa. We exchanged the training and testing datasets for the purposes. That is, the model trained with the dataset for estimating the engagement state was used to predict the help-seeking state, and the model trained for estimating the help-seeking state was used to predict the engagement states. The results of the cross-examination are shown in Table 7. They show that the AUCs of all feature sets are never higher than 0.5. This proves that the estimations of the engagement and help-seeking states do not share the same processes related to facial expressions and head poses.

Table 7. AUC Results of cross-validation by exchanging training dataset and testing dataset.

	Feature sets		
	Basic AUs	Head Pose	Co-occurring AUs
Training: Engagement Testing: Help-seeking	0.45	0.46	0.48
Training: Help-seeking Testing: Engagement	0.47	0.47	0.48

There has been an ongoing debate on whether engagement leads to learning performance. Some studies reported a correlation between learning performance and engagement (Chen, 2017; Galikyan & Admiraal, 2019; Xie et al., 2019), while others reported no significant correlation (Li et al., 2021; Whitehill et al., 2014). We explored this question with the present data. Among the participants in the current study, Pearson's correlation between scores and the mean of engagement levels is not significant ($t(7)=1.51$, $r=0.56$, $p=0.58$). However, Pearson's correlation coefficient of the engagement and help-seeking state is significant ($t(15607)=-8.13$, $r=-0.065$, $p<0.05$) in all 15609 frames, which we picked up for estimating the help-seeking state. The results suggested that the help-seeking state is related to high engagement. In addition, research showed that learners' attention is greater in a real class when they can use an interactive device to respond to questions (Bunce et al., 2010). We can further explore what kinds of interaction technology affect behaviors more. However, the correlation results might be influenced by the amount of samples. This remains the issue for future research to explore other factors influencing learning performance.

There are two major issues that still need to be addressed for future studies in the current study. First, we trained machine learning models by pooling all participants'

data, which did not consider individual variations. If there are large variations among participants, as pointed out previously (Kato et al., 2022a, 2022b; Miao et al., 2023; Sato et al., 2022; Shioiri et al., 2021), customizing a model for each individual is more appropriate. Although the estimations for group data worked well, an individual model still needs to be analyzed since there might be large individual differences. Second, we used six feature statistics, i.e., mean, median, standard deviation, minimum, maximum, and range. The results of the SHAP analysis showed that mean, max, min, and median are more important than range or standard deviation. Although using all statistics is likely effective, a model with one or two statistical features works similarly well because these statistics are dependent, or a model with more complex time series features (Christ et al., 2018) could show better estimation.

Chapter 4: Study 2: Cultural comparison on estimating learners' engagement and help-seeking behaviors by facial and head features

1. Introduction

Estimating learners' mental states, including the engagement and help-seeking state, is important. Teachers manage their classes by observing students' engagement levels or detecting when they need support in an online or real classroom (Atif et al., 2020). On the other hand, learners can monitor themselves by knowing whether they are engaging in learning or can benefit by reducing their hesitation to ask a question. Although some researchers used learning system, such as Massive Online Open Course (MOOC) or intelligent tutoring system (ITS) with electroencephalography (EEG) and behavioral logs to estimate learners' mental states (Chaouachi et al., 2019; Kim et al., 2023; Lin & Kao, 2018), the invasive method of using a camera to capture students' facial expressions and head posture is widely used (Kato et al., 2022a; Miao et al., 2023; Sümer et al., 2021). The rising of invasive and automated methods without additional devices to estimate the learners' mental state is beneficial to education. Their applications have the potential to be able to implement in real-classroom and e-learning environments.

However, the research revealed that facial behaviors are different from culture to culture from large-scale data (McDuff et al., 2017). Therefore, the models to estimate the learners' mental state by their facial expressions might be unstable due to cultural differences.

The current study used an intelligent tutoring system which was also used in the previous studies (Wang, Nagata, et al., 2023), and the same experiment settings were used on different learners in another country. The current study aims to estimate the learners' mental states by using their facial expressions and compare them to a previous study (Wang, Hatori, et al., 2023). The learners from Japan and from Taiwan were compared.

2. Research Review

2.1 Mental states estimation by facial expressions

During learning, learners' mental states are changing all the time, and many researchers tried to estimate different kinds of mental states to improve and support learners learning. The facets of the mental states were stressed in different research. For example, a study focused on the mental states, such as attention, comprehension, and

stress, to develop an adaptive learning system (Kim et al., 2023). Another study focus on the mental states, including concentration, confusion, frustration, and boredom, to develop a system which can support students learning (Peng & Nagao, 2021).

Among many kinds of the mental state, they can be classified to positive and negative ones. For estimating the positive mental states, it is widely believed that the engagement/concentration state is essential. A systematic review pointed out that the journal articles about engagement estimation in educational context rapidly grew in around 2020s (Hasegawa et al., 2020). On the other hand, negative mental states are also estimated, but they have been divided into different meanings and word, such as confusion, frustration, boredom, stress, difficulty, etc (Hasegawa et al., 2020; Kim et al., 2023; Peng & Nagao, 2021). On top of that, the main purpose to estimate the negative mental states is to infer learners' need and support them during learning. Therefore, trying to understand when learners need help by estimating their help-seeking state is essential to reach this goal (Wang, Hatori, et al., 2023).

To be specific, the definition of engagement from Fredricks et al. (2004) is widely used in many research (Karimah & Hasegawa, 2022). Behavioral engagement describe learners' participation in learning task, behaviors of attention and concentration, and interest in activities. In contrast, help-seeking is the state of confusion, difficulty or frustration when a learner needs support. The current study aims to estimate the two mental states to understand the learning process. We believed that understanding learners' engagement and help-seeking state is beneficial to learners because it would be like an ideal teacher who can observe students' engagement to adjust the teaching strategies and detect whether they need help or not to provide them with appropriate education.

As for the methods to estimate the mental states, many tools and data were used in research, such as EEG data (Desai et al., 2020; Lin & Kao, 2018), facial images (Bosch & D'Mello, 2021; el Kaliouby & Robinson, 2005; Hasegawa et al., 2020; Kaliouby & Robinson, 2004; Kato et al., 2022a; Miao et al., 2022; Sümer et al., 2021), blink rate (Ren et al., 2019) or multimodal data, including heart rate and audio data (Al-Alwani, 2016; Kim et al., 2023; Monkarezi et al., 2017; Peng & Nagao, 2021). Among these data, the facial images can be obtained by a camera, which allows researchers to set a camera in front of learners without disturb them by setting external devices on their body. Besides, thanks to the development of computer vision, using facial images to estimate mental state by machine learning models is convenient and effective. The current study undertook this approach to estimate learners' mental state.

2.2 Cultural differences from face and their learning behavior

Cultural differences might be related to races. For example, a previous study, which

tested on Asian, Caucasian and chimpanzees, indicated that human's perception of faces exist other-race and other-species discrimination (Dahl et al., 2014). However, a human judger can adopt the culture to reduce the discrimination on recognizing human's emotions from different culture. For example, a study showed that Chinese living in Australia has higher ability to differentiate the emotions from Caucasian's face than Chinese living in mainland China (Prado et al., 2014). Nowadays, a society with diversity is common, and culture is related to but not the same as the nationality or the race of human.

Hofstede (2001) described culture in five dimensions: individualism, long-term orientation, masculinity, power distance index, and uncertainty avoidance. A large scale study on facial behavior of about 750,000 people from 12 countries suggested that the facial expression of brow furrowing expressed more on high individualism countries than low individualism countries (high collectivism) (McDuff et al., 2017). A meta-analysis study related to e-learning also indicated that individualist or collectivist culture influence learners' subjective norms, self-efficacy, and their perceived usefulness of e-learning platform (Zhao et al., 2021).

According to Hofstede (2001), the individualism index for Taiwanese is 17, which is one standard deviation lower than the average of all cultures the study investigated, and that for Japanese is 46, which is approximately equal to the average. A previous study showed that the facial expressions and perception between Taiwanese and Japanese have different patterns (Kaminosono et al., 2022).

The current study aims to compare the faical expressions between two different cultures. Researchers pointed out that machine's judgement of face expression would have bias due to the unbalance training data of different races (Sham et al., 2023), and they added more diverse data into the training dataset to make the AI model reduce the bias. However, lack of study compared the machine learning results with people who come from two or more cultures but same race. The current study compared Taiwanese people and Japanese people, which are all East Asian. Therefore, the bias of machine learning model can be reduced since the faical appearance should not have big different.

On top of that, the individualism index between the two culture suggested that the two cultures are different. We hypothesized that Japanese people will be more expressive than Taiwanese people. Furthermore, the expression of eyebrow furrowing, which is related to AU04, should also be more important in Japan's dataset than Taiwan's dataset. The facial expression differences affect by culture not by races are expected to be observed in our experiment.

2.3 Current Study

The current study used facial expressions and head pose features to analyze the

engagement state and help-seeking state. In order to compare the cultural differences, the method of estimating the mental states was followed the previous study. Besides, if the estimations of the mental states need different features from Taiwan's and Japan's data, it will be possible to use the facial features to identify the nationality. Thus, we also used facial features to classify the nationality. The specific research questions are as followings:

RQ1: How well is the machine learning model to estimate the engagement states and the help-seeking states by using facial expressions on Taiwan's participants?

RQ2: What are the differences between Japan's participants and Taiwan's participants by using SHAP analysis?

3. Methods

3.1 Participants

We have recruited 21 students (18 females and 3 males) from National Taiwan University. Their average age is 22 (standard deviation = 2.93). The students all voluntarily participated in our experiment. In addition, they don't have any Linguistics Olympiad experience and don't major in linguistics, literature, foreign languages, and other related fields. But their majors are different, including information sciences, psychology, etc. To preserve anonymity, random number codes in our dataset replaced all participants' names and data filenames.

The results of estimating the engagement state and the help-seeking states of Japanese participants that we want to compare in this study were published in another journal paper (Wang, Hatori, et al., 2023). In this study, we report analyses of Taiwanese participants and culture comparison with data from Japanese participants and Taiwanese participants.

3.2 Materials

The current study used the same website as in the previous study, but the website can be shown in Japanese or traditional Chinese depending on participants' mother language. The web cameras are installed on the top of the screen, and we execute a video recorder which made by a python program in the setting of 640×480 pixels if resolution with 20 Hz of sampling rate.

The problem-solving task is a problem from the International Olympiad of Linguistics in 2018, and we used the official translation of the problem in Japanese and traditional Chinese. They need to analysis the pair of a scarce language and their mother language, and then to answer the translation questions. We used this problem because it is necessary to have any prior knowledge to solve it and the cues to solve the problem are all in the problem itself.

During problem-solving, several interactive functions are provided on the website. Participants can highlight the word and click the hint buttons if they need help. If a button is clicked, the hint will be shown until the same button is clicked again. The participants answer the question by typing their answer on the blanks shown on the website. The website is like a virtual exam, once the participants want to hand in their answer sheet, they can click the submit button on the bottom of the website.

3.3 Procedures of the Experiment

In order to compare with the previous study, the current study followed the same experimental procedures of them. The participants followed the instructions of the webpages, including introduction of the Linguistic Olympiad competition, the instructions and rules, the basic information investigation, the practice session, and the experiment session.

The introduction informs the participants that there is no worry about the linguistic knowledge. The instructions and rules are written on the website and the experimenter still explains in verbal, and then helps participants turn on a web camera which can take their facial videos. After the participant provides their basic information, such as their age and gender, they can confirm the instructions and start the practice session. The practice session allows the participants to get familiar with the functions of the website and the rules of Linguistic Olympiad. If they have no problem, they can continue to the experimental session to finish answering the questions.

3.4 Mental States Categorization

This study estimates two mental states: the engagement and help-seeking states, which is the same as the previous study. The two mental states were estimated separately since we believe that the learning states can be complex and the different mental states can be overlapped on each other. For example, a learner can engage in learning but need help at the same time.

The engagement state was annotated by four labelers whose nationalities are different to reduce the ground-truth bias (Renier et al., 2021). The guidelines of judging the engagement levels are the same as the previous study and the rules followed another previous study (Whitehill et al., 2014). They used an annotation software, VGG Image Annotator (VIA) (Dutta & Zisserman, 2019), to complete their annotation work.

Labelers rate the videos continuously but they were able to pause the video if necessary. They gave the engagement level under the instruction of “How engaged does the participants *appear to be*” from 1 (not engaged at all) to 4 (very engaged). The inter-rater reliability measured by Fleiss Kappa was $\kappa = 0.24$, which was calculated by equation [1] and [2] (Fleiss, 1971). The extent of agreement among the n raters for the i th subject was indexed by the proportion of agreeing pairs out of all the $n(n - 1)$

possible pairs of assignments. Then, the overall extent of agreement was measured by the mean of all P_i , where N is all pairs. The results suggested a fair agreement of judgment across labelers (Landis & Koch, 1977). The annotation scores were averaged over four labelers in every frame. But due to the imbalanced amount of four level's data and the consistency of the compared study, the data was divided into high or low levels of engagement by the threshold at the score of 3 (high ≥ 3 and low < 3). The high and low engagement bins were 83198 (47.22%) and 92976 (52.78%).

$$P_i = \frac{1}{n(n-1)} \sum_{i=1}^k n_{ij}(n_{ij} - 1) \quad [1]$$

$$\bar{P} = \frac{1}{N} \sum_{i=1}^N P_i \quad [2]$$

On the other hand, the help-seeking state was obtained by every click of the hint button. Since the same website was used this study and the previous study, the definition of timing is also the same. The help-seeking state is the 4 to 0 seconds before they clicking the hint button, but we discarded the 1 to 0 seconds since their clicking behavior might influence their facial and head data. There are 1526 (44.37%) samples of “help-seeking state” and 1913 (55.63%) samples of “working state”, which we randomly picked up other 3-second intervals. The samples of the help-seeking state was less than the working state because some timings of clicking the hint buttons are too close, but the overlapped frames would not be counted repeatedly.

3.5 Feature Engineering

The features from facial video are extracted by OpenFace 2.0 (Baltrusaitis et al., 2018). It detects a face from every video frame, and then mark the boundaries of eyes, eyebrows, and mouth as landmarks. The degree of facial muscle activities are concluded to Action Units (AUs) (Ekman et al., 1978) by analyzing the position changes of facial landmarks. OpenFace is able to extracted 18 AUs by their presence (0 and 1) and strength (0 to 5, except for AU28). The description of 18 AUs are explained in Table 8. Besides, OpenFace also detect the head's position in three axes and rotation angles: pitch, yaw, and row. The gaze data is also detected by OpenFace, including the direction vector, radians in world coordinates for the left and right eye, and the left-right and up-down angles.

Table 8 Description of 18 AU features that can extract by OpenFace

AU	Description	AU	Description
1	Inner Brow Raiser	14	Dimpler

2	Outer Brow Raiser	15	Lip Corner Depressor
4	Brow Lowerer	17	Chin Raiser
5	Upper Lid Raiser	20	Lip stretcher
6	Cheek Raiser	23	Lip Tightener
7	Lid Tightener	25	Lips part
9	Nose Wrinkler	26	Jaw Drop
10	Upper Lip Raiser	28	Lip Suck
12	Lip Corner Puller	45	Blink

AUs and head pose features are composed of three feature sets, which were also used in the previous studies. The details of feature sets are summarized in Table 9. The statistics and the distribution of a feature are calculated in 0.5-second time window.

Table 9 The summary table of the three feature sets

Name	Raw Value	Statistics or Equations	Total Features
Basic AUs	intensity of 17 AUs and presence of 18 AUs	mean, median, standard deviation, minimum, maximum, and range	$(17+18) \times 6 = 210$ features
Head Pose	three coordination of head (x,y,z axes) and three rotation angles (pitch, yaw, row)	mean, median, standard deviation, minimum, maximum, and range	$(3+3) \times 6 = 36$ features
Co-Occurring	similarity of every AU pair from the 17 AU intensities	Jensen-Shannon divergence equation	$17 \times (17-1) / 2 = 136$ features
Gaze	gaze direction vector in world coordinates for the left and right eyes, the direction in radians averaged for both eye, and the left-right and up-down angles	mean, standard deviation	$8 \times 2 = 16$ features

3.6 Machine Learning and SHAP analysis

The previous study suggested that the Light Gradient Boosting Machine (LightGBM) is more effective than a shallow machine learning model, support vector machine

(SVM). LightGBM is built in a gradient-boosting framework that uses tree-based learning algorithms and is known as a fast method for training (Ke et al., 2017). The current study trained LightGBM for each feature set (Basic AUs, Head Pose, and Co-occurring AUs) to predict the engagement and help-seeking states separately. The models were evaluated by 5-fold cross-validation, which divided 80% data for training and 20% data for testing. The data are pooling from all participants and randomly selected to divide into training or testing group.

Following the previous studies (Bosch & D'Mello, 2021; Li et al., 2021), we used Area Under the curve of Receiver Operating Characteristic, the F_1 score, and the rate of correct judgement (accuracy). The ROC curve shows the performance of a classification model at all classification threshold. The straight line connecting (0,0) and (1,1) in the graph of ROC curve showed a random classification results. In contrast, the line connecting (0,1) and (1,1) showed a perfect classification results. Therefore, the AUC (Area Under the Curve) varies between 0.5 (random classification) to 1 (perfect classification). The chance level of AUC score is 0.5. The F_1 score is the harmonic mean of precision and recall. Accuracy is the proportion of the frames classified in the correct label in all classified frames.

The comparison are explained by Shapley Additive exPlanations (SHAP) analysis (Lundberg et al., 2019). SHAP is an explainable AI tool to help researchers and engineers to examine the machine learning model and has been widely used (Bai et al., 2023; Ikeda et al., 2022; Miao et al., 2023). The strong advantage of SHAP analysis is to estimate the important features ranking by its algorithm (Belle & Papantonis, 2021). In order to compared with the previous study, we focused on the Basic AUs and Head Pose feature sets. The SHAP values were calculated by selecting the split with highest AUC result among 5-fold cross-validations. The SHAP figures, including a bar plot and a scatter plot, are generated to help us study the contribution of AU and head pose features associated with the engagement state and help-seeking state. The bar plots showed the mean absolute value of features' SHAP value. The length of the bar showed the effect of a feature on the estimation. The scatter plots showed the SHAP value distribution of the features, with each point representing the SHAP value of an estimation on mental states.

4. Estimating Mental State with Taiwan's data

The results are divided into three sessions for the three research question from session 4 to session 6. Firstly, we examine the results of estimating the engagement and help-seeking states by using facial expression on Taiwan's participants. Secondly, we compared the data from Taiwan's with Japan's data. Finally, as the classification results were different, we used Action Unit to classify nationalities.

4.1 Behavioral results

The average time to complete the problem was 51.8 minutes (3109 seconds) (the longest was 87.5 minutes; the fastest was 18.4 minutes). The average score is 7.07 points (the highest was 8.5; the lowest was 4.5) with 10 questions. If an answer is almost correct but has some grammar error, it will be counted as 0.5 points. The average clicks of the hint buttons is 35.9 times (the most was 77 times; the least was 7 times).

The correlations between the completion time, score, and the clicks of the hint buttons are calculated and the matrix are shown in Table 10. The correlation results showed that the completion time and the clicks of the hint buttons has positive significant correlation, which indicates that the longer the learners use the website, the more frequency they tend to use the hint buttons. But the completion time, and score has no significant correlations.

Table 10 Pearson's correlation coefficient between behavioral variables

	Completion time	Score	Hint button clicks
Completion time	-	-	-
Score	0.216 ($t(19)=0.96, p=0.34$)	-	-
Hint button clicks	0.542 ($t(19)=2.81, p<0.05$)	0.301 ($t(19)=1.38, p=0.18$)	-

4.2 Classification of the engagement state

The performance of classifying the engagement states by LightGBM is summarized in Fig. 29 and Table 12. The results for the feature sets were evaluated by F_1 , AUC, and accuracy. The feature sets includes Basic AUs feature, Head Pose features, Co-occurring features, and Gaze features. In total, nine kinds of feature sets were trained by machine learning models. These values are the average values of the five-fold cross-validation. The F_1 was between 0.58 to 0.73, the AUC was between 0.67 to 0.82, and the accuracy was between 0.64 to 0.82. The results of AUCs are higher than the chance level ($t(2)= 7.635, p<0.05$). The ROC curves showed that the line are departed from the straight line between (0,0) and (1,1).

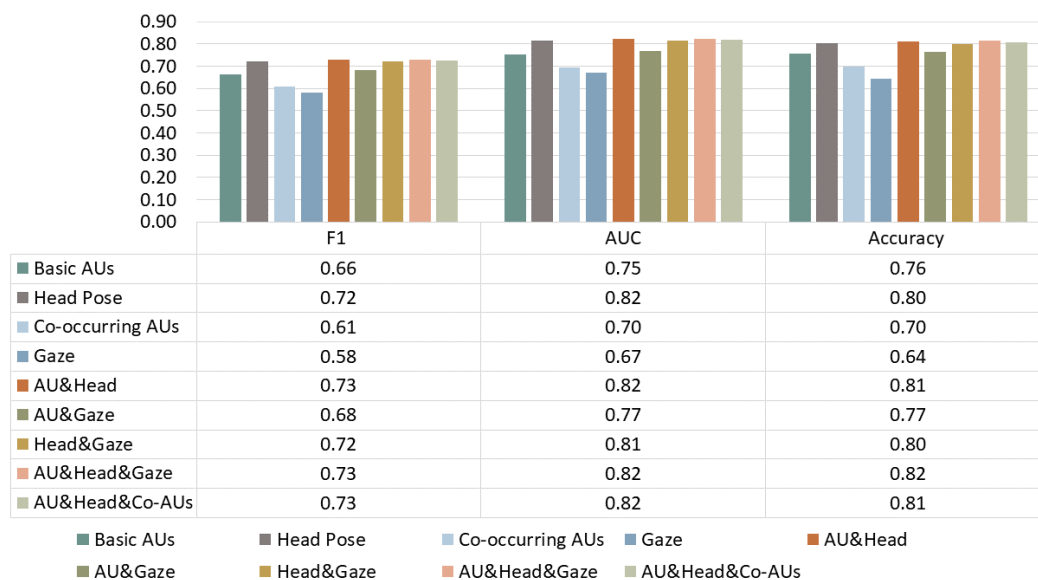
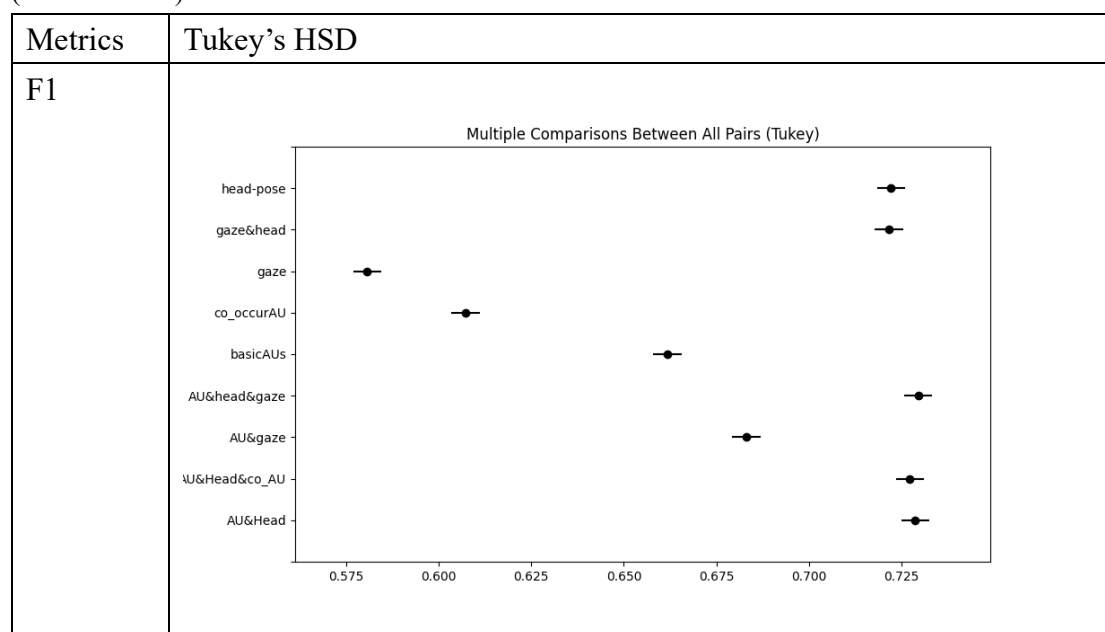


Fig. 29 The results of estimating engagement states in Taiwan’s participants by LightGBM

As for the comparison between these 9 feature sets, significant differences were on their F1 ($F(8,36)=1235.88, p<0.001$), AUC ($F(8,36)=1499.69, p<0.001$) and Accuracy ($F(8,36)=1666.81, p<0.001$). The multiple comparison by Tukey’s HSD also conducted to examine the differences. The detail of the results was shown in Table 11, which suggested that the performances of the gaze feature, co-occurring AU features were worse than others. On top of that, if two more kinds of feature set combine with each other, a higher performance will get.

Table 11 The Results of multiple comparison by Tukey’s HSD when estimate the engagement state (Taiwan’s data)



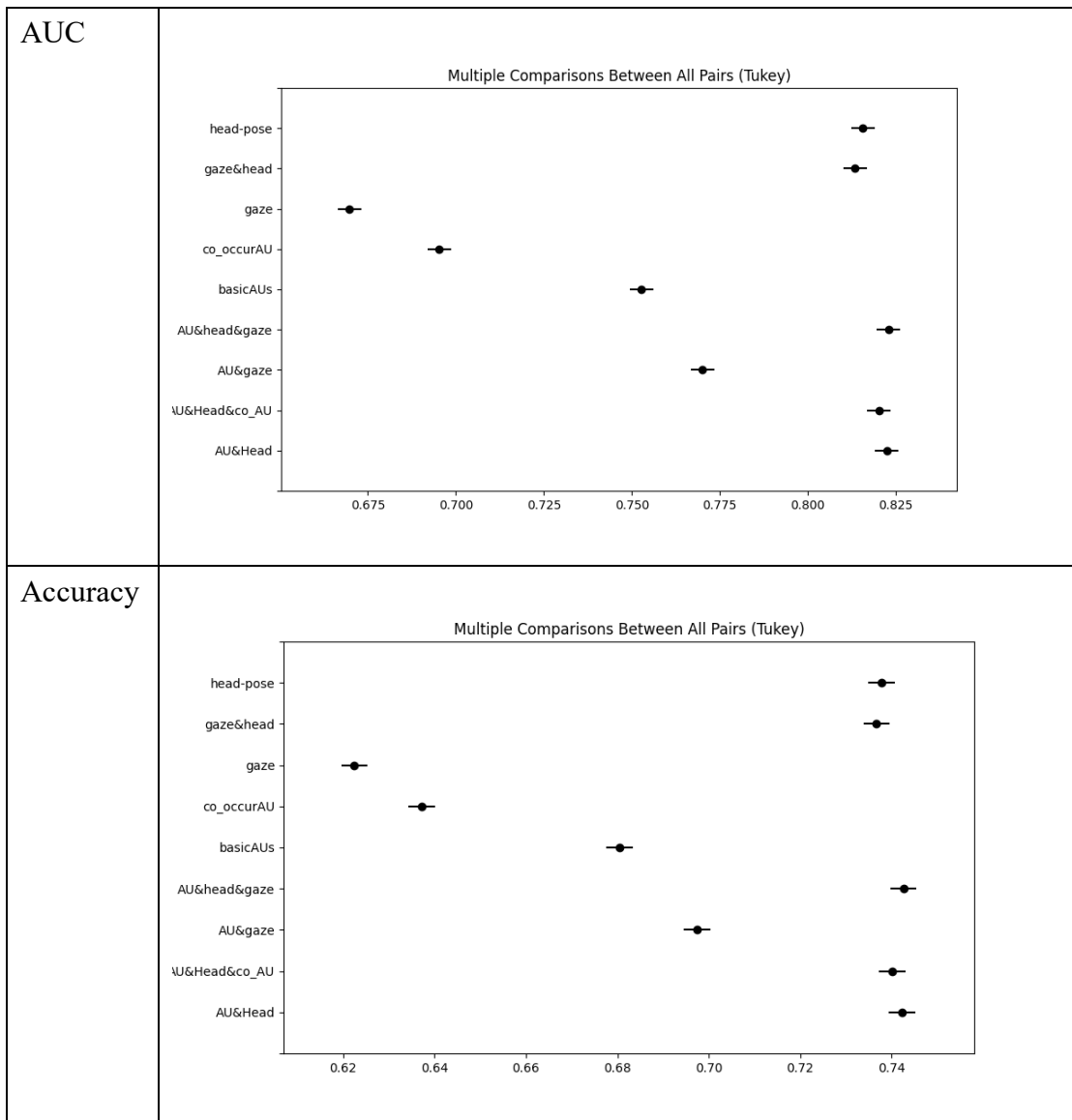
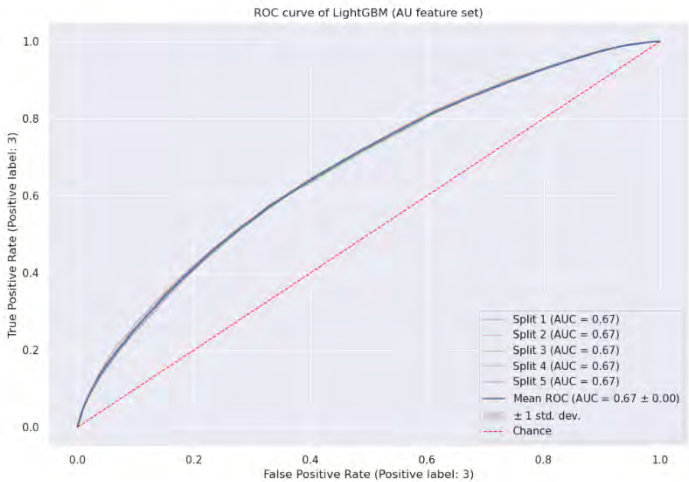
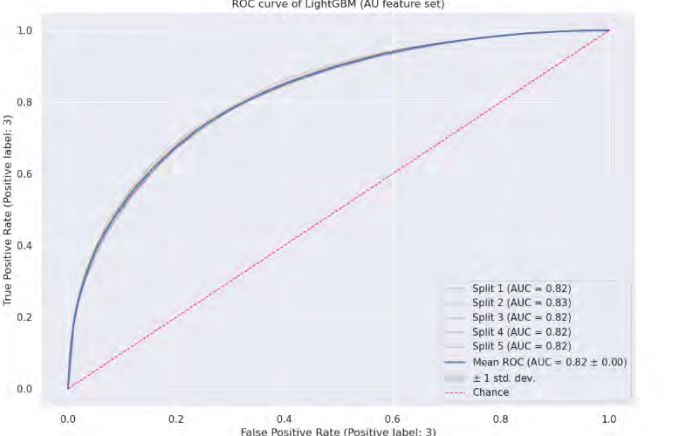
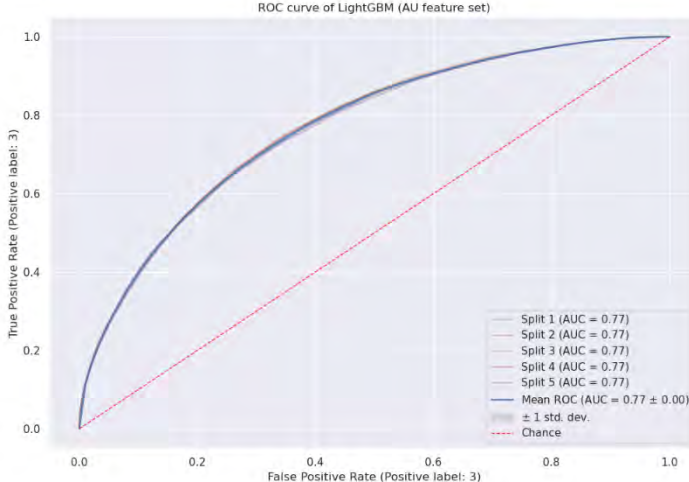


Table 12 ROC results of the engagement classification by LightGBM on multiple feature sets

Feature sets	ROC curve
Basic AUs	
Head Pose	
Co-occurring AUs	

<p>Gaze</p>	
<p>Basic AUs& Head Pose</p>	
<p>Basic AUs&Gaze</p>	

<p>Head Pose & Gaze</p>	
<p>Basic AUs&Head&Gaze</p>	
<p>Baisc AUs&Head&Co-AUs</p>	

In order to understand the facial expression and the head pose, here shows the SHAP

analysis of the Basic AU feature set and the Head Pose feature set. The SHAP analysis estimated the importance of the features in the LightGBM model with the best AUC score from the 5-fold cross-validation with the Basic AUs feature set (Figure 2) and Head Pose feature set (Figure 3). The left panel shows absolute SHAP values, which indicate the total contribution value from all samples. The right panel of Figure 2 shows the points corresponding to a SHAP value of each estimation.

The top ten important features in Basic AU feature set are shown in the figure, which suggested that the top three important features are “AU05_c_mean”, “AU04_c_mean”, and “AU04_r_max”, which are related to facial expressions of “upper lid raiser” and “brow lowerer”. Those are facial expressions focus on the upper part of the face.

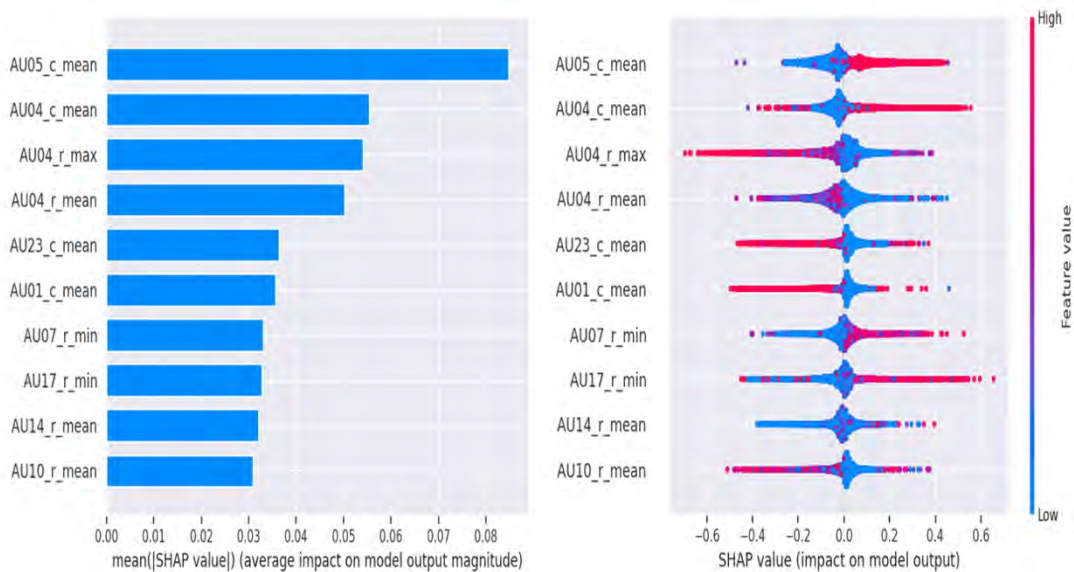


Figure 2 SHAP summary plots of LightGBM classifying engagement state by the Basic AU feature set. The left panel showed the mean absolute SHAP value indicating the total contribution of AU features and the right panel showed the points corresponding to a value of an estimation on engagement. The suffix “_c” indicates the presence of the AU; “_r” indicates the intensity of the AU; “_min” indicates minimum; “_max” indicates maximum.

In addition, the SHAP analysis also estimated the important features from Head Pose feature set trained by LightGBM model. The split with best AUC from the 5-fold cross-validation. The Figure 3 shows the top 10 important features. The top three important features are “pose_Ty_min”, “pose_Tz_min”, and “pose_Tx_range”, which suggested that the position of head was more important than the rotation of head.

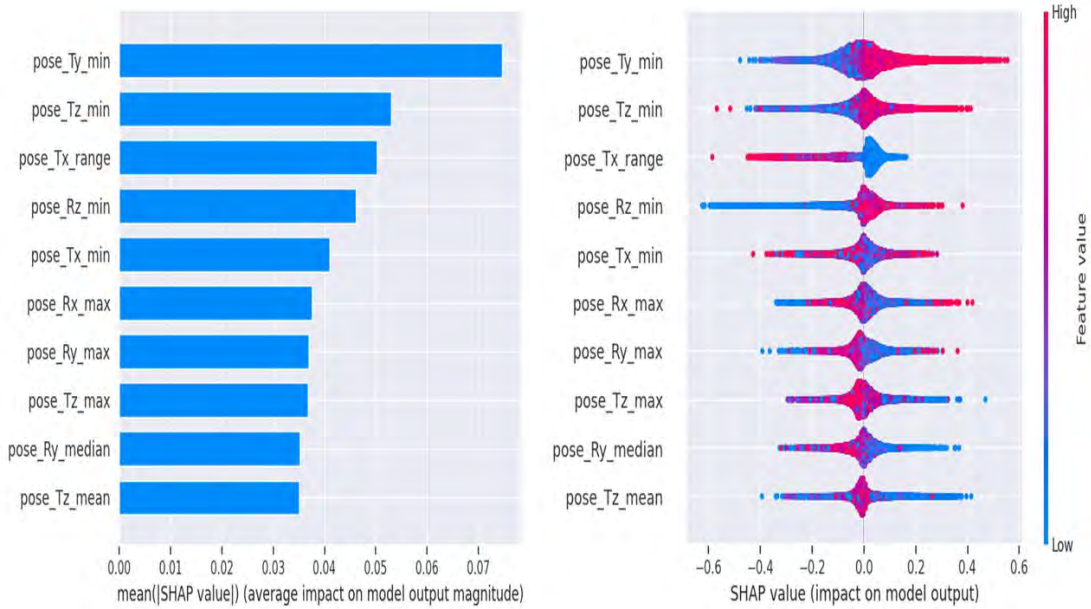


Figure 3 SHAP summary plots of LightGBM classifying engagement state by the Head Pose feature set. The left panel showed the mean absolute SHAP value indicating the total contribution of Head Pose features and the right panel showed the points corresponding to a value of an estimation on engagement. The suffix “Tx/Ty/Tz” indicates the location of the head with respect to the camera in millimeters; the suffix “Rx/Ry/Rz” indicates rotation that in radians around X,Y, Z axes, which can be seen as pitch, yaw, and roll. The suffix also represents the descriptive statistics we calculated; “min” indicates minimum; “max” indicates maximum.

4.3 Classification of the help-seeking state

The performance of classifying the help-seeking states by LightGBM is summarized in Fig. 30 and Table 13. The results for the nine feature sets (Basic AUs, Head Pose, Co-occurring AUs, Gaze, Basic AUs & Head Pose, Basic AUs & Gaze, Head Pose & Gaze, Basic AUs & Head Pose & Gaze, and Basic AUs & Head & Co-occurring AUs) were evaluated by F_1 , AUC, and accuracy. These values are the average values of the five-fold cross-validation. The F_1 was between 0.69 to 0.93, the AUC was between 0.78 to 0.98, and the accuracy was between 0.70 to 0.93.. The ROC curves showed that the line are departed from the straight line between (0,0) and (1,1).

According to one-way ANOVA analysis conducting on every metrics, there were significant differences across the nine feature sets on F1 score ($F(8,36)=343.09$, $p<0.001$), AUC($F(8,36)=319.14$, $p<0.001$), and accuracy($F(8,36)=260.15$, $p<0.001$). The multiple comparison by Tukey’s HSD are shown in Table 14. The results suggested that the performance of Co-occurring AUs and Gaze feature sets were worse than other feature sets, but the performance of Head Pose feature sets was better than others. The fusion dataset which contains head pose features also performed well than others.

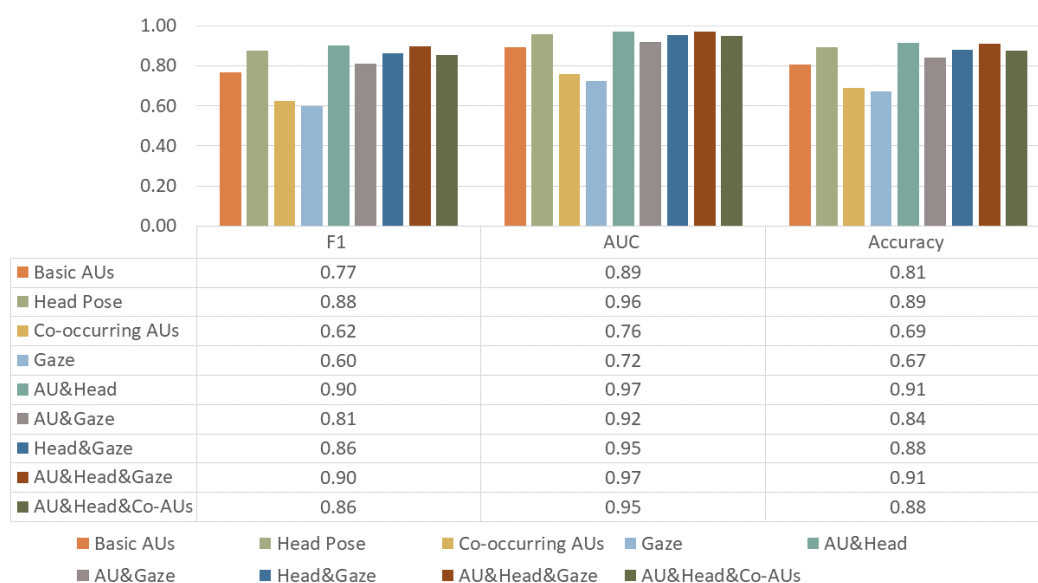
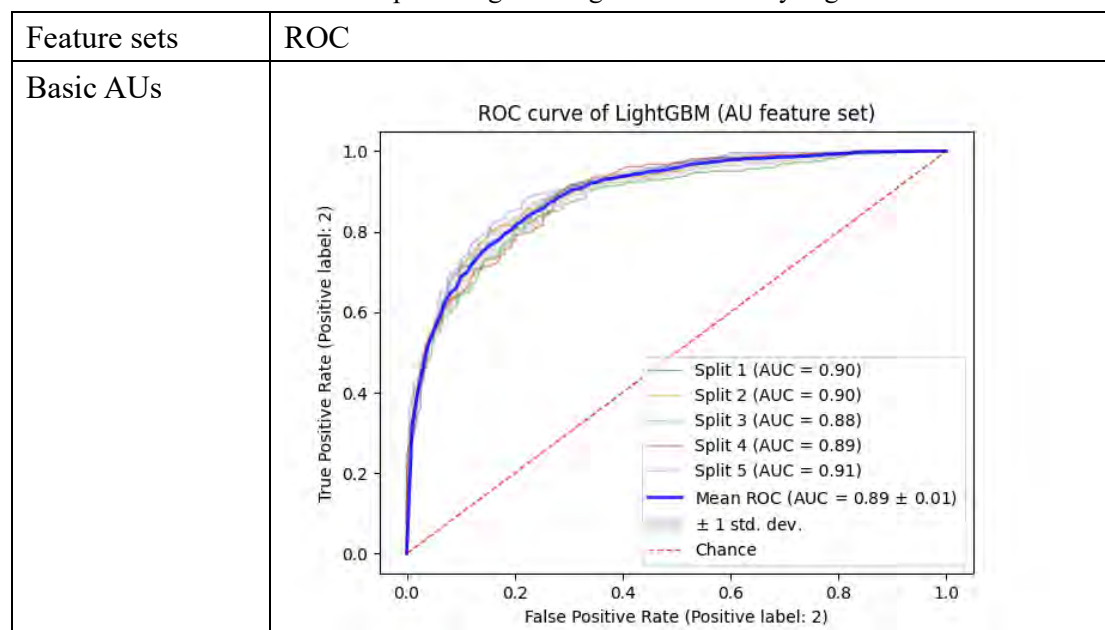
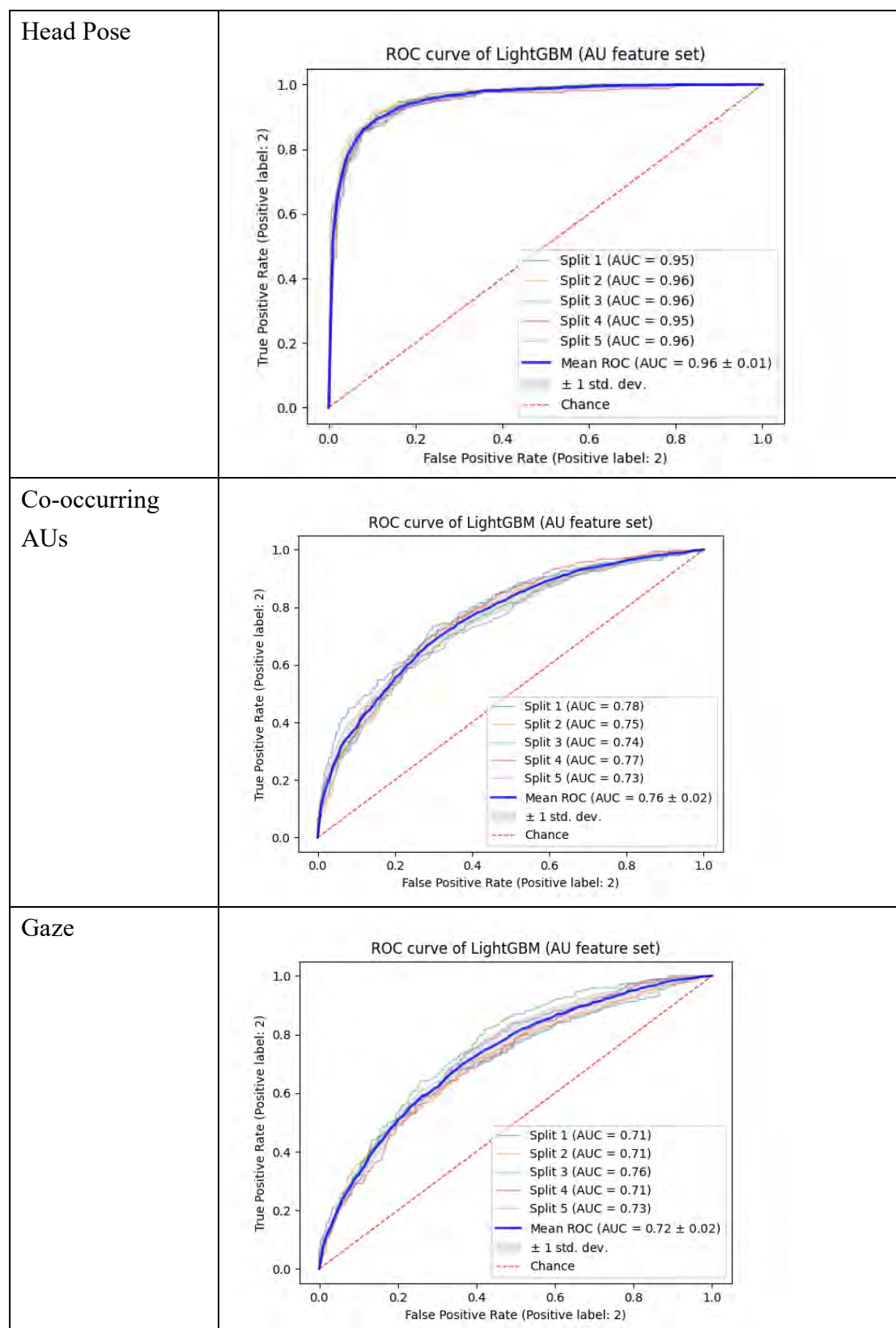


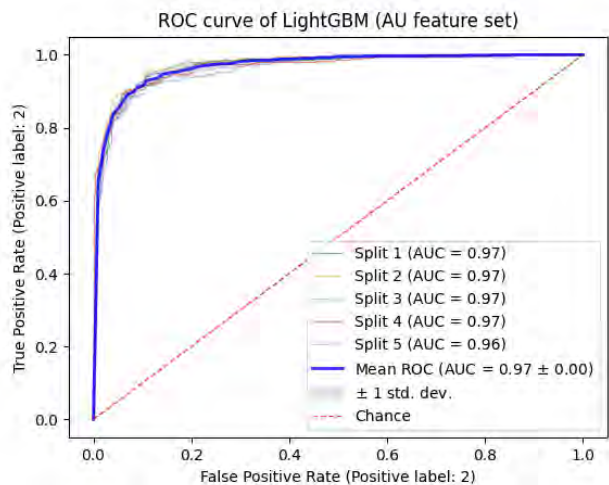
Fig. 30 The results of estimating help-seeking states in Taiwan’s participants by LightGBM

Table 13 ROC curve results of help-seeking/working classification by LightGBM

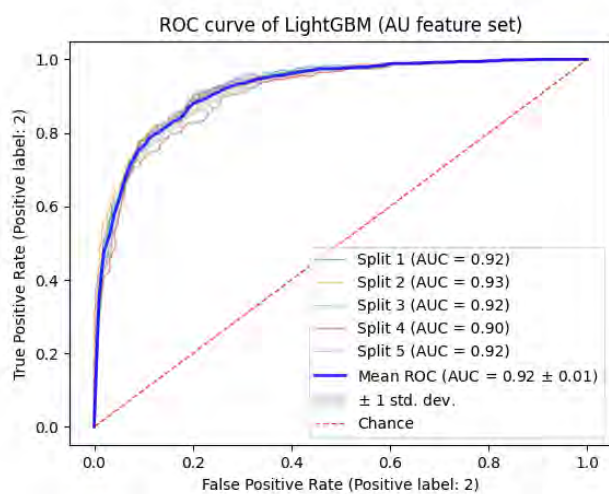




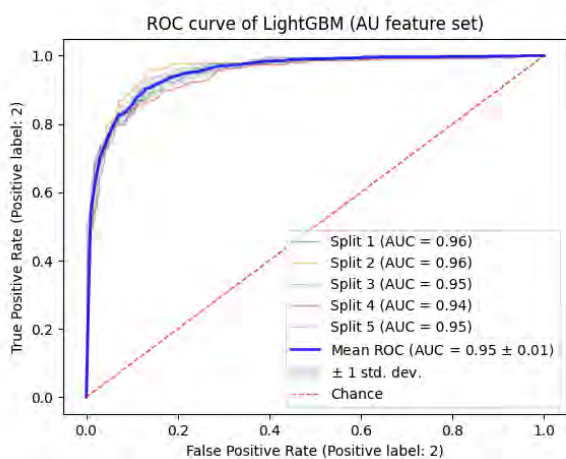
AU&Head



AU&Gaze



Head&Gaze



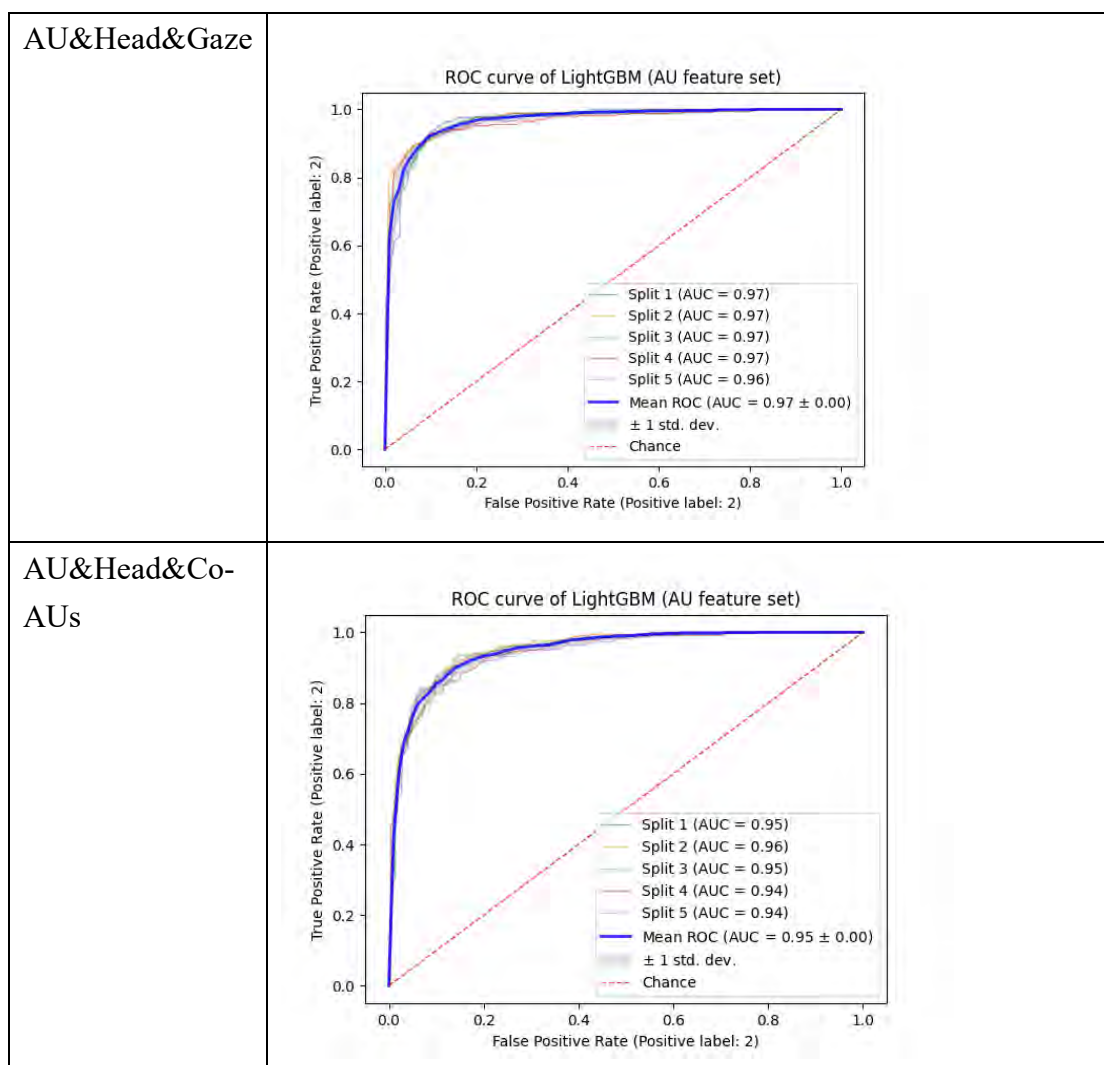
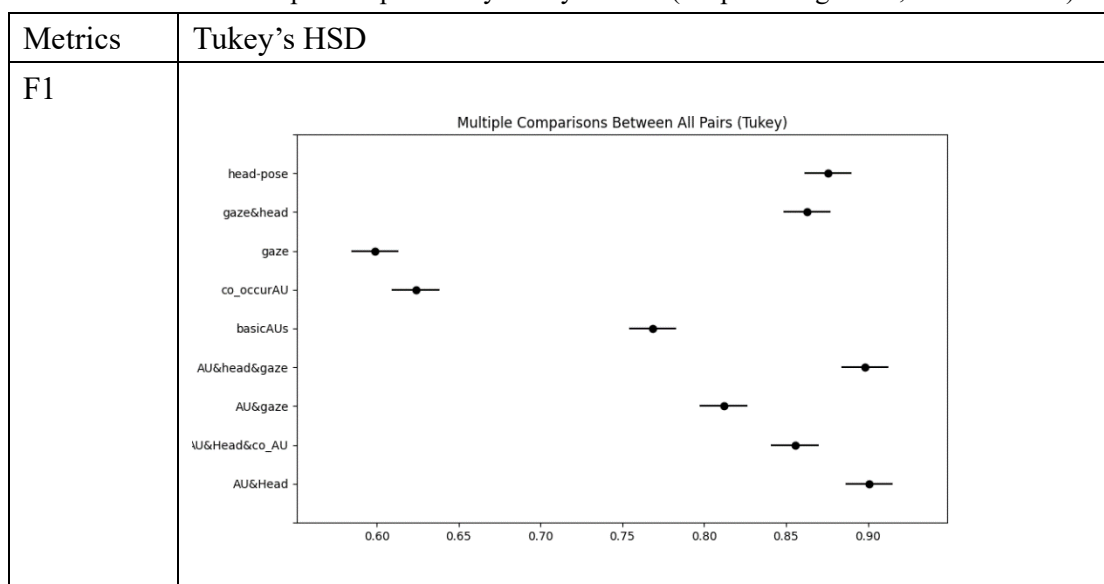
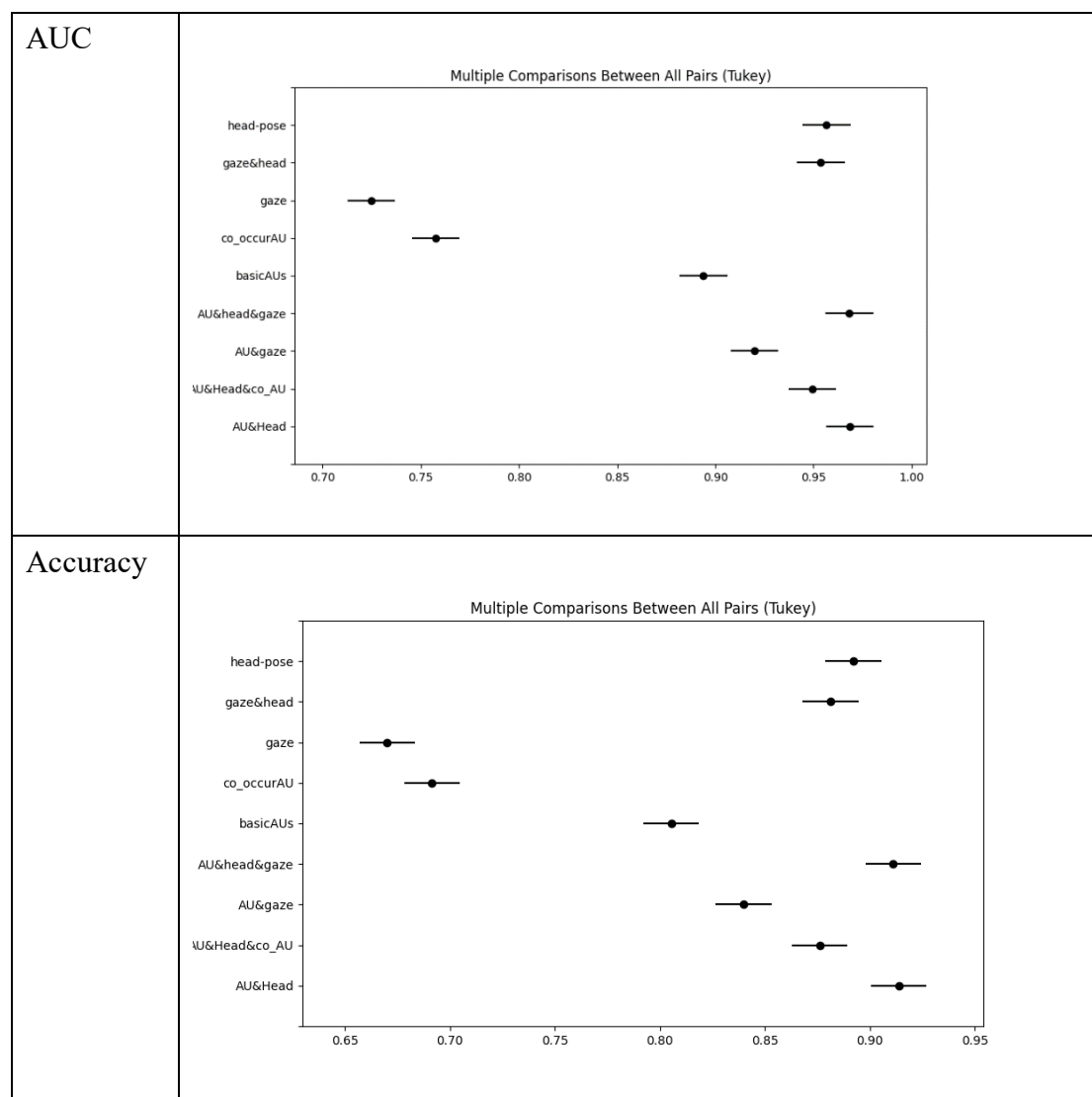


Table 14 Results of multiple comparison by Tukey’s HSD. (Help-seeking states, Taiwan’s data)





The SHAP analysis estimated the importance of the features in the LightGBM model with the best AUC score from the 5-fold cross-validation with the Basic AUs feature set. The left panel of Figure 4 shows absolute SHAP values, which indicate the total contribution value from all samples. The right panel of Figure 4 shows the points that corresponding to the SHAP value of each estimation. The SHAP results shows that “AU25_r_min”, “AU15_c_mean”, and “AU25_r_std” are the top three important AU features in the model within the top ten important features which are shown in the bar plot. The AU25 is related to the facial expression of “lips part”, and the AU15 indicates the facial expression of “lip corner depressor”, which are around the mouth.

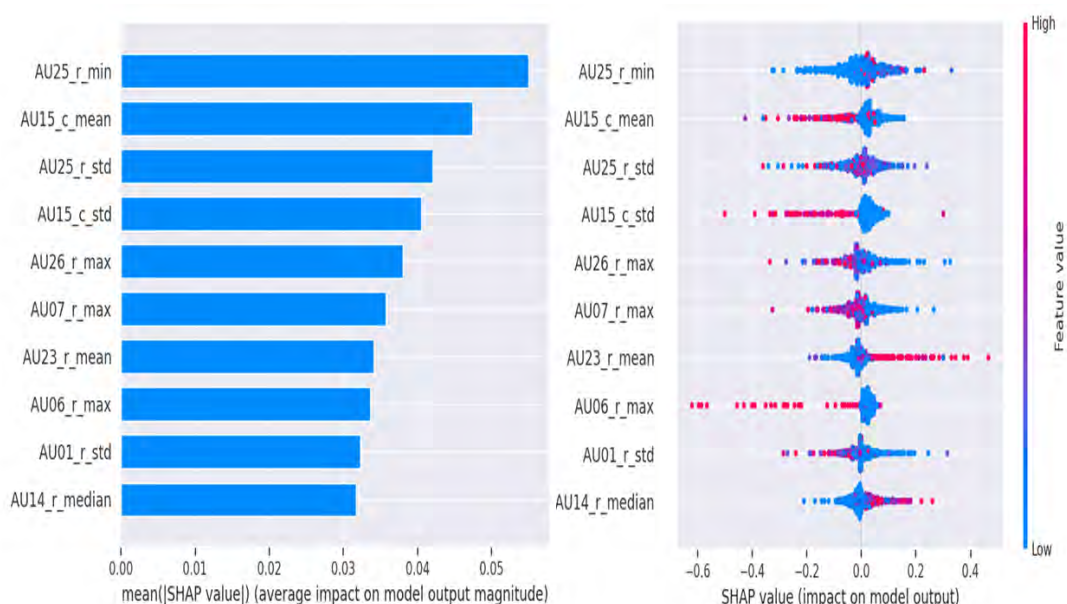


Figure 4 SHAP summary plots of LightGBM classifying help-seeking state by the Basic AU feature set. The left panel showed the mean absolute SHAP value indicating the total contribution of AU features and the right panel showed the points corresponding to a value of an estimation on the help-seeking state. The suffix of the features indicate the descriptive statistics we calculated; “min” indicates minimum; “max” indicates maximum; “std” means standard deviation. Besides, “c” indicates the presence of the AU; “r” indicates the intensity of the AU.

In addition, the important features of Head Pose feature set were also calculated by SHAP analysis. The split trained by LightGBM model with the best AUC score from the 5-fold cross-validation was estimated. The Figure 5 showed the top ten important features. It suggested the top three important features were “pose_Tx_min”, “pose_Rz_min”, and “pose_Rz_max”, which were related to the rotation of the head pose.

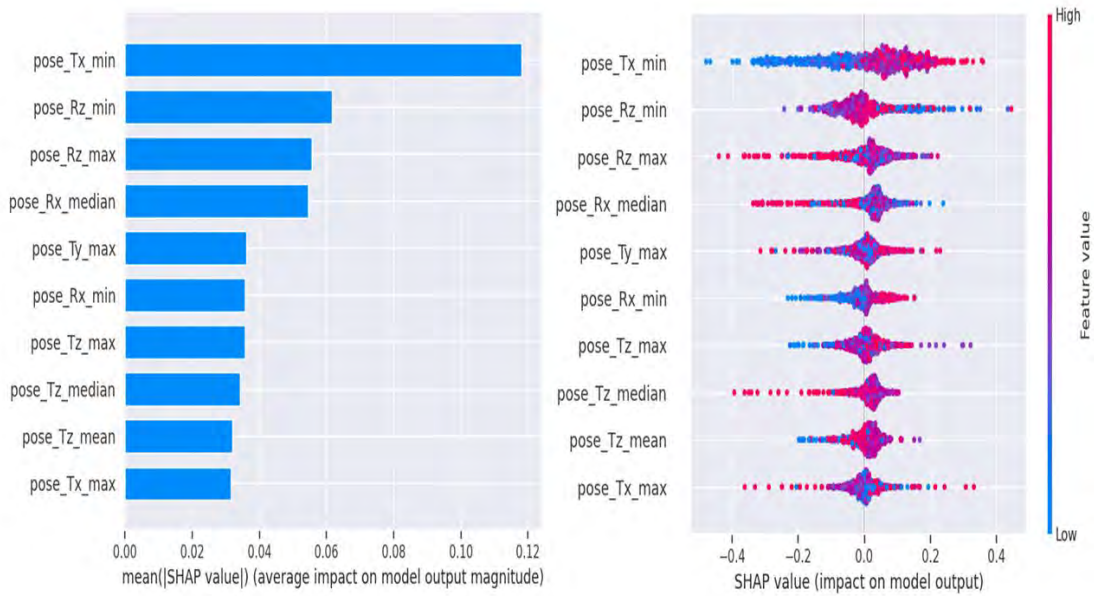


Figure 5 SHAP summary plots of LightGBM classifying help-seeking state by the Head Pose feature set. The left panel showed the mean absolute SHAP value indicating the total contribution of Head Pose features and the right panel showed the points corresponding to a value of an estimation on the help-seeking state. The suffix “Tx/Ty/Tz” indicates the location of the head with respect to the camera in millimeters; the suffix “Rx/Ry/Rz” indicates rotation that in radians around X,Y, Z axes, which can be seen as pitch, yaw, and roll. The suffix also represents the descriptive statistics we calculated; “min” indicates minimum; “max” indicates maximum.

5. Comparison between Taiwan’s data and Japan’s data

The Japan’s data comes from the previous study (Wang, Hatori, et al., 2023). In this session, firstly, we compared the behavioral data, including the completion time, score, and the times of clicking the hint button, between the Japan’s and Taiwan’s participants. Secondly, we compared the classification results of the engagement and help-seeking state and the results of the SHAP analysis for Basic AUs and Head Pose feature sets.

5.1 Comparison of Behavioral Results

The website we used in the current study has two language version: Japanese and traditional Chinese. We compare their behavioral data between Japan’s data and Taiwan’s data to validate that if the website has any bias on the language or not. Besides, the Japan’s data are the same participants of the previous study.

The completion time is not significantly different ($t(28)=0.80$, $p=0.40$) between Japanese (45.99 ± 16.07 min) and Taiwanese (51.82 ± 18.51 min). The score is not significantly different ($t(28)=0.12$, $p=0.89$) between Japanese (7.44 ± 1.04) and Taiwanese (7.07 ± 1.02). The clicks of the hint buttons is not significantly different ($t(28)=0.89$, $p=0.12$) between Japanese (34.77 ± 20.12) and Taiwanese (35.86 ± 20.84).

In sum, the similar behavior results between Japanese and Taiwanese suggest that the problem we used and the hints we designed were fair to the two groups of participants.

5.2 Comparison of the Classification Results

Firstly, we compared the metrics of the classification results of Taiwan's and Japan's data, including F_1 , AUC, and accuracy. The overall performance of the classification of the engagement state showed that the classifiers for Japan's data are significantly better than Taiwan's data ($t(8)=3.14, p<0.05$). In contrast, the overall performance of the classification of the help-seeking states has no significant difference between Taiwan's data and Japan's data ($t(8)=1.63, p=0.12$). The significant difference of result on estimating engagement might be caused by the sample size.

Secondly, comparing the SHAP analysis of Taiwan's and Japan's data, the rankings of important features were different. The top ten important features are shown in Table 15 and Table 16. The Kendall's tau was calculated to measure the association between the features of the two top 10 rankings. For the Basic AU feature set, there is no significant association between Japan and Taiwan's ranking of the features estimating engagement ($\tau_c = 0, p = 1.0$), and there is also no significant association between Japan and Taiwan's ranking of the features estimating the help-seeking state ($\tau_c = -0.43, p = 0.09$). For the Head Pose feature set, there is no significant association between Japan and Taiwan's ranking of the features estimating engagement ($\tau_c = -0.31, p = 0.22$), and there is also no significant association between Japan and Taiwan's ranking of the features estimating the help-seeking state ($\tau_c = 0.18, p = 0.51$).

Table 15 Rankings of Basic AU features for estimating the mental states.

	Japan	Taiwan
Engagement	AU02_c_mean	AU05_c_mean
	AU23_c_mean	AU04_c_mean
	AU04_r_mean	AU04_r_max
	AU12_r_min	AU04_r_mean
	AU14_r_min	AU23_c_mean
	AU25_r_max	AU01_c_mean
	AU04_r_max	AU07_r_min
	AU07_r_median	AU17_r_min
	AU45_r_mean	AU14_r_mean
	AU23_r_mean	AU10_r_mean
help-seeking	AU04_r_median	AU25_r_min
	AU23_r_means	AU15_c_mean

AU14_r_min	AU25_r_std
AU26_r_max	AU15_c_std
AU20_r_std	AU26_r_max
AU23_r_max	AU07_r_max
AU20_r_range	AU23_r_mean
AU45_r_mean	AU06_r_max
AU23_r_std	AU01_r_std
AU04_r_min	AU14_r_median

Table 16 Rankings of Head Pose features for estimating the mental states.

	Japan	Taiwan
Engagement	pose_Tx_min	pose_Ty_min
	pose_Tz_min	pose_Tz_min
	pose_Ty_max	pose_Tx_range
	pose_Tz_max	pose_Rz_min
	pose_Rx_min	pose_Tx_min
	pose_Tz_median	pose_Rx_max
	pose_Rz_min	pose_Ry_max
	pose_Rx_max	pose_Tz_max
	pose_Ty_min	pose_Ry_median
	pose_Ry_max	pose_Tz_mean
help-seeking	pose_Ty_min	pose_Tx_min
	pose_Tz_mean	pose_Rz_min
	pose_Tx_min	pose_Rz_max
	pose_Tx_std	pose_Rx_median
	pose_Ry_min	pose_Ty_max
	pose_Rx_max	pose_Rx_min
	pose_Tz_min	pose_Tz_max
	pose_Ty_mean	pose_Tz_median
	pose_Ry_mean	pose_Tz_mean
	pose_Tx_max	

6. Discussion

In this study, we found that the estimation of the help-seeking states had cultural differences in the SHAP contribution values. The extended participants of this study revealed the cultural effect on models. Furthermore, the estimation of the help-seeking states was related to the AUs that around the mouth, while the estimation of the engagement states was related to the AUs taht around the eyes.

The cultural differences on the facial expression focused on around the mouth. In this study, the SHAP values showed the contributions of each feature when estimating the help-seeking states were different between Japan's and Taiwan's people. The values revealed the AU features such as "AU25 (lips part)" were contributed more in Taiwan's estimation but AU features such as "AU23 (lip tightener)" were contributed more in Japan's estimation. A previous study showed that Taiwanese facial expressions on the mouth were more dramatic (e.g. show their tooth) than Japanese (Kaminosono et al., 2022). Opening one's lips was a more dramatic expression than tightening one's lips. The differences of the expression related to AU25 and AU23 might indicated the different ways to express the mental states when the learners face difficulty in the different cultures. However, this results was different from the previous study showed the cultural differences could be observed on the facial expression of eye brow furrow (McDuff et al., 2017). The results showed that the upper part of face had cultural effect.

On the other hand, the features around upper face were important for estimating the engagement states. Although the difference between Japan's and Taiwan's data was not observed, the common features, including "AU_07_mean", "AU04_r_max", and "AU04_r_mean", were focus on the expression near the eyes. The AU07 (lid tightener) and AU04 (brow lowerer) should be important indicators when estimating the engagement states. However, McDuff et al. (2017) found the differences on the eyebrow between different cultures, but they used advertisements on television to the participants. Their task is different from our study. In this study, we used a serious task that asked participants to complete it. A previous study showed that when people were watching serious videos, they smile less, lower their eyebrow, and apart or depress their lips (McDuff & Kaliouby, 2017). To be specific, a study, which focused on estimating learners' engagement state when learners were learning with online videos, revealed that AU09 (nose wrinkle, but also shows slight eyebrow furrow and lip raiser) was an important feature (Miao et al., 2023). In addition, AU01, AU04, AU14, AU17, AU23, AU25, AU45 were selected in estimating the engagement state in a previous study (Li et al., 2021), which were highly overlapped with our important features. There were many different results from the previous studies since the tasks were different, but in summary, the participants serious attitude and their engagement were reflected by their eyebrow and lips.

As for the help-seeking estimation, the features that around lower face were important in both cultures. As we mentioned on the above paragraph, people watched serious contents with less smile and lips will apart or depress. Another study also revealed that when people felt unsure, their lips would apart (el Kaliouby & Robinson, 2005). The common features of both cultures, including “AU26_r_max” and “AU20_r_mean”, were related to the facial expression of jaw dropping and lip stretching. The help-seeking states is the state of predicting learners’ difficulties. The results revealed that before they take the action to ask, the facial expression of the lower part would express more at first.

The current study used SHAP value to estimate the importance of the features and indicated the cultural differences by using the SHAP values. As an explanation tool for machine learning model, it helped us analyze the contribution of the features. To best of our knowledge, few studies used the SHAP analysis on the mental state estimations. In contrast, most of the research that estimate the mental states used metrics including F1, AUC and accuracy to evaluate the machine learning models (Li et al., 2021; Sümer et al., 2021; Whitehill et al., 2014). But the contributions of every feature were not quantified. In this study, the SHAP values were used to explained the results and we found what were the important facial features when estimating the mental state of engagement and help-seeking.

Another important discovery of this current study is to compare the learners from more than a single cultural background, and found the cultural differences were exist when interacting with an ITS. Participants in previous studies that estimate students’ mental state usually from single culture (Desai et al., 2020; Hasegawa et al., 2020; Pellet-Rostaing et al., 2023; Peng & Nagao, 2021). However, it should be noted that when the system was implemented into education, the data should be adapted to different cultures. Therefore, the system can be improved and be benefit to learners.

Chapter 5: Inter-person Learning models on Estimating the mental states

1. Introduction

The below machine learning models were based on intra-person learning, which make a same person's data be in the training dataset and the testing dataset at the same time. Although this thesis used 5-fold cross-validation to estimate a stable results, the machine still in somewhat "know" the person's features since it has seen that in the training phrase before the testing phrase. Besides, if we want to identify the nationality by the facial expressions, we cannot split one person's data into pieces. The inter-person learning method is needed for more flexible estimations.

Therefore, the current thesis used inter-person learning to conduct the analysis. This chapter aims to estimate the mental states, including the engagement states and the help-seeking states, by using inter-person learning. Furthermore, the identification of nationality also conducted by inter-person learning in this chapter.

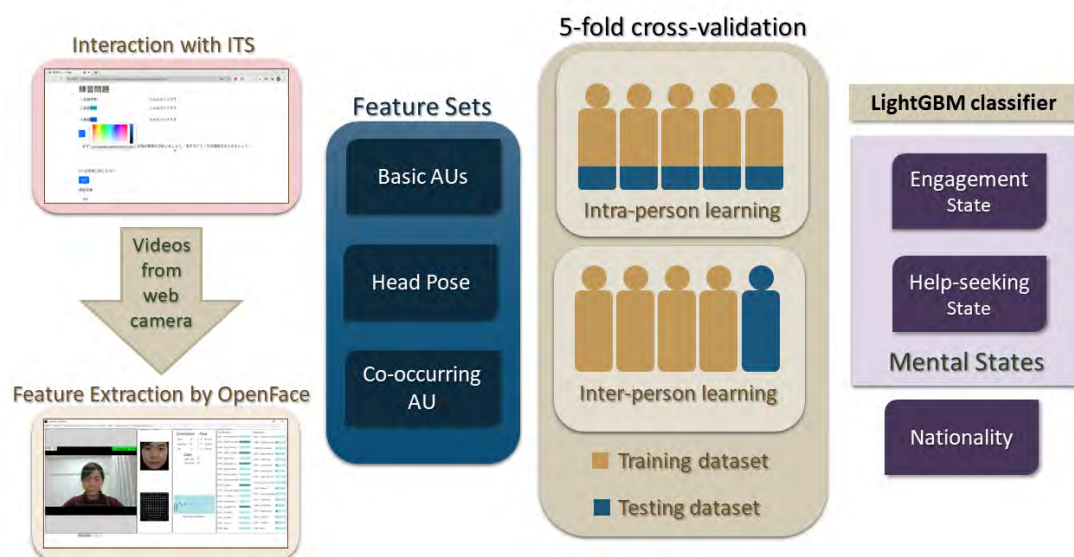


Fig. 31 The framework of intra-person learning and Inter-person learning. This chapter focuses on inter-person learning. By this approach, we can examine that whether the machine learning model can apply how it learned from facial expression and head pose to a new person or not.

2. Methods

2.1 Participants and materials

The participants are 9 Japan's students from Tohoku University and 21 Taiwan's

students from National Taiwan University. They were the same as the chapter 3 and chapter 4. The website of the experiment and videos taken during the experiments were also the same.

2.2 Feature Engineering

The features from facial video are extracted by OpenFace 2.0 (Baltrusaitis et al., 2018). It detects a face from every video frame, and then mark the boundaries of eyes, eyebrows, and mouth as landmarks. The degree of facial muscle activities are concluded to Action Units (AUs) (Ekman et al., 1978) by analyzing the position changes of facial landmarks. OpenFace is able to extract 18 AUs by their presence (0 and 1) and strength (0 to 5, except for AU28). The description of 18 AUs are explained in Table 8. Since the individual differences on the facial expression might influence the model estimation, in this chapter the values of AU were normalized. The z score of each value in the sample, relative to the sample mean and standard deviation was computed by SciPy API.

Besides, OpenFace also detect the head's position in three axes and rotation angles: pitch, yaw, and roll. There are 6 indexes of head information.

Table 17 Description of 18 AU features that can extract by OpenFace

AU	Description	AU	Description
1	Inner Brow Raiser	14	Dimpler
2	Outer Brow Raiser	15	Lip Corner Depressor
4	Brow Lowerer	17	Chin Raiser
5	Upper Lid Raiser	20	Lip stretcher
6	Cheek Raiser	23	Lip Tightener
7	Lid Tightener	25	Lips part
9	Nose Wrinkler	26	Jaw Drop
10	Upper Lip Raiser	28	Lip Suck
12	Lip Corner Puller	45	Blink

AUs, head pose, and gaze features are composed of the unimodal feature sets, which were also used in the previous studies. The details of feature sets are summarized in Table 9. The statistics and the distribution of a feature are calculated in 0.5-second time window. Besides, we also trained the model by using multimodal feature set, which contains the “AU+gaze”, “AU+Head pose”, “AU+Head+Gaze”, and “AU+Head+Co-AU”. The details of the SHAP analysis were showed in appendix.

Table 18 The summary table of the unimodal feature sets

Name	Raw Value	Statistics or Equations	Total Features
Basic AUs	intensity of 17 AUs and presence of 18 AUs	mean, median, standard deviation, minimum, maximum, and range	$(17+18) \times 6 = 210$ features
Head Pose	three coordination of head (x,y,z axes) and three rotation angles (pitch, yaw, row)	mean, median, standard deviation, minimum, maximum, and range	$(3+3) \times 6 = 36$ features
Co-Occurring	similarity of every AU pair from the 17 AU intensities	Jensen-Shannon divergence equation	$17 \times (17-1) / 2 = 136$ features
Gaze	gaze direction vector in world coordinates for the left and right eyes, the direction in radians averaged for both eye, and the left-right and up-down angles	mean, standard deviation	$8 \times 2 = 16$ features

2.3 Machine Learning and SHAP analysis

In intra-person learning, the same person's data will be contained in both training and test datasets. But if the model needs to be applied on new person that it never see, the inter-person learning should be applied.

Five kinds of inter-person learning are conducted, including:

- (1) training on 20 Taiwan's participants and testing on a Taiwan's participant, which validated for 21 times on each participant,
- (2) training on 8 Japan's participants and testing on a Japan's participant, which validated for 9 times on each participant,
- (3) training on 21 Taiwan's participants and testing on a Japan's participant, which validated for 9 times on each participant,
- (4) training on 9 Japan's participants and testing on a Taiwan's participant, which validated for 21 times on each participant,
- (5) training on 29 Taiwan and Japan's participants and testing on a participant, which validated for 30 times on each participant.

The Light Gradient Boosting Machine (LightGBM) was used to estimate the above five divisions of the training and testing dataset, since in the previous studies the LightGBM classifiers suggested better performances than the Support Vector Machine (SVM) classifiers. In addition, the LightGBM is so fast that all validations of inter-person learning model can be conducted effectively since there are many times of learning need to be conducted.

Following the previous studies (Bosch & D'Mello, 2021; Li et al., 2021), we used Area Under the curve of Receiver Operating Characteristic, the F_1 score, and the rate of correct judgement (accuracy). The ROC curve shows the performance of a classification model at all classification threshold. The straight line connecting (0,0) and (1,1) in the graph of ROC curve showed a random classification results. In contrast, the line connecting (0,1) and (1,1) showed a perfect classification results. Therefore, the AUC (Area Under the Curve) varies between 0.5 (random classification) to 1 (perfect classification). The chance level of AUC score is 0.5. The F_1 score is the harmonic mean of precision and recall. Accuracy is the proportion of the frames classified in the correct label in all classified frames.

The comparison are explained by Shapley Additive exPlanations (SHAP) analysis (Lundberg et al., 2019). SHAP is an explainable AI tool to help researchers and engineers to examine the machine learning model and has been widely used (Bai et al., 2023; Ikeda et al., 2022; Miao et al., 2023). The strong advantage of SHAP analysis is to estimate the important features ranking by its algorithm (Belle & Papantonis, 2021).

The SHAP values were calculated by every time of the testing. In order to understand the facial expression and the head pose, this chapter focused on the Basic AUs and Head Pose feature sets. The SHAP figures, including a bar plot and a scatter plot, are generated to help us study the contribution of AU and head pose features associated with the engagement state and help-seeking state. The bar plots showed the mean absolute value of features' SHAP value. The length of the bar showed the effect of a feature on the estimation. The scatter plots showed the SHAP value distribution of the features, with each point representing the SHAP value of an estimation on mental states. The value of every SHAP value also calculated to compared by t test.

3. Results

The results we reported here only contains the features of Basic AU and Head Pose. On top of that, other features, including Gaze, were also used for prediction. In order to explain and compared the features between Action Units and Head Pose, the results showed here were unimodal models. More details about multimodal models were shown in the Appendix, which represents other SHAP analysis in detail.

3.1 Estimation on the engagement state

The results of the above dataset combinations are estimated by Basic AU feature sets, and the detail is shown in Table 19. The metrics, including accuracy, AUC, and F1 score have not significantly different between the five training and testing datasets.

Table 19 The results of inter-person learning (after z-score) on estimating engagement states by the Basic AUs feature set

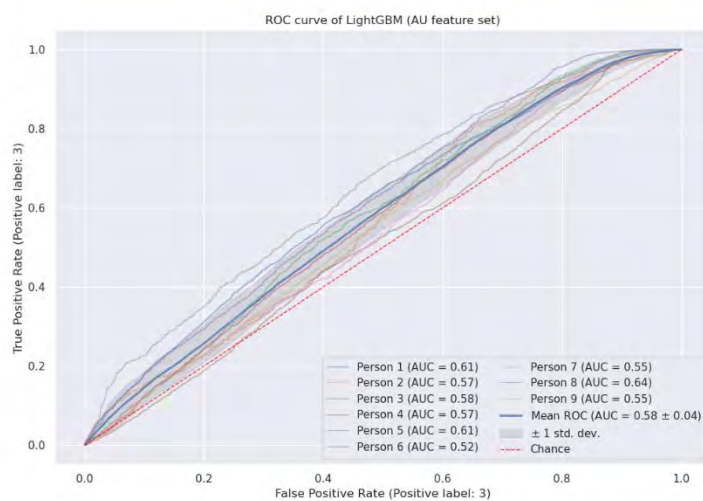
	n	Accuracy	AUC	F1
Training: Taiwan	21	0.52±0.07	0.57±0.06	0.49±0.16
Testing: Taiwan				
Training: Japan	9	0.52±0.08	0.58±0.04	0.49±0.22
Testing: Japan				
Training: Taiwan	9	0.49±0.07	0.56±0.07	0.45±0.19
Testing: Japan				
Training: Japan	21	0.53±0.08	0.54±0.13	0.58±0.13
Testing: Taiwan				
Training: JP+TW	30	0.52±0.07	0.57±0.17	0.49±0.17
Testing: JP+TW				
ANOVA results		F(4,76)=0.45, p=0.77	F(4,76)=1.47, p=0.22	F(4,76)=1.4, p=0.24

The ROC curves describe the details of every person's testing results, which are shown as following Table 20.

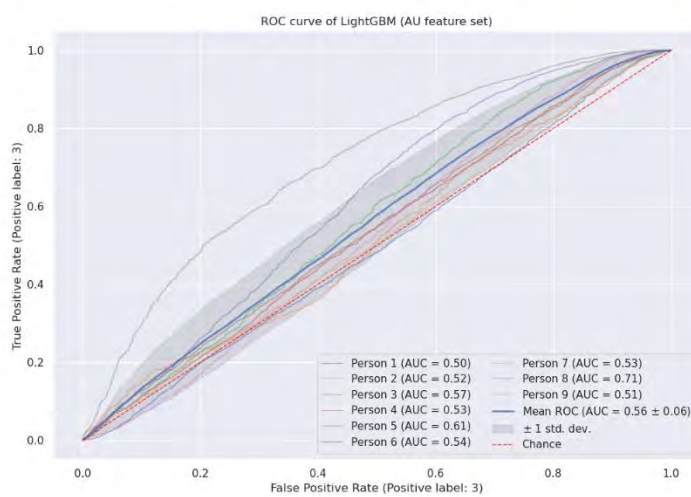
Table 20 The graphs of ROC when estimating the engagement states by Basic AU feature set

Datasets	Graphs of ROC
Training: Taiwan Testing: Taiwan	<p>ROC curve of LightGBM (AU feature set)</p> <p>True Positive Rate (Positive label: 3)</p> <p>False Positive Rate (Positive label: 3)</p> <p>Person 1 (AUC = 0.53) Person 13 (AUC = 0.56) Person 2 (AUC = 0.70) Person 14 (AUC = 0.51) Person 3 (AUC = 0.49) Person 15 (AUC = 0.58) Person 4 (AUC = 0.55) Person 16 (AUC = 0.51) Person 5 (AUC = 0.46) Person 17 (AUC = 0.53) Person 6 (AUC = 0.64) Person 18 (AUC = 0.54) Person 7 (AUC = 0.53) Person 19 (AUC = 0.54) Person 8 (AUC = 0.63) Person 20 (AUC = 0.65) Person 9 (AUC = 0.58) Person 21 (AUC = 0.58) Person 10 (AUC = 0.60) Mean ROC (AUC = 0.57 ± 0.06) Person 11 (AUC = 0.66) ± 1 std. dev. Person 12 (AUC = 0.55) Chance</p>

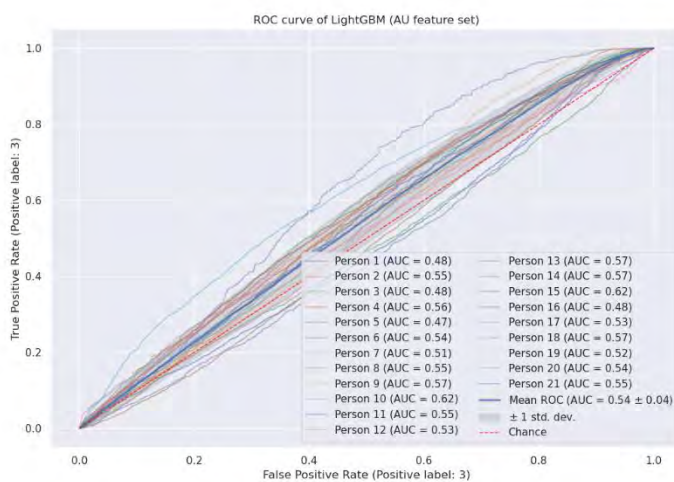
Training: Japan
Testing: Japan

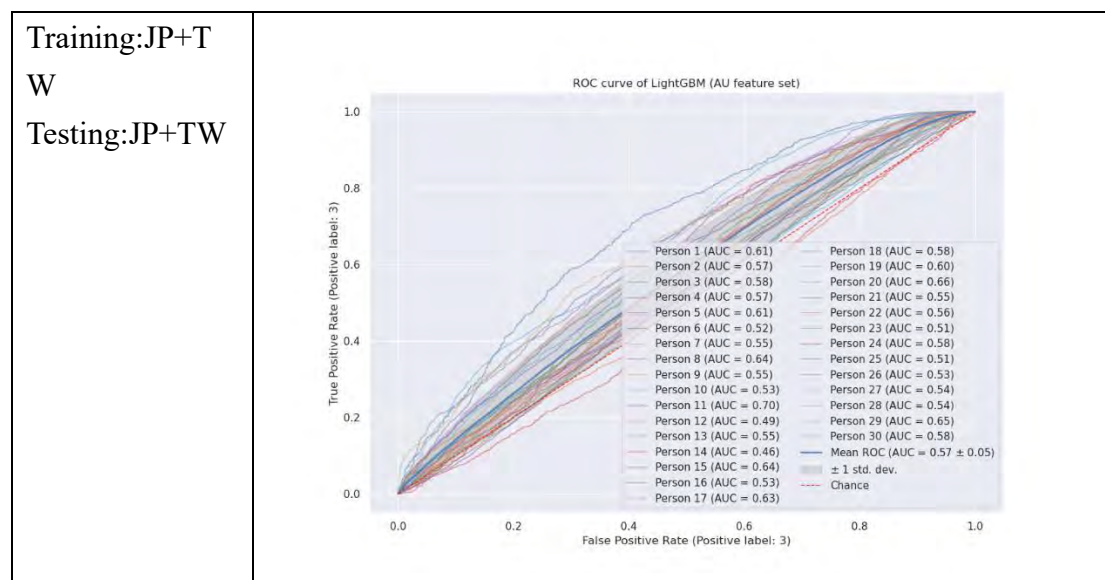


Training:
Taiwan
Testing: Japan



Training: Japan
Testing: Taiwan





3.2 Estimation on the help-seeking state

The results of the above dataset combinations are estimated by Basic AU feature sets, and the detail is shown in Table 21. The metrics, including accuracy, AUC, and F1 score have not significantly different between the five training and testing datasets.

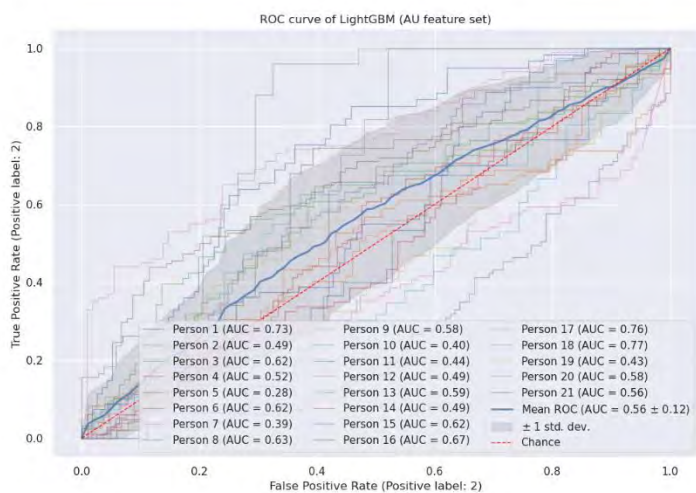
Table 21 The results of inter-person learning (after z-score) on estimating help-seeking states by the Basic AUs feature set

	n	Accuracy	AUC	F1
Training: Taiwan Testing: Taiwan	21	0.54±0.10	0.55±0.13	0.42±0.14
Training: Japan Testing: Japan	9	0.54±0.08	0.55±0.12	0.39±0.22
Training: Taiwan Testing: Japan	9	0.54±0.11	0.57±0.17	0.47±0.13
Training: Japan Testing: Taiwan	21	0.55±0.10	0.56±0.15	0.46±0.16
Training: JP+TW Testing: JP+TW	30	0.55±0.10	0.55±0.12	0.41
ANOVA results		f=0.05, p=0.99	f=0.02, p=0.99	f=0.29, p=0.88

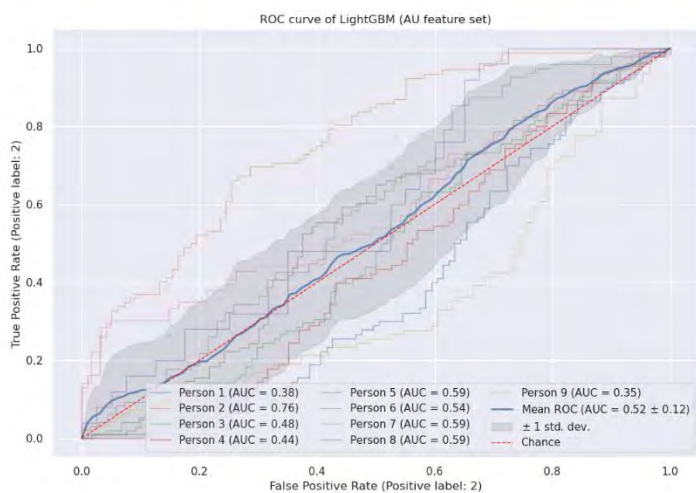
Table 22 The graphs of ROC when estimating the help-seeking states by Basic AU feature set

Datasets	Graphs of ROC
----------	---------------

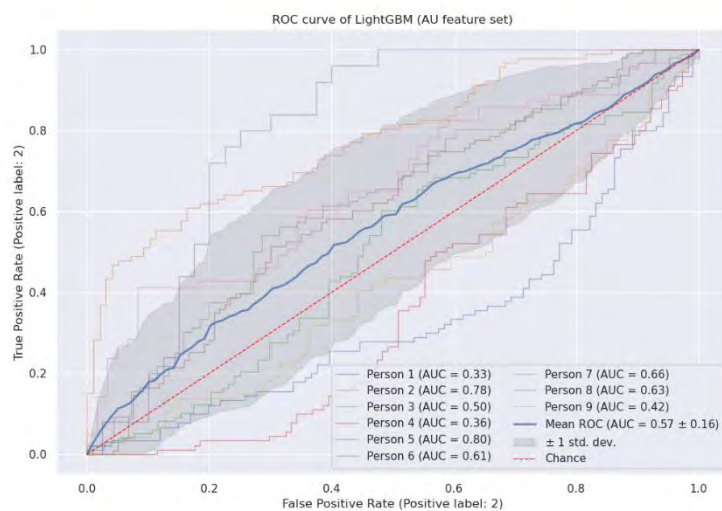
Training:
Taiwan
Testing: Taiwan



Training: Japan
Testing: Japan



Training:
Taiwan
Testing: Japan



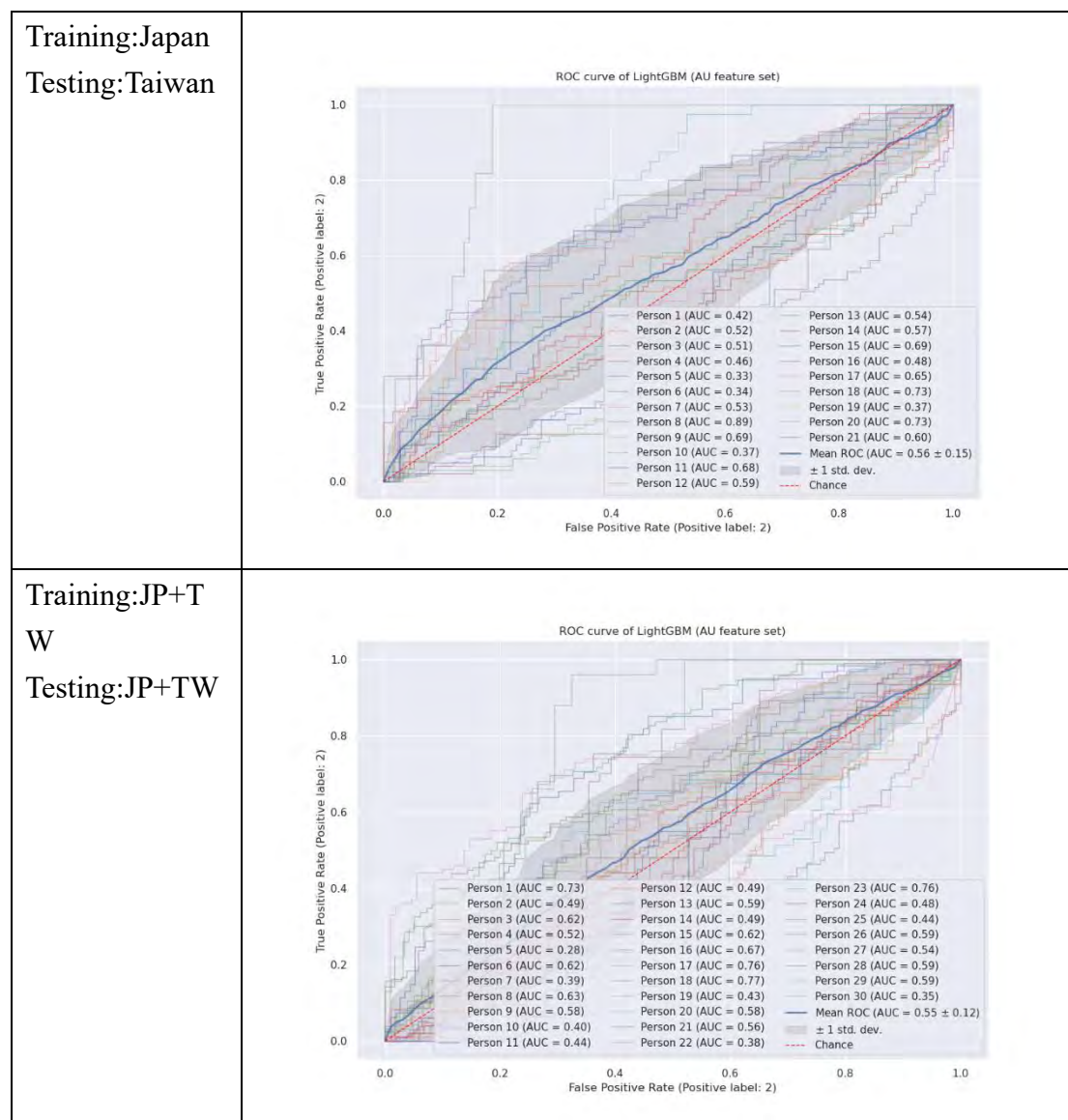
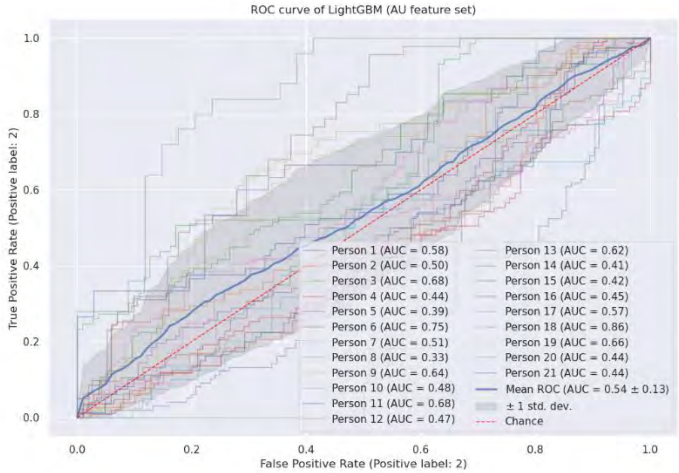
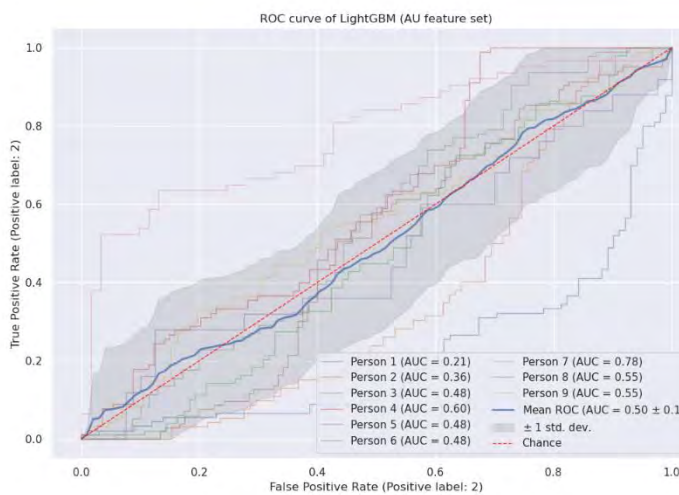


Table 23 The results of inter-person learning (after z-score) on estimating help-seeking states by the Head Pose feature set

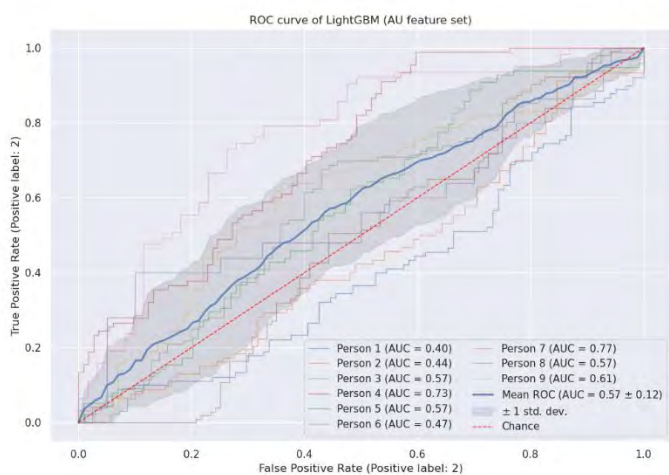
	n	Accuracy	AUC	F1
Training: Taiwan Testing: Taiwan	21	0.55±0.08	0.54±0.13	0.32±0.22
Training: Japan Testing: Japan	9	0.53±0.10	0.50±0.15	0.35±0.20
Training: Taiwan Testing:Japan	9	0.57±0.08	0.57±0.12	0.38±0.17
Training:Japan Testing:Taiwan	21	0.57±0.15	0.54±0.18	0.35±0.27

Training:JP+TW	30	0.55±0.09	0.53±0.14	0.33±0.21
Testing:JP+TW				
ANOVA results		F(4,76)=0.28, p=0.89	F(4,76)=0.29, p=0.88	F(4,76)=0.21, p=0.92

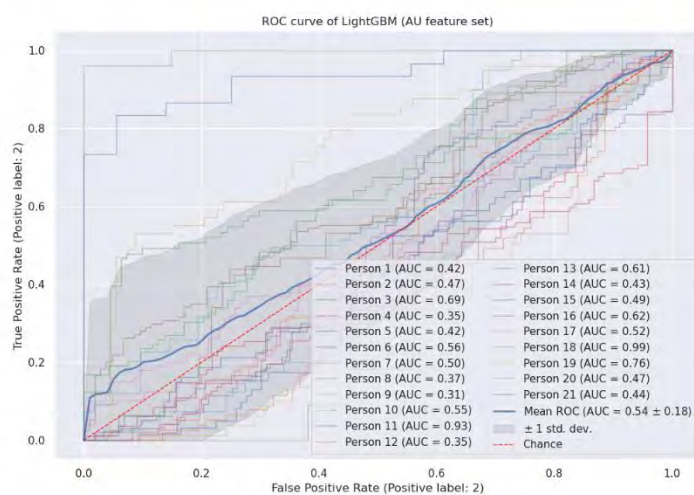
Table 24 The graphs of ROC when estimating the help-seeking states by Head Pose feature set

Datasets	Graphs of ROC
Training: Taiwan Testing: Taiwan	
Training: Japan Testing: Japan	

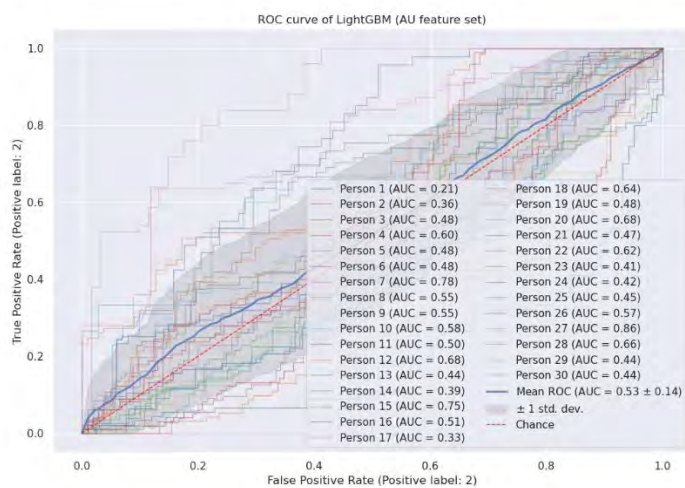
Training:
Taiwan
Testing:Japan



Training:Japan
Testing:Taiwan



Training:JP+T
W
Testing:JP+TW



3.3 Nationality Classification by Action Units

As the important feature of estimating the mental states were different between Taiwan's and Japan's participants, it suggested that facial expression might be also different from these two cultures. Therefore, we used Action Units features to identify the face is from a Taiwan's participant or a Japan's participant.

The data of facial expression were from 21 Taiwan's participants and 9 Japan's participants. The prediction was the nationality of the participant. The descriptive statistics of the AUs, including mean, median, standard deviation, minimum, maximum, and range, were calculated in 60-second time window.

The training dataset was from 29 participants' and the testing dataset was from the other 1 participant. The validations ran 30 times so that every participant has been in testing dataset. The SHAP analysis followed by every validation, and totally generated 30 times of ranking on important features. The framework of dividing training and testing datasets is shown in Figure 6.

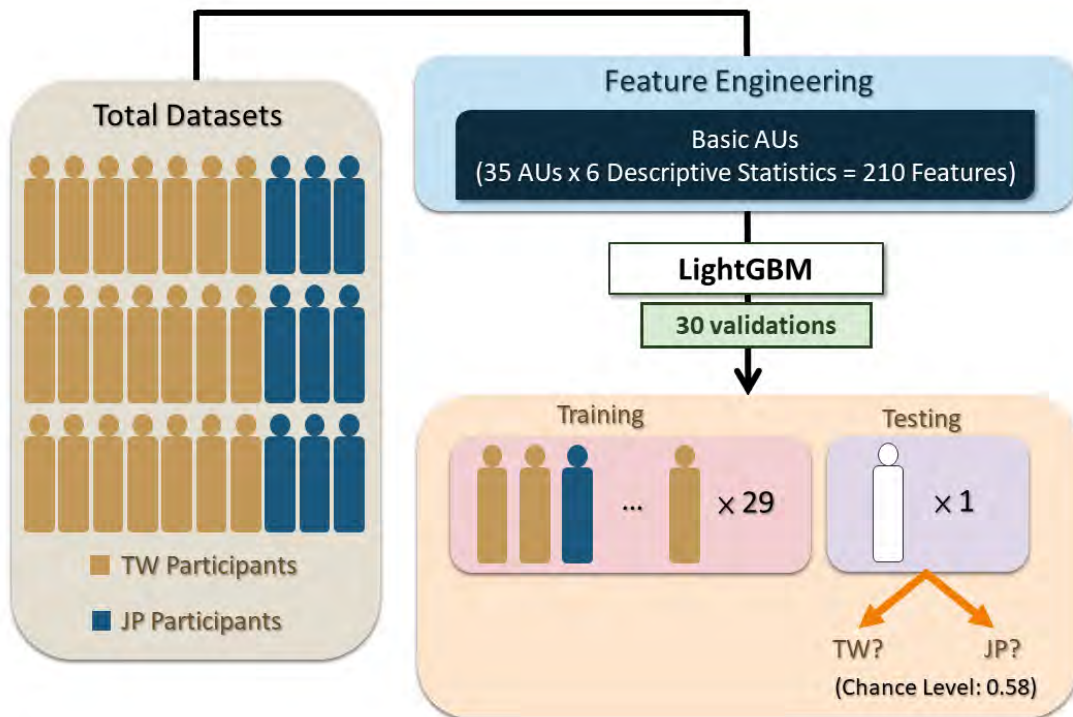


Figure 6 The framework of training and testing nationality identification.

The chance level of correct answer was 58% (9% for Japan and 49% for Taiwan). The mean accuracy of the 30-time predictions of testing datasets was 0.70, and 70% (21 times) of predictions' accuracies exceed 0.58, which was higher than the chance level. As for the SHAP analysis results, the most frequently ranked features are showed in Table 25. The frequency was counted by excluding the models which accuracy was

lower than 0.58. Besides, the feature which was most frequently ranked in the top 1 was “AU10_r_mean”, which indicate the facial expression of “lupper lip raiser”.

Table 25 The top 10 frequently ranked features when identify the nationality.

Features	Counts	Description of the related AU
AU05_c_std	16	upper lid raiser
AU23_c_mean	14	lip tightener
AU04_r_median	10	brow lowerer
AU07_r_median	10	lid tightener
AU12_r_mean	10	lip corner puller
AU02_c_mean	9	outer brow
AU45_r_range	9	blink
AU10_r_mean	9	upper lip raiser
AU45_r_std	8	blink
AU23_c_std	7	lip tightener

4. Discussion

This chapter examined the inter-person learning on the mental states classification both on the Japan and Taiwan’s participants. Overall, the results showed that it is hard to make the machine learn from inter-person. Although the Action Units data was computed in z score, the individual differences still cannot be deleted.

However, an interesting fact in the estimation of the engagement states showed that the ROC curves in only Taiwanese dataset or only Japanese dataset were mostly gathered on the above of the chance level line, but this phenomenon cannot be seen in the mix dataset or cross testing and training dataset. Although statistically we cannot find the significant differences on the metrics between the two datasets, it still can be inferred that the two cultures have different facial expression in some extent.

On the other hand, the estimation of the help-seeking states seems to be more difficult than the estimation of the engagement states. However, some people were easier to be estimated than others. According to the results of ROC curve, some people have high AUC score than others. Although the model has never seen the testing person’s face, the testing performance still showed that their help-seeking behavior can be identified. Surprisingly, the Head Pose feature sets can even showed the AUC=0.99 results when we trained on Japan’s datasets but test on Taiwan’s participants.

The results showed that, the facial expressions of the engagement have more culture

differences than the ones of the help-seeking states. This suggested that the help-seeking behavior has more potentials on generalizing to other cultures. Therefore, for application on other system and scenario, the estimation of help-seeking is more applicable than the estimation of the engagements.

In addition, the privilege of this thesis is that two cultures were investigated. To the best of my knowledge, there was limited research related to educational technology compared the cultural differences on the facial expression. Since the students in different cultures might express their emotion and cognition process in different way, it is essential to study the cultural differences on the mental states. The SHAP analysis showed in the Appendix also suggested that, the explanations of every prediction model were various. We split the dataset as one person to the testing dataset, and the others to the training dataset. As a result, every model can explain the prediction of that person which is in the testing dataset. The method of splitting the training and testing dataset should be noted that it is one of a possible explainable AI approach.

Therefore, we further examined that the differences on the facial expression between the two cultures. We tested the nationality classification by using AUs. The result is higher than the possibility of correct answer. The result suggested that the facial expression is different from two cultures and the movements of facial muscle are useful for identifying the nationality. But the results were based on the small samples, the generalization of this classification still should be noted.

Chapter 6: Apply Machine Learning Methods to Questionnaire

Datasets

Abstract

Statistic tests and machine learning play different roles on data analysis. This chapter describes applications of LightGBM and SHAP analysis to a large number of questionnaire datasets (610 respondents) about effect of income and educational degree on optimistic bias and social trust. The questionnaire test was conducted to analyze using an ANOVA model. The current chapter discovered that the regression of LightGBM predicts the dependent variables with RMSE between 0.61-0.89. SHAP analysis had the advantage of explaining the several independent variables' effects simultaneously, and the SHAP analysis revealed the effect of the relationships between income and educational degree and optimistic bias and social trust, which were not revealed by the ANOVA. Based on these findings, we discussed the potential usage and advantage of SHAP analysis on large amounts of the features which can overcome the limitation of the complexity of ANOVA more than four factors.

1. Introduction

In this chapter, we used LightGBM and SHAP analysis to learn the data which conducted for the investigation of information, cognition and prevention behaviors of facing COVID-19 in Taiwan in Yueh et al. (2022). The research has 610 valid participants' data from a questionnaire and analyzes with four-way ANOVA to investigate the relationship between variables.

We used the demographic variables, including gender, age, occupation, living area, income and education degree, as the features of the machine learning model, which were treated as independent variables. On the other hand, the predictors of the machine learning model were optimistic bias, social trust, information credibility, personal protective measures, avoidance of human contact and immune system strengthening, which were regarded as the dependent variables.

2. Methods

2.1 Participants

The number of valid respondents were 610 and all of them were Taiwanese, the average age was 44.01 years. They were selected from 709 respondents, rejecting 99 respondents whose age was unreasonable (older than 100 years and younger than 7) or whose responses have errors or blanks. The questionnaire test was conducted during the beginning phase of the pandemic, and before the COVID-19 vaccine is invented. At that time, the COVID-19 was an unknown disease and was spreading very quickly.

2.2 Independent Variables (Features)

There were 6 demographic variables. Gender was divided into male or female; living area was divided into northern Taiwan and other areas. Age was the real number of the respondents. Occupation was divided into government employee, healthcare and other, since one of the purpose of the research was to identify the behavior of people who comes from the public, health, or medical industries in the context of the COVID-19 pandemic. Income (monthly) was divided into less than 50,000 NT dollars and Over 50,000 NT dollars this threshold approximately equals the average monthly income in Taiwan. Education degree was divided into “bachelor degree and under” and “graduate school degree and above”.

2.3 Dependent Variables (Output)

The dependent variables includes optimistic bias, social trust, information credibility, protective measures, avoiding human contact, and strengthening one’s immune system. The datasets used for this current study was from Yueh et al., 2022. The detail of the questionnaire was shown in Table 26.

In order to compare the machine learning and ANOVA, we followed the procedure of the previous study. Therefore, we built the model separately by every dependent variable. In addition, the value of optimistic bias was calculated by subtracting “You think others may be infected by COVID-19” to “You think you yourself may be infected by COVID-19”. Because the definition of optimistic bias is the perceived likeliness of infection between self and others, the subtraction between these two items can reflect the extent of optimistic bias. On the other hand, other variables, including social trust, information credibility etc., were calculated by their mean of the items from every facet. Those variables were investigated by 6-point scale or 4-point scale.

Table 26 The descriptive statistics results of every variables (n = 610) (Resources: page 5, Yueh et al., 2022)

Items	Mean	SD
Optimistic Bias (1–6, strongly disagree to strongly agree) ($\alpha = 0.8451$)		

Items	Mean	SD
You think you yourself may be infected by COVID-19	3.34	0.90
You think <u>your neighbors</u> or colleagues may be infected by COVID-19	3.53	0.80
You think <u>others</u> may be infected by COVID-19	3.96	0.80
Social Trust (1–6, strongly disagree to strongly agree) ($\alpha = 0.9113$)		
I think the government is credible to adopt policy on COVID-19	4.77	1.01
I think the government is correct to adopt policy on COVID-19	4.73	1.01
I think the government should develop long-term plans for COVID-19	5.05	0.92
I think the government can solve problems related to COVID-19	4.67	1.01
Information Credibility (1–6, strongly disagree to strongly agree) ($\alpha = 0.7731$)		
Information about COVID-19 from family members and friends is credible	3.57	0.86
Information about COVID-19 from newspapers, television and radio is credible	3.99	0.86
Information about COVID-19 on the Internet and social media is credible	3.55	0.87
Information about COVID-19 from research institutes is credible	4.68	0.89
Information about COVID-19 from the government is credible	4.84	1.00
Personal Protective Measures (1–4, rarely to always) ($\alpha = 0.6627$)		
Wear a mask	3.11	0.85
Take eye protection measures	2.22	1.08
Wash your hands frequently with soap	3.47	0.68
Avoid touching your eyes, nose, and mouth	3.11	0.82
Avoiding Human contact (1–4, rarely to always) ($\alpha = 0.7710$)		
Avoid close contact with other people	3.07	0.79
Avoid crowded places	3.28	0.67
When I feel ill, I distance myself from others	3.59	0.57
If you feel ill, immediately notify the person in charge of the epidemic, such as a doctor or a neighbor	2.98	0.97
Avoid taking public transportation	2.90	0.98
Avoid entering physical shops and shop online instead	2.69	0.95
Avoid unnecessary travel	3.57	0.72
Strengthening one's immune system (1-4, rarely to always) ($\alpha = 0.7262$)		
Do more exercise	2.17	0.58
Balance nutrition and consume nutritional supplements	2.32	0.58
Keep positive emotions	2.18	0.58
Sleep enough	2.22	0.60

Items	Mean	SD
Drink more water	2.40	0.64
Buy masks and 75% alcohol online	2.30	0.92
Buy tissue paper and wet napkins	2.10	0.66

2.4 Machine learning

The features for machine learning model were the same as the ANOVA model in the previous study but we added two variables (income and education degree) to make more complex model. The detail of variables was introduced in the previous sections.

The machine learning model used in this chapter was LightGBM (Light Gradient Boosting Machine)(Ke et al., 2017), the same one used in the previous chapters. The previous chapters used the classification estimator, while this analysis used the regression estimator since the values of the dependent variables were collected by 4- or 6-point scale. The outcome of the model was evaluated by the R^2 value and RMSE (Root Mean Squared Error). The equations of R^2 and RMSE were shown in (1) and (2) in the following, where y_i is the true value from questionnaire response, \hat{y}_i is the prediction value from machine learning model, and \bar{y} is the average of the true values. n is the numbers of data.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}} \quad (2)$$

In this chapter, the R^2 value are calculated by calling the function from official LightGBM. The best possible score of R^2 is 1.0 and if the model is arbitrarily worse, the function will provide us with negative value. On the other hand, the RMSE are calculated manually by *numpy* library, a library from Python since the official LightGBM did not have RMSE functions. The validation of the machine learning model used 5-fold cross-validation. The training dataset was 80% and testing dataset was 20%. Overall, the framework of the input features and dependent variables was shown in Fig. 32.

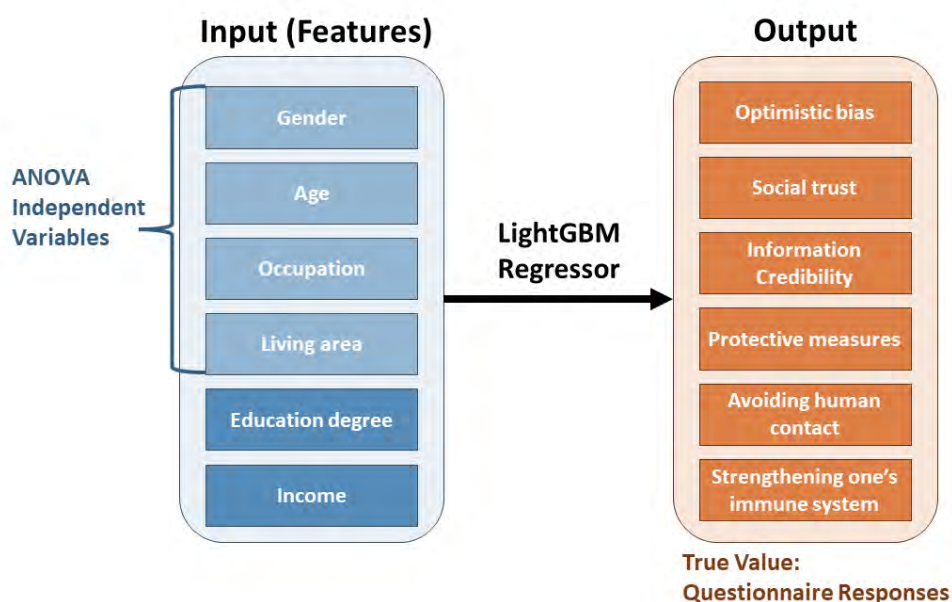


Fig. 32 The framework of the machine learning analysis. Note: The four-way ANOVA model used gender, age, occupation and living area as independent variables to predict the six dependent variables. In this chapter, two independent variables were expanded and the predictions were conducted by the LightGBM model.

After machine learning of every split, the importance of the features are explained by SHAP analysis (Lundberg et al., 2019). We used SHAP analysis to deal with the six independent variables at the same time. The SHAP values were calculated by every time of the testing. The bar plots showed the mean absolute value of features' SHAP value. The length of the bar showed the effect of a feature on the estimation. The scatter plots showed the SHAP value distribution of the features, with each point representing the SHAP value of an estimation on the target variable. The decision plot showed the cumulative SHAP value and interaction values. The interaction value is a tensor of all pairs of SHAP value. The bar plots and the scatter plots are the same as previous analysis, and we used an additional SHAP analysis for the interaction.

3. Results

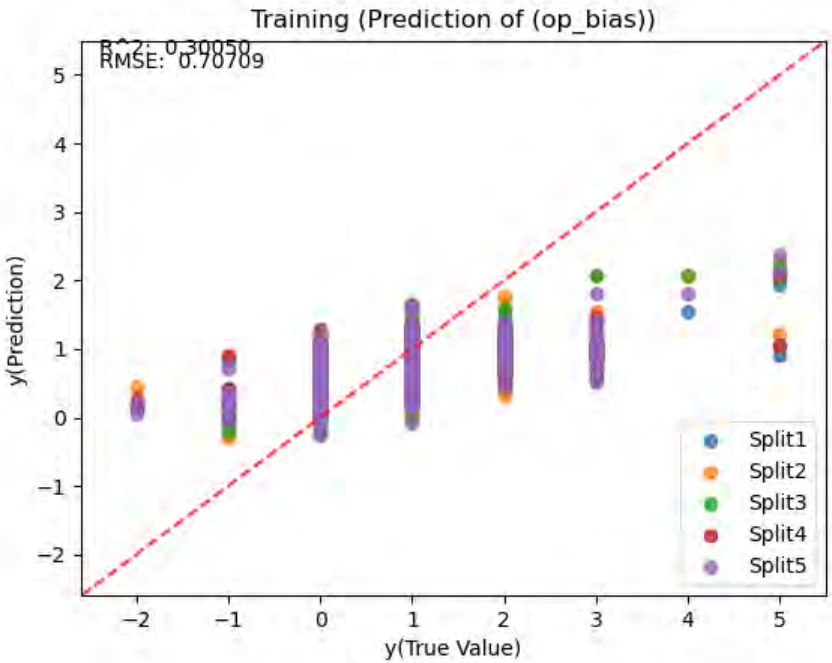
3.1 Regression by LightGBM

The LightGBM regressor was used to predict the 6 dependent variables. The independent variables are demographic variables, including gender, occupation, living area, age, educational degree, and income (The first four variables were used in the

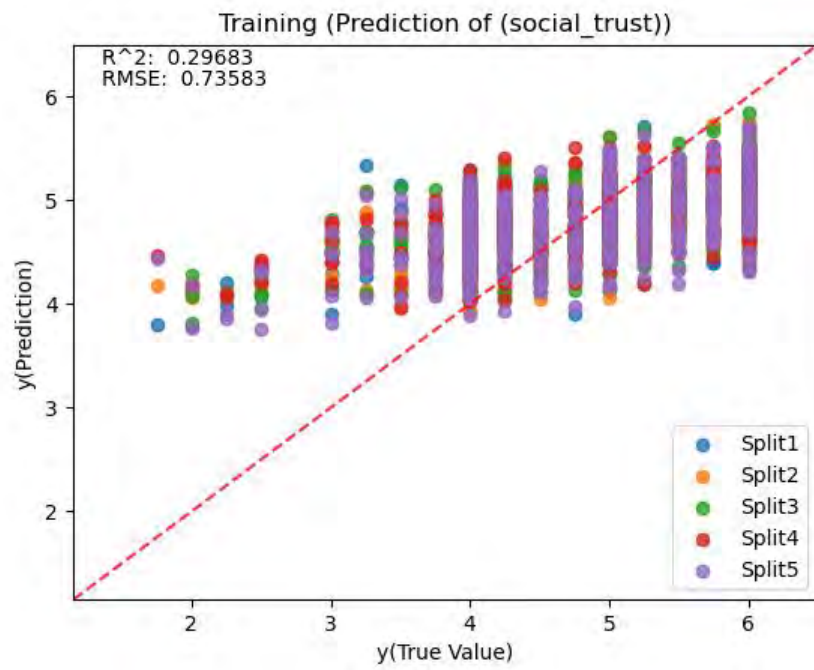
ANOVA model). The dependent variables predicted optimistic bias, social trust, information credibility, protective measures, avoiding human contact, and strengthening one’s immune system.

The R^2 were between 0.30 and 0.36, and RMSEs were between 0.32 and 0.73. The details of prediction results were shown in Table 27 for all cases. Although all weak correlation coefficients were not very high, the guidance of social science research used R-square suggests that a R-squared that is between 0.10 and 0.50 is acceptable when some of the explanatory variables are statistically significant (Ozili, 2023). In the current chapter, we calculate the importance of the features and the previous study also revealed that the demographic variables were useful for predicting the targets. Therefore, the results of the regressions were all acceptable.

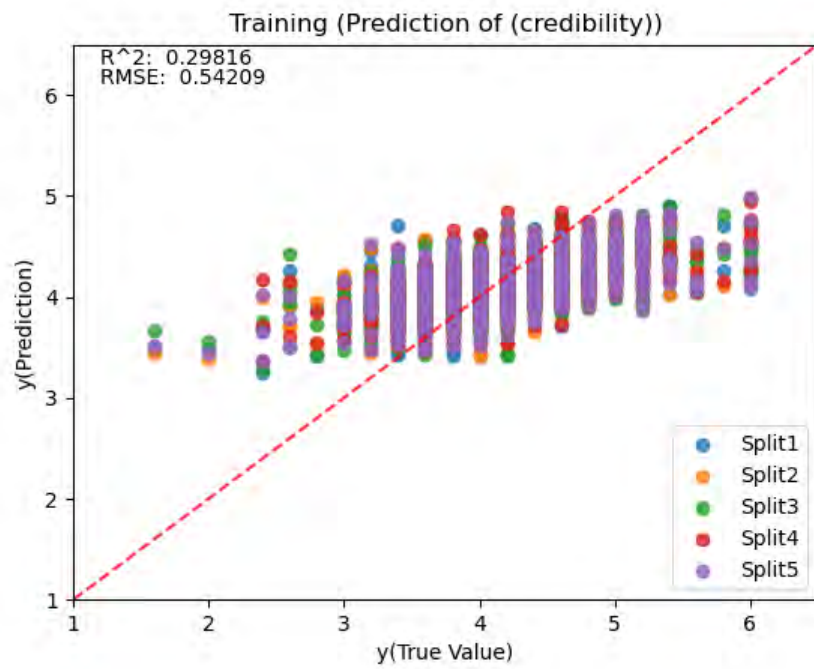
Table 27 The scatter plots of true value and predictive value of all dependent variables.

Dependent Variables	Scatter plot of true value (x-axis) and predictive value (y-axis).
Optimistic bias	 <p>The scatter plot displays the relationship between the true value (x-axis) and the predicted value (y-axis) for the dependent variable 'Optimistic bias'. The plot is titled 'Training (Prediction of (op_bias))'. The x-axis is labeled 'y(True Value)' and ranges from -2 to 5. The y-axis is labeled 'y(Prediction)' and ranges from -2 to 5. A red dashed diagonal line represents the identity function (y=x). The data points are clustered around this line, indicating a positive correlation. The points are color-coded by split: Split1 (blue), Split2 (orange), Split3 (green), Split4 (red), and Split5 (purple). The plot includes the following statistics: $R^2: 0.30050$ and $RMSE: 0.70709$.</p>

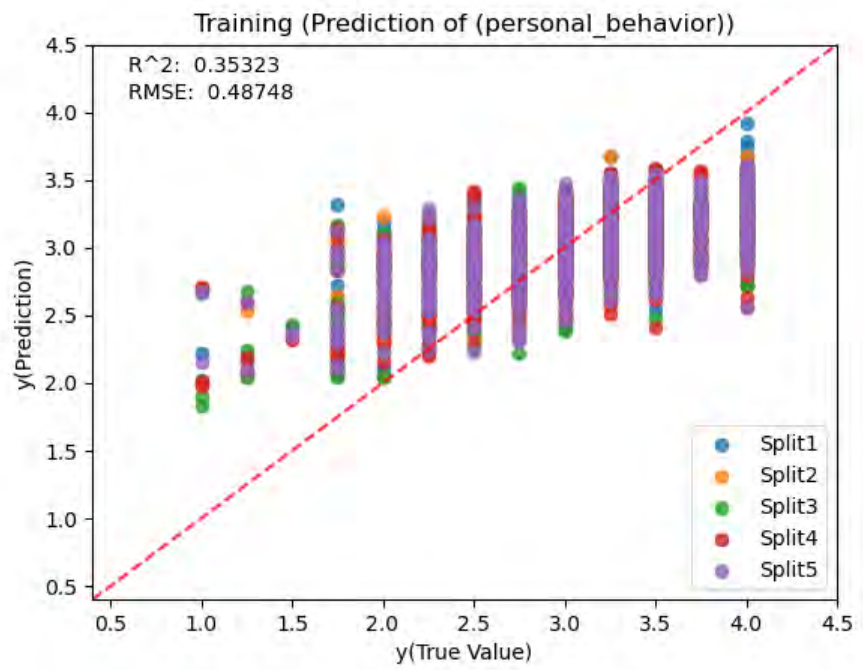
Social trust



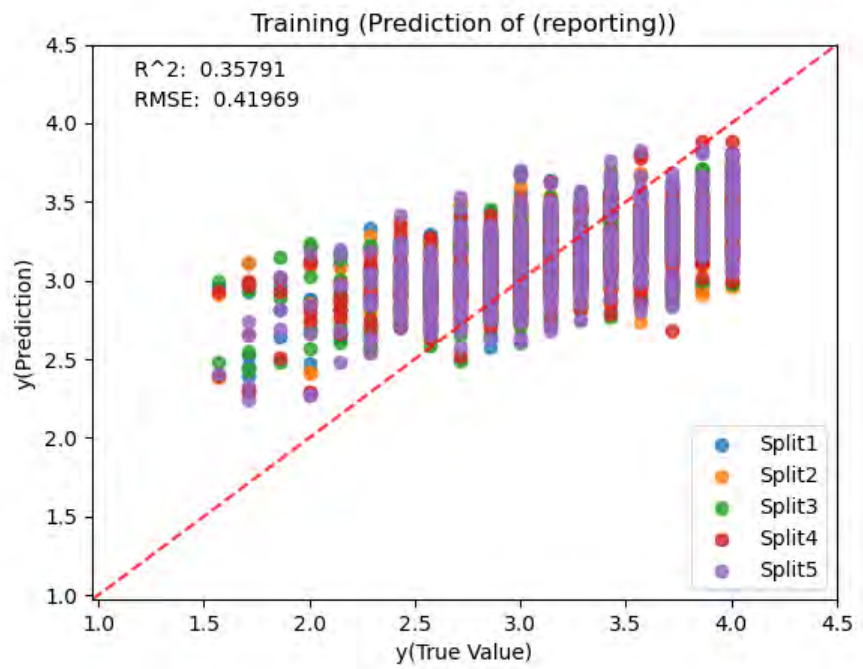
Information credibility

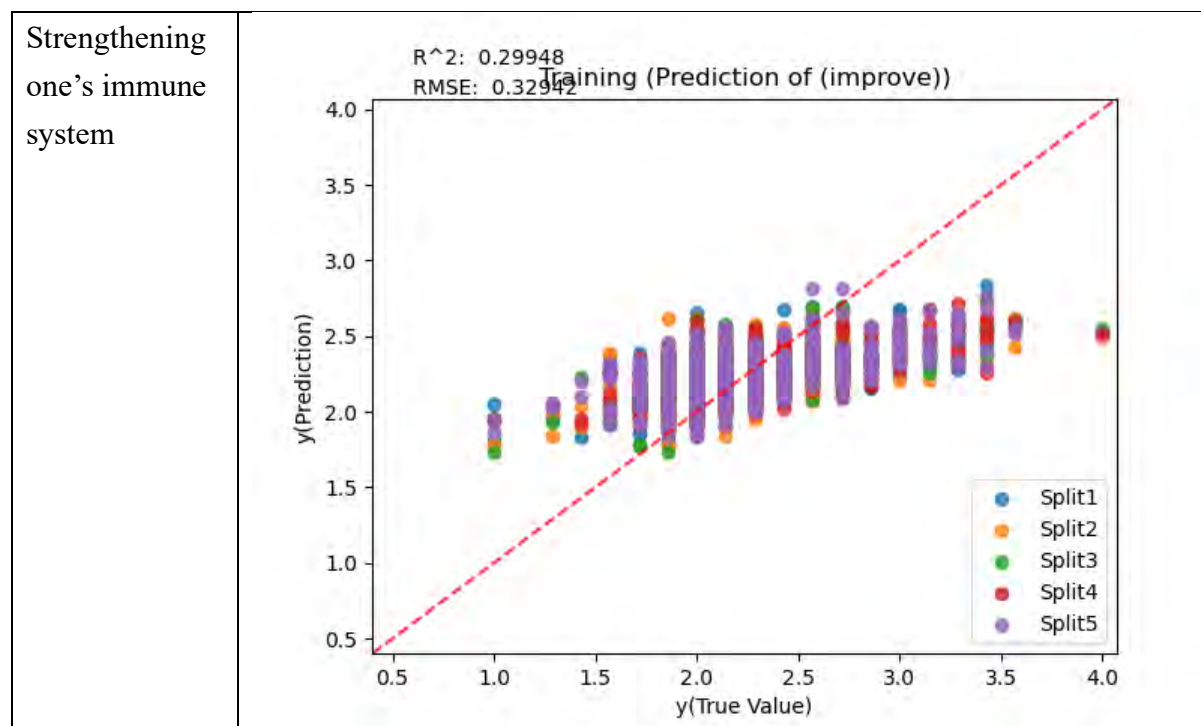


Protective measures



Avoiding human contact





3.2 SHAP Analysis

SHAP analysis shows contribution of each factors separately while no information of combination effect of multiple factors is available. Every SHAP summary plot was a representative graph from the lowest RMSE results of the 5-fold cross-validations. The details of scatter plots and bar plots of SHAP analysis were shown in Table 28. The education, income, gender, and area have two categories, and the feature of occupation has three categories. Therefore, the color of the points indicated the categories of the feature value. The summary table of points color was shown in Table 29. Overall, the feature of age is the most important feature in all predictions of dependent variables. In some cases, some demographic variables suggested some pattern and distribution to determine the prediction. For example, some features showed that the blue points and red points are separately distributed along the SHAP value axis. The pattern of distribution was often found in the second or the third important feature.

Table 28 The SHAP summary plot and SHAP bar plot.

Dependent Variables	SHAP summary plot	SHAP bar plot
Optimistic bias		
Social trust		
Information credibility		
Protective measures		
Avoiding human contact		

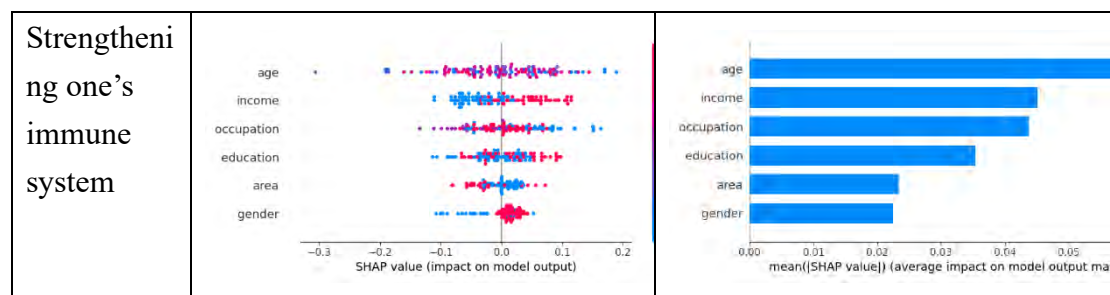


Table 29 The meaning of colors of feature value in SHAP scatter plot.

	Blue	Purple	Red
Gender	Male	-	Female
Age	Lower		Higher
Occupation	Government employee	Healthcare	Others
Living area	Northern	-	Other area
Education	Undergraduate or below	-	Graduate School or above
Income	Lower than NT50,000 dollar	-	Higher than NT 50,000 dollar

To investigate interactions of features we used SHAP decision plots. The plots here explained the predictions from the datasets with main effects and interactions. In the decision plot, the bottom of the plot is the starting value for each prediction. In total, there should be $6 \times (6+1) / 2 = 21$ features. The center of the x-axis is the expected value of SHAP values, and the y-axis list the model's features. The importance of the features was in descending order. In the decision plots, each line showed the cumulative SHAP values, and showed how LightGBM made the decision to give the value of the prediction.

All results of the SHAP decision plot suggest that the model of predicting optimistic bias can be determined by “age”, “age*occupation”, “gender*age”, “education”, and “age*income”, etc. In contrast, the variations of other variables, including social trust, information credibility, protective measures, avoiding human contact, and strengthening one's immune system, were not significantly separated by each features. But “occupation” “income”, “education” were important features when it was used as one feature or was calculated by interaction value.

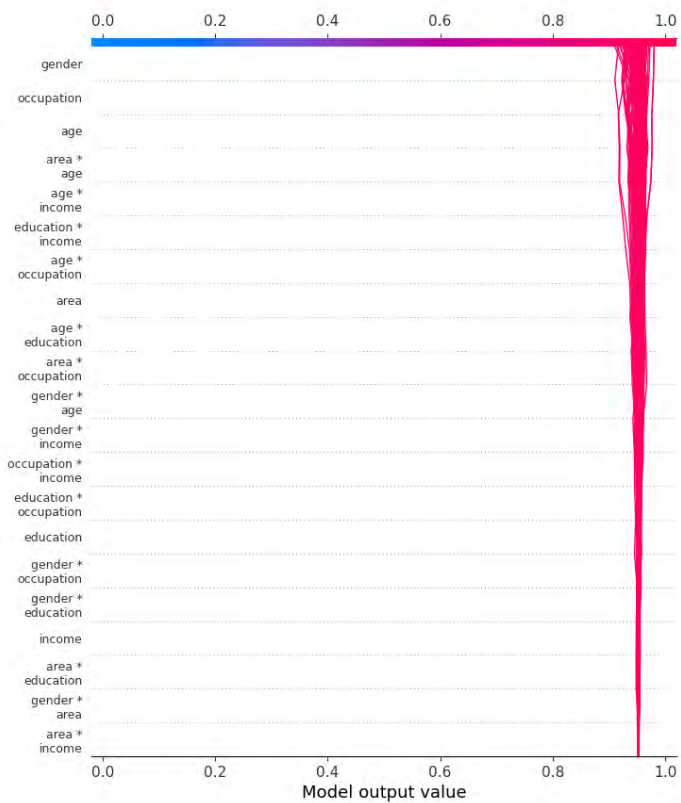
Table 30 The SHAP Decision Plots of each Dependent variables

Dependent Variables	SHAP Decision Plots
<p>Optimistic bias</p>	
<p>Social trust</p>	

Information credibility



Protective measures



Avoiding human contact



Strengthening one's immune system



4. Discussion

In this chapter, we used LightGBM and SHAP to analyze the datasets from a questionnaire related to COVID-19. LightGBM and SHAP can handle independent variables as many as needed, which is different from ANOVA which handle less than five factors in general. The SHAP analysis provides information of interaction among all combinations of more than four factors.

The SHAP analysis applied to the questionnaire test showed that the education degree and the income were important features in predicting the variables of *optimistic bias* and *social trust*. The higher education degree might have an effect on optimistic bias. People with higher education are more likely to believe that they have less possibility to encounter negative events than others. The results can be further inferred that people who are in higher social status (higher educational degree and higher income) have more self-efficacy to control their life, which reflects that they tend to overestimate themselves and do not believe in the authorities overall.

When predicting optimistic bias, the four-way ANOVA table (Table 31) suggested that there was an interaction on “Gender × Age” and a significant main effect on “Occupation”. The results were compared with the SHAP analysis of the prediction of optimistic bias. The SHAP analysis showed that the top three important features were age, education degree, and occupation, and that the important interactions showed by SHAP included “age × occupation”, “gender × age”, and “age × income”. The SHAP showed that the interaction of “gender × age” is important as ANOVA showed. Furthermore, the SHAP analysis revealed that education and income were also important for predicting optimistic bias. This suggests that there are advantages of using machine learning and SHAP tools to understand more complicated factors or features.

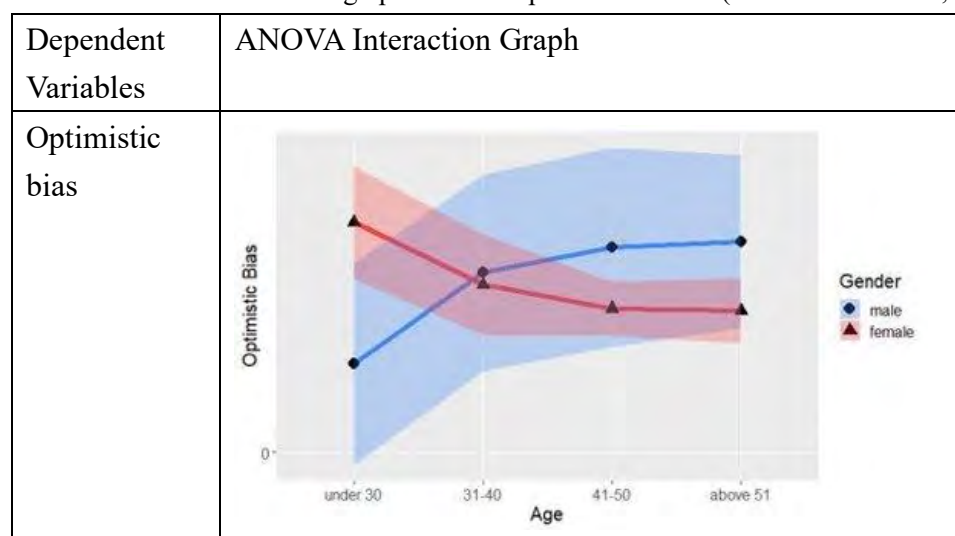
Interaction of ANOVA analysis indicates influences of combination of the multiple factors for predictions. Other interaction results from the previous study showed in the Table 3233. As we can see, the complexity of factors will limit the application of ANOVA model. Although the previous study tried to use four-way ANOVA to analyze the independent variables as much as possible, the interaction effects of four factors were hardly to understand by human intuition. Another possible approach is to use machine learning to select features before regression analysis (Kukowski et al., 2021). But Kukowski et al. (2021)’s results showed that, the perceived risk and age displayed nonsignificant weak to zero associations with health-protective behavior. In contrast, our results found that age is an important feature for machine learning prediction. This is a consequence of benefits of machine learning analysis.

Table 31 Four-way analysis of variance of optimistic bias with gender, region, age and occupation

Source	SS	df	MS	F	Sig.
Gender	0.3556	1	0.3556	0.5034	0.4783
Region	0.0111	1	0.0111	0.0156	0.9005
Age	1.3705	3	0.4568	0.6467	0.5853
Occupation	5.9289	2	2.9645	4.1962	0.0155*
Gender × Region	0.0802	1	0.0802	0.1135	0.7364
Gender × Age	7.7278	3	2.5759	3.6463	0.0126*
Region × Age	1.5275	3	0.5092	0.7207	0.5399
Gender × Occupation	0.7772	2	0.3886	0.5501	0.5772
Region × Occupation	1.4429	2	0.7214	1.0212	0.3608
Age × Occupation	2.3753	6	0.3959	0.5604	0.7619
Gender × Region × Age	1.8611	3	0.6204	0.8782	0.4521
Gender × Region × Occupation	2.8567	2	1.4283	2.0218	0.1334
Gender × Age × Occupation	4.6905	6	0.7817	1.1066	0.3570
Region × Age × Occupation	1.0775	6	0.1796	0.2542	0.9576
Gender × Region × Age × Occupation	5.9174	5	1.1835	1.6752	0.1386
Error	397.7351	563	0.7065		

Note: $p < .05$ * Source: Page 7, Yueh et al., 2022

Table 32 ANOVA interaction graph of each dependent variable (Source: Yueh et al., 2022)



<p>Social trust</p>	<p>Gender</p> <ul style="list-style-type: none"> ● male ▲ female
<p>Information credibility</p>	<p>(no significant interaction effect)</p>
<p>Protective measures</p>	<p>Gender</p> <ul style="list-style-type: none"> ● male ▲ female
<p>Avoiding human contact</p>	<p>Gender</p> <ul style="list-style-type: none"> ● male ▲ female
<p>Strengthening one's immune system</p>	<p>Gender</p> <ul style="list-style-type: none"> ● male ▲ female

5. Conclusion

Extending previous research, this chapter apply the machine learning methods on a questionnaire datasets. The advantages of the machine learning methods are having more independent variables to predict a single dependent variable, using more features for regression, and being able to explore interaction of explaining machine learning model. Some potential combination of interaction were discovered in the SHAP decision tree plot. The limitation of the current work is that the questionnaire did not include much variables. This can be further explored in the future research to add more demographic or psychological traits variables to predict people's belief, intention, behavior, etc. The contribution of analyzing the questionnaire about society's problem is critical for policy maker to determine the promotion on public health, the higher sensitivity to detect the effect is an important advantage of SHAP analysis for the data obtained here.

Chapter 7: General Discussions

1. Estimation of the mental states

No matter dealing with Japan's or Taiwan's data, this thesis examined two kinds of mental states during learners learning in the experiment. One is that we classify their high and low engagement states using data labeled by their appearance. The other is that we classify students' help-seeking states, which indicates the 3-second intervals before they inquire about a hint and the moment they are just working on problem-solving without particular inquiry behavior. We simulated an intelligent tutoring system that we can access on the Internet and conduct our experiments like a real Linguistic Olympiad. Namely, our problem-solving task is simulating reality environments, and we expect our results will benefit ITS research to develop an automatic system with artificial intelligence (AI).

We conducted the experiment for Japan's and Taiwan's students; the behavioral data showed that they don't have significant differences, including their time of completing the problem, score, and times of clicking the hint buttons. This indicates that our webpage design would not discriminate for different students, and both Japanese and traditional Chinese versions are similar for students to use. In addition, for overall data, "times of clicking the hint buttons" has a significant correlation with "score", which suggested that our design of hint and website interaction can improve student learning to some extent. We got similar results that a previous study also showed that principle-based hints could improve student learning (Alevan et al., 2016a).

As for the machine learning part, one of our strengths is that we used OpenFace to extract features used for machine learning. This approach is used in several studies and applications (Amos et al., 2016; Bosch & D'Mello, 2021; Li et al., 2021), and OpenFace is possible to apply to experiments that need to extract facial features in real-time. Results from this study revealed that facial features from recorded videos were effective indicators for classifying engagement and help-seeking states. The overall prediction of the accuracy is higher than 70%, which is higher than previous studies (Bosch & D'Mello, 2021; Li et al., 2021).

The first mental states we estimated is the engagement level by students' appearance. Comparing all feature sets and two machine learning models, our results showed that the performance of using the head pose feature set is better than the other feature sets, and co-occurring feature set and gaze feature set were not as good as the others. Our result is similar to a previous study (Bosch & D'Mello, 2021), which suggests that the Basic AUs feature set is more effective than others since it is simpler and estimates first-order expressions of a single facial muscle. As for the co-occurring AUs feature

set, all pairs of AUs are calculated, and therefore without theory-based, some unnecessary pairs are still calculated. Furthermore, research indicates that many combinations of AUs are found, and combinations of AUs can be two, three, or more (Zhi et al., 2020). Therefore, it is also worth considering instead of pair combination used in the co-occurring method. However, we found that the head poses feature set is also effective at estimating engagement level; this result supports a previous study (Li et al., 2021), which indicates that the head pose features are good indicators since it usually shows the learners are concentrating or thinking when they tilt their head (el Kaliouby & Robinson, 2005). Overall, the indicators we used to train the machine learning model effectively estimate the engagement state when students are working on the problem-solving task of linguistics.

On the other hand, the second mental state is about predicting the help-seeking state that a student wants to inquire about some help. LightGBM showed that the three feature sets effectively classify help-seeking and working states, but SVM only sometimes worked. Therefore, in the Chapter 4 to 5, we focused on LightGBM and the expanded feature sets were also trained and tested by LightGBM. The results of the all feature sets are similar to engagement classification, which showed that the head poses feature set is better than the basic AUs feature set, but co-occurring AUs feature set is not an adequate indicator. In addition, the combination of features including head pose would performed better than others. The SHAP values are also calculated to investigate further what facial features are important to detect a student needing help. The results showed that the brow and lower part of the face are more important. A previous study showed that when individuals feel unsure, their lips will move apart, and the shape of their brows is also changed (el Kaliouby & Robinson, 2005). Another research showed that people tend to furrow their eyebrows and their lips depressing when watching videos related to banking, fuel, pharmaceuticals, etc., rather than videos about pet care, entertainment, or baby care that will let them smile more (McDuff & Kaliouby, 2017). Although those previous studies didn't use AUs to estimate their datasets, their results showed that when people are unsure about something or watching more serious videos, they smile less, lower their brows, and apart or depress their lips, which are expressions related to AU04(brow lowerer), AU25(lips part), AU26(jaw drop) and other AUs around lips.

Furthermore, the novelty of the current study is that we not only estimate students' engagement state when they are solving a challenging task like a real competition, but we also predict their mental states that they require some help and hint before they take action (click a hint button) to seek for help. A previous study indicated that research about students' engagement state has several different opinions that the level of engagement is not always correlated to high learning performance. Therefore, exploring

more mental states, such as students' decision-making processes, is necessary (Li et al., 2021). The help-seeking state we have tested is a mental state that can apply to instructional interventions since it indicates when a student needs a hint. The times of their clicks and score also have a positive correlation, and it can be expected that for the ITS application, our model can be used on the system with immediate feedback to provide hints before students click a button, which can improve their learning performance.

However, both in estimation on the engagement state and the help-seeking state showed that Gaze features were not as useful as other features. This result is different from a previous study that revealed the gaze data is effectively to estimate students' engagement (Li et al., 2021). Furthermore, in our current data, even if the combination feature set contains gaze features cannot be useful for the mental state estimation since the SHAP analysis suggested that the importance of gaze features were less than other features. On the other hand, as this current study explored learners' facial features, including AUs and head pose, to estimate the two mental states with machine learning methods, other features such as the upper body are still needed to explore further. For example, a study showed that body motion data also effectively estimate students' engagement (Anzalone et al., 2015). There should be some criteria to determine what kinds of features are essential.

2. Machine learning

As for machine learning training, we found that the overall performance of LightGBM is better and faster than the SVM model since its indexes and training time are higher and shorter. The results of the current thesis can contribute to the ITS application that implemented facial expression analyses to predict students' mental states and provide them with learning support accordingly. In addition, the operational definition in the current study of the help-seeking state is the 3 seconds interval before clicking a hint button, and the working state is the data from other randomly-chosen 3 seconds intervals. It might be so arbitrary that three seconds interval is the best – though we have tried 3 seconds, 10 seconds, and 15 seconds beforehand.

Moreover, we not only trained the machine learning model by pooling all participants' data together (intra-person learning), but also trained the model by inter-person learning approach. Although the results showed that inter-person model is unstable, there were still some successful estimations since their AUC scores were good when we further take a look on the detail of ROC curve results. The applications of inter-person learning make sure that our model can be generalized to more people. The inter-person learning has a potential to implement into a real ITS.

As for the statistics of features, we used six functions, including mean, median, standard deviation, minimum, maximum, and range, but they are hard to explain intuitively their meaning when they are possible to estimate students' mental state. However, the results of SHAP in the current study showed that mean, max, min, and median are more likely to be the top important statistics of features. Therefore, choosing those explainable statistics to make features is more effective for research.

Furthermore, in Chapter 6, we explored the application of machine learning method on questionnaire datasets. We used LightGBM regressor to predict the variables which were the same as four-way ANOVA in the previous study. The results of LightGBM regression showed that using those independent variables can predict the dependent variables. The SHAP analysis was used to determine the contribution of the features for machine learning. The results revealed some interaction effects and important features which were not revealed by ANOVA model. Therefore, we believe that the machine learning approach can be more useful when dealing with big data and more large amount of features.

In this thesis, the machine learning approach is used for estimating participants' facial videos and questionnaire datasets. In the future, it should be consider that, we can mix up the research methods to understand human's behavior by their appearance and their self-report response. It is crucial to explore a mental state by using both approach.

3. Cultural differences and cross training and testing

Last but not least, we didn't find notable cultural differences in behavioral data between Japan's and Taiwan's data. But the differences on the facial expression were showed on their data. The important features which were analyzed by SHAP were different between the two cultures.

Furthermore, this thesis developed the method to train and test the data. In Chapter 3 to 5, to compare the two dataset, we tried to exchange the training dataset and the testing dataset by cultures or by mental state. Besides, when we switched the data set of the intra-person learning, the AUC score also near the chance level. For example, the detail of estimating the help-seeking states was shown on Table 33.

Table 33 ROC results of cross training and testing on estimating the help-seeking states

Feature sets	Training: Taiwan's data Testing: Japan's data	Training: Japan's data Testing: Taiwan's data
--------------	--	--

<p>Basic AUs</p>		
<p>Head Pose</p>		
<p>Co-occurring AUs</p>		
<p>Gaze</p>		
<p>Basic AUs & Head Pose</p>		

Basic AUs & Gaze		
Head Pose & Gaze		
Basic AUs & Head Pose & Gaze		
Basic AUs & Head Pose & Co-occurring AUs		

4. Future Issues and Limitation

Automated detection of students' mental states during learning will become more common in the future since the growth of e-learning and ITS applications. This thesis provides a piece of evidence that facial expression helps estimate students' engagement state when they are learning with a problem-solving task. Besides, a notable contribution of this study is that students' interactive behavior of inquiry for a hint can be predicted via their facial expressions. This contribution can be expected to

implement in an ITS system to allow a system to automatically detect students' mental state and provide them with learning support adapted to their needs. Our data showed that the lightGBM model is more effective at estimating mental states than the SVM model. Besides, we also summarized that different parts of a human's face would have different remarkable meanings for models to predict engagement or help-seeking state, which are the upper face and lower face, separately. Previous research showed that, the perception of the upper part and lower part of the face have different mechanisms, which indicates that the facial expression can be judged by parts not the whole (Chen & Chen, 2010). Although the action units are indicated the facial muscle's movements, the classifications and the features analysis of separate action units still have their theoretical meanings.

For further investigation, although this thesis tried nine types of feature set to train the machine learning model. The way to selecting the features still remains since we did not choose the features by their weight but by their types. Furthermore, although we found that LightGBM is fast and effective to train a big amount of data, other machine learning model should also be considered in the future.

On the other hand, the limitation of questionnaire dataset is that the numbers of independent variables were insufficient. Originally, the design of questionnaire is for ANOVA analysis. Due to the minimum complexity an ANOVA model can deal with, more than four independent variables are too many to be understood by human intuition. Therefore, the questionnaire did not contains many question about demographic variables. However, since in the current thesis we revealed that educational degree and income played a curcial role in prediction, more variables about social status, such as job title or position, might also important for prediction. Therefore, we can further investigate more variables to utilized the advantages of the machine learning approach.

Besides, cultural differences between countries should be further tested by collecting more data. For example, the participants can be recruited from countries with high individualism cultures, and the learning task could include other problem-solving tasks or materials. Specifically, future work implementing the pre-trained model on real-time learning tasks with more participants is worth considering.

Chapter 8: Conclusions

The studies presented in this thesis investigated how the mental states influences learners learning in an e-learning environment. By designing a webpage for learners to solve a linguistic problem and recording the facial video via a web camera, it was possible to extract the features for machine learning, which allow us to estimate learner's mental states by using computer vision tools. Furthermore, the machine learning tools can be applied to a questionnaire dataset.

1. Major Findings

1.1 Mental states classification by facial videos

The mental states classification includes two mental states: engagement states and help-seeking states. The binary classification showed that the head pose features are useful for classifying both states. In addition, the multimodal features have more powerful performance than unimodal features. Therefore, we suggested that the machine learning method for classifying the mental states should utilized the facial videos, since the action units, head pose, gaze data can be extracted by the videos.

1.2 Cultural differences between Japan and Taiwan

The comparison between Japanese and Taiwanese allows us to examine the facial expression from similar appearances but different cultural contexts. Previous studies usually conducted an experiment with participants from single cultures. This study recruited the participants from two countries and we found that the estimation of engagement state shared some common features between the two cultures. In contrast, the classification of the help-seeking state had different important features. We used AI explanation tool to find the differences. The help-seeking behavior might be caused by cultural context, since in different culture, people ask questions in different ways.

1.3 Machine learning on inter-person model

Typically, studies on classifying mental states during learning used intra-person models to train the machine learning model. This study discussed the differences between intra-person and inter-person models. If the inter-person model showed good results, it can be said that the model has a potential to generate to others who are not "seen" by machine. However, the comparison results showed that the inter-person models were unstable. The individual variances are too big to learn by machine learning model. Therefore, for future application, the machine should learn all individuals data to get more accurate results, or more data from participants should be collected.

1.4 Machine learning applied on a questionnaire dataset

We tested the machine learning model on a questionnaire dataset which related to the mental states and behaviors when people facing COVID-19. The results showed that machine learning tool help us to find more information from the variables. In addition, the machine learning can deal with more features than a typical statistic method, such as ANOVA, can do.

2. Concluding Remarks

In summary, we revealed that facial videos are useful for estimating two mental states: engagement states and help-seeking states. The cultural differences between Japan and Taiwan showed that the estimation of engagement states shared more common features but help-seeking states leads to more different behaviors between these two cultures. Furthermore, we explored the inter-person models to compare with the intra-person model. We found that the classifications of inter-person models were less accurate than the intra-person models, which indicated that the personal data is critical to application and the individual variance is existing. Last but not least, we applied the machine learning method to a questionnaire datasets, and we found important features which could not be find by a typical statistical method. These finding of the thesis provide important insights on e-learning and educational data mining.

Appendix

Appendix A. The Manual of annotation

The manual of annotation is written in both Japanese and English.


Manual_for_annotation(for_public).md

7/26/2023

Instruction for annotation アノテーションの手順

Steps

1. Open "**via_video_annotator.html**" by your browser. (recommend: Google Chrome) It doesn't need Internet connection.

2. click the button  to open the project "**via_project_for_engagement.json**".

3. If you opened the project, please choose a video file to start annotating.

4. We have 4 levels of engagement, please copy these labels of level and paste them:


- 4_VeryHigh
- 3_High
- 2_Low
- 1_VeryLow

Please check your annotator is like:



6. If you are well prepared, you can start annotation. Please use your keyboard to do the annotation work, it will be easier for you.

- use up↑ and down↓ to choose the right track of engagement value
- press 'a' when you find the moment when engagement level is changing.
- press 'space' can pause or play the video. (If you want to take a break, you can pause the video by that. You can rest at anytime.)
- You don't need to worry about the lasting time, we will do that after we get your annotation file.

7. After finishing one video, please click this button  to export and download your annotation results.

- select export format: Only Temporal Segments as CSV

- Select Export Format:

- click the button: Export
- Please change the file name and make it be the same as the video file name also include labeller's name.

The following paragraphs introduce the criterion of annotating. Besides, the first 5 to 10 minutes of every video are usually the preparation of the experiment (e.g., conversations between a participant and an experimenter are recorded). You can use that period of videos to practice annotating. (In data analysis, the videos taken before the experiment will be discarded.)

ラベル付けの標準を紹介します。ちなみに、すべてのビデオの最初から5~10分間くらい実験を準備する時間なので(被験者は実験操作者と話したりするとか)、この頃のアノテーションは練習としてやってください。(実際の分析は、この実験開始前の動画を抜けます。)

Besides, here are some explanations of the engagement levels. (reference: Whitehill, 2014)

エンゲージメントレベルについて、説明は以下になります。(引用先: Whitehill, 2014)

Please use the same criterion when you annotate the videos. Although the engagement level might differ for every person, the criterion should still be on overall people, not on a particular person. Therefore, one person's average engagement level might be higher or lower than others.

レベルのラベルを付けている作業は、全被験者ら同じような標準でラベルを付けてください。人によって集中力は違いますが、本研究の目的として、個体のエンゲージメントのレベルの判断ではなく、全個体の通用のエンゲージメントのレベルの判断が欲しいです。故に、1人の被験者のエンゲージメントラベル結果は高いレベルばかりか低いレベルばかりか傾いている場合もあります。

level 4: Very high engagement

student could be "commended" for his/her level of engagement in task. he/she seems enjoy the task and want to figure out something. keep in deeply thinking

レベル4: 非常に高いエンゲージメント

称賛されられるエンゲージメント。タスクに夢中している。何か発見したい顔している。細かいことや奥まで考えている。テンション高い顔。

level 3: High engagement

student requires no admonition to "stay on task"

レベル3: 高いエンゲージメント

言わなくてもタスクに集中し続けそうな感じ。特に夢中していないけどタスクを考えている。

level 2: Low engagement

eyes barely open, clearly not "into" the task, feel bored or not very enjoy in task

レベル2: 低いエンゲージメント

目はほとんど開いていない、タスクに入っていない。退屈を感じて、タスクを楽しんでいない。

level 1: Very low engagement

e.g., looking away from computer and obviously not thinking about task, eyes completely closed, listen to someone beside he/she (not focus on task)

レベル1：非常に低いエンゲージメント

例えば、コンピューター以外のところを見る。明らかに目の前のタスクについて考えていない。目を完全に閉じている。周りの人の話を聞いている（実験の説明の段階だけど、スクリーンを見てるのに、スクリーンに集中してない。つまり、ターゲット以外に集中する）。

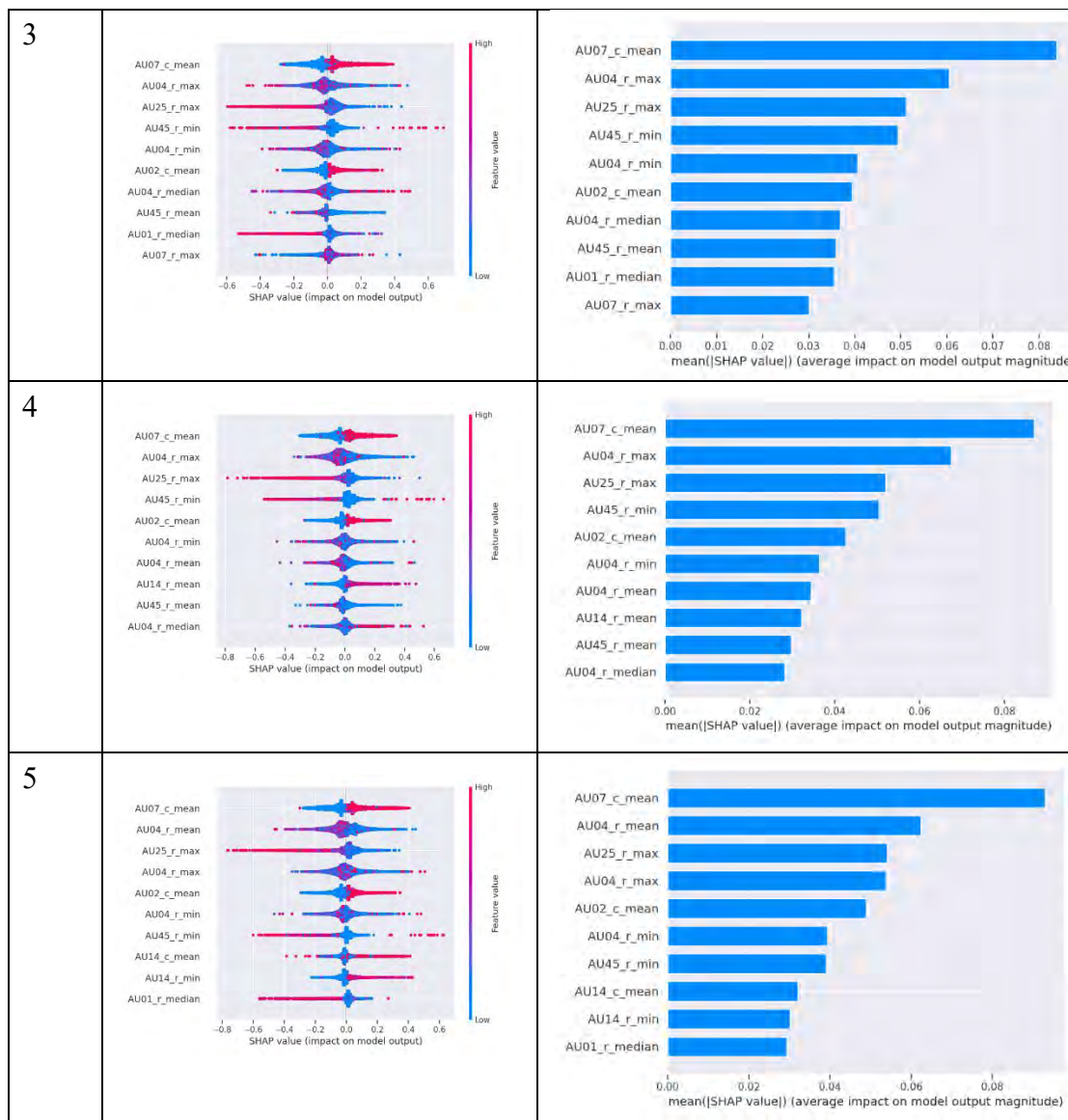
Appendix B. The Results of SHAP analysis

B-1. Intra-person learning results of estimation of engagement states

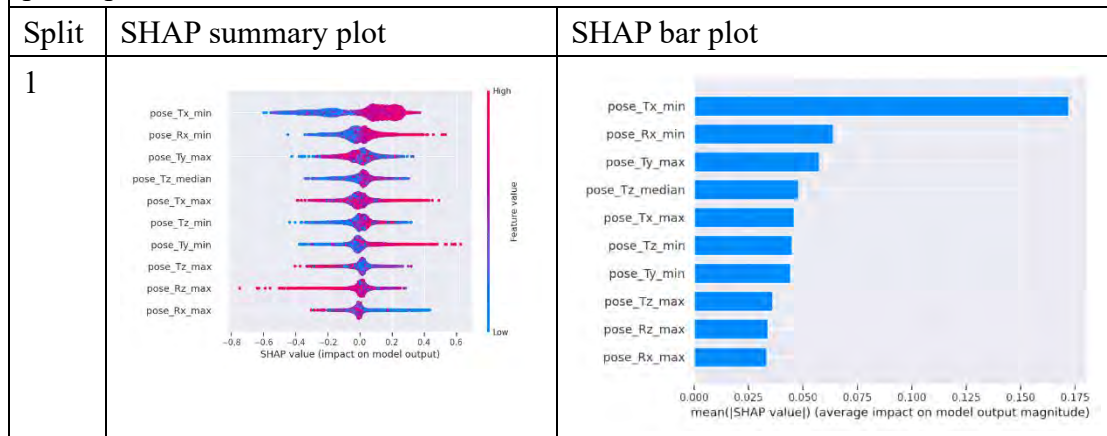
The intra-learning were conducted in 5-fold cross-validation. This part showed all SHAP analysis results in summary plot and bar plot made by the SHAP library. However, because the data was too large to show, here we only provided the examples from Basic AUs feature set and Head Pose feature set.

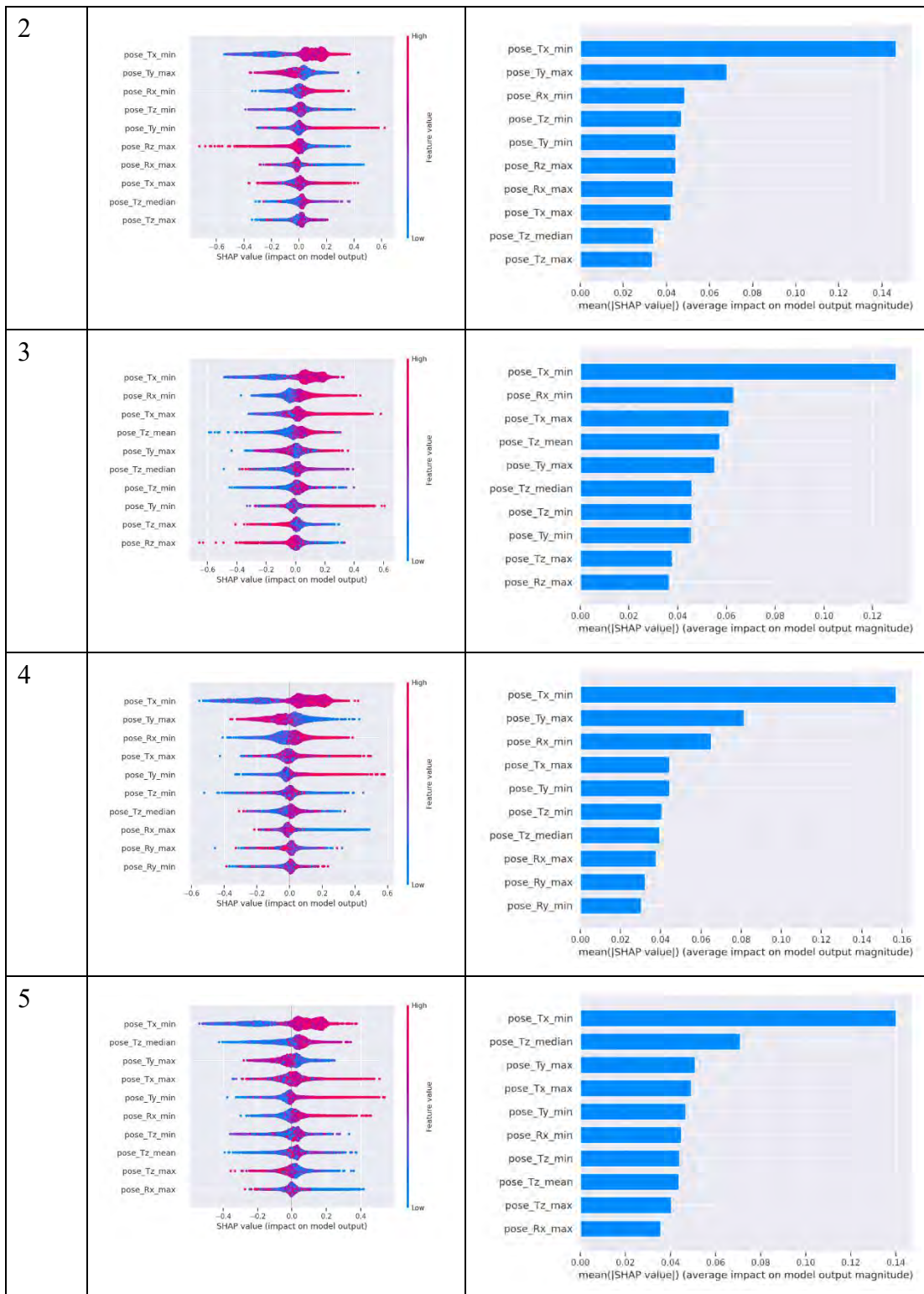
B-1-1 Japan's data (Training on Japan's data; Testing on Japan's data)

The results of classifying the engagement states by Basic AUs features in Japan's participants.		
Split	SHAP summary plot	SHAP bar plot
1		
2		



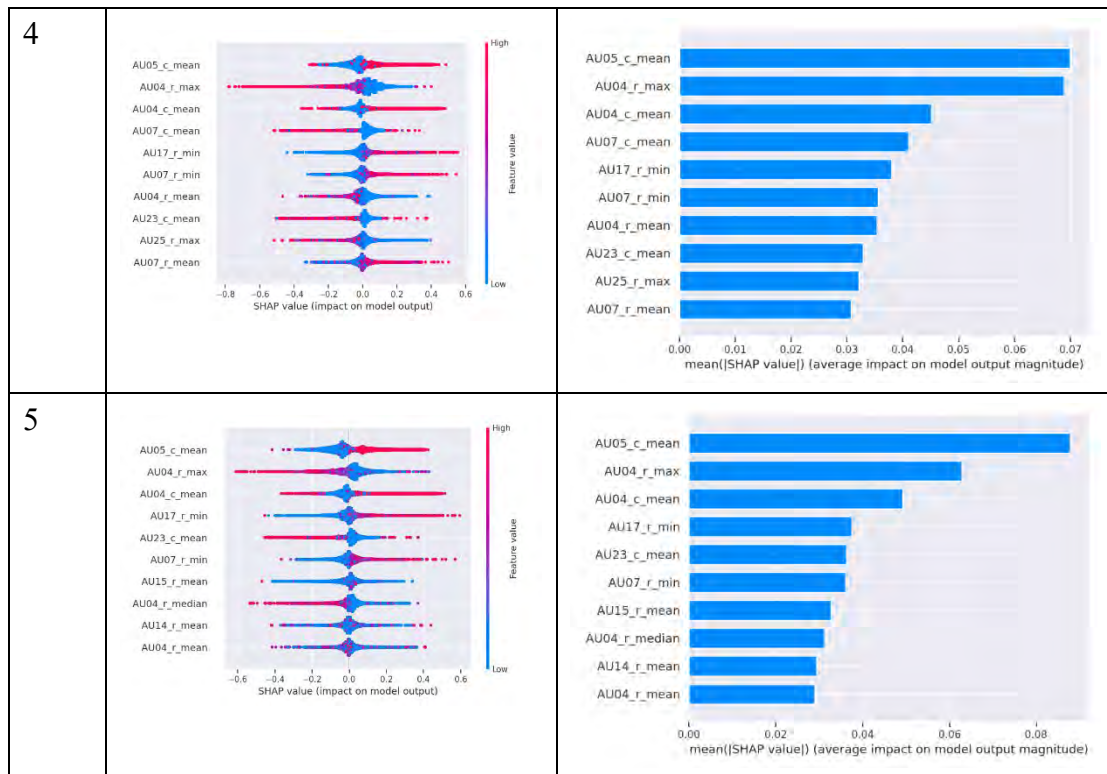
The results of classifying the engagement states by Head Pose feature set in Japan's participants.



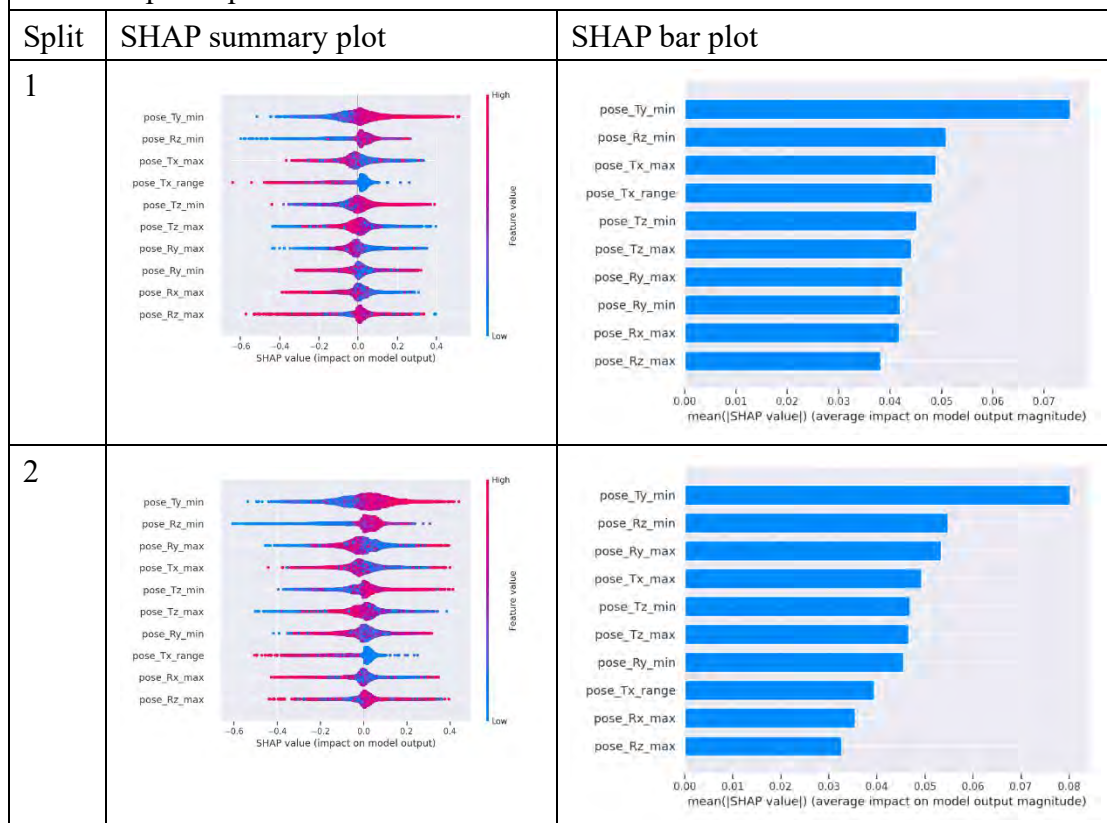


B-1-2 Taiwan's data (Training on Taiwan's data; Testing on Taiwan's data)

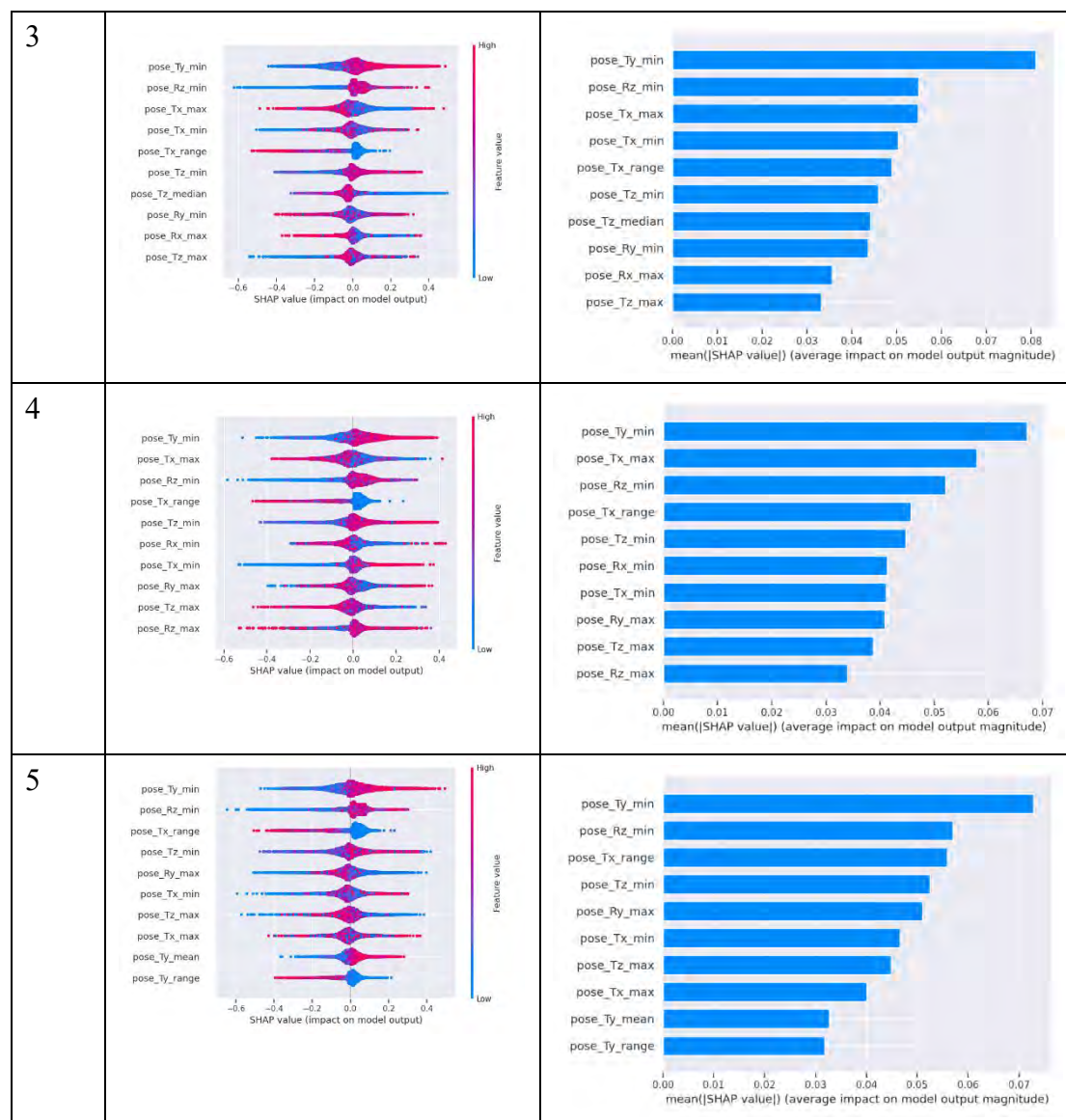
The results of classifying the engagement states by Basic AUs features in Taiwan's participants.		
Split	SHAP summary plot	SHAP bar plot
1		
2		
3		



The results of classifying the engagement states by Head Pose feature set in Taiwan's participants.



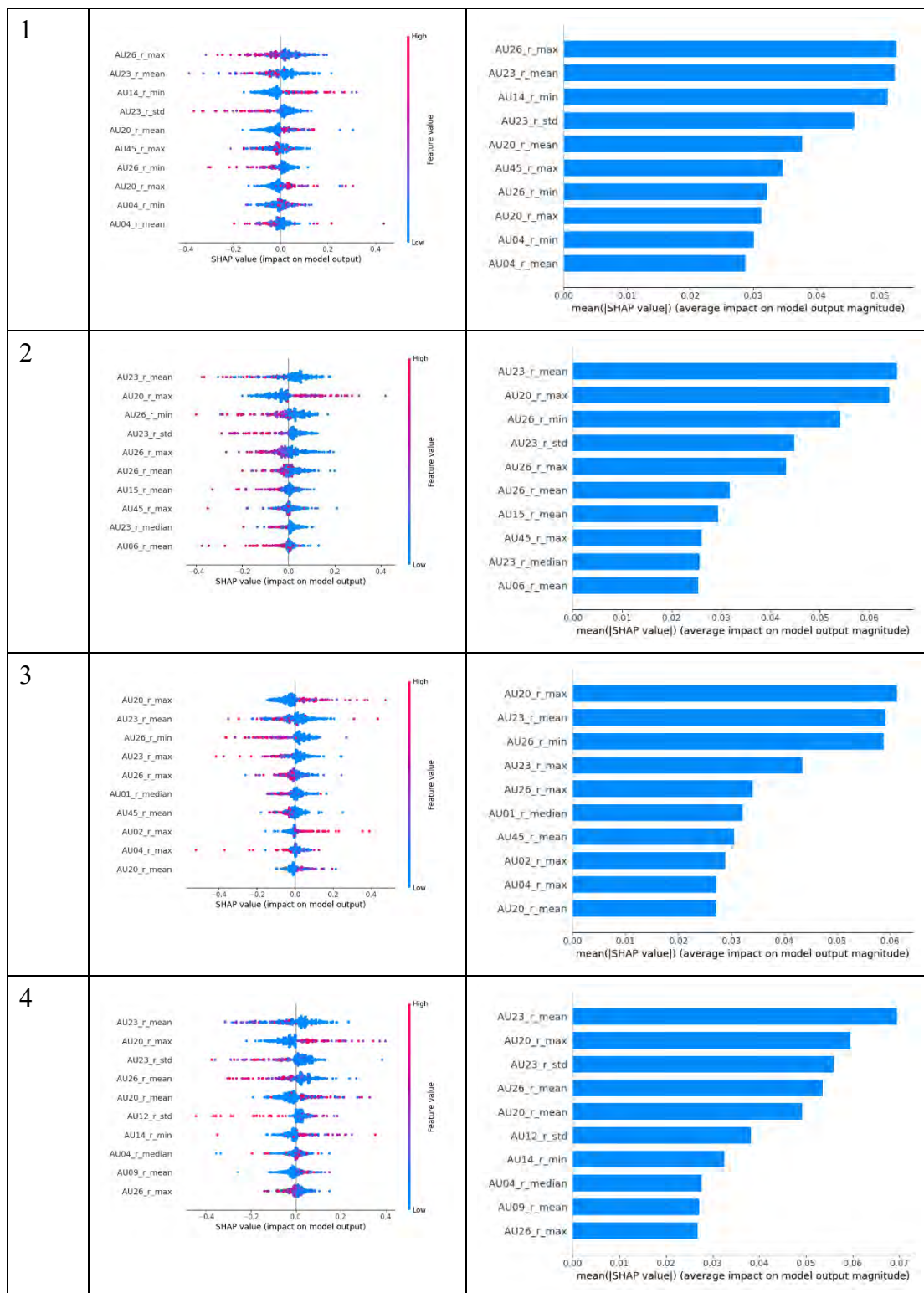
Appendix



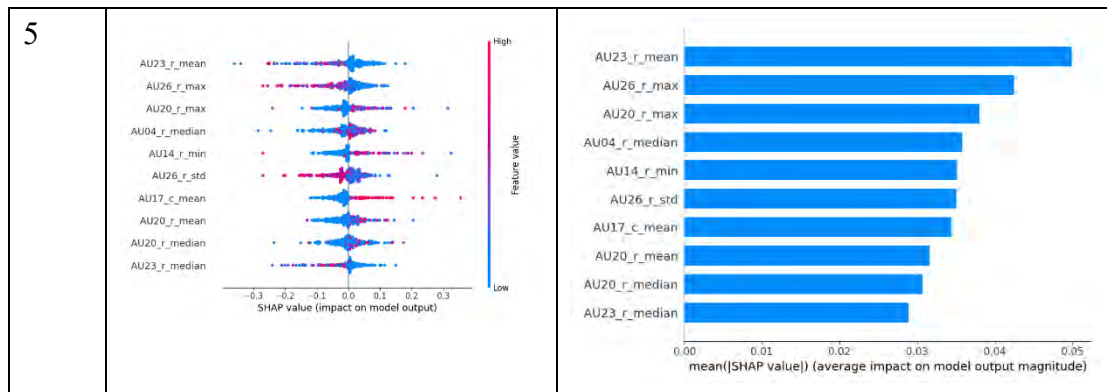
B-2. Intra-person learning results of classifying help-seeking states

B-2-1 Japan's data (Training on Japan's data; Testing on Japan's data)

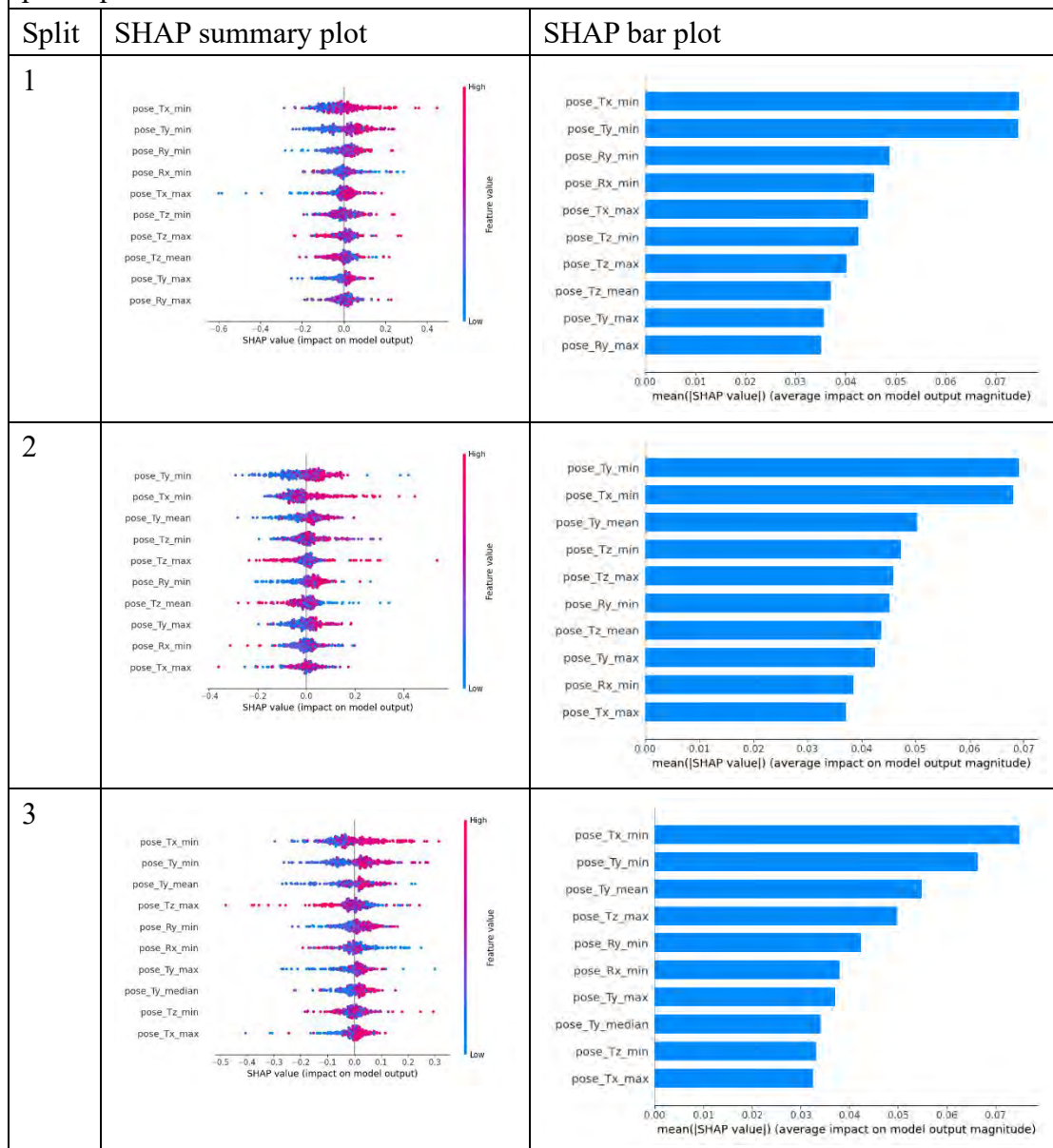
<p>The results of classifying the engagement states by Basic AUs feature set in Japan's participants.</p>		
<p>Split</p>	<p>SHAP summary plot</p>	<p>SHAP bar plot</p>

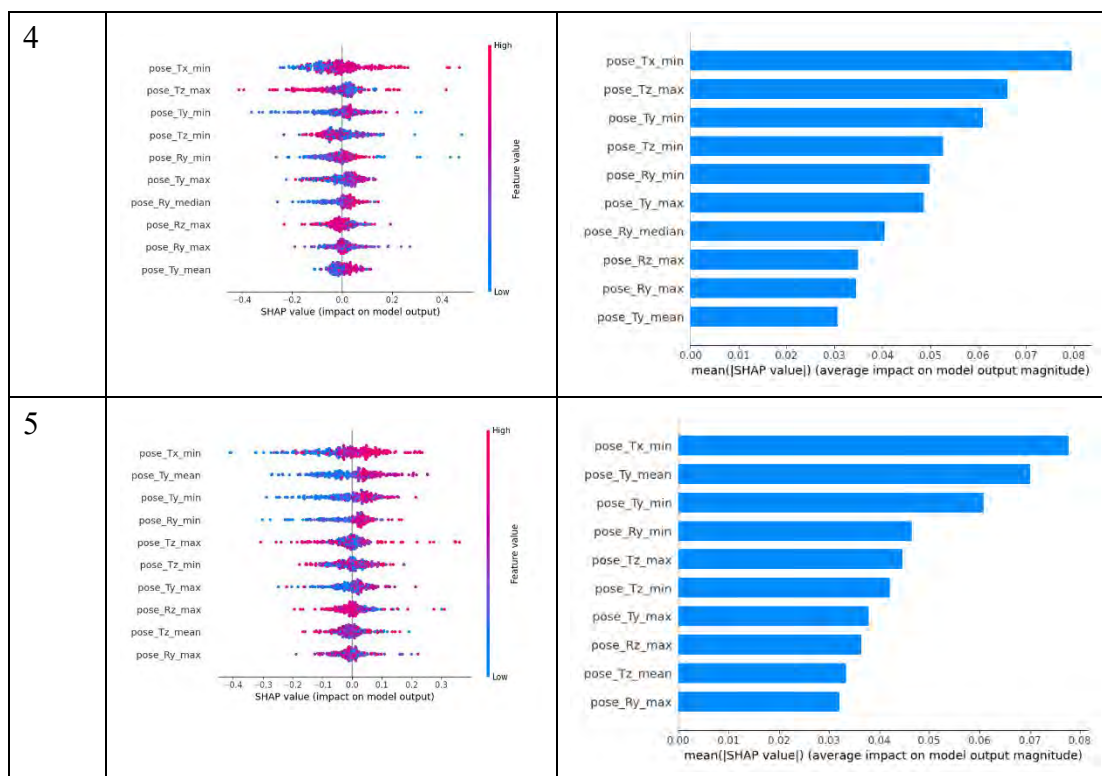


Appendix

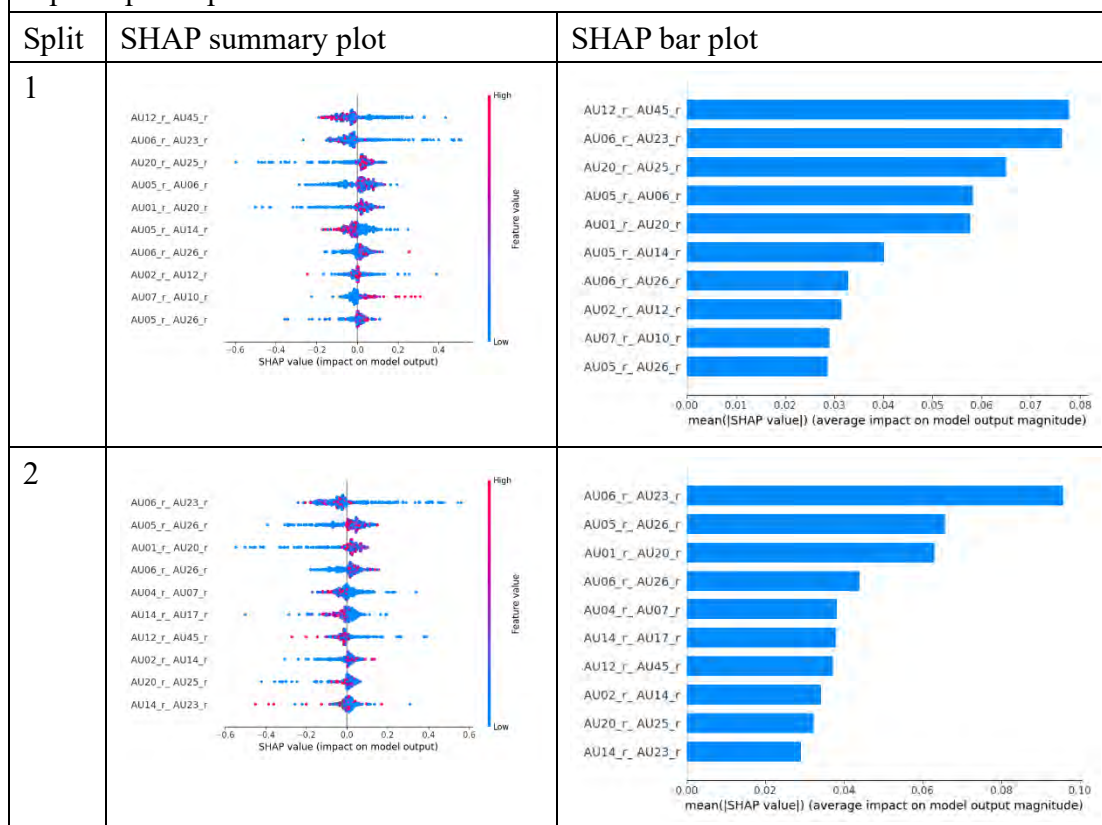


The results of classifying the engagement states by Head Pose feature set in Japan's participants.

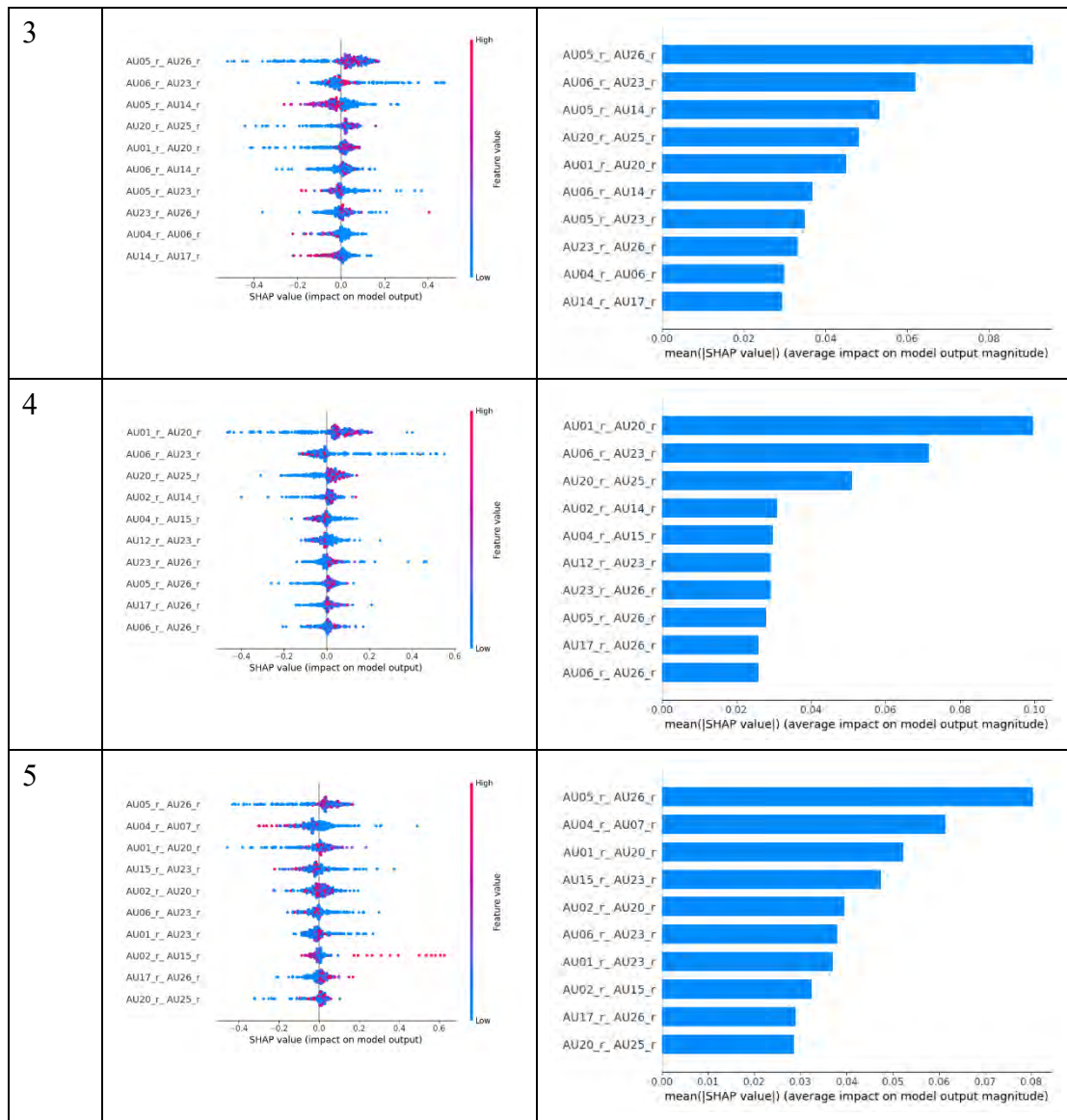




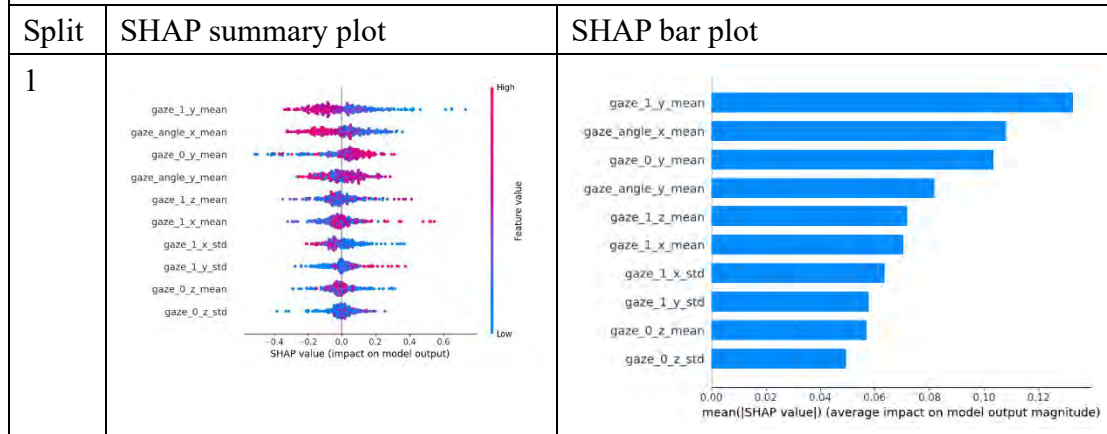
The results of classifying the engagement states by Co-occurring AUs feature set in Japan's participants.



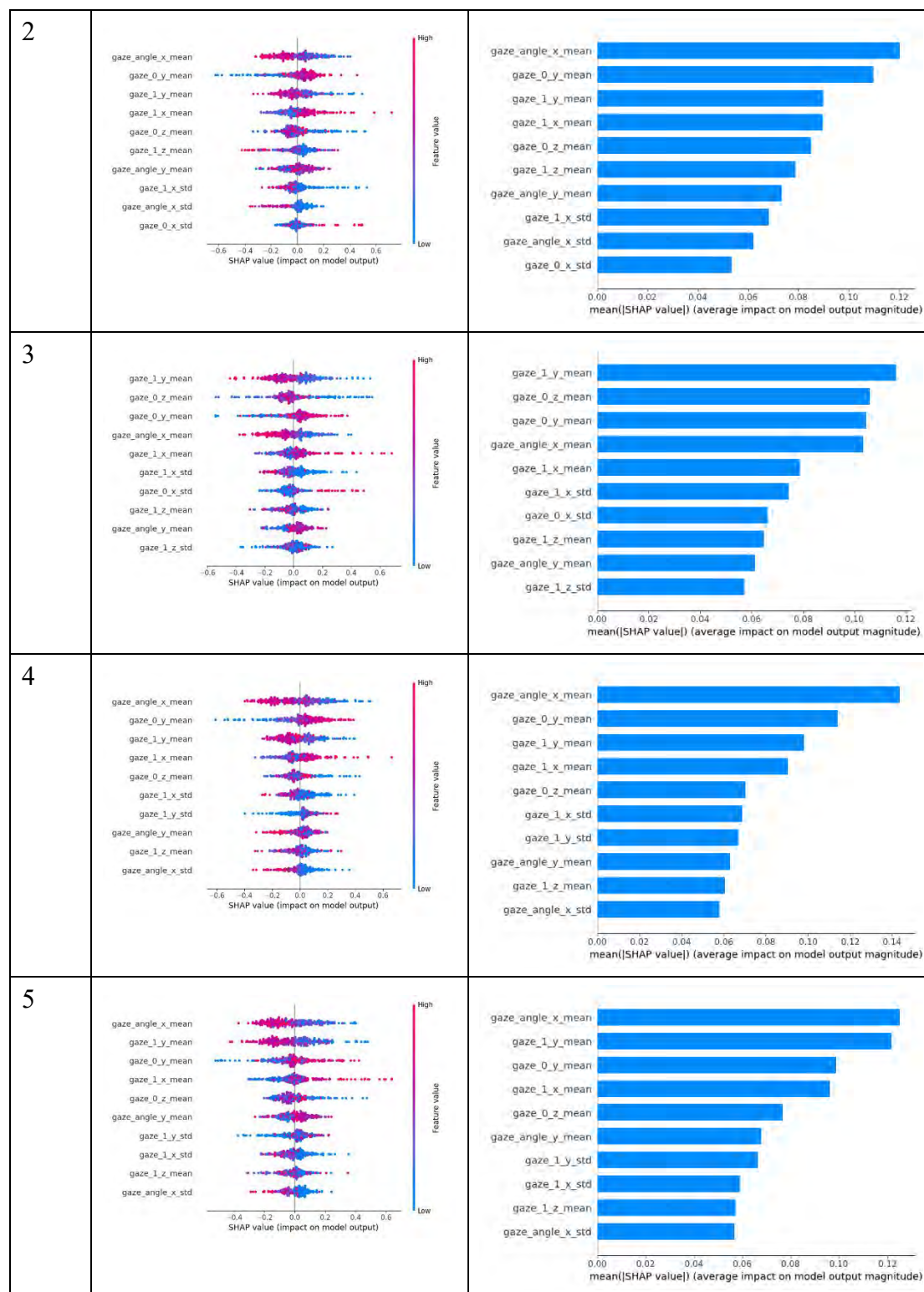
Appendix



The results of classifying the engagement states by Gaze feature set in Japan's participants.



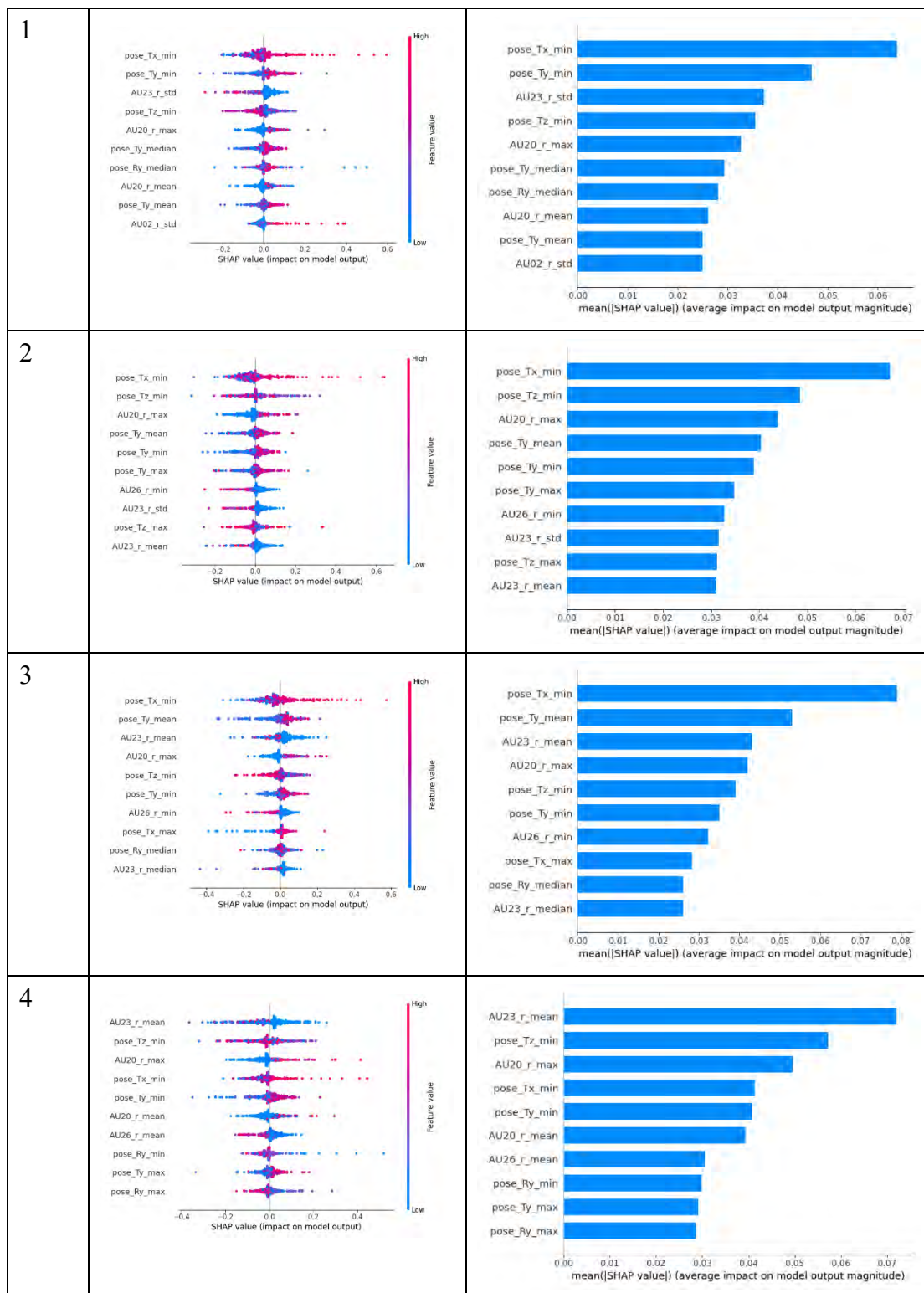
Appendix

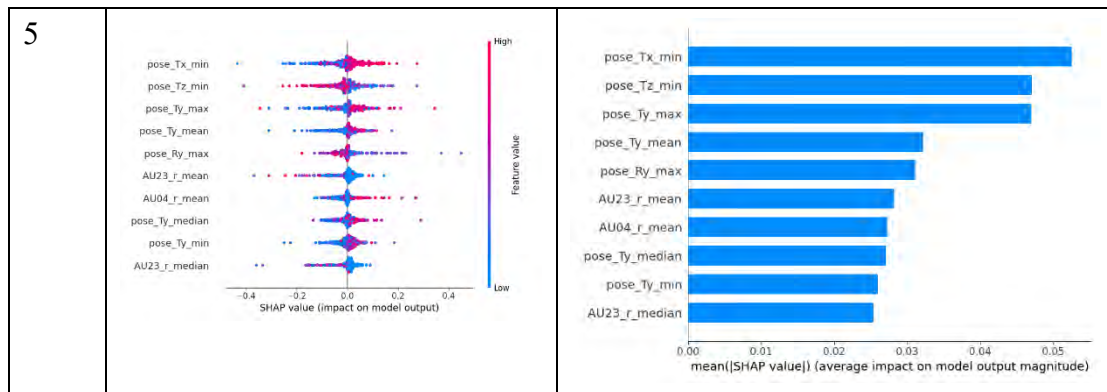


The results of classifying the engagement states by Basic AUs & Head Pose feature set in Japan's participants.

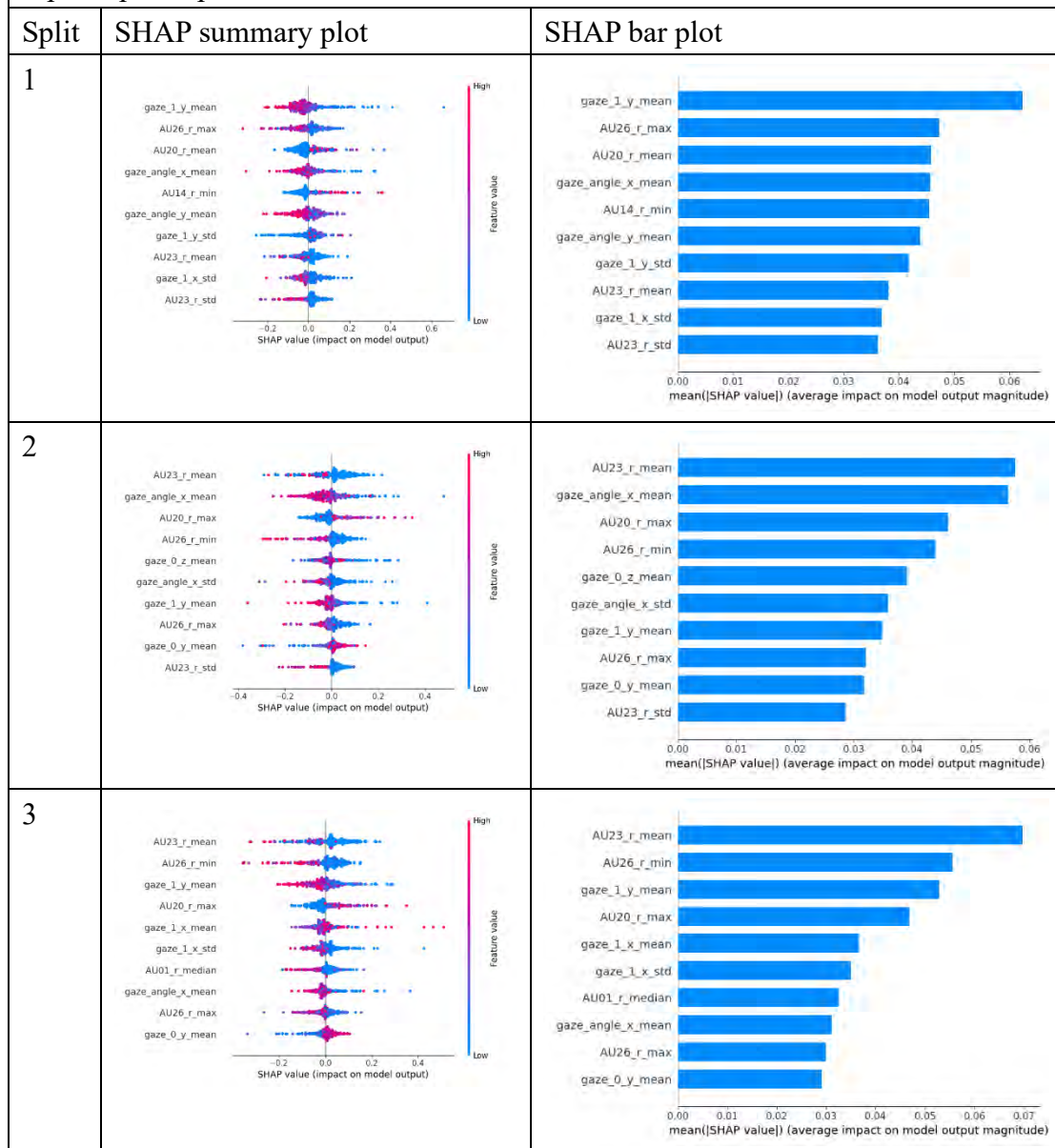
Split	SHAP summary plot	SHAP bar plot
-------	-------------------	---------------

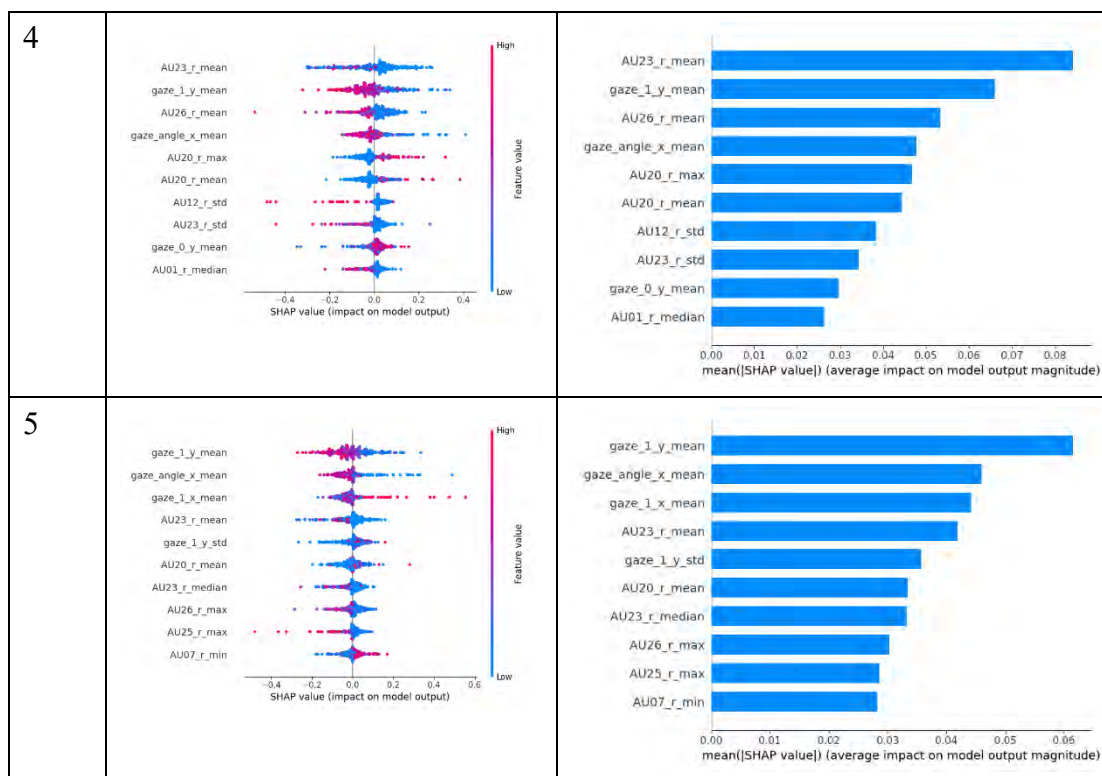
Appendix



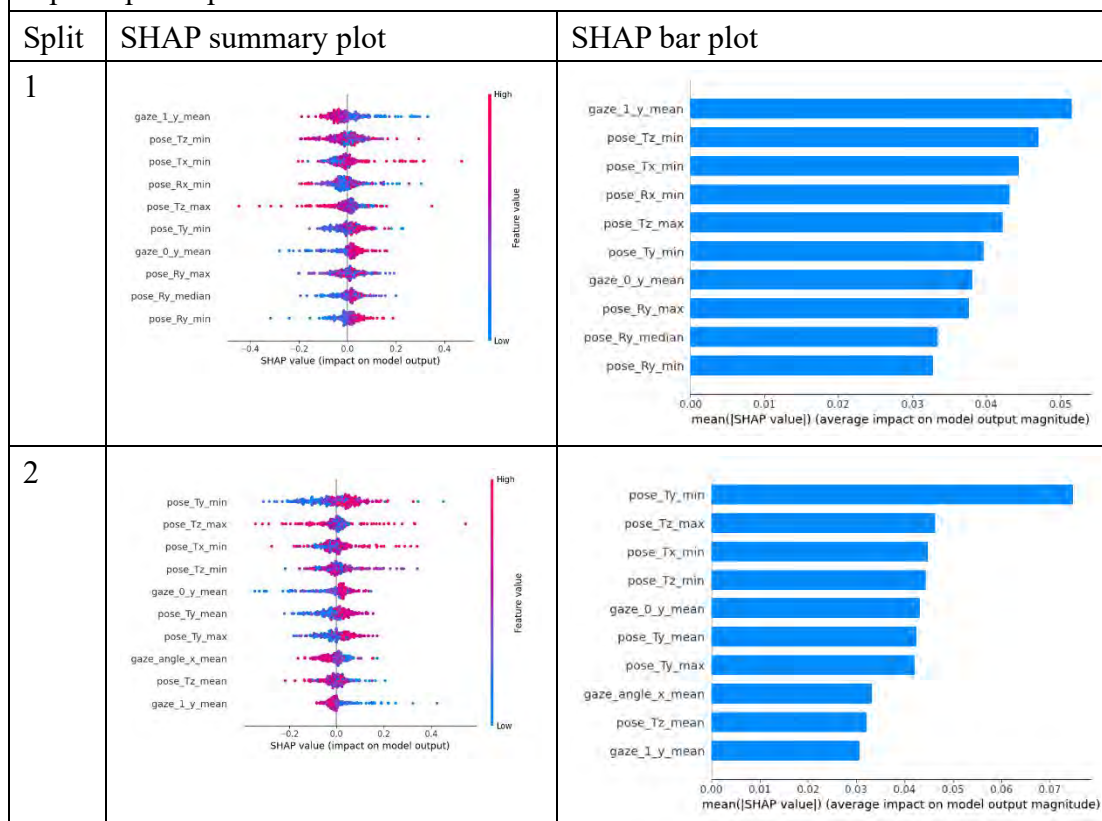


The results of classifying the engagement states by Basic AUs & Gaze feature set in Japan's participants.

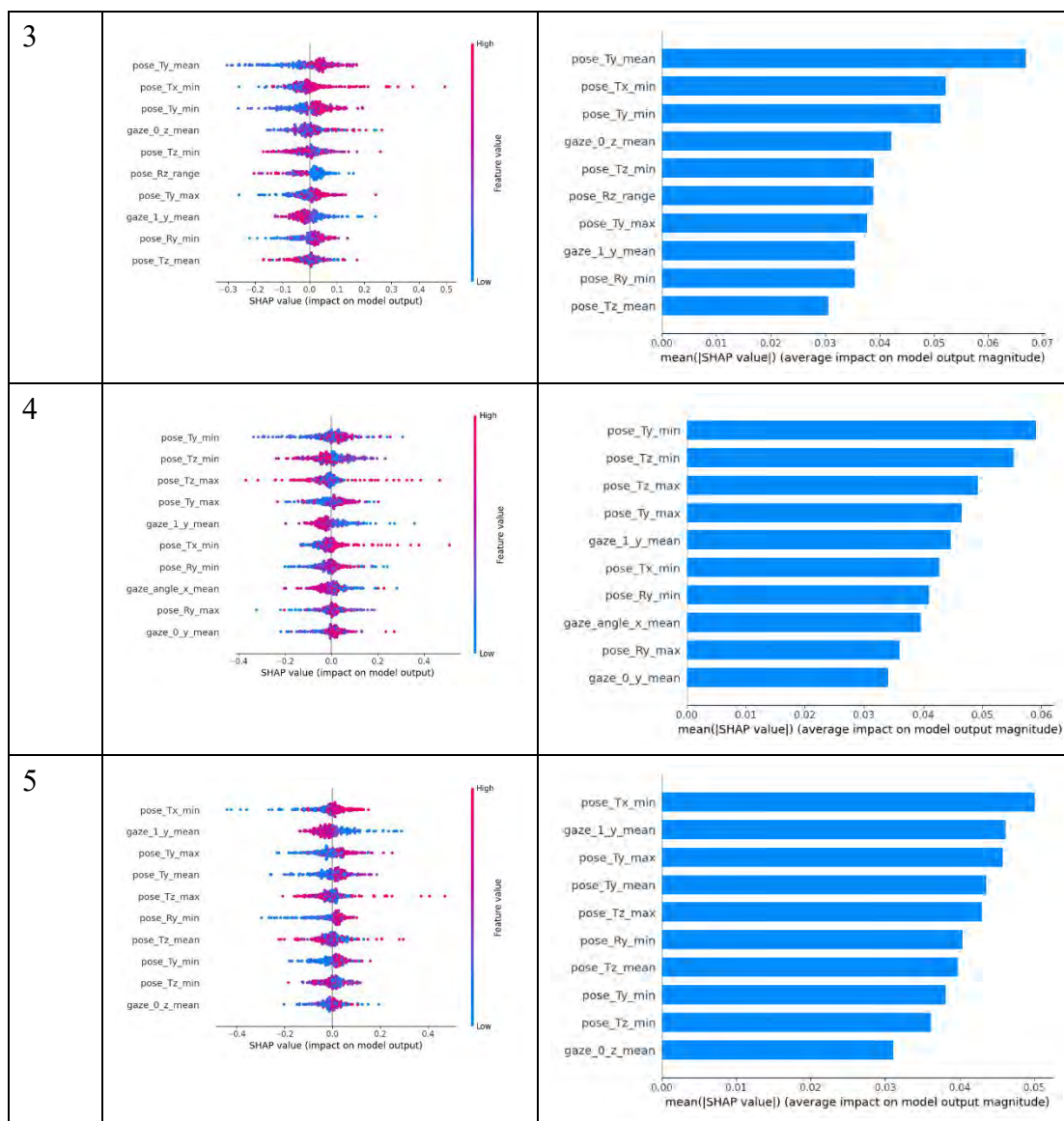




The results of classifying the engagement states by Head Pose & Gaze feature set in Japan's participants.



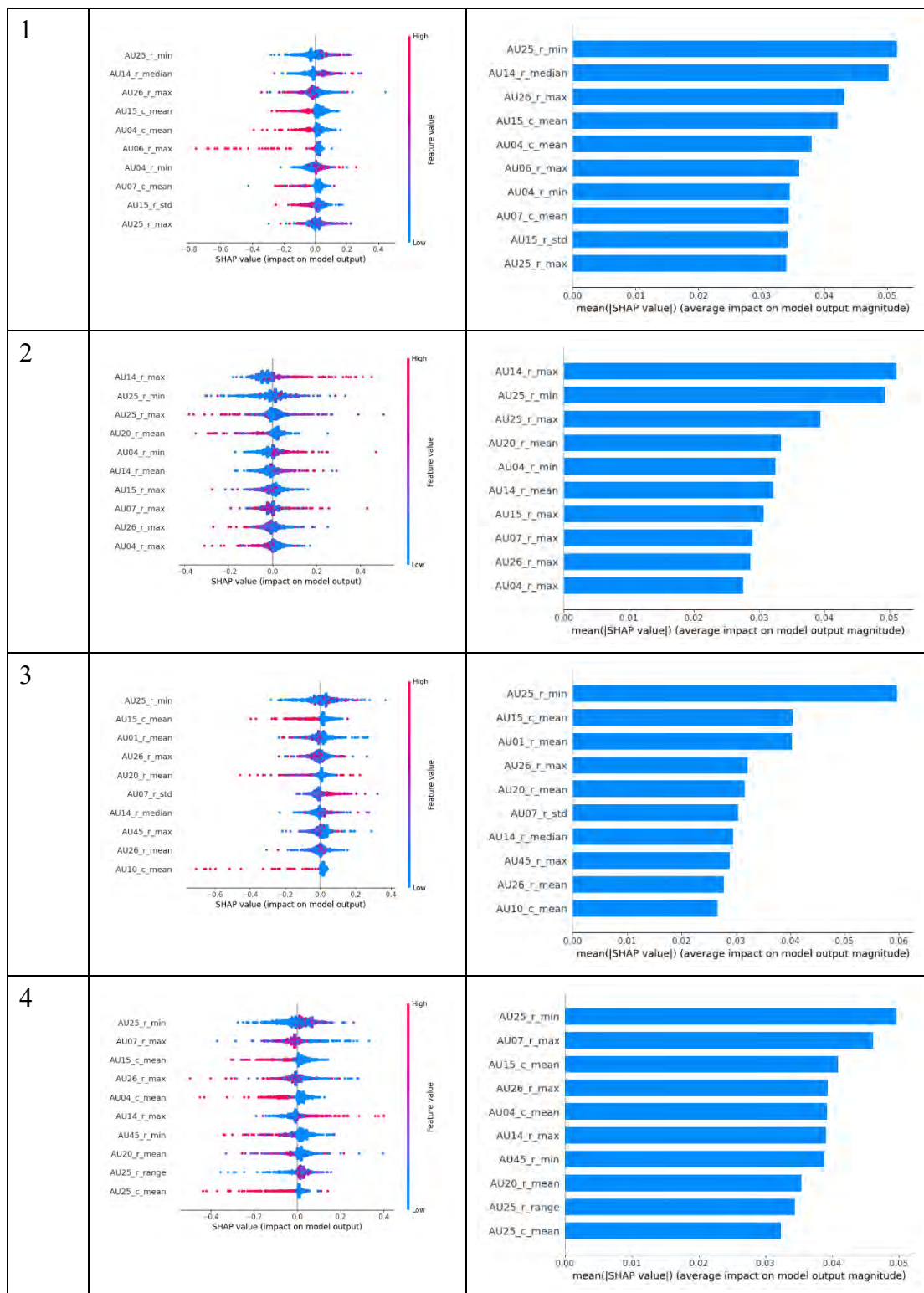
Appendix

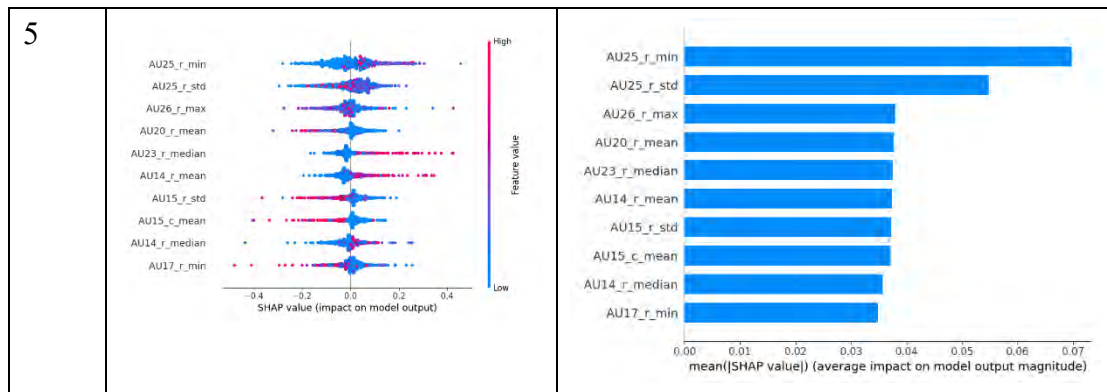


B-2-2 Taiwan's data (Training on Taiwan 's data; Testing on Taiwan's data)

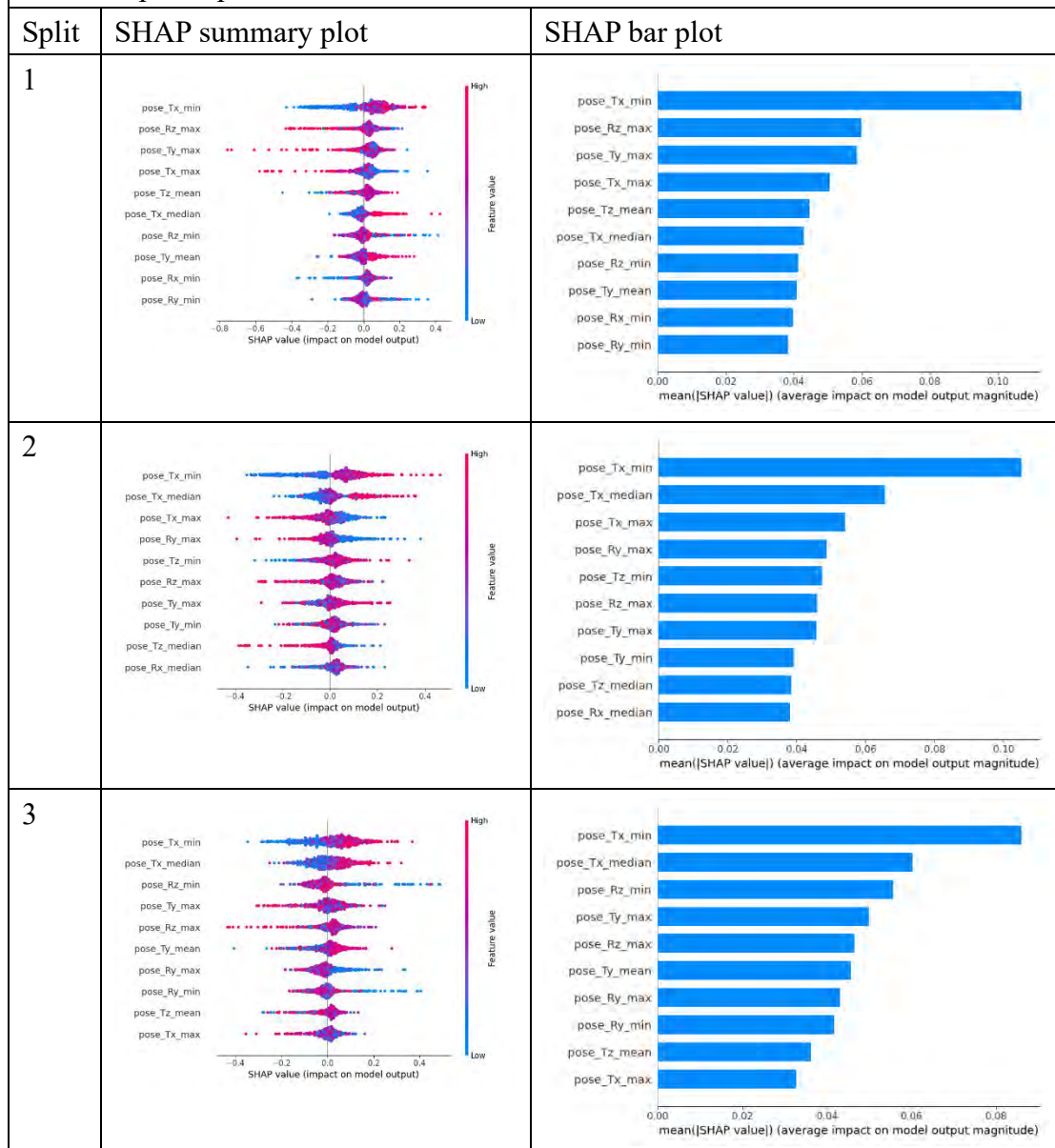
<p>The results of classifying the help-seeking states by Basic AUs feature set in Taiwan's participants.</p>		
<p>Split</p>	<p>SHAP summary plot</p>	<p>SHAP bar plot</p>

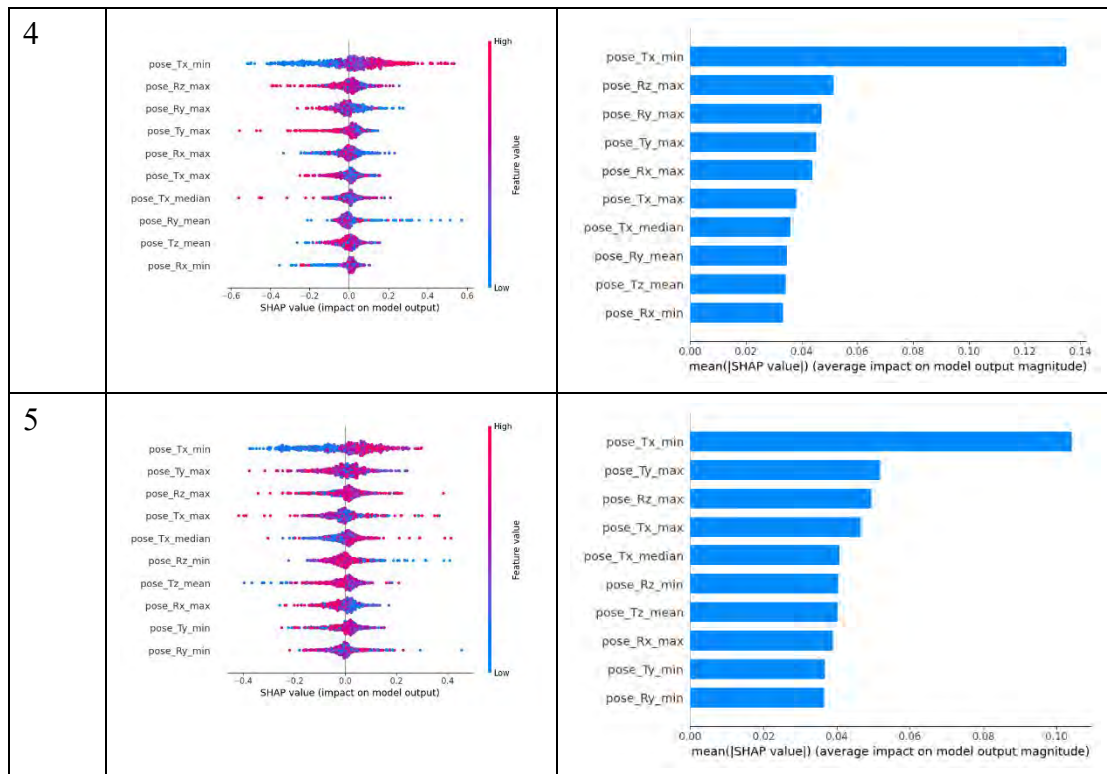
Appendix



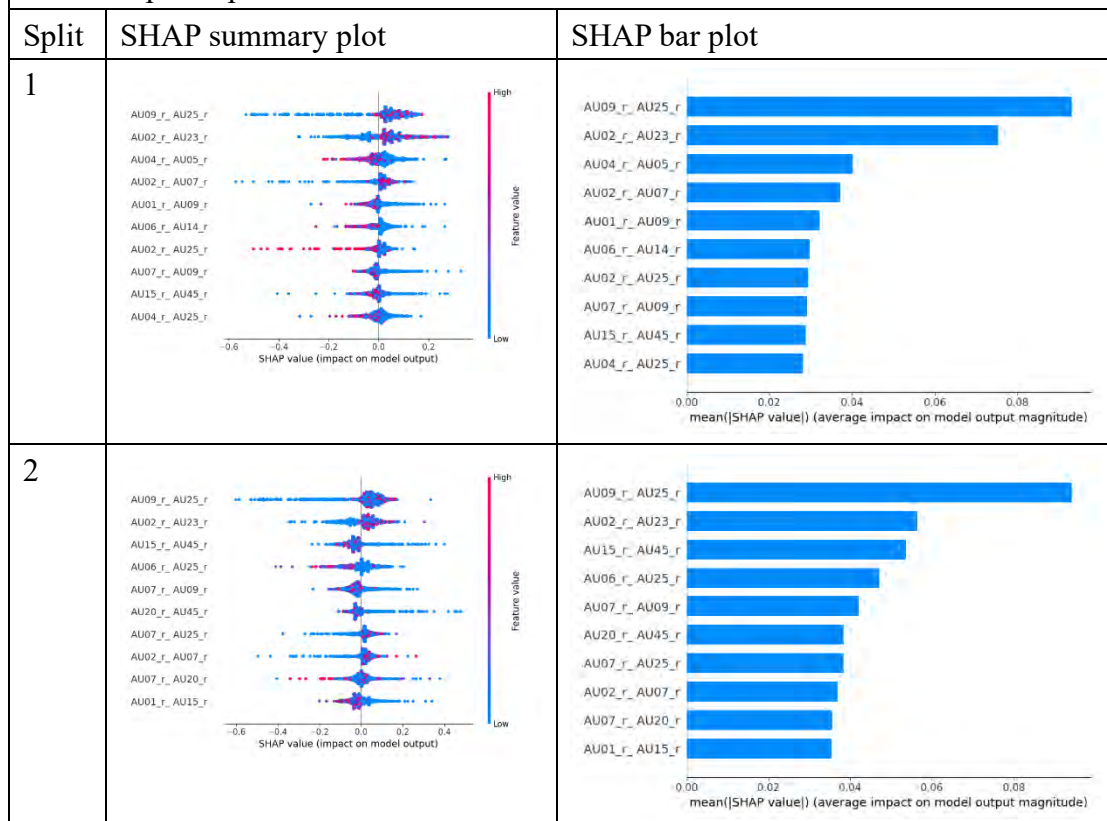


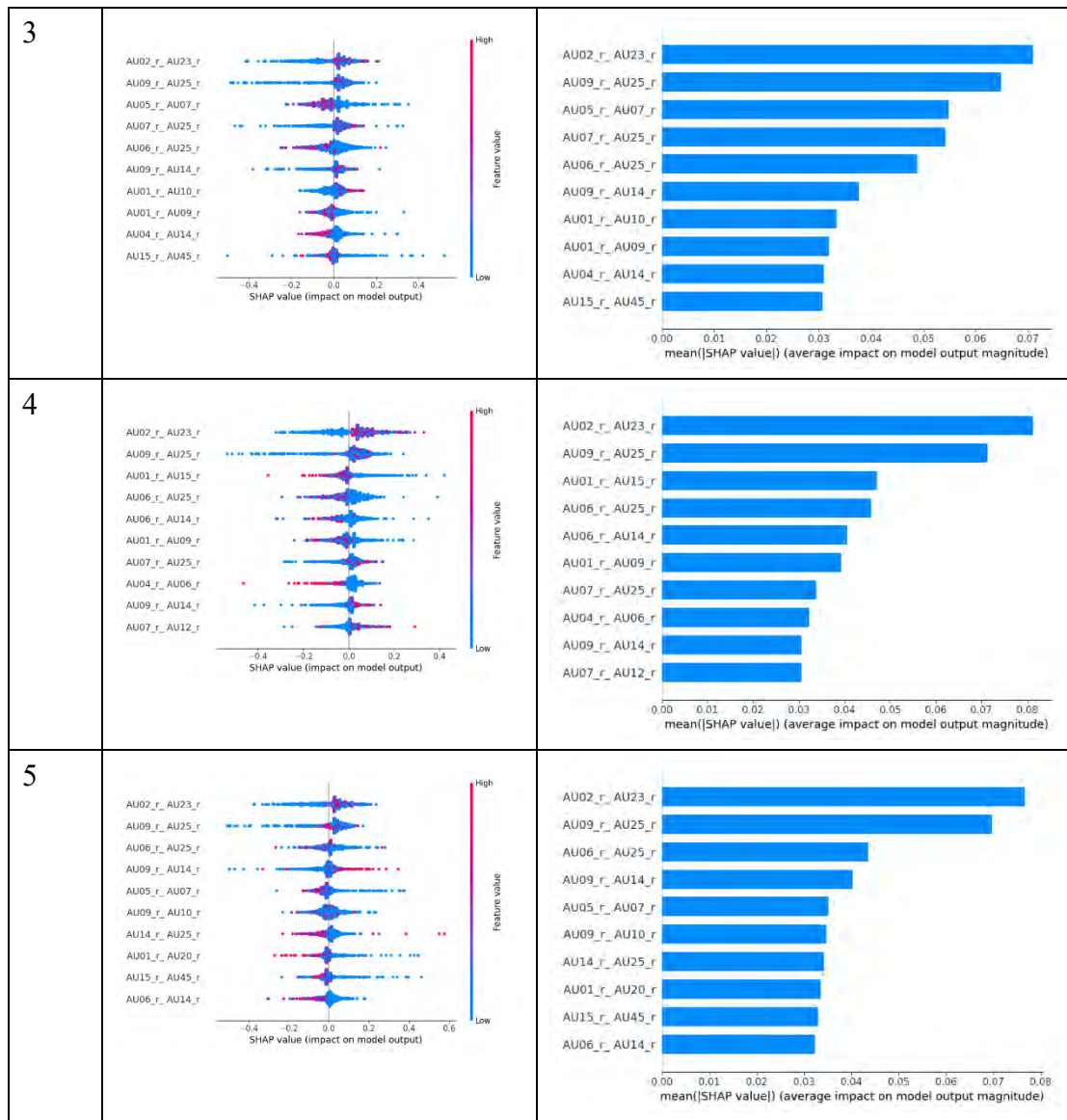
The results of classifying the help-seeking states by Head Pose feature set in Taiwan's participants.



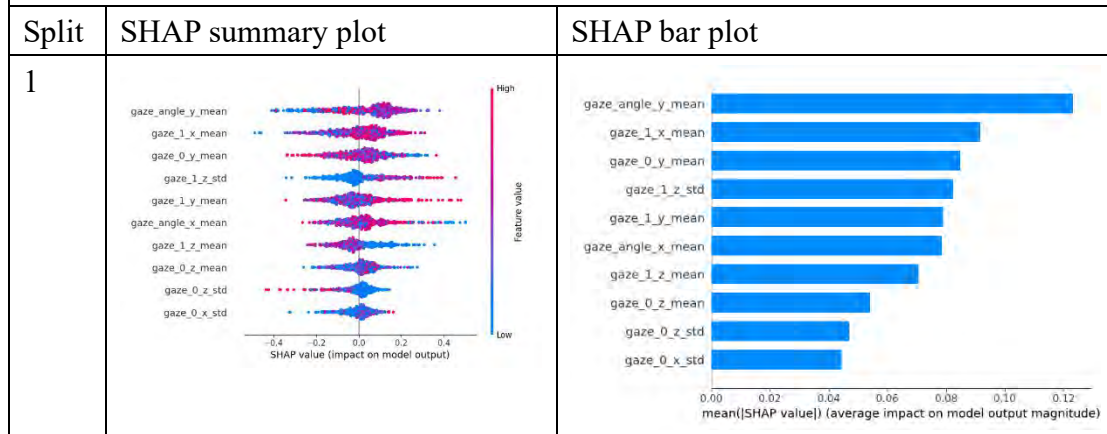


The results of classifying the help-seeking states by Co-occurring AUs feature set in Taiwan's participants.

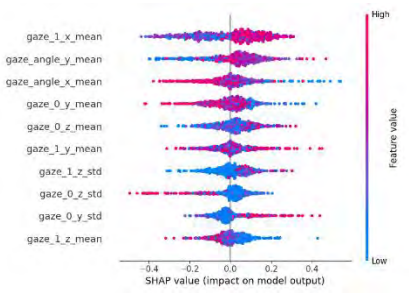
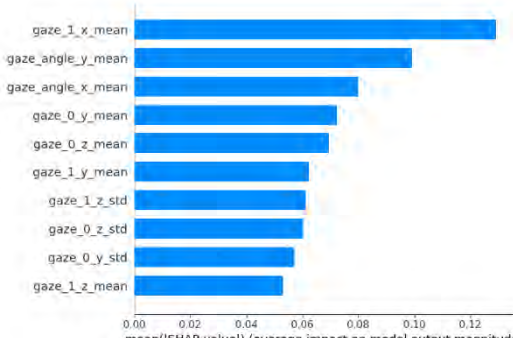
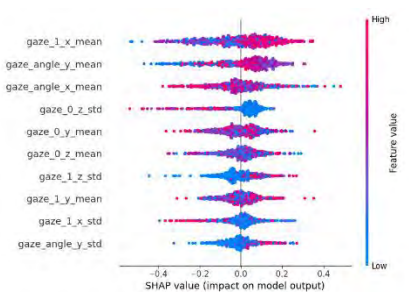
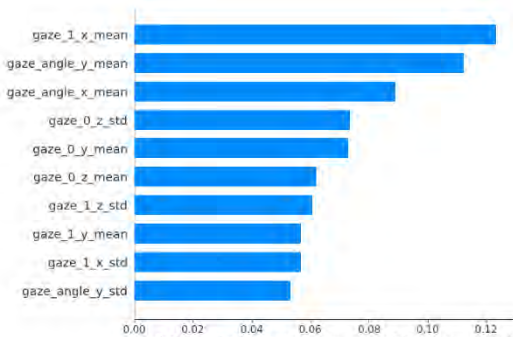
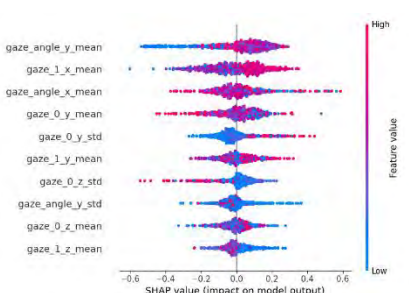
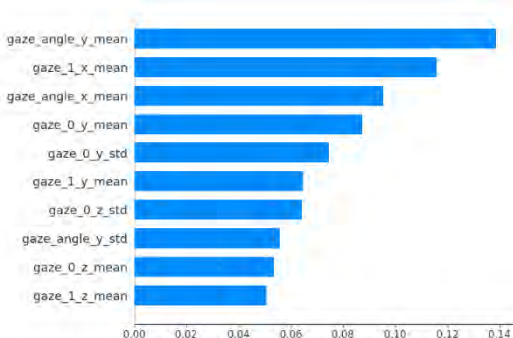
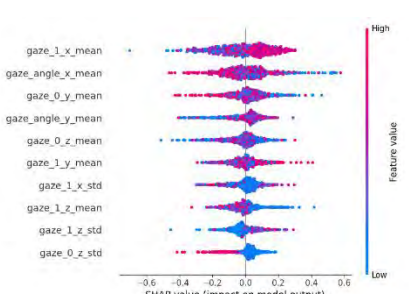
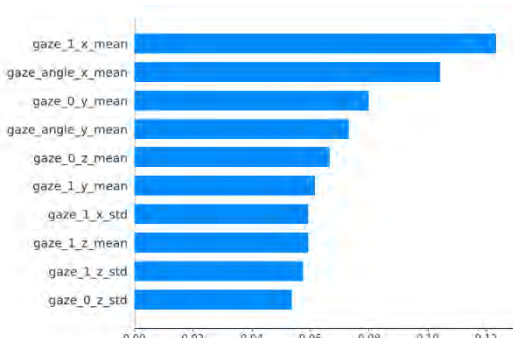




The results of classifying the help-seeking states by Gaze feature set in Taiwan's participants.



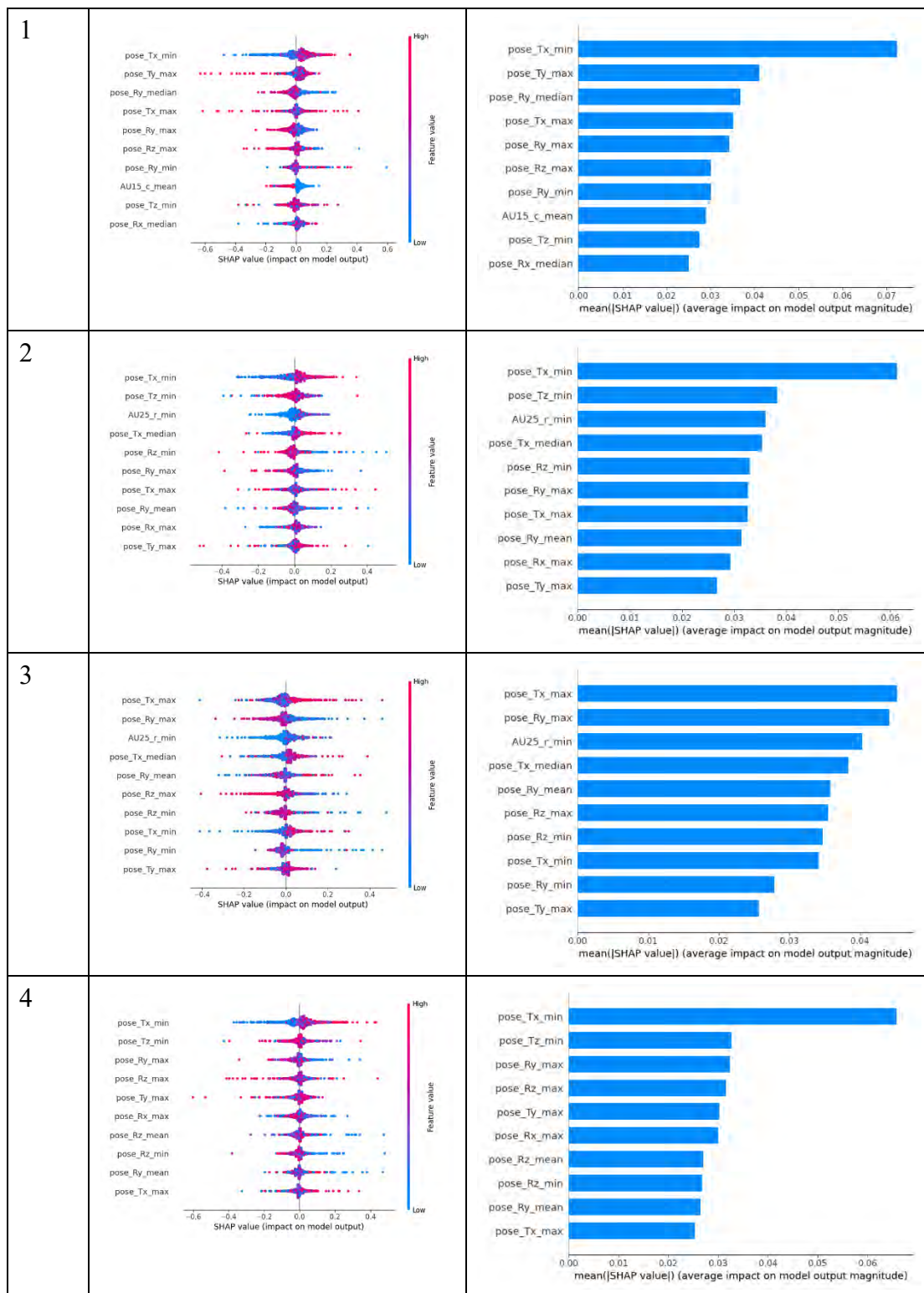
Appendix

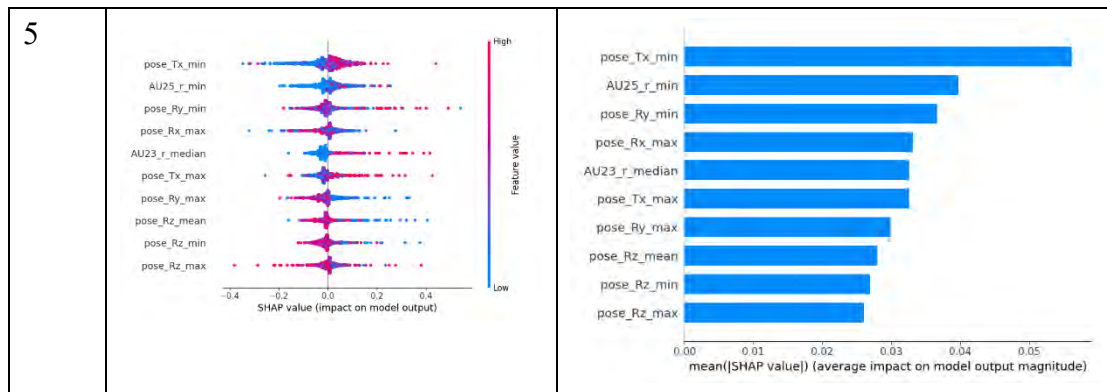
2		
3		
4		
5		

The results of classifying the help-seeking states by Basic AUs & Head Pose feature set in Taiwan's participants.

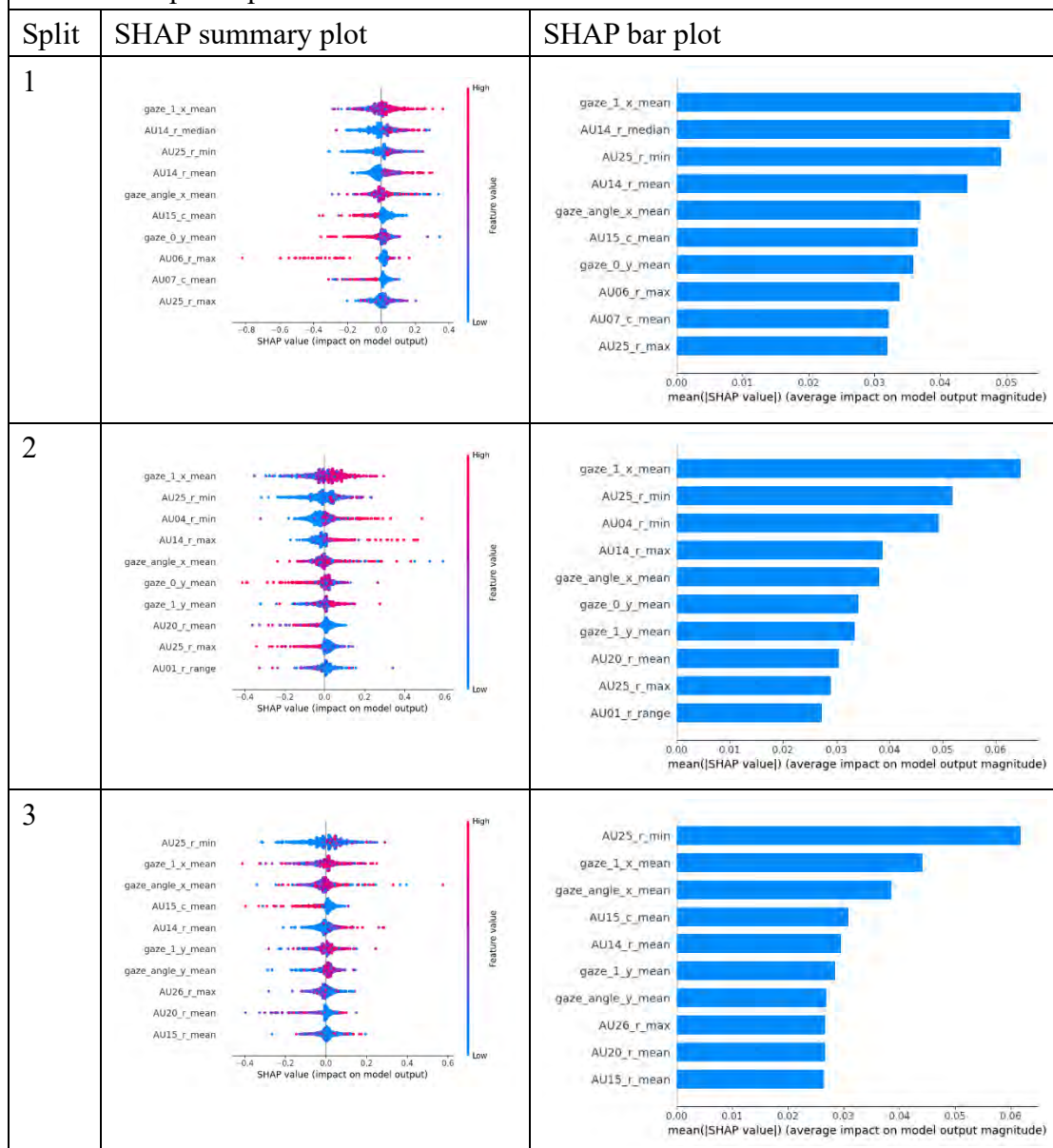
Split	SHAP summary plot	SHAP bar plot
-------	-------------------	---------------

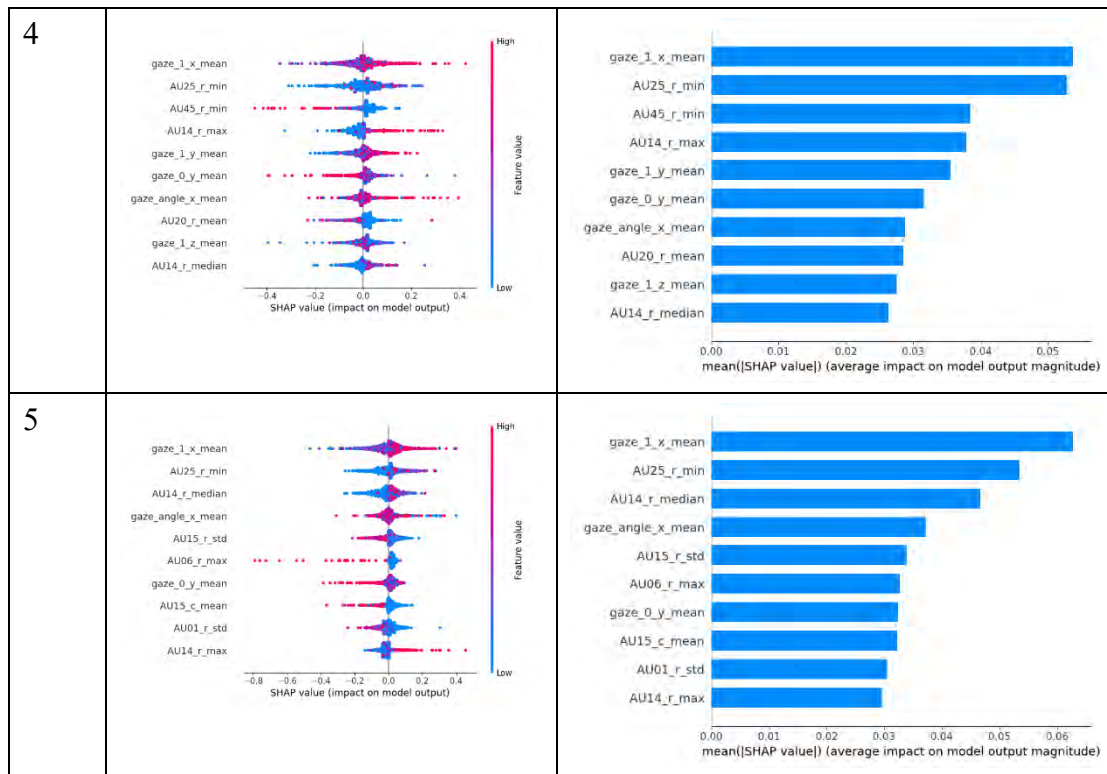
Appendix



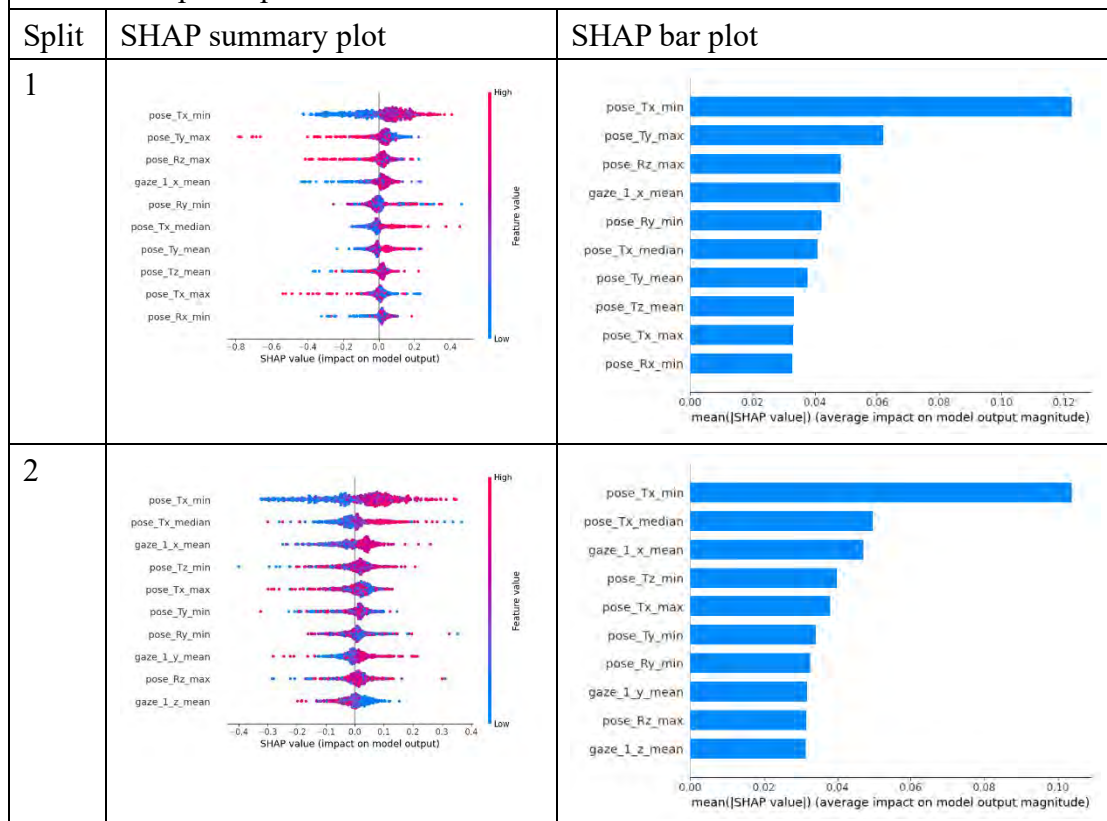


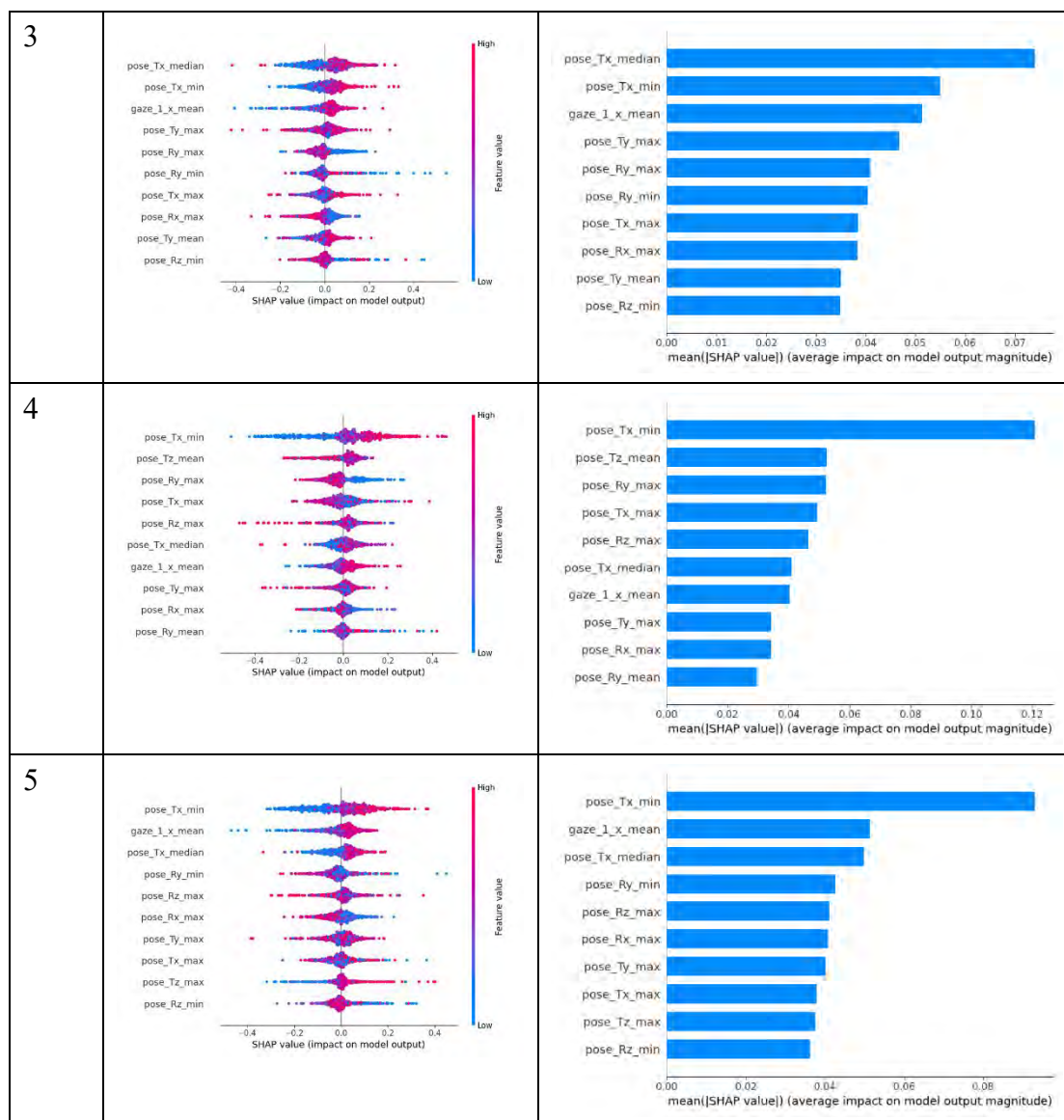
The results of classifying the help-seeking states by Basic AUs & Gaze feature set in Taiwan's participants.





The results of classifying the help-seeking states by Head Pose & Gaze feature set in Taiwan's participants.





B-3. Inter-person learning results of estimation of engagement states

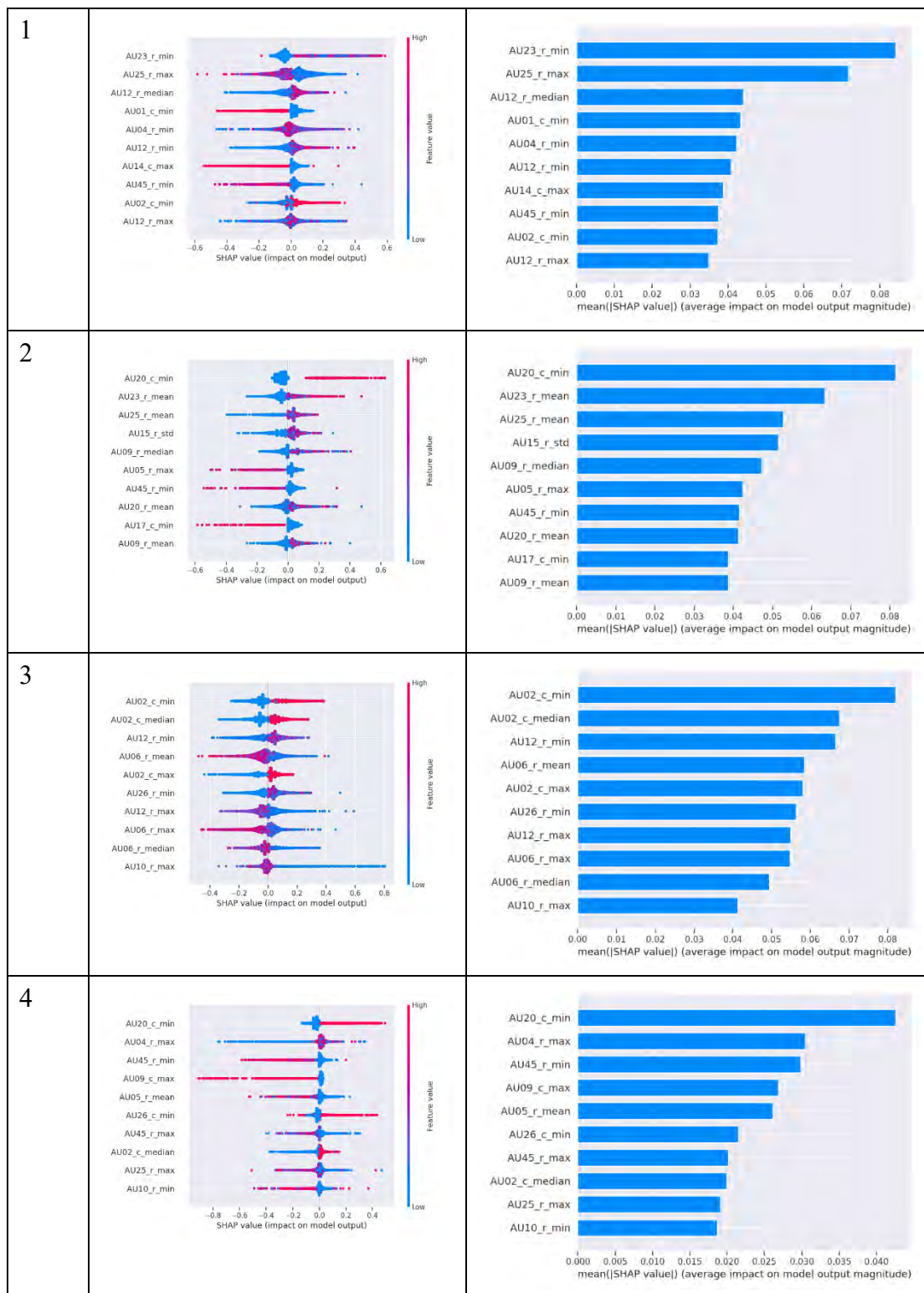
The five types of model are conducted on inter-person learning, but here only showed the detail of the first two:

- (1) Training: Japan’s data; Testing: Japan’s data
- (2) Training: Taiwan’s data; Testing: Taiwan’s data

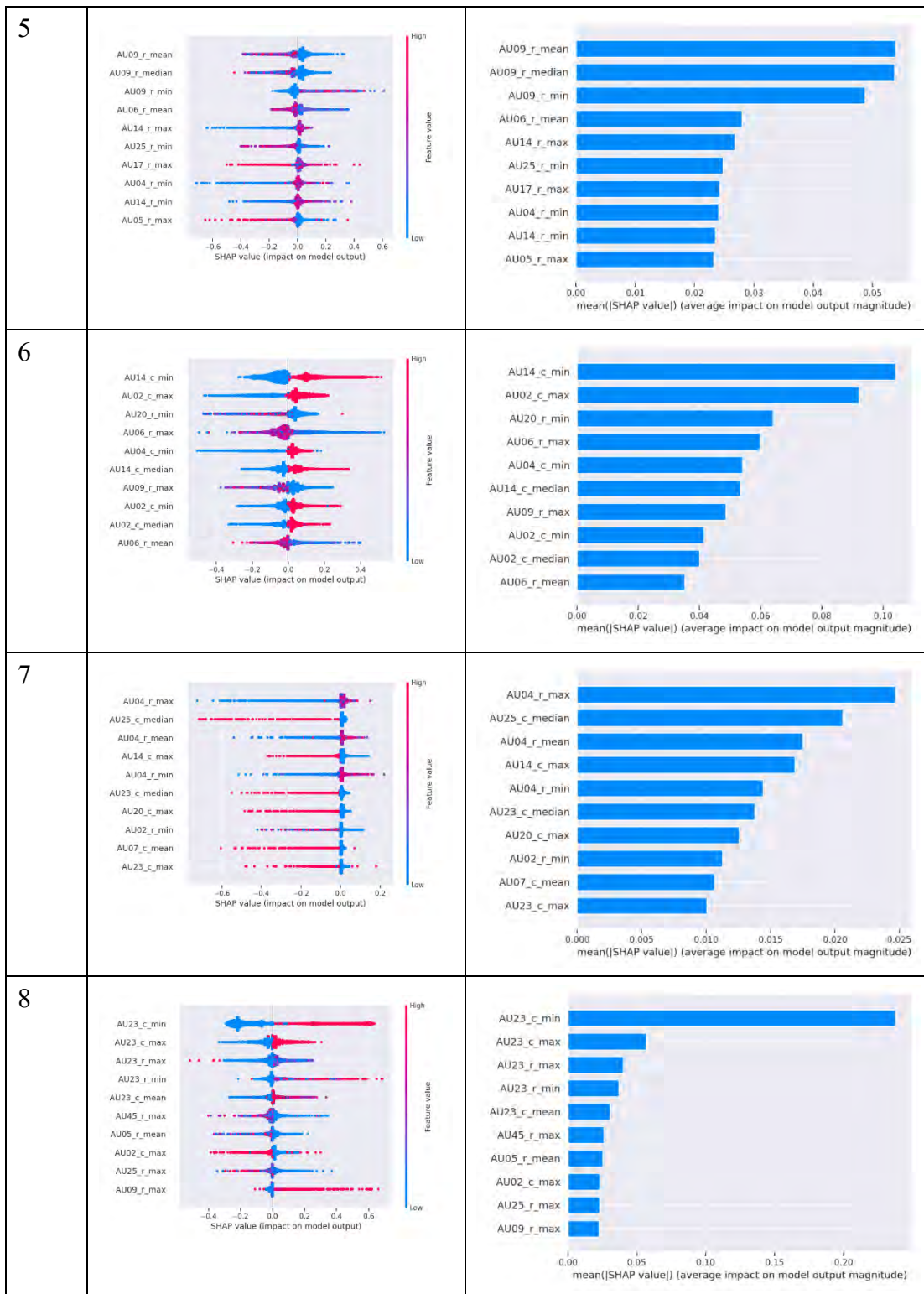
B-3-1 Training on Japan’s data; Testing on Japan’s data

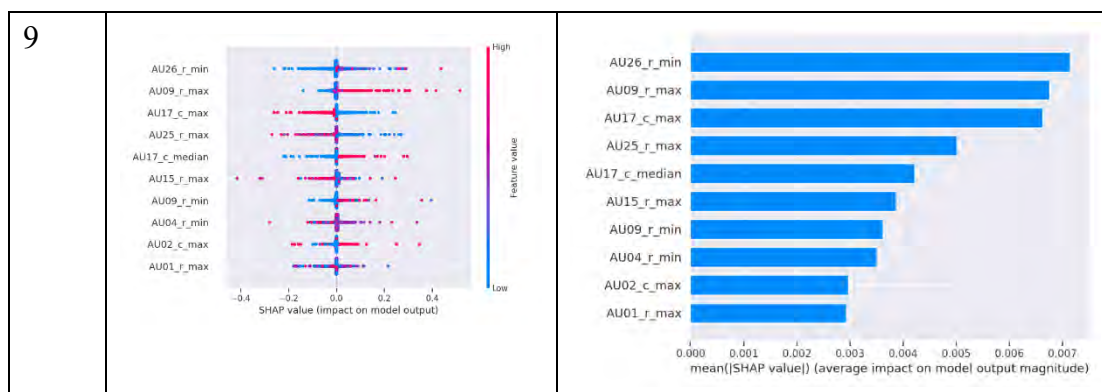
<p>The results of classifying the engagement states by Basic AUs feature set in Japan’s participants.</p>		
<p>Split</p>	<p>SHAP summary plot</p>	<p>SHAP bar plot</p>

Appendix

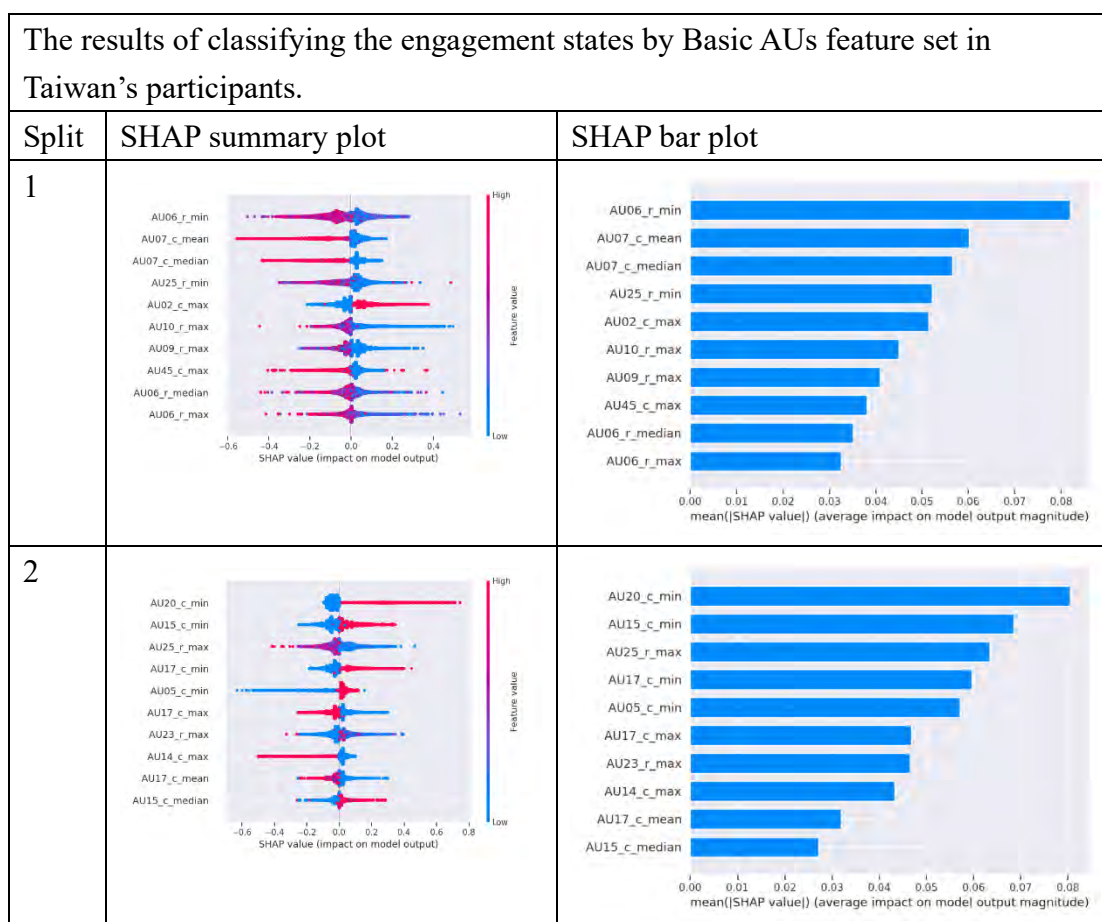


Appendix

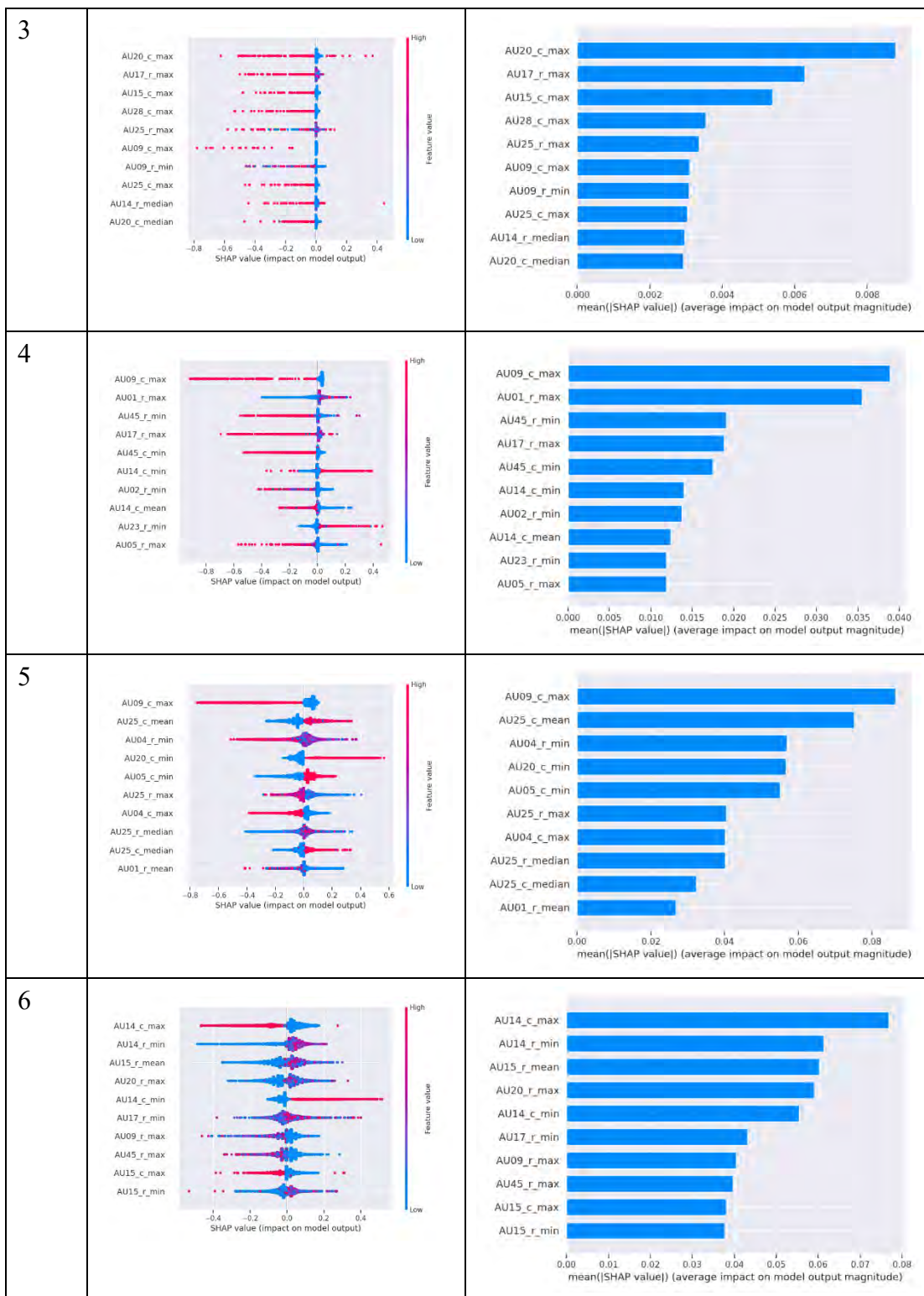


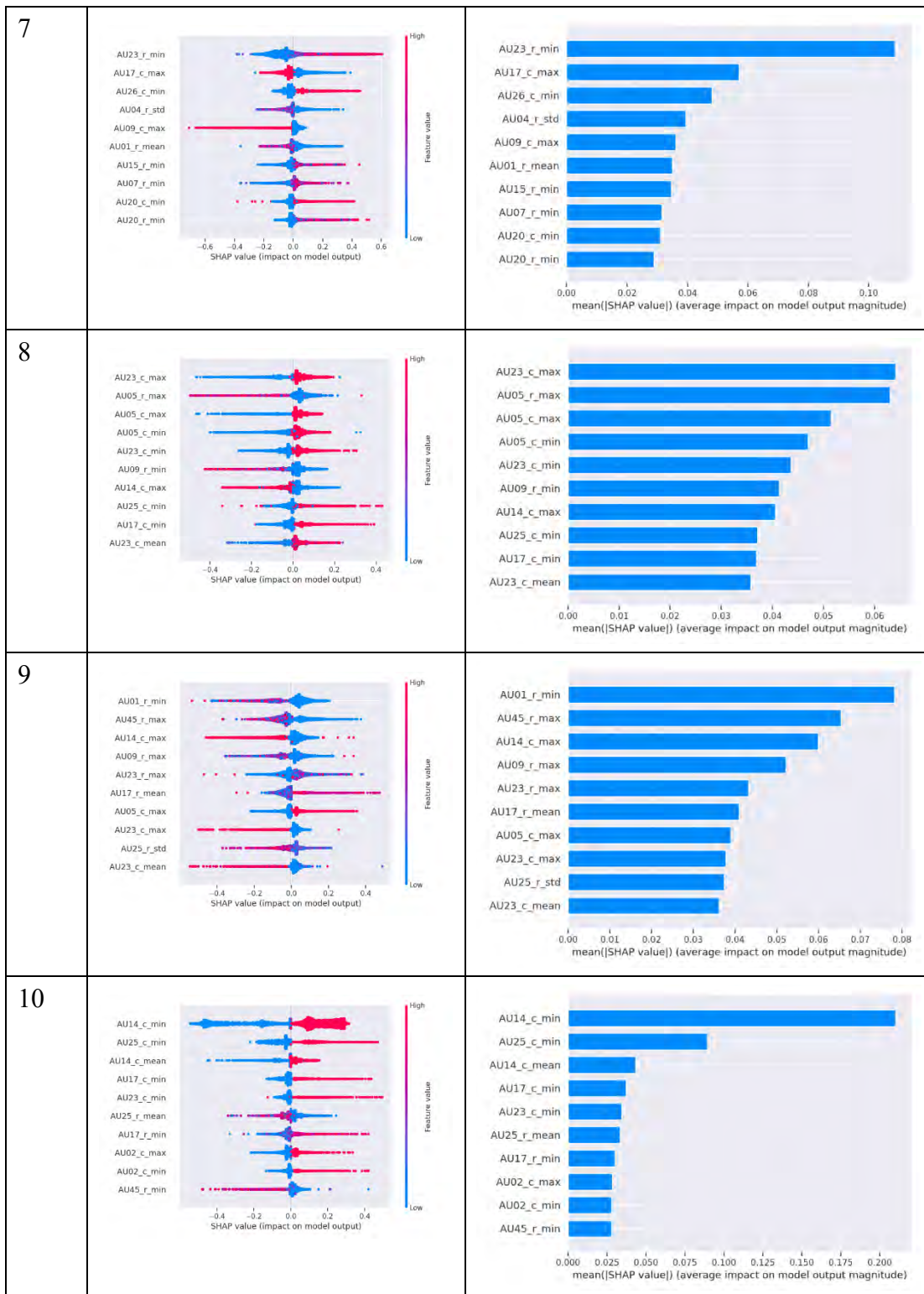


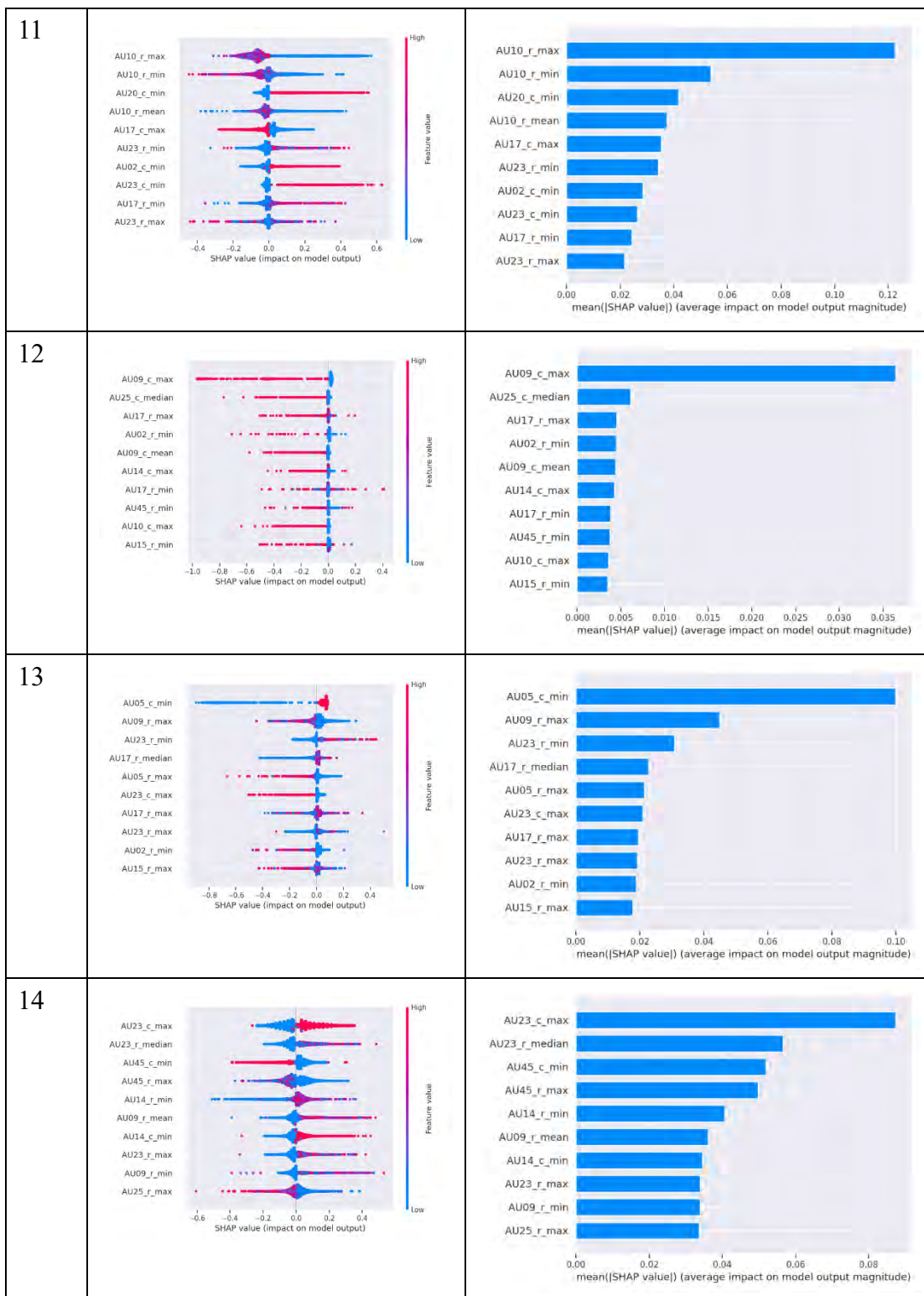
B-3-2 Training on Taiwan's data; Testing on Taiwan's data

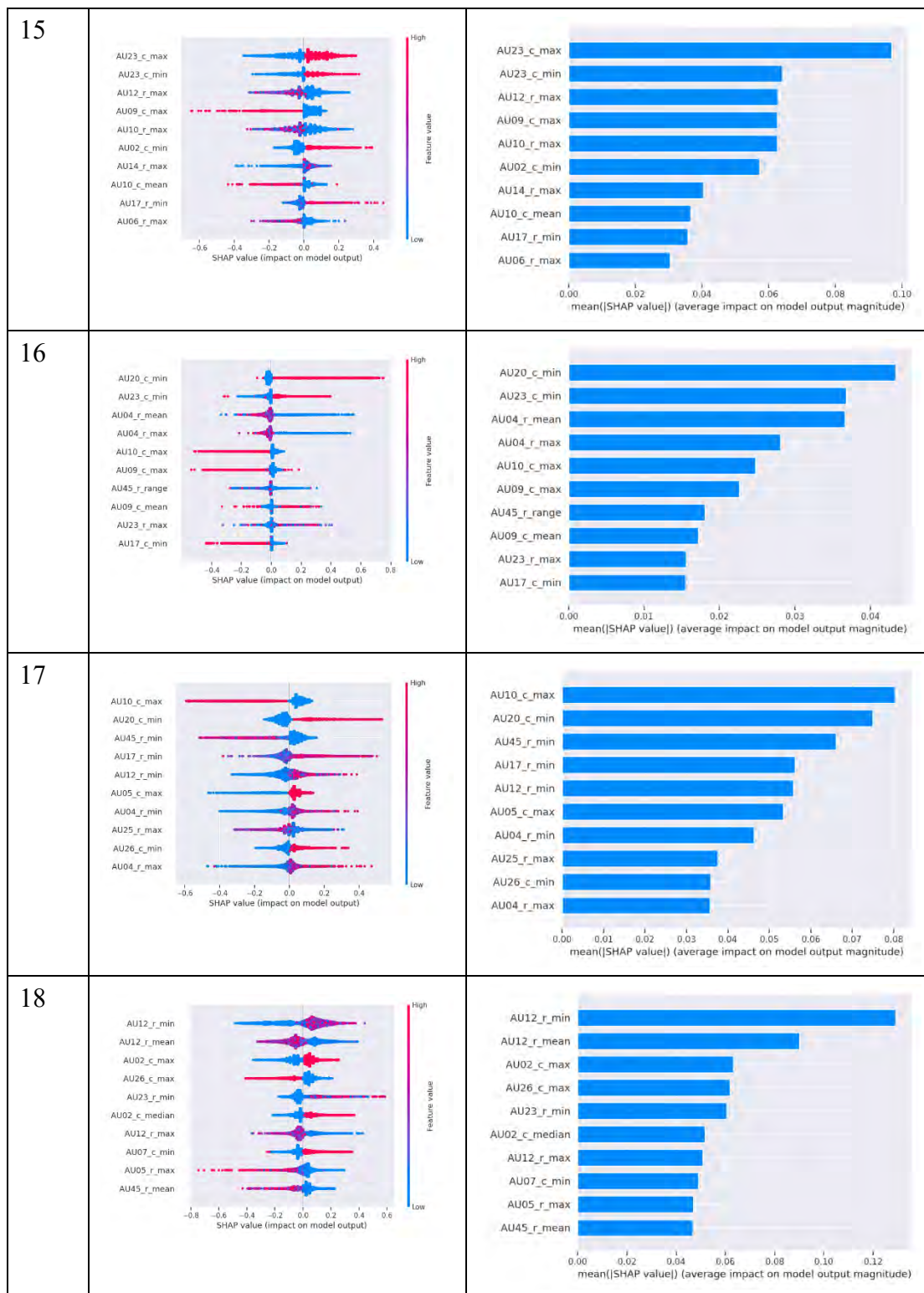


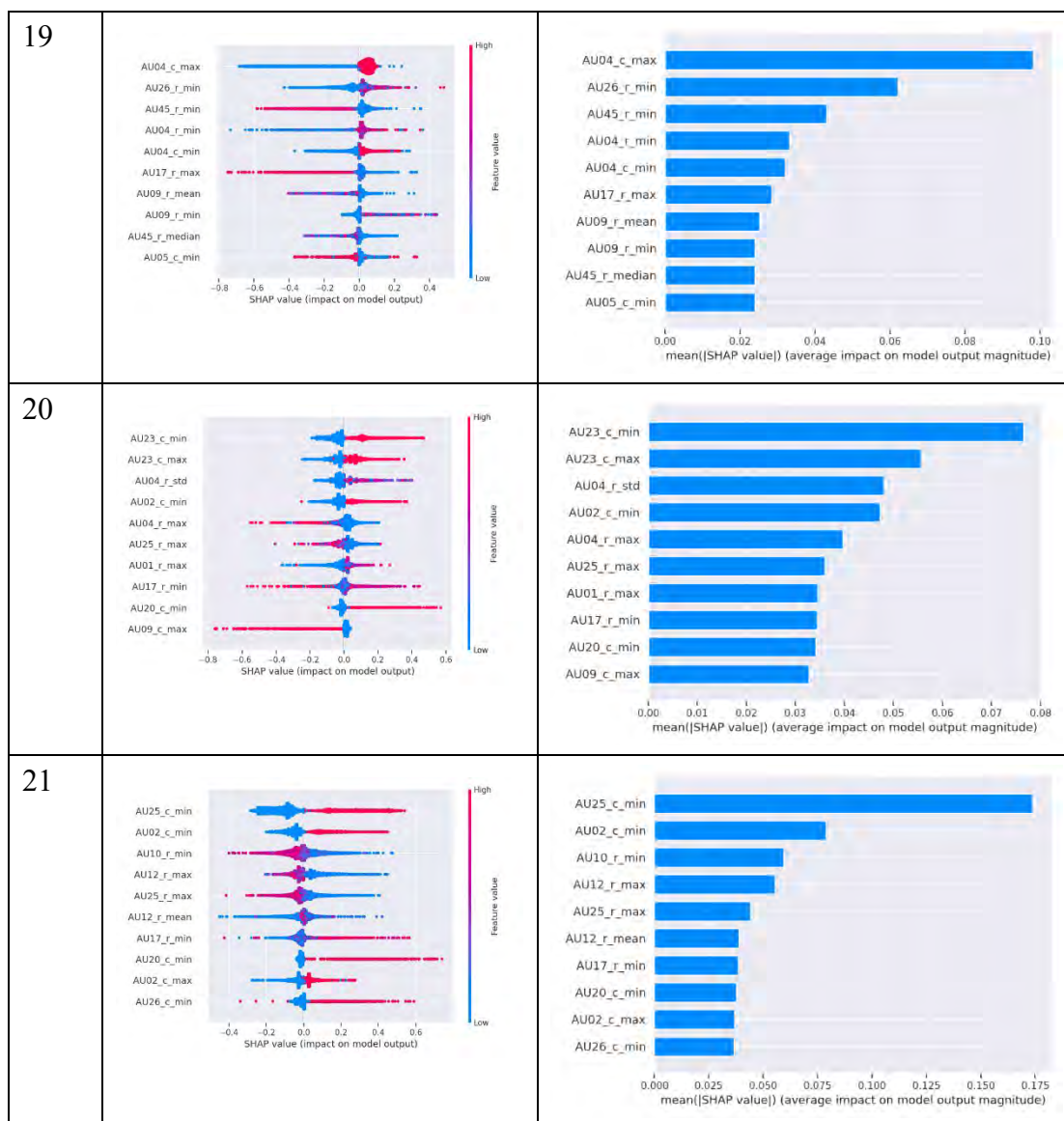
Appendix



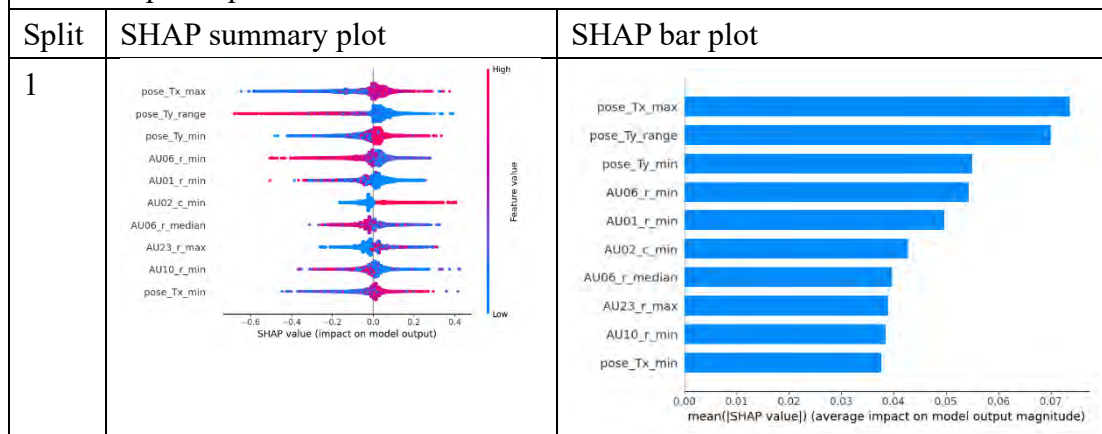




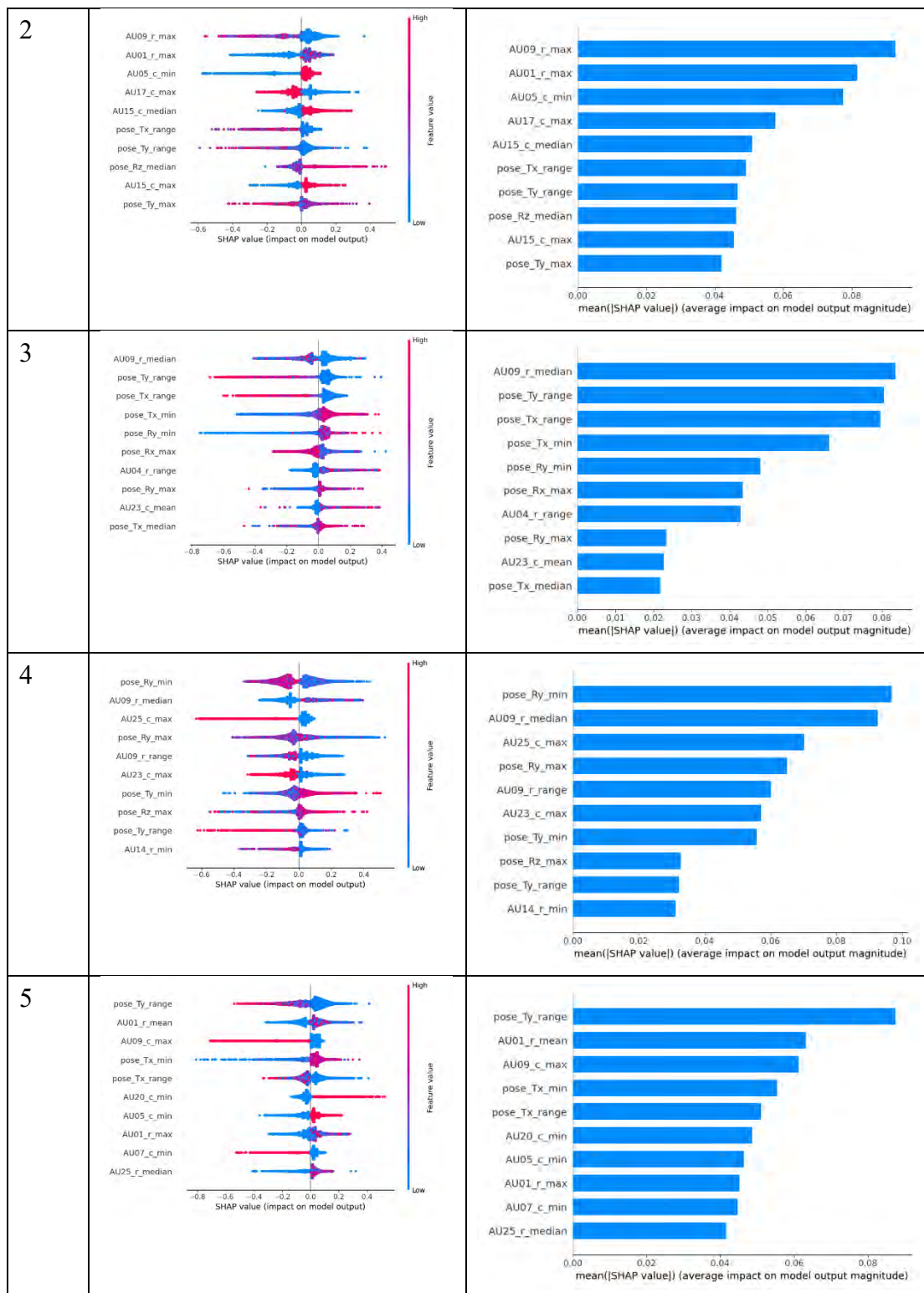




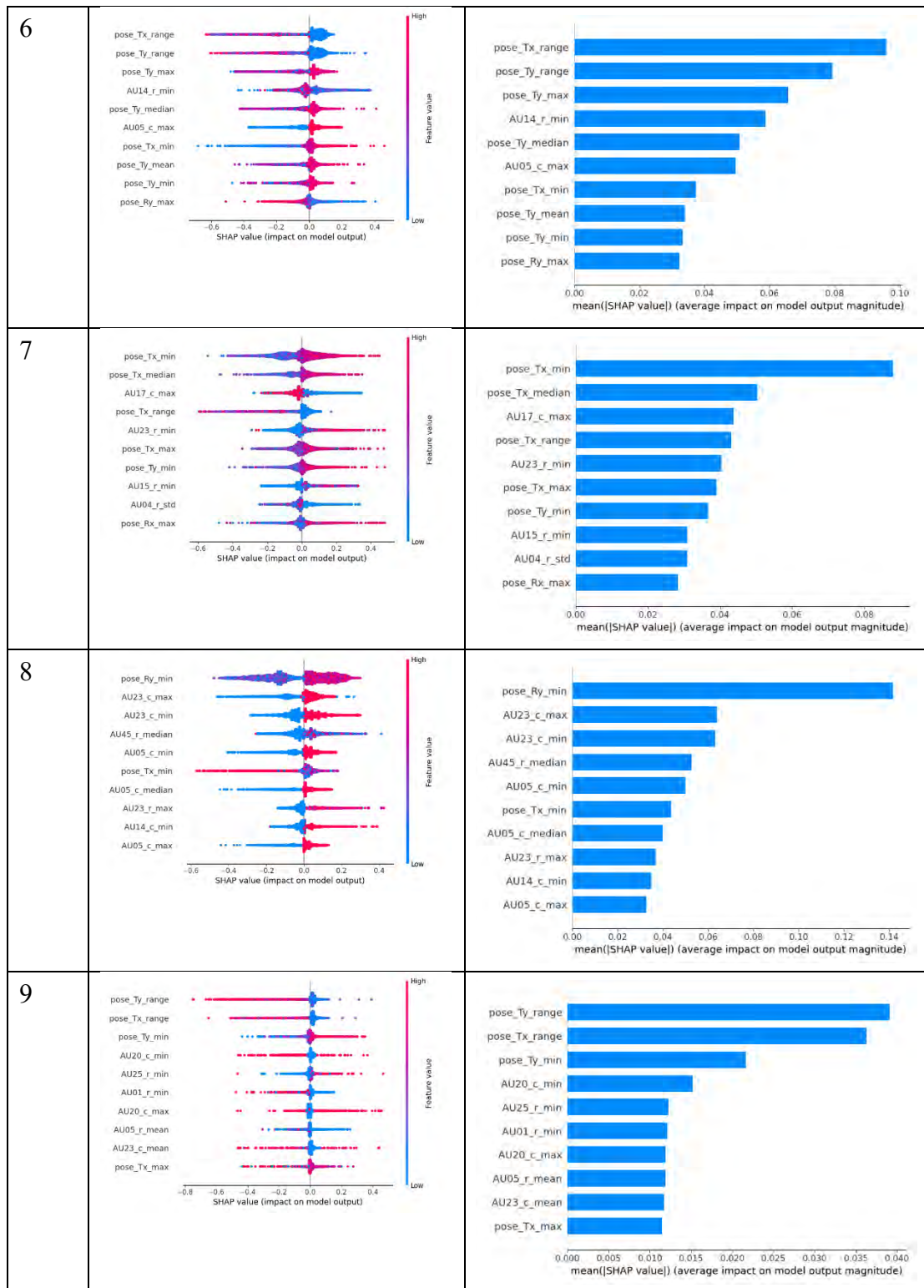
The results of classifying the engagement states by “AU+Head Pose” feature set in Taiwan’s participants.

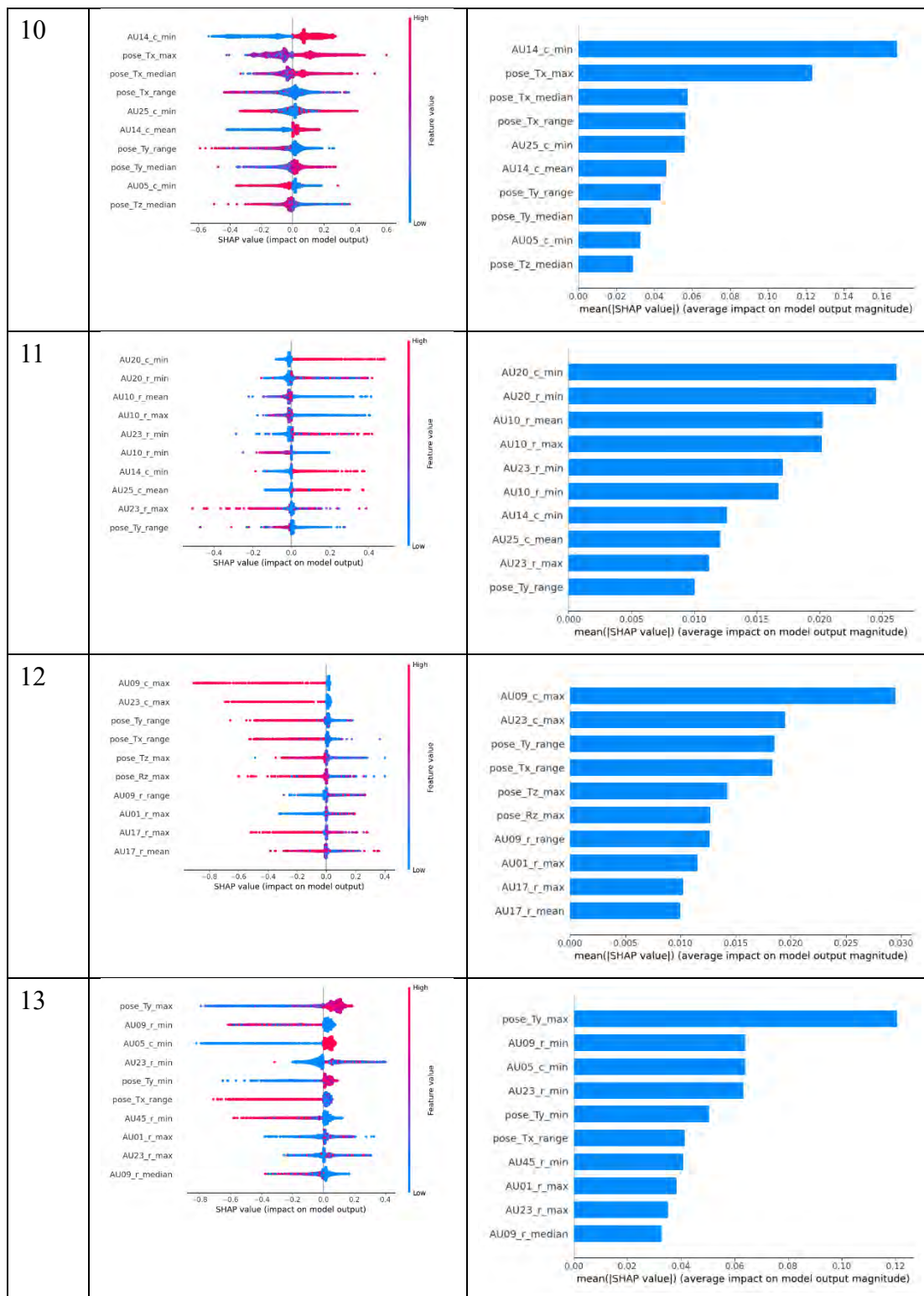


Appendix

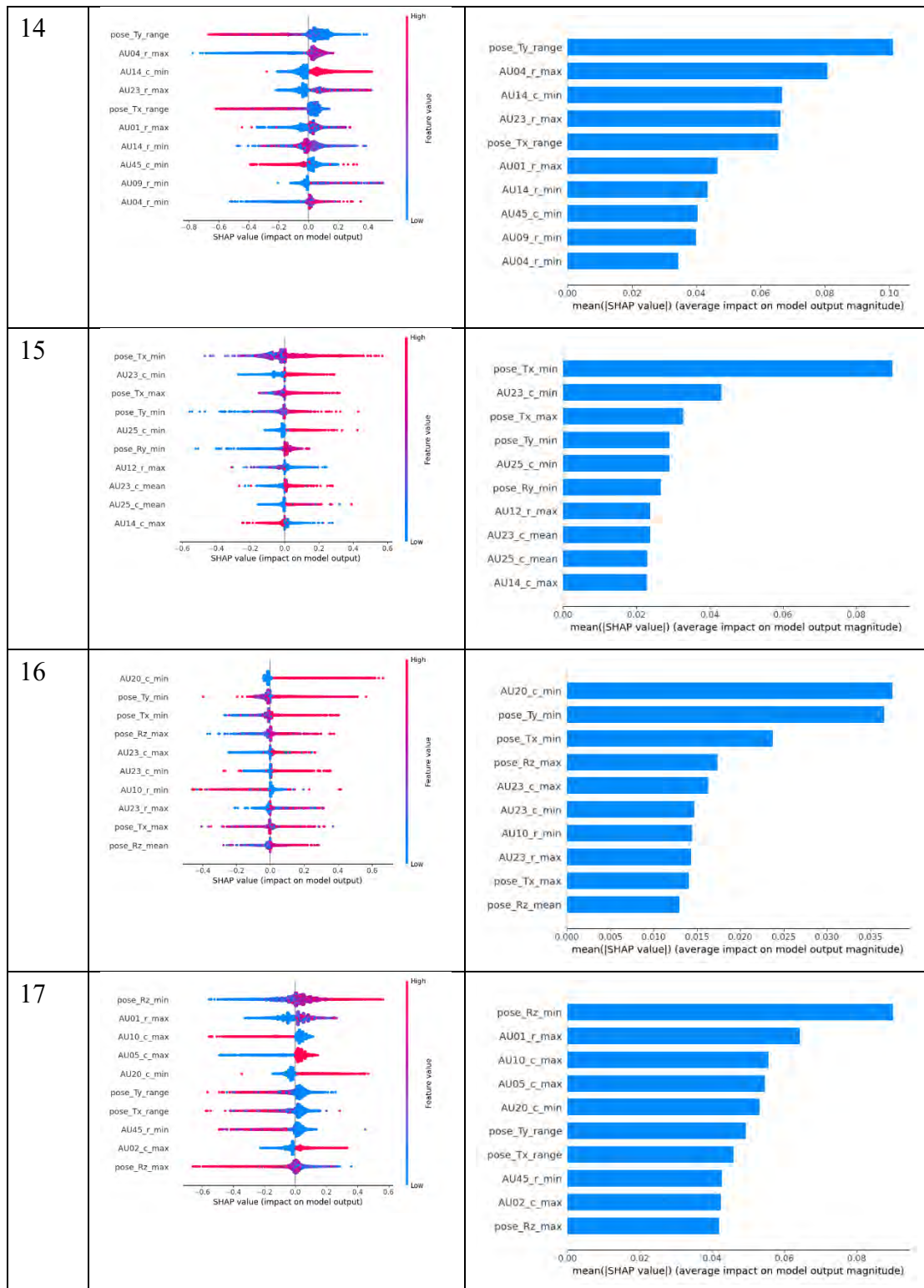


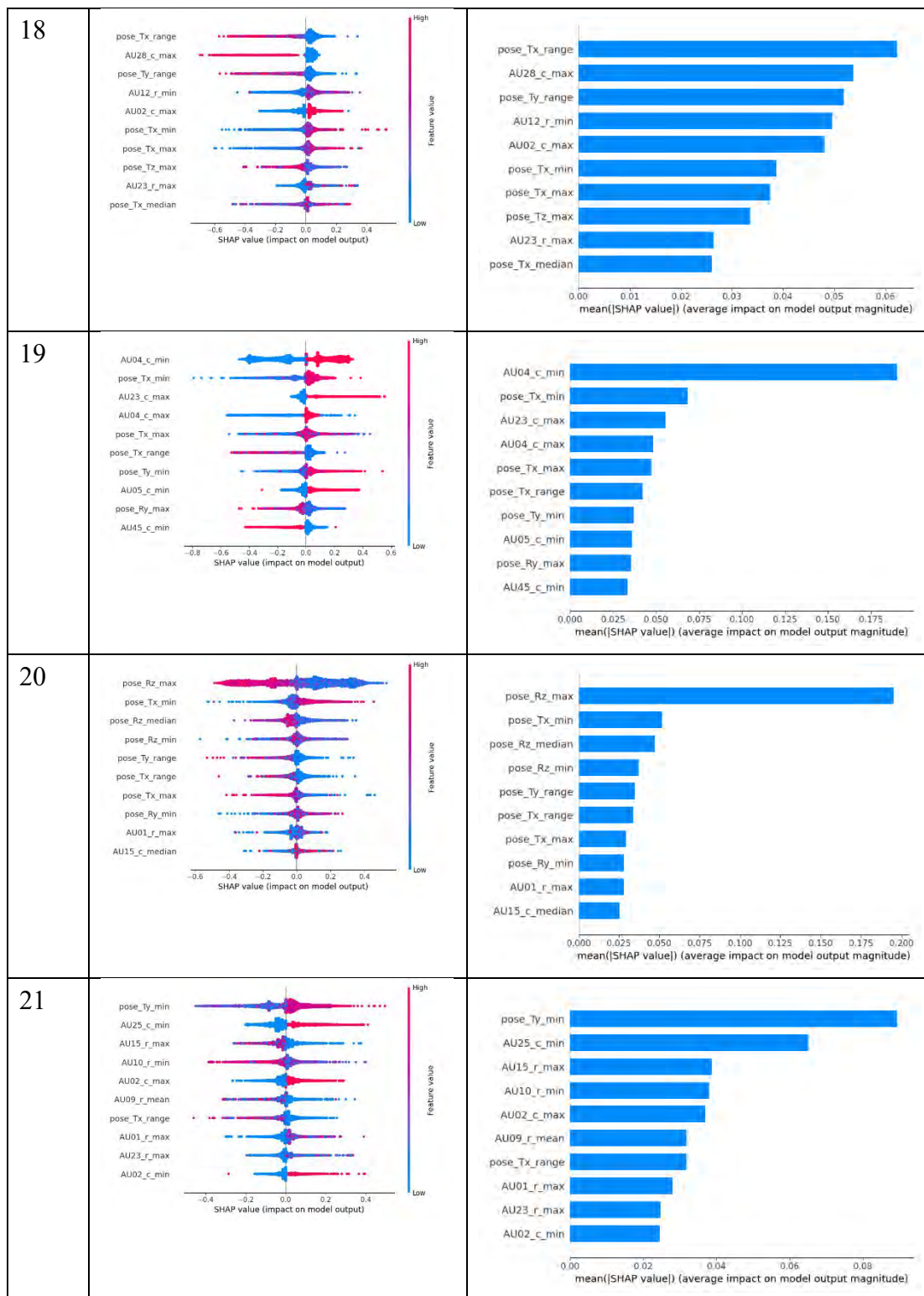
Appendix





Appendix





B-4. Inter-person learning results of classifying help-seeking states

The five types of model are conducted on inter-person learning, but here only showed

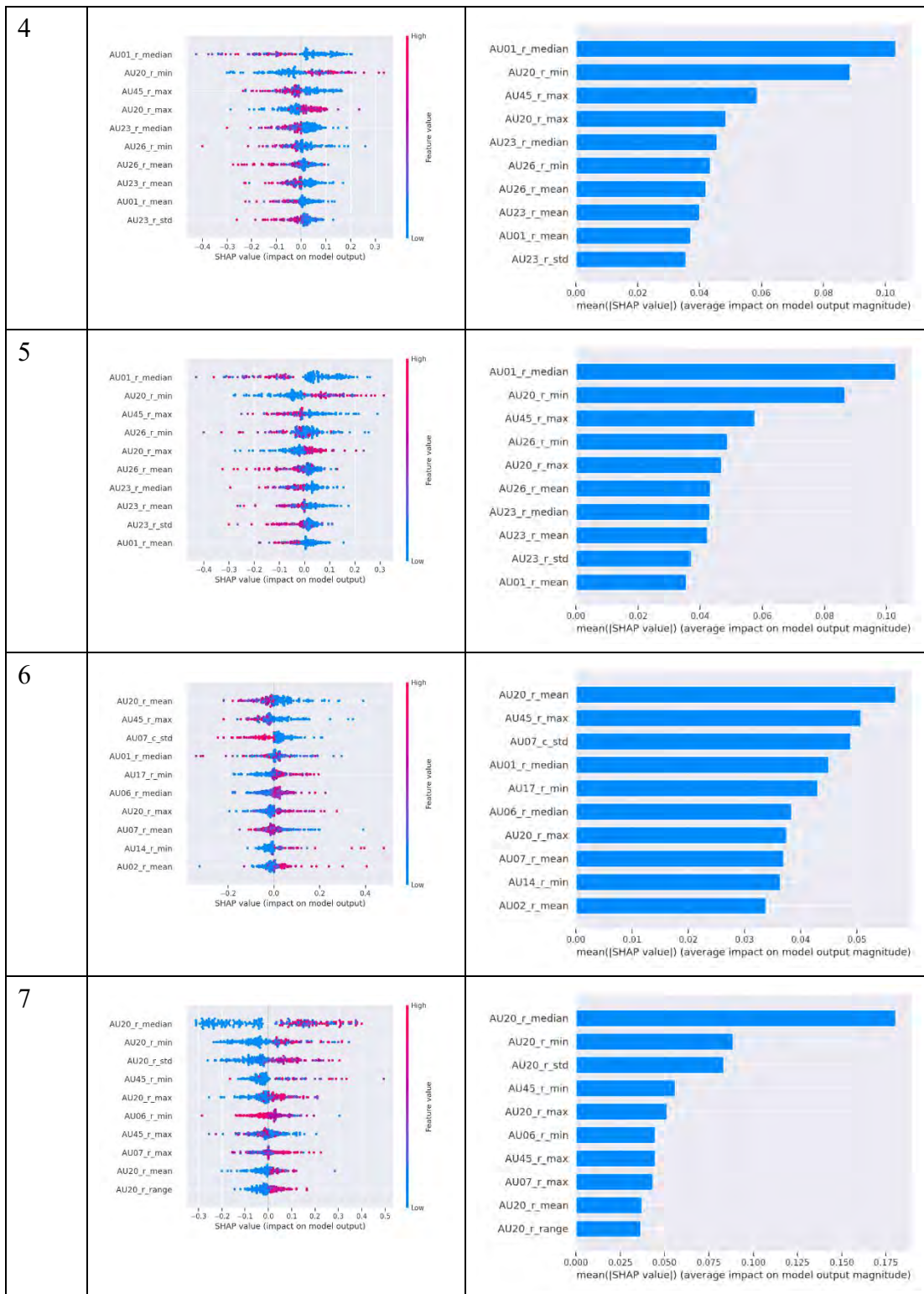
the detail of the first two:

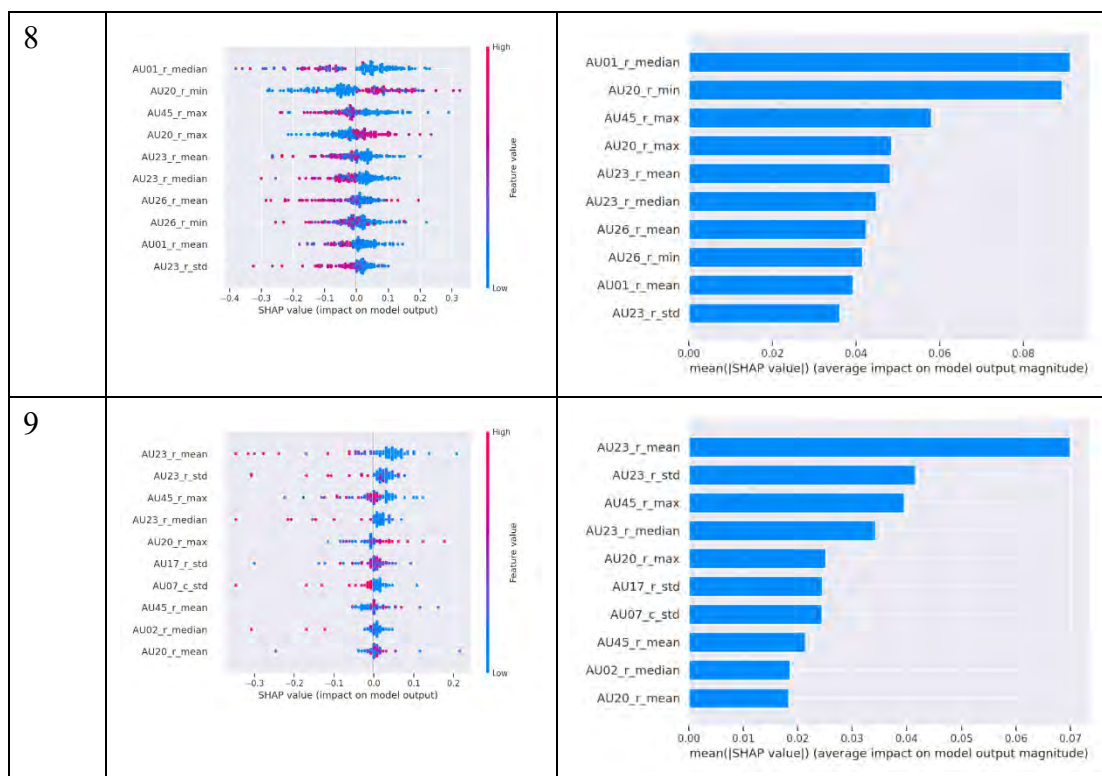
(1) Training: Japan's data; Testing: Japan's data

(2) Training: Taiwan's data; Testing: Taiwan's data

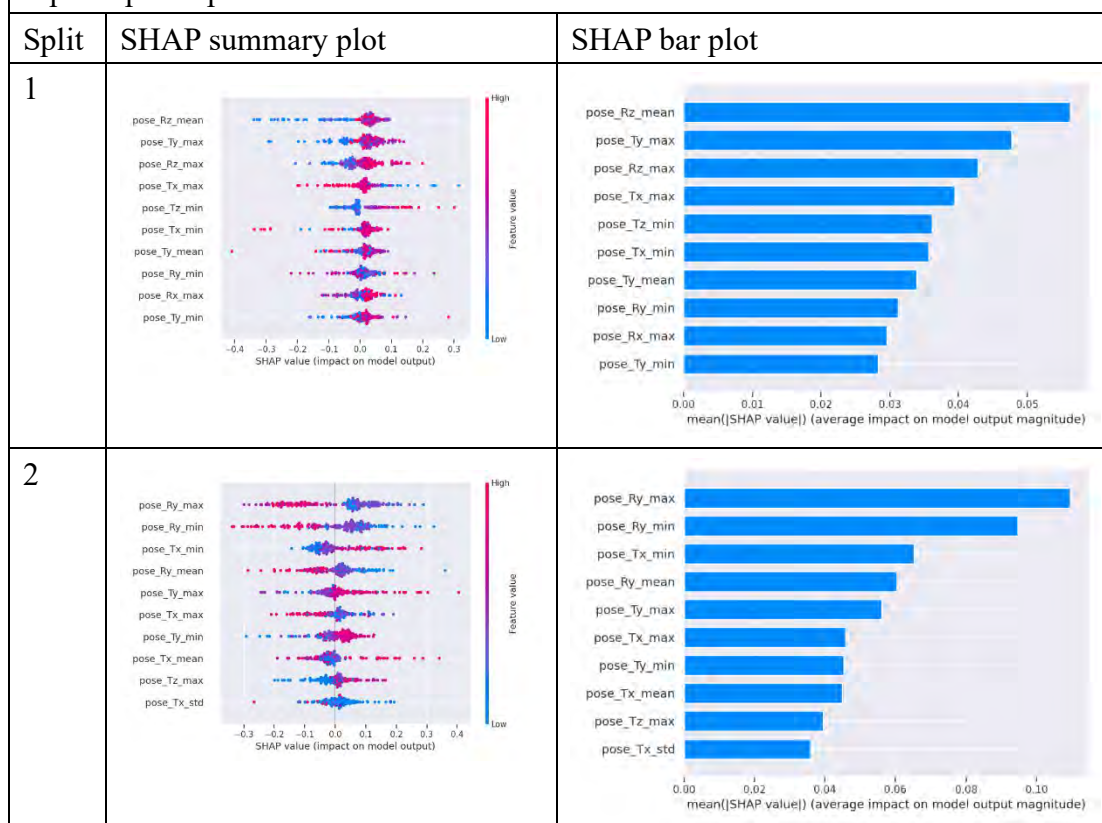
B-4-1 Training on Japan's data; Testing on Japan's data

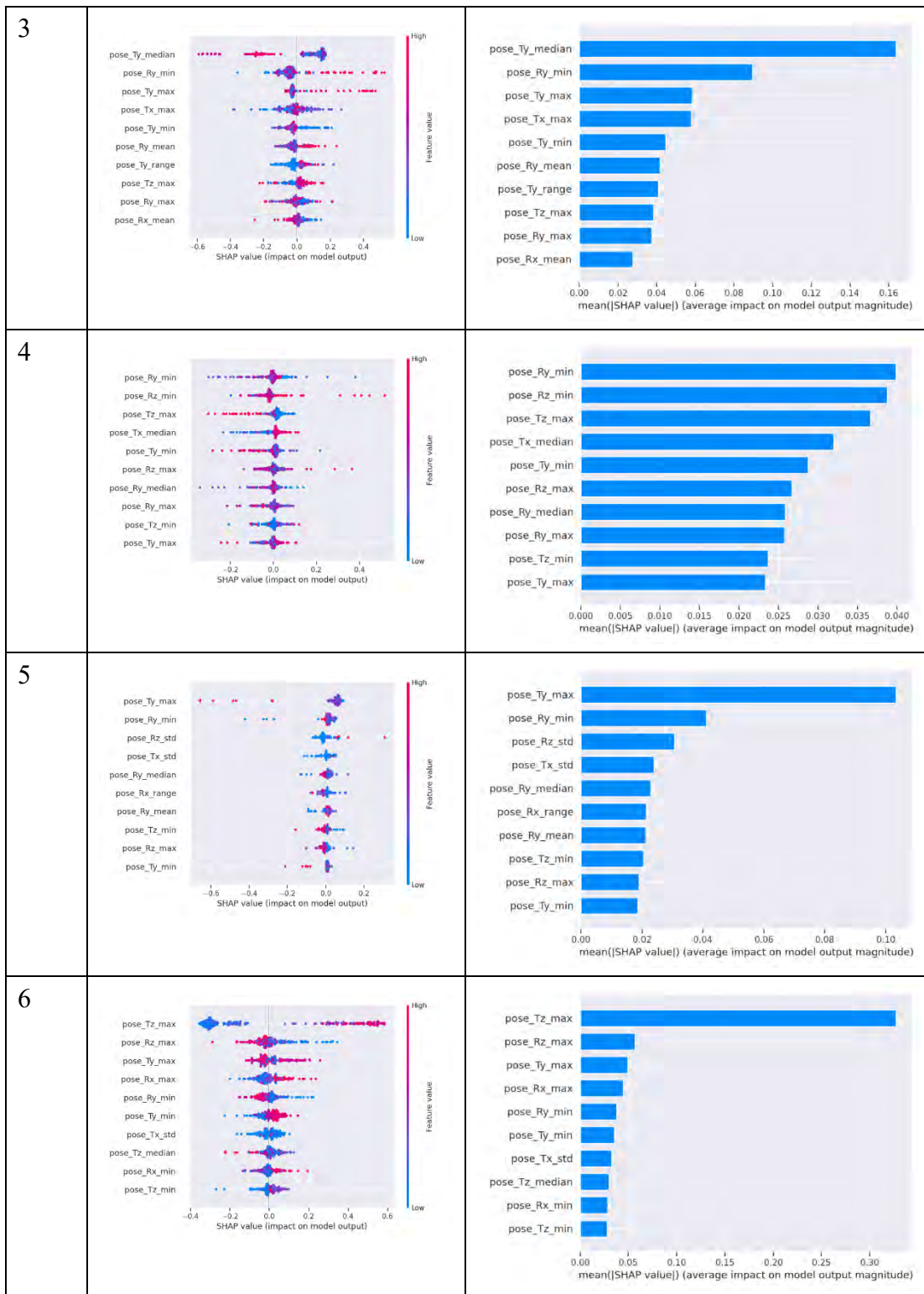
The results of classifying the help-seeking states by Basic AUs feature set in Japan's participants.		
Split	SHAP summary plot	SHAP bar plot
1		
2		
3		



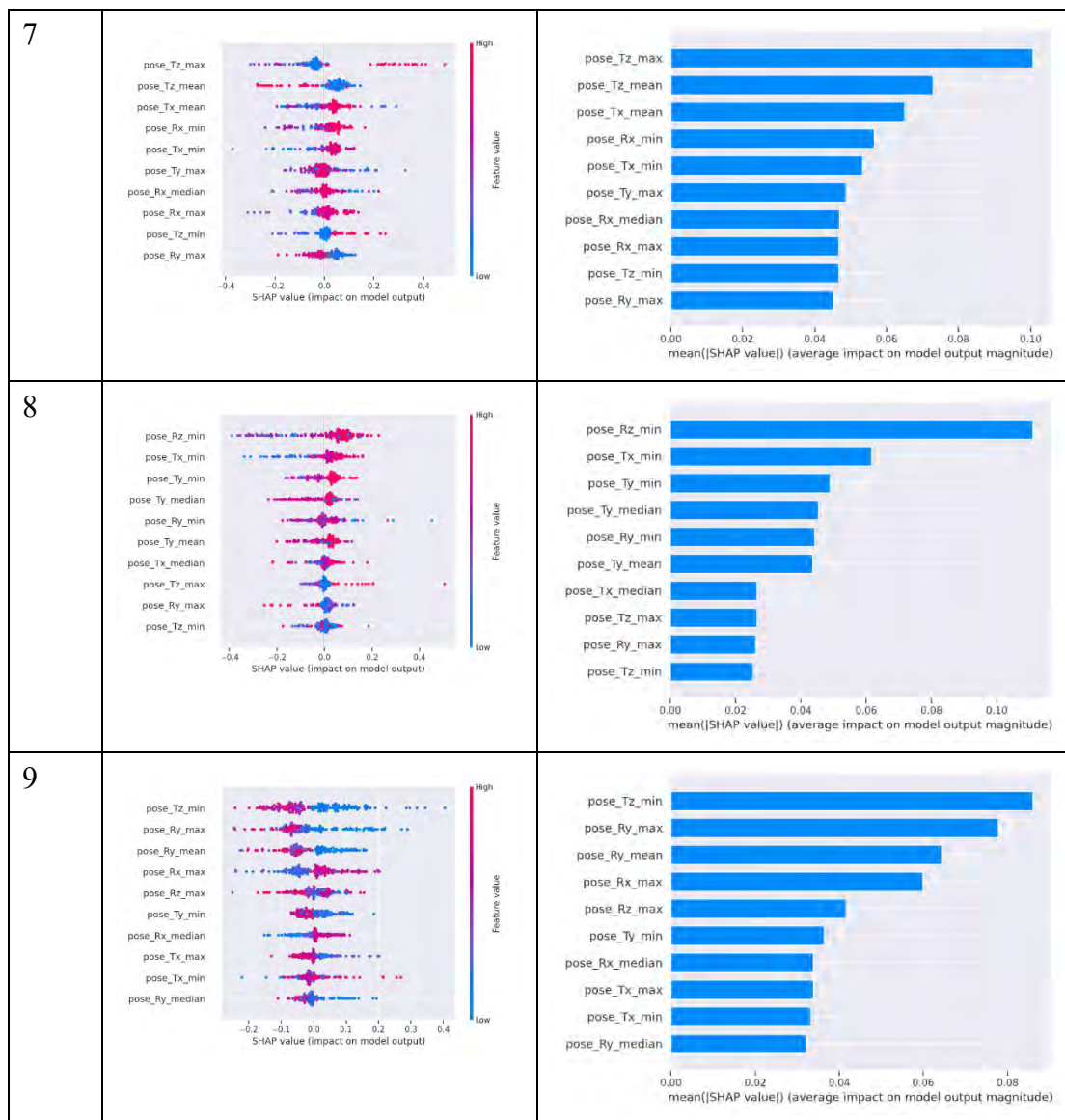


The results of classifying the help-seeking states by Head Pose feature set in Japan's participants.





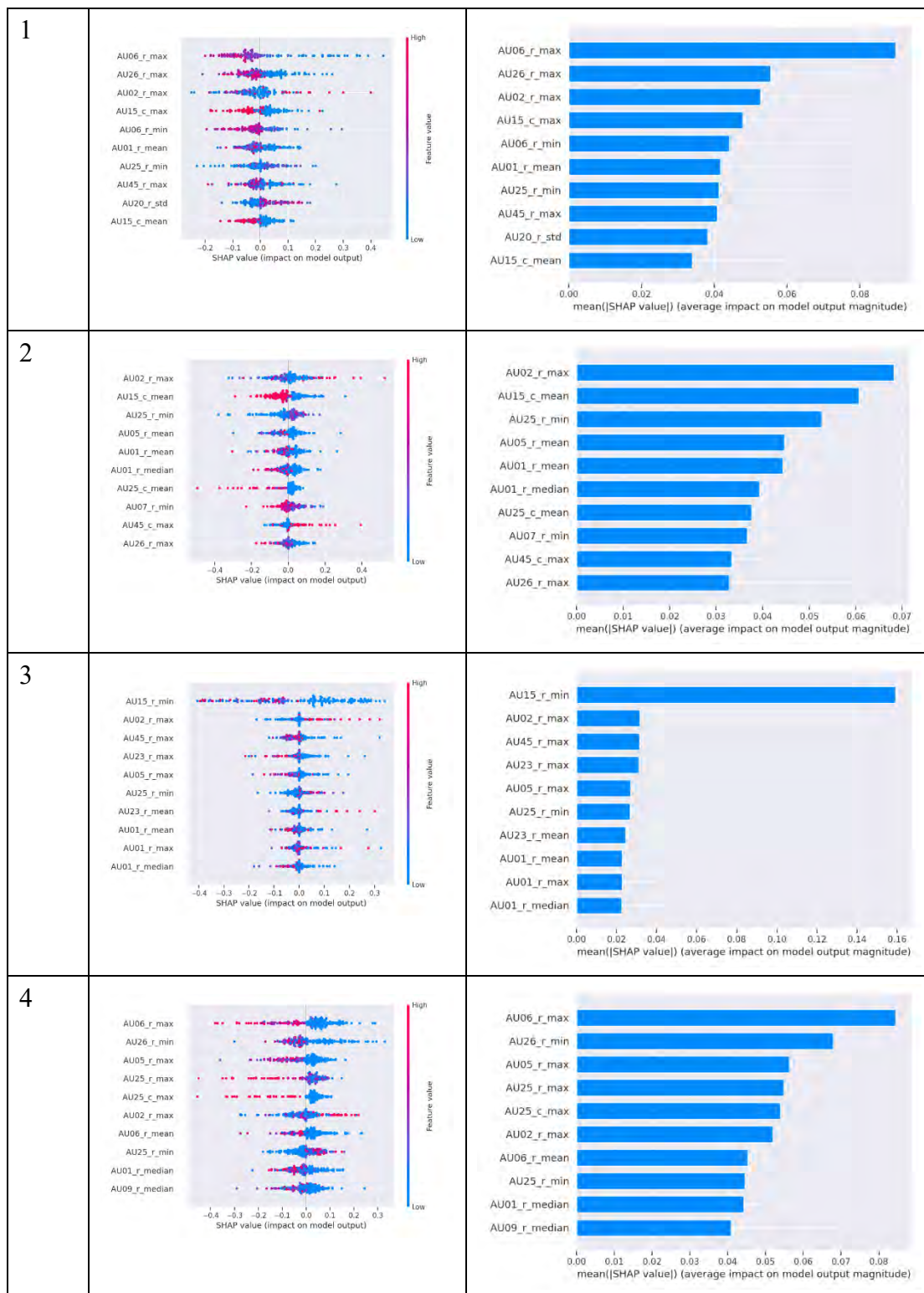
Appendix

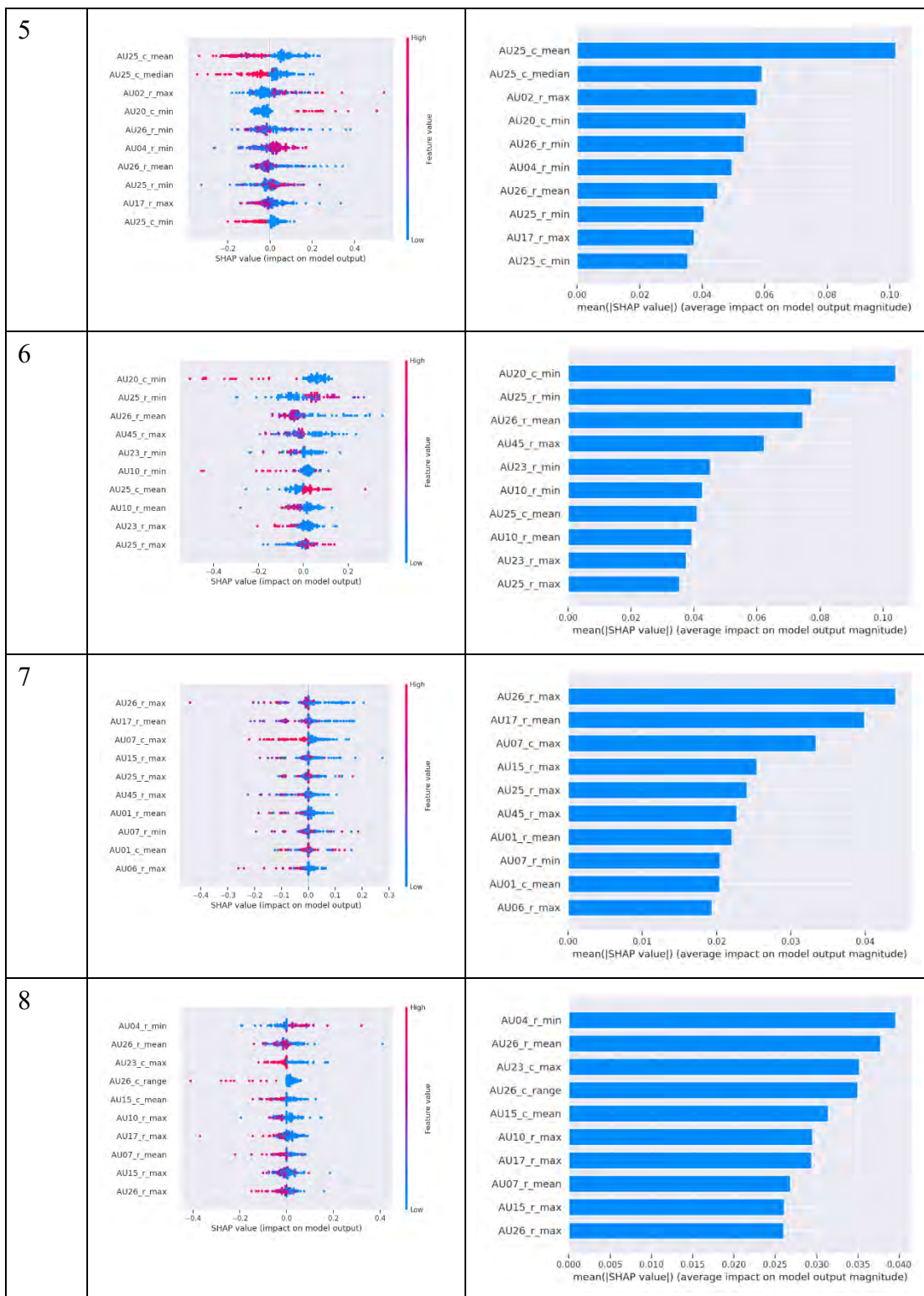


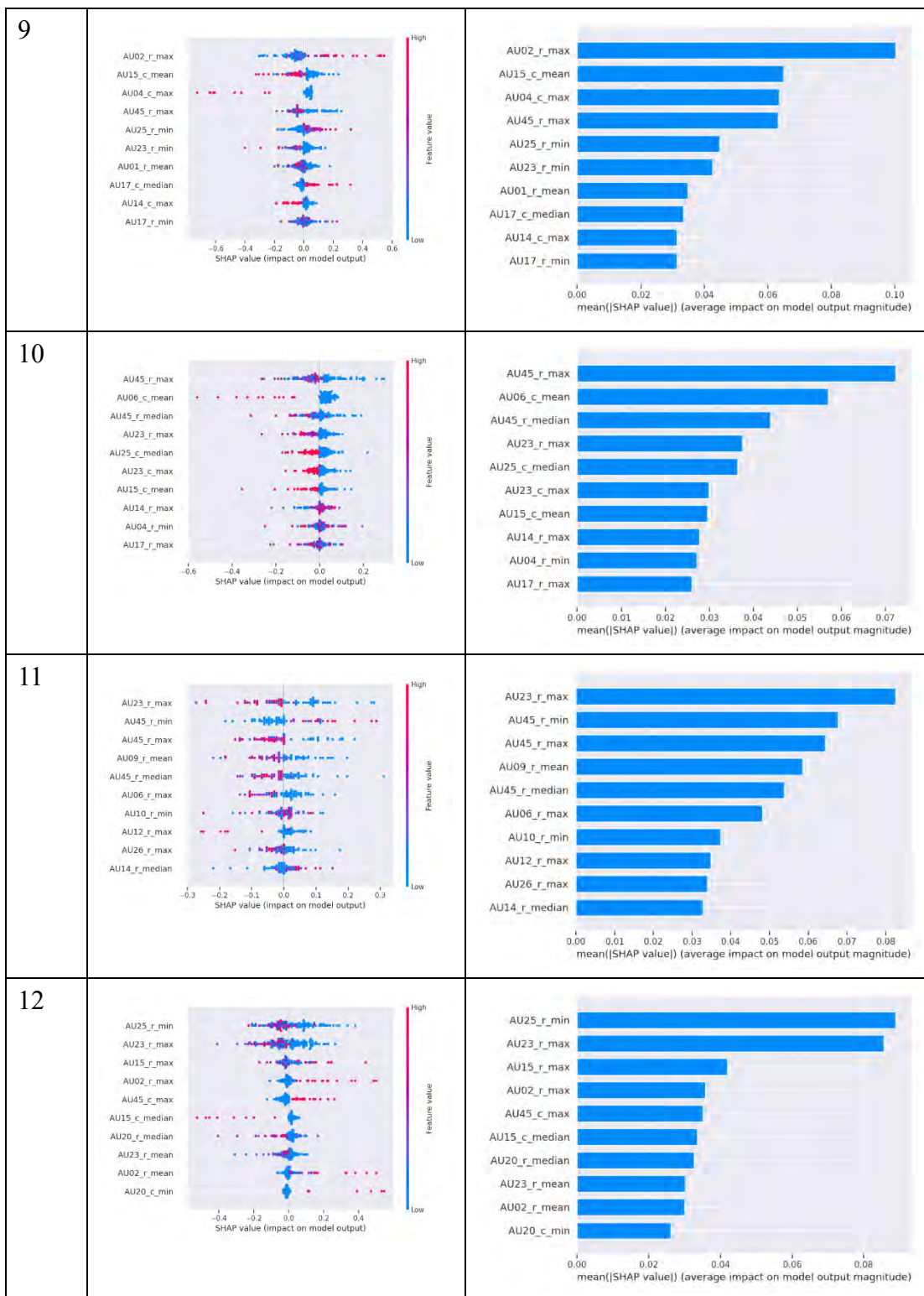
B-4-2 Training on Taiwan's data; Testing on Taiwan's data

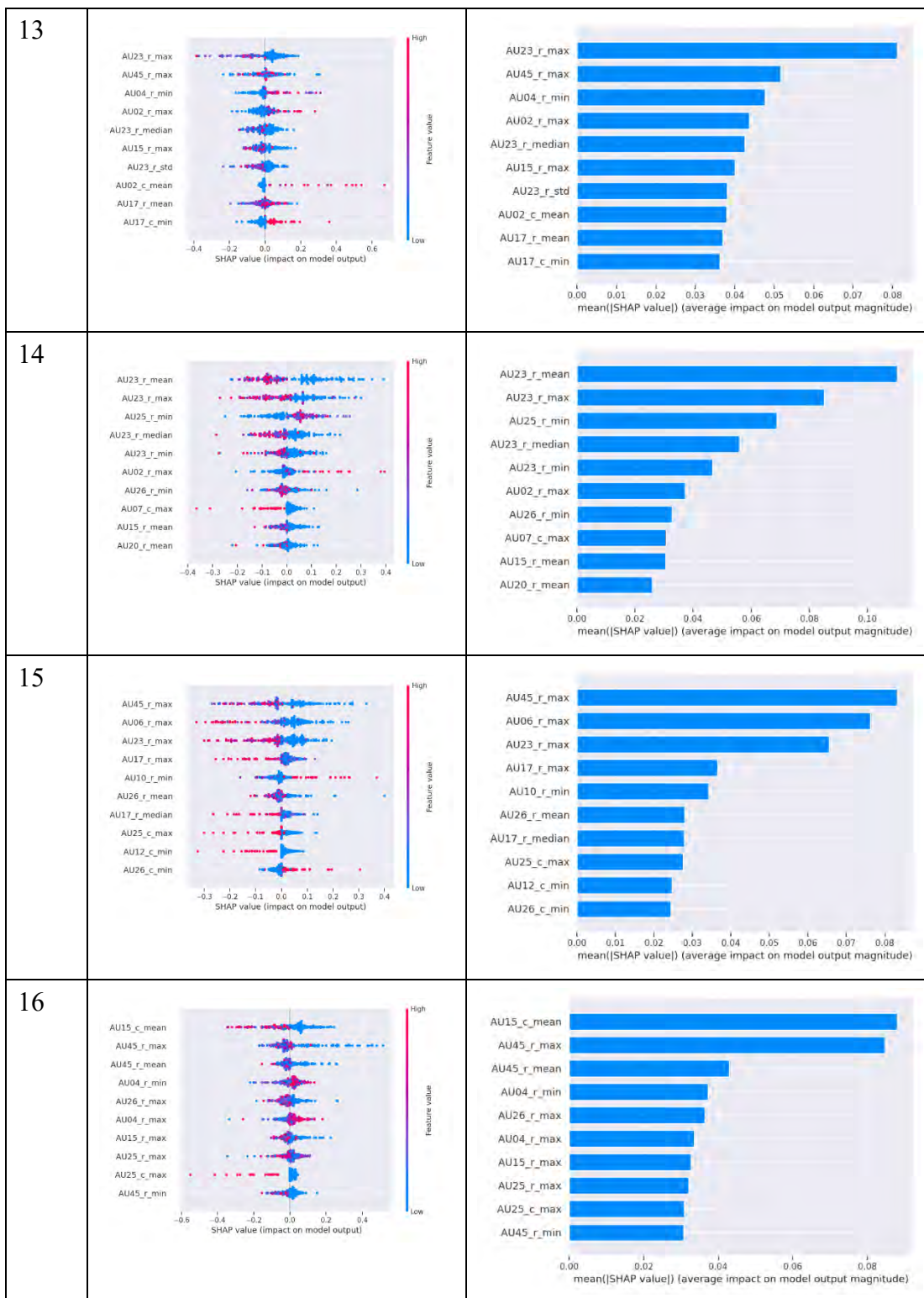
<p>The results of classifying the help-seeking states by Basic AUs feature set in Taiwan's participants.</p>		
<p>Split</p>	<p>SHAP summary plot</p>	<p>SHAP bar plot</p>

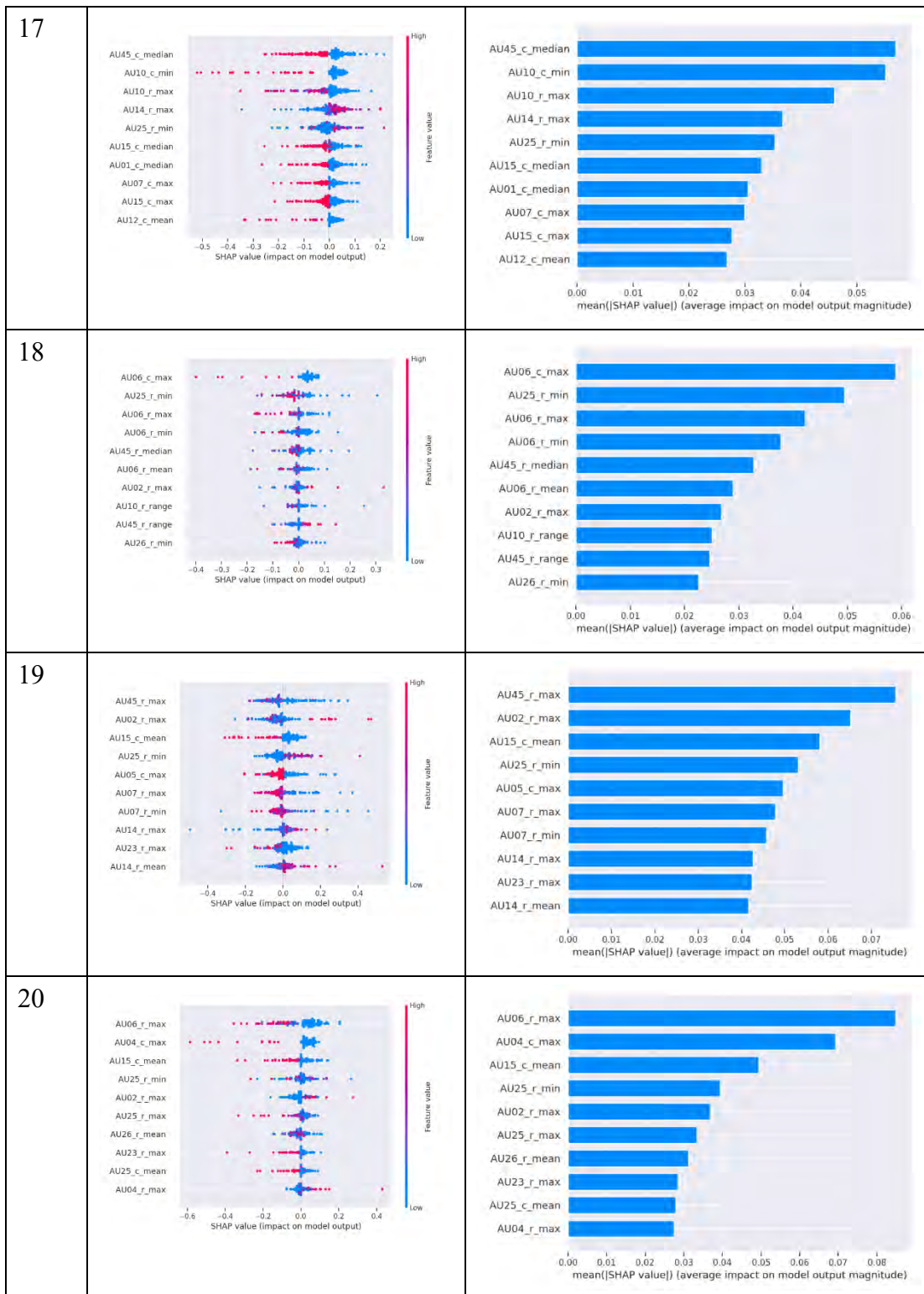
Appendix

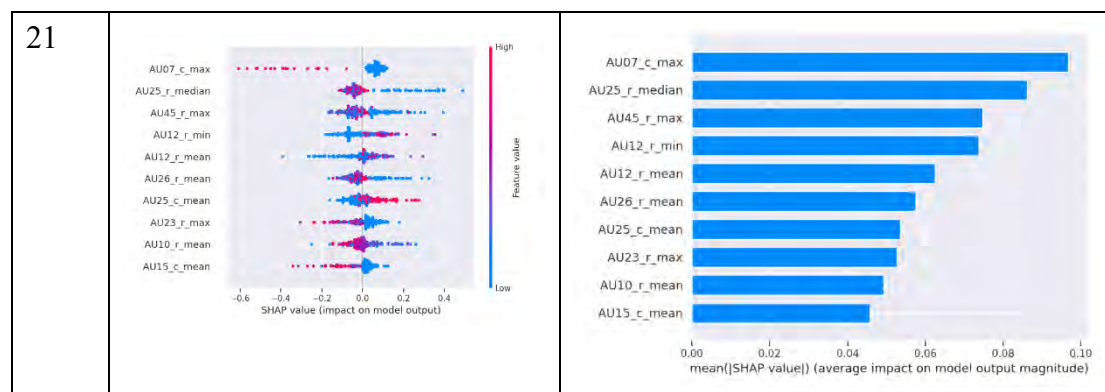




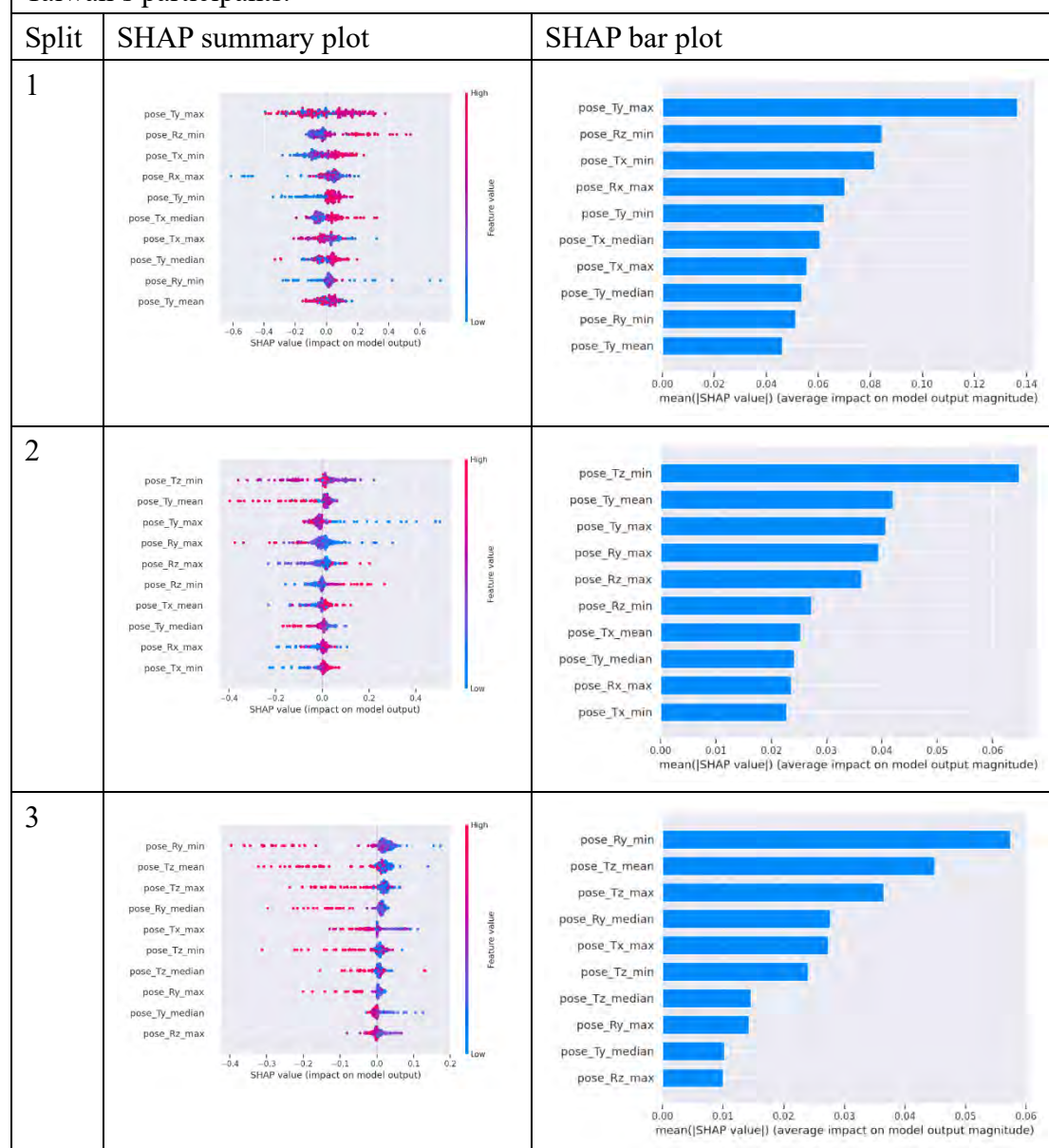


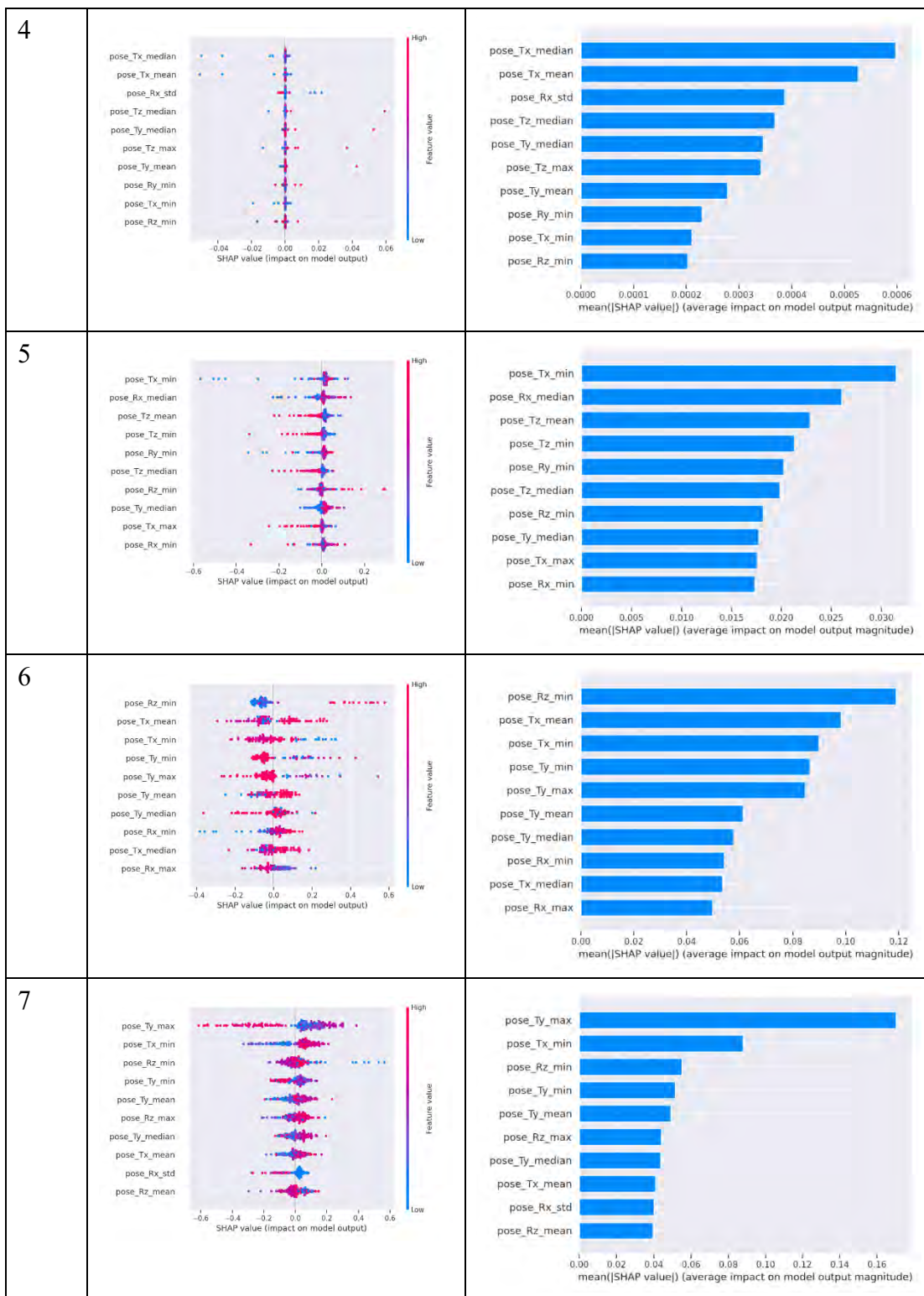




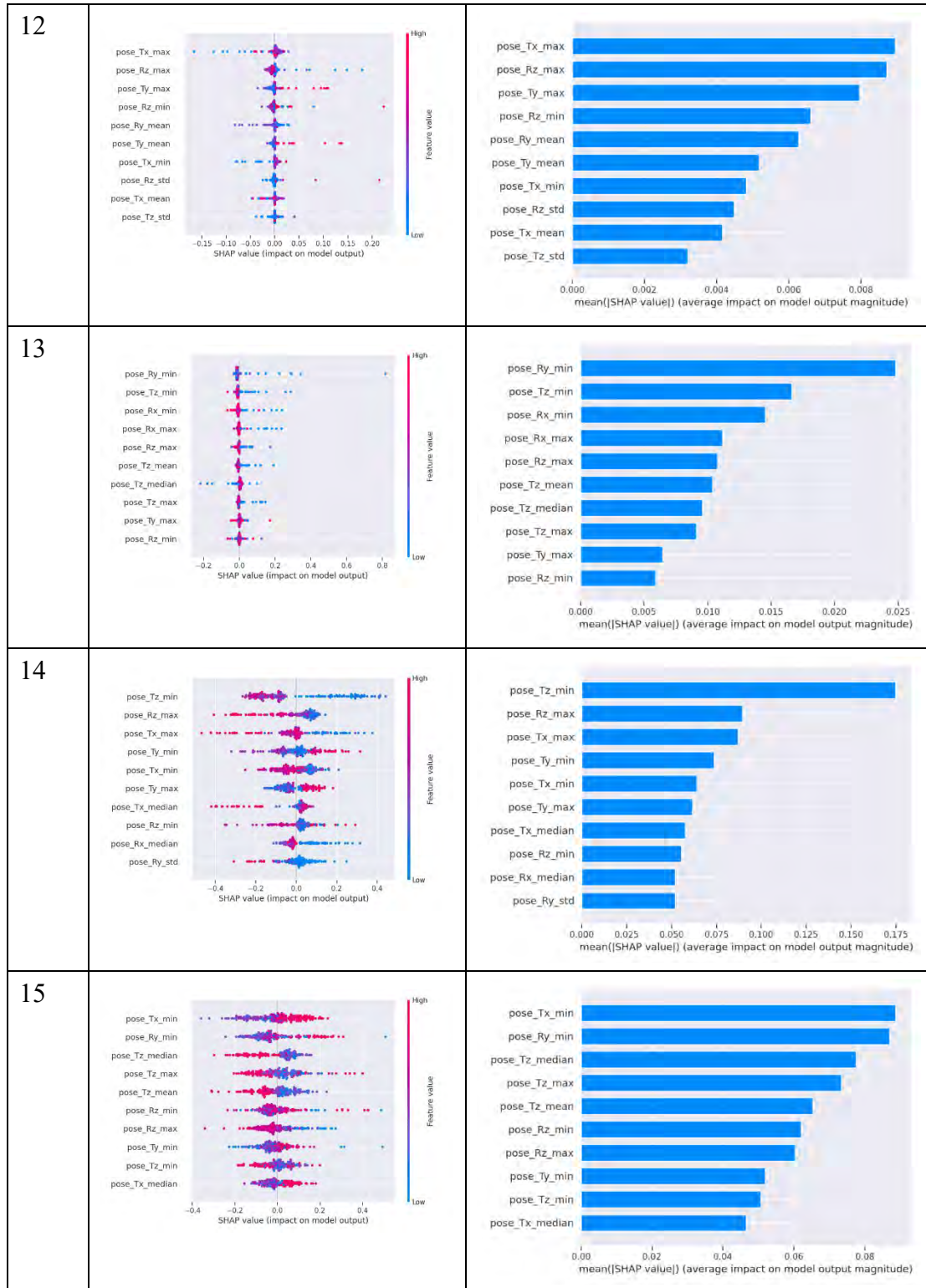


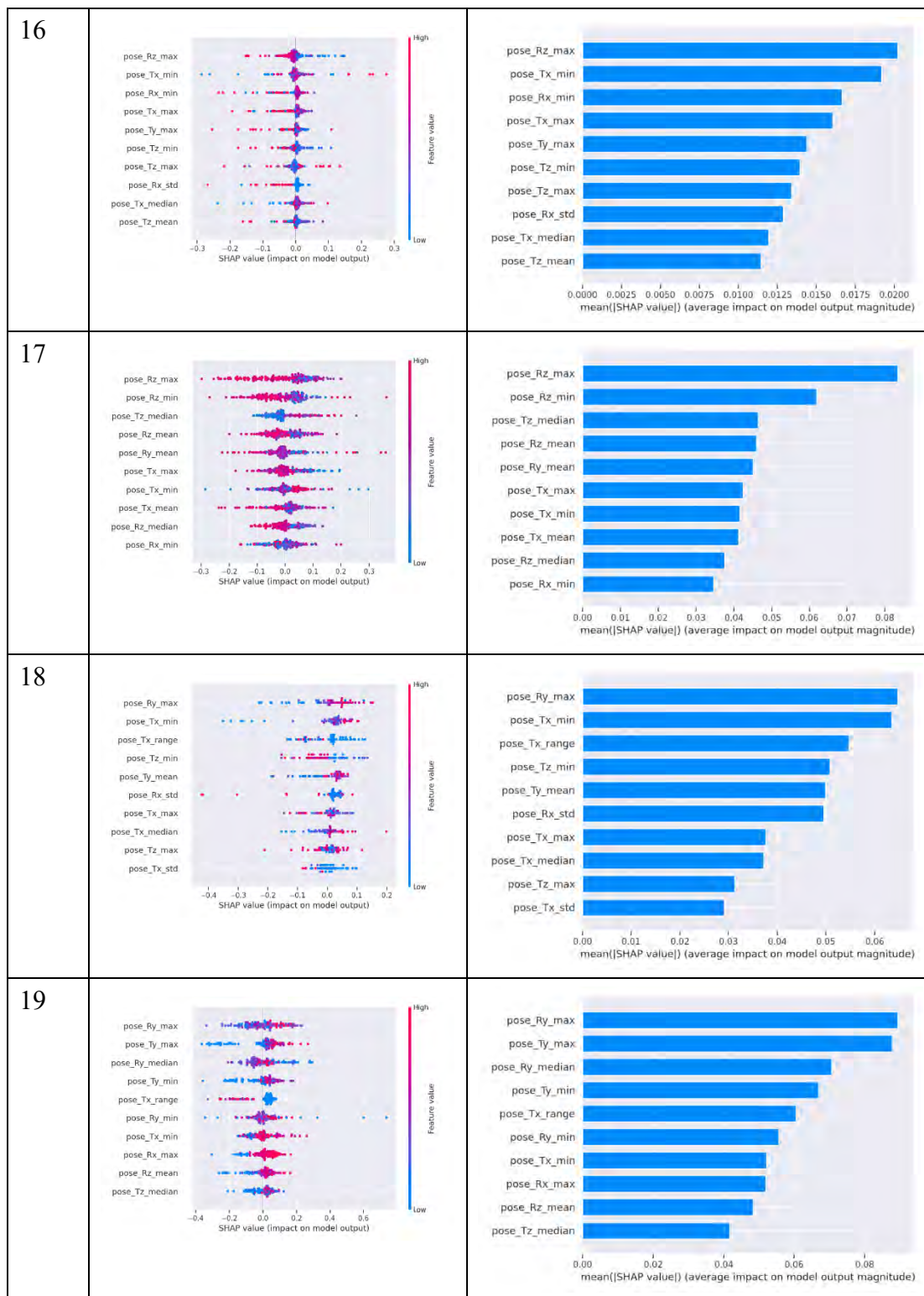
The results of classifying the help-seeking states by Head Pose feature set in Taiwan's participants.



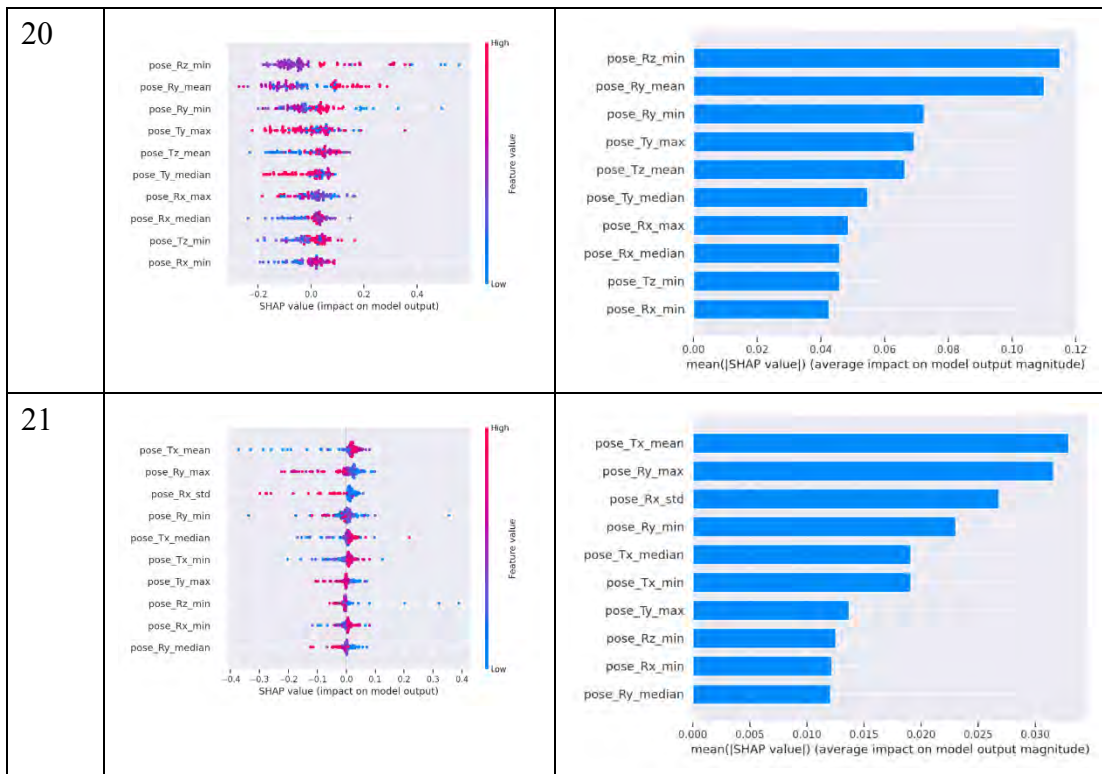


<p>8</p>	<p>pose_Tx_min pose_Rz_max pose_Ry_min pose_Tx_max pose_Ty_max pose_Tx_mean pose_Tx_std pose_Tz_mean pose_Rx_max pose_Tz_median</p> <p>SHAP value (impact on model output)</p> <p>feature value</p>	<p>pose_Tx_min pose_Rz_max pose_Ry_min pose_Tx_max pose_Ty_max pose_Tx_mean pose_Tx_std pose_Tz_mean pose_Rx_max pose_Tz_median</p> <p>mean(SHAP value) (average impact on model output magnitude)</p>
<p>9</p>	<p>pose_Rz_max pose_Tx_min pose_Rx_max pose_Ry_max pose_Ry_min pose_Ty_max pose_Ry_median pose_Tx_mean pose_Tx_max pose_Tx_median</p> <p>SHAP value (impact on model output)</p> <p>feature value</p>	<p>pose_Rz_max pose_Tx_min pose_Rx_max pose_Ry_max pose_Ry_min pose_Ty_max pose_Ry_median pose_Tx_mean pose_Tx_max pose_Tx_median</p> <p>mean(SHAP value) (average impact on model output magnitude)</p>
<p>10</p>	<p>pose_Ty_max pose_Rz_min pose_Rx_min pose_Tz_max pose_Rz_max pose_Tx_min pose_Rz_mean pose_Tx_mean pose_Rx_std pose_Rx_mean</p> <p>SHAP value (impact on model output)</p> <p>feature value</p>	<p>pose_Ty_max pose_Rz_min pose_Rx_min pose_Tz_max pose_Rz_max pose_Tx_min pose_Rz_mean pose_Tx_mean pose_Rx_std pose_Rx_mean</p> <p>mean(SHAP value) (average impact on model output magnitude)</p>
<p>11</p>	<p>pose_Rx_max pose_Rz_max pose_Tz_min pose_Ry_min pose_Tz_max pose_Rx_min pose_Ry_mean pose_Ty_max pose_Ry_max pose_Ty_median</p> <p>SHAP value (impact on model output)</p> <p>feature value</p>	<p>pose_Rx_max pose_Rz_max pose_Tz_min pose_Ry_min pose_Tz_max pose_Rx_min pose_Ry_mean pose_Ty_max pose_Ry_max pose_Ty_median</p> <p>mean(SHAP value) (average impact on model output magnitude)</p>





Appendix



References

- Al-Alwani, A. (2016). A Combined Approach to Improve Supervised E-Learning using Multi-Sensor Student Engagement Analysis. *American Journal of Applied Sciences*, 13(12). <https://doi.org/10.3844/ajassp.2016.1377.1384>
- Aleven, V., & Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Intelligent Tutoring Systems, Proceedings* (Vol. 1839, pp. 292-303). <Go to ISI>://WOS:000171336100033
- Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016a). Help Helps, But Only So Much: Research on Help Seeking with Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education*, 26(1), 205-223. <https://doi.org/10.1007/s40593-015-0089-1>
- Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016b). Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 26, 205-223.
- Aleven, V. A. W. M. M., & Koedinger, K. R. (2002). An effective metacognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, 26(2), 147-179. https://doi.org/https://doi.org/10.1207/s15516709cog2602_1
- Amaro, R. (2016). *Teaching computational linguistics* Proceedings of II World Congress on Computer Science, Engineering and Technology Education, Castelo Branco, Portugal.
- Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2).
- Anzalone, S. M., Boucenna, S., Ivaldi, S., & Chetouani, M. (2015). Evaluating the Engagement with Social Robots. *International Journal of Social Robotics*, 7(4), 465-478. <https://doi.org/10.1007/s12369-015-0298-7>
- Atif, A., Richards, D., Liu, D., & Bilgin, A. A. (2020). Perceived benefits and barriers of a prototype early alert system to detect engagement and support ‘at-risk’ students: The teacher perspective. *Computers & Education*, 156, 103954. <https://doi.org/https://doi.org/10.1016/j.compedu.2020.103954>
- Ayouni, S., Hajje, F., Maddeh, M., & Al-Otaibi, S. (2021). A new ML-based approach to enhance student engagement in online environment. *PloS one*, 16(11), e0258788. <https://doi.org/10.1371/journal.pone.0258788>
- Bai, R., Lam, J. C. K., & Li, V. O. K. (2023). What dictates income in New York City? SHAP analysis of income estimation based on Socio-economic and

References

- Spatial Information Gaussian Processes (SSIG). *Humanities and Social Sciences Communications*, 10(1), 60. <https://doi.org/10.1057/s41599-023-01548-7>
- Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. (2018, 15-19 May 2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018),
- Bartholomé, T., Stahl, E., Pieschl, S., & Bromme, R. (2006). What matters in help-seeking? A study of help effectiveness and learner-related factors. *Computers in Human Behavior*, 22(1), 113-129. <https://doi.org/https://doi.org/10.1016/j.chb.2005.01.007>
- Belle, V., & Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning [Review]. *Frontiers in Big Data*, 4, Article 688969. <https://doi.org/10.3389/fdata.2021.688969>
- Betto, I., Hatano, R., & Nishiyama, H. (2023). Distraction detection of lectures in e-learning using machine learning based on human facial features and postural information. *Artificial Life and Robotics*, 28(1), 166-174. <https://doi.org/10.1007/s10015-022-00809-z>
- Bosch, N., & D'Mello, S. K. (2021). Automatic Detection of Mind Wandering from Video in the Lab and in the Classroom. *IEEE Transactions on Affective Computing*, 12(4), 974-988. <https://doi.org/10.1109/TAFFC.2019.2908837>
- Broadbent, J. (2017). Comparing online and blended learner's self-regulated learning strategies and academic performance. *The Internet and Higher Education*, 33, 24-32. <https://doi.org/https://doi.org/10.1016/j.iheduc.2017.01.004>
- Chaouachi, M., Jraidi, I., Lajoie, S. P., & Frasson, C. (2019). Enhancing the learning experience using real-time cognitive evaluation [Article]. *International Journal of Information and Education Technology*, 9(10), 678-688, Article 1287. <https://doi.org/10.18178/ijiet.2019.9.10.1287>
- Chen, M.-Y., & Chen, C.-C. (2010). The contribution of the upper and lower face in happy and sad facial expression classification. *Vision research*, 50(18), 1814-1823. <https://doi.org/https://doi.org/10.1016/j.visres.2010.06.002>
- D'Mello, S., Picard, R. W., & Graesser, A. (2007). Toward an Affect-Sensitive AutoTutor. *IEEE Intelligent Systems*, 22(4), 53-61. <https://doi.org/10.1109/MIS.2007.79>
- Dahl, C. D., Rasch, M. J., & Chen, C.-C. (2014). The other-race and other-species effects in face perception – a subordinate-level analysis [Original Research]. 5. <https://doi.org/10.3389/fpsyg.2014.01068>
- Davenport Huyer, L., Callaghan, N. I., Dicks, S., Scherer, E., Shukalyuk, A. I., Jou, M., & Kilkeny, D. M. (2020). Enhancing senior high school student

References

- engagement and academic performance using an inclusive and scalable inquiry-based program. *npj Science of Learning*, 5(1), 17.
<https://doi.org/10.1038/s41539-020-00076-2>
- de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47(1), 1-12.
<https://doi.org/10.3758/s13428-014-0458-y>
- Desai, R., Porob, P., Rebelo, P., Edla, D. R., & Bablani, A. (2020). EEG Data Classification for Mental State Analysis Using Wavelet Packet Transform and Gaussian Process Classifier. *Wireless Personal Communications*, 115(3), 2149-2169. <https://doi.org/10.1007/s11277-020-07675-7>
- Dragon, T., Arroyo, I., Woolf, B. P., Bursleson, W., el Kaliouby, R., & Eydgahi, H. (2008, 2008//). Viewing Student Affect and Learning through Classroom Observation and Physical Sensors. *Intelligent Tutoring Systems*, Berlin, Heidelberg.
- Dutta, A., & Zisserman, A. (2019). The VIA Annotation Software for Images, Audio and Video. Proceedings of the 27th ACM International Conference on Multimedia, Nice, France.
- Edyburn, D. J. E. T. R., & Development. (2021). Transforming student engagement in COVID-19 remote instruction: a research perspective. *Educational Technology Research and Development*, 69, 113-116.
- Ekman, P., Friesen, W. V. J. E. P., & Behavior, N. (1978). Facial action coding system.
- el Kaliouby, R., & Robinson, P. (2005). Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In B. Kisačanin, V. Pavlović, & T. S. Huang (Eds.), *Real-Time Vision for Human-Computer Interaction* (pp. 181-200). Springer US. https://doi.org/10.1007/0-387-27890-7_11
- Elbawab, M., & Henriques, R. (2023). Machine Learning applied to student attentiveness detection: Using emotional and non-emotional measures. *Education and Information Technologies*. <https://doi.org/10.1007/s10639-023-11814-5>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378-382. <https://doi.org/10.1037/h0031619>
- Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research*, 74(1), 59-109. <https://doi.org/10.3102/00346543074001059>
- Gasevic, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *Internet and Higher Education*, 28, 68-84.

References

- <https://doi.org/10.1016/j.iheduc.2015.10.002>
- Graesser, A. C. (2016). Conversations with AutoTutor Help Students Learn. *International Journal of Artificial Intelligence in Education*, 26(1), 124-132. <https://doi.org/10.1007/s40593-015-0086-4>
- Graesser, A. C., Cai, Z., Morgan, B., & Wang, L. (2017). Assessment with computer agents that engage in conversational dialogues and trialogues with learners. *Computers in Human Behavior*, 76, 607-616. <https://doi.org/https://doi.org/10.1016/j.chb.2017.03.041>
- Graesser, A. C., Hu, X., Nye, B. D., VanLehn, K., Kumar, R., Heffernan, C., . . . Baer, W. (2018). ElectronixTutor: an intelligent tutoring system with multiple learning resources for electronics. *International journal of STEM education*, 5(1), 1-21.
- Ha, Y., & Im, H. (2020). The Role of an Interactive Visual Learning Tool and Its Personalizability in Online Learning: Flow Experience. *Online Learning*, 24(1), 205-226.
- Hasegawa, S., Hirako, A., Zheng, X., Karimah, S. N., Ota, K., & Unoki, T. (2020). *Learner's Mental State Estimation with PC Built-in Camera Learning and Collaboration Technologies*. Human and Technology Ecosystems: 7th International Conference, LCT 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II, Copenhagen, Denmark. https://doi.org/10.1007/978-3-030-50506-6_12
- Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. sage.
- Hong, J.-C., Liu, X., Cao, W., Tai, K.-H., & Zhao, L. (2022). Effects of Self-Efficacy and Online Learning Mind States on Learning Ineffectiveness during the COVID-19 Lockdown [Article]. *Journal of Educational Technology & Society*, 25(1), 142-154. <https://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=155008666&lang=ja&site=ehost-live>
- Hudson, R., & Sheldon, N. (2013). Linguistics at school: The UK Linguistics Olympiad. *Language and Linguistics Compass*, 7(2), 91-104. <https://doi.org/https://doi.org/10.1111/lnc3.12010>
- Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. *Computational Intelligence and Neuroscience*, 2018, 6347186. <https://doi.org/10.1155/2018/6347186>
- Ikeda, T., Cooray, U., Hariyama, M., Aida, J., Kondo, K., Murakami, M., & Osaka, K.

References

- (2022). An Interpretable Machine Learning Approach to Predict Fall Risk Among Community-Dwelling Older Adults: a Three-Year Longitudinal Study. *Journal of General Internal Medicine*, 37(11), 2727-2735.
<https://doi.org/10.1007/s11606-022-07394-8>
- Jonassen, D. H., Carr, C., & Yueh, H.-P. J. T. (1998). Computers as mindtools for engaging learners in critical thinking. *43(2)*, 24-32.
- Kaliouby, R. E., & Robinson, P. (2004, 27 June-2 July 2004). Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. 2004 Conference on Computer Vision and Pattern Recognition Workshop,
- Kaminosono, S., Tseng, C.-H., Chen, C.-C., & Shioiri, S. (2022, September 5-7). *Cutural comparison between Taiwan and Japan's facial expressions* Vision Society of Japan 2022 Summer Meeting (VSJ), Kanazawa City, Ishikawa, Japan.
- Karimah, S. N., & Hasegawa, S. (2022). Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods. *Smart Learning Environments*, 9(1), 31.
<https://doi.org/10.1186/s40561-022-00212-y>
- Kato, H., Takahashi, K., Hatori, Y., Sato, Y., & Shioiri, S. (2022a). *Prediction of engagement from facial expressions: Effect of dynamic factors* The 18th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP-2022), Kitakyushu, Japan.
- Kato, H., Takahashi, K., Hatori, Y., Sato, Y., & Shioiri, S. (2022b). *Prediction of Engagement from Temporal Changes in Facial Expression*. World Conference on Computers in Education (IFIP WCCE, 2022), Hiroshima, Japan.
- Kawamura, R., & Murase, K. (2020). Concentration Estimation in E-Learning Based on Learner's Facial Reaction to Teacher's Action. Proceedings of the 25th International Conference on Intelligent User Interfaces Companion,
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T.-Y. J. A. i. n. i. p. s. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *30*.
- Kim, H., Chae, Y., Kim, S., & Im, C. H. (2023). Development of a Computer-Aided Education System Inspired by Face-to-Face Learning by Incorporating EEG-Based Neurofeedback Into Online Video Lectures. *IEEE Transactions on Learning Technologies*, 16(1), 78-91.
<https://doi.org/10.1109/TLT.2022.3200394>
- Kouahla, M. N., Boughida, A., Chebata, I., Mehenaoui, Z., & Lafifi, Y. (2022). Emorec: a new approach for detecting and improving the emotional state of learners in an e-learning environment. *Interactive Learning Environments*, 1-19. <https://doi.org/10.1080/10494820.2022.2029494>

References

- Kukowski, C., Bernecker, K., & Brandstätter, V. (2021). Self-Control and Beliefs Surrounding Others' Cooperation Predict Own Health-Protective Behaviors and Support for COVID-19 Government Regulations: Evidence From Two European Countries. *Social Psychological Bulletin*, *16*(1), 1-28.
<https://doi.org/10.32872/spb.4391>
- Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159 - 174.
- Lerche, T., & Kiel, E. (2018). Predicting student achievement in learning management systems by log data analysis. *Computers in Human Behavior*, *89*, 367-372.
<https://doi.org/https://doi.org/10.1016/j.chb.2018.06.015>
- Li, S., Lajoie, S. P., Zheng, J., Wu, H., & Cheng, H. (2021). Automated detection of cognitive engagement to inform the art of staying engaged in problem-solving. *Computers & Education*, *163*, 104114.
<https://doi.org/https://doi.org/10.1016/j.compedu.2020.104114>
- Lin, F.-R., & Kao, C.-M. (2018). Mental effort detection using EEG data in E-learning contexts. *Computers & Education*, *122*, 63-79.
<https://doi.org/https://doi.org/10.1016/j.compedu.2018.03.020>
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., . . . Lee, S.-I. J. a. p. a. (2019). Explainable AI for trees: From local explanations to global understanding.
- McDuff, D., Girard, J. M., & Kaliouby, R. e. (2017). Large-Scale Observational Evidence of Cross-Cultural Differences in Facial Behavior. *Journal of Nonverbal Behavior*, *41*(1), 1-19. <https://doi.org/10.1007/s10919-016-0244-x>
- McDuff, D., & Kaliouby, R. e. (2017). Applications of Automated Facial Coding in Media Measurement. *IEEE Transactions on Affective Computing*, *8*(2), 148-160. <https://doi.org/10.1109/TAFFC.2016.2571284>
- Miao, R., Kato, H., Hatori, Y., Sato, Y., & Shioiri, S. (2022). *Analysis of facial expressions for the estimation of concentration on online lectures*. World Conference on Computers in Education (IFIP WCCE, 2022), Hiroshima, Japan.
- Miao, R., Kato, H., Hatori, Y., Sato, Y., & Shioiri, S. (2023). Analysis of facial expressions to estimate the level of engagement in online lectures. *IEEE Access*, 1-1. <https://doi.org/10.1109/ACCESS.2023.3297651>
- Mills, C., Bosch, N., Graesser, A., & D'Mello, S. (2014, Jun 05-09). To Quit or Not to Quit: Predicting Future Disengagement from Reading Patterns. *Lecture Notes in Computer Science* [Intelligent tutoring systems, its 2014]. 12th International Conference on Intelligent Tutoring Systems (ITS), Honolulu, HI.
- Monkaresi, H., Bosch, N., Calvo, R. A., & Mello, S. K. D. (2017). Automated

References

- Detection of Engagement Using Video-Based Estimation of Facial Expressions and Heart Rate. *IEEE Transactions on Affective Computing*, 8(1), 15-28. <https://doi.org/10.1109/TAFFC.2016.2515084>
- Mousavinasab, E., Zarifsanaiey, N., Kalhori, S. R. N., Rakhshan, M., Keikha, L., & Saeedi, M. G. (2021). Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, 29(1), 142-163. <https://doi.org/10.1080/10494820.2018.1558257>
- O'Brien, H. L., Roll, I., Kampen, A., & Davoudi, N. (2022). Rethinking (Dis)engagement in human-computer interaction. *Computers in Human Behavior*, 128, 107109. <https://doi.org/https://doi.org/10.1016/j.chb.2021.107109>
- O'Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1), 50-69. <https://doi.org/https://doi.org/10.1002/asi.21229>
- Ozili, P. K. (2023). The acceptable R-square in empirical modelling for social science research. In *Social research methodology and publishing results: A guide to non-native english speakers* (pp. 134-143). IGI Global.
- Pellet-Rostaing, A., Bertrand, R., Boudin, A., Rauzy, S., & Blache, P. (2023). A multimodal approach for modeling engagement in conversation [Original Research]. 5. <https://doi.org/10.3389/fcomp.2023.1062342>
- Peng, S., & Nagao, K. (2021). Recognition of Students' Mental States in Discussion Based on Multimodal Data and its Application to Educational Support. *IEEE Access*, 9, 18235-18250. <https://doi.org/10.1109/ACCESS.2021.3054176>
- Prado, C., Mellor, D., Byrne, L. K., Wilson, C., Xu, X., & Liu, H. (2014). Facial emotion recognition: A cross-cultural comparison of Chinese, Chinese living in Australia, and Anglo-Australians. *Motivation and Emotion*, 38(3), 420-428. <https://doi.org/10.1007/s11031-013-9383-0>
- Ren, P., Ma, X., Lai, W., Zhang, M., Liu, S., Wang, Y., . . . Xu, X. (2019). Comparison of the Use of Blink Rate and Blink Rate Variability for Mental State Recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(5), 867-875. <https://doi.org/10.1109/TNSRE.2019.2906371>
- Renier, L. A., Schmid Mast, M., Dael, N., & Kleinlogel, E. P. (2021). Nonverbal Social Sensing: What Social Sensing Can and Cannot Do for the Study of Nonverbal Behavior From Video. *Frontiers in Psychology*, 12, 606548. <https://doi.org/10.3389/fpsyg.2021.606548>
- Risko, E. F., Anderson, N., Sarwal, A., Engelhardt, M., & Kingstone, A. (2012).

References

- Everyday Attention: Variation in Mind Wandering and Memory in a Lecture. *Applied Cognitive Psychology*, 26(2), 234-242.
<https://doi.org/https://doi.org/10.1002/acp.1814>
- Roll, I., Alevan, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 21(2), 267-280.
<https://doi.org/https://doi.org/10.1016/j.learninstruc.2010.07.004>
- Ryan, A. M., Ryan, A. M., Gheen, M. H., & Midgley, C. (1998). Why do some students avoid asking for help? An examination of the interplay among students' academic efficacy, teachers' social-emotional role, and the classroom goal structure. *Journal of educational psychology*, 90(3), 528-535.
<https://doi.org/10.1037/0022-0663.90.3.528>
- Ryan, A. M., Ryan, A. M., & Pintrich, P. R. (1997). "Should I ask for help?" The role of motivation and attitudes in adolescents' help seeking in math class. *Journal of educational psychology*, 89(2), 329-341. <https://doi.org/10.1037/0022-0663.89.2.329>
- Sümer, Ö., Goldberg, P., D'Mello, S., Gerjets, P., Trautwein, U., & Kasneci, E. (2021). Multimodal Engagement Analysis From Facial Videos in the Classroom. *IEEE Transactions on Affective Computing*, 14(2), 1012-1027.
<https://doi.org/10.1109/TAFFC.2021.3127692>
- Sato, Y., Horaguchi, Y., Vanel, L., & Shioiri, S. (2022). Prediction of Image Preferences from Spontaneous Facial Expressions. *Interdisciplinary Information Sciences*, 28(1), 45-53. <https://doi.org/10.4036/iis.2022.A.02>
- Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). *Education and Information Technologies*, 28(7), 8299-8333. <https://doi.org/10.1007/s10639-022-11536-0>
- Sham, A. H., Aktas, K., Rizhinashvili, D., Kuklianov, D., Alisininoglu, F., Ofodile, I., . . . Anbarjafari, G. (2023). Ethical AI in facial expression analysis: racial bias. *Signal, Image and Video Processing*, 17(2), 399-406.
<https://doi.org/10.1007/s11760-022-02246-8>
- Shioiri, S. (2022). New Informatics Paradigm to Manage Quality and Value of Information. *Interdisciplinary Information Sciences*, 28(1), iv-iv.
<https://doi.org/10.4036/iis.2022.A.00>
- Shioiri, S., Sato, Y., Horaguchi, Y., Muraoka, H., & Nihei, M. (2021, 22-28 May 2021). Quali-Informatics in the Society with Yotta Scale Data. 2021 IEEE International Symposium on Circuits and Systems (ISCAS),
- Son, N. H., Takahata, Y., Goto, M., Tanaka, T., Ohsuga, A., & Matsumoto, K. J. P. o. t.

References

- I. C. (2020). Estimating the Concentration of Students from Time Series Images. *69*, 224-229.
- Tang, K.-Y., Chang, C.-Y., & Hwang, G.-J. (2021). Trends in artificial intelligence-supported e-learning: a systematic review and co-citation network analysis (1998–2019). *Interactive Learning Environments*, 1-19.
<https://doi.org/10.1080/10494820.2021.1875001>
- Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98-110. <https://doi.org/10.1016/j.chb.2018.07.027>
- Wang, F. H. (2019). On prediction of online behaviors and achievement using self-regulated learning awareness in flipped classrooms. *International Journal of Information Education Technology*, *9*(12), 874-879.
- Wang, G.-Y. (2020). *Emulating empathy in emotional support system to test performance on collaborative learning*. [National Taiwan University]. Taipei, Taiwan.
- Wang, G.-Y., Hatori, Y., Sato, Y., Tseng, C.-H., & Shioiri, S. (2023). Predicting learners' engagement and help-seeking behaviors in an e-learning environment by using facial and head pose features. [preprint], Available at SSRN: <http://dx.doi.org/10.2139/ssrn.4600003>.
- Wang, G.-Y., Nagata, H., Hatori, Y., Sato, Y., Tseng, C.-H., & Shioiri, S. (2023, July 20-22). Detecting Learners' Hint Inquiry Behaviors in e-Learning Environment by using Facial Expressions. Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23). Copenhagen, Denmark.
- Whitehill, J., Serpell, Z., Lin, Y. C., Foster, A., & Movellan, J. R. (2014). The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions. *IEEE Transactions on Affective Computing*, *5*(1), 86-98.
<https://doi.org/10.1109/TAFFC.2014.2316163>
- Xiao, J., Jiang, Z., Wang, L., & Yu, T. (2022). What can multimodal data tell us about online synchronous training: Learning outcomes and engagement of in-service teachers. *Frontiers in Psychology*, *13*, 1092848.
<https://doi.org/10.3389/fpsyg.2022.1092848>
- Yueh, H.-P., Wang, G.-Y., & Lee, T. S.-H. (2022). The cognition, information behaviors, and preventive behaviors of Taiwanese people facing COVID-19. *Scientific Reports*, *12*(1), 16934. <https://doi.org/10.1038/s41598-022-20312-6>
- Zhao, Y., Wang, N., Li, Y., Zhou, R., & Li, S. (2021). Do cultural differences affect users' e-learning adoption? A meta-analysis. *52*(1), 20-41.
<https://doi.org/10.1111/bjet.13002>
- Zhi, R., Liu, M., & Zhang, D. (2020). A comprehensive survey on automatic facial

References

action unit analysis. *The Visual Computer*, 36(5), 1067-1093.
<https://doi.org/10.1007/s00371-019-01707-5>