Children State Anxiety Detection: BSDCSA Dataset and Detection Algorithms

Thesis for the Master Artificial Intelligence

Author: Shaoya Ren

First Supervisor: Prof. dr. Remco Veltkamp

Second Examiner: Dr. Chao Zhang

Utrecht University

Universiteit Utrecht

The Netherlands

July, 2021

Preface

This thesis is for the Master Artificial Intelligence and the project of Alliance Program HUMAN-AI. The HUMAN-AI program was established by the Utrecht University (UU), the Eindhoven University of Technology (TU/e) and the University Medical Center Utrecht (UMCU), which centers on improving the transparency of AI decision-making and the autonomy of human in the process of decision making. Participants in this program are required to propose their own research direction and form three-person teams to complete one project. Therefore, Ning Fang from UU majoring in Psychology, Laura Bijl from TU/e majoring Industrial Design and I (Shaoya Ren) from UU majoring in Artificial Intelligence work together in this program. Our supervisors are dr. Chao Zhang from UU, Prof. dr. Remco Veltkamp from UU and dr. Supraja Sankaran from TU/e. The aim of our project is to design an affective Socially Assistive Robot (SAR), which will be an assistant for children (10-12 years old) with generalized anxiety disorder (GAD). On the one hand, anxiety disorder is prevalent among children. On the other hand, socially assistive robots (SARs) can provide promising help in improving children's mental health and alleviating clinical symptoms. The SAR detects whether children are under anxiety through their voices and provide the Cognitive behavioural therapy (CBT) method to relieve their anxiety.

There are mainly three aspects of this SAR, corresponding to our majors. Laura is responsible for explainable AI, which helps children understand the decision-making process of SAR. This is an important aspect of human's autonomy, which makes the decision-making process of SAR more transparent. Ning is responsible for the whole psychology part, including the design of the data collection experiment and the implementation of CBT. My part mainly focuses on the state anxiety detection algorithms. I also collected the dataset for this project with Ning. The SAR applied in this research is Zenbo, which is developed by ASUS. Zenbo has a professional development platform for developers, Zenbolab, and convenient APIs. It has been proved that children are easy to accept Zenbo, as one part of the data collection experiment was conducted with Zenbo.

This thesis will introduce the part I am responsible for in this project. The first target of this thesis is to describe the BSDCSA dataset, which is composed of collected data. The second target is to exhibit children's state anxiety detection algorithms. The third target of this thesis is to propose and answer research questions related to data and anxiety detection algorithms.

Abstract

Anxiety disorders in children have been long ignored by researchers, psychologists and parents. On the one hand, children's anxiety symptoms cannot be found and paid attention to in time. On the other hand, the diagnosis and treatment for children with anxiety disorders mainly depend on therapists' experience, which may be affected by subjective and lack of daily observation. A method or system that detects and monitors children's state anxiety based on the machine learning method can solve these two problems. However, due to the lack of data, there is little research in this area. This research is a part of the project from the HUMAN-AI program. This research will propose a new dataset, Bilingual Speech Dataset for Children's State Anxiety (BSDCSA), for fellow research in this field. In this thesis, a basic flow of data processing for this dataset and state anxiety recognition models are exhibited. As state anxiety is a kind of emotion, the models trained for common emotions are referred to. The performances of models are evaluated and compared through evaluation metrics. As BSDCSA contains two kinds of labels, anxiety level and anxiety label, regression models and classification models are both trained. For anxiety state classification, the easy ensemble from under-sampling and threshold moving are applied to solve the problem of unbalanced data. Boosting algorithms, GBDT (Gradient boosting decision tree) and XGB (XGBoost), with easy ensemble and threshold moving gain the highest TPR (True Positive Rate), TNR (True Negative Rate), F1 score and AUC. For anxiety level prediction, boosting algorithms, GBR and XGBR, gain the lowest RMSE, MAE and highest R^2 score. The effectiveness of anxiety labels will be discussed. The research questions about the acoustic parameters most related to children's state anxiety, the change of common acoustic parameters with the increase of anxiety and the influence of sociodemographic characteristics on anxiety and acoustic parameters will be analyzed. This research aims to lay data and algorithm foundations for further projects and offer recommendations for future study in children's state anxiety recognition field. Limitations and future works will also be discussed.

Keywords: *State Anxiety of Children, Speech Anxiety Analysis, Acoustic Features, Speech processing, Machine Learning;*

Contents

1 Introduction

1.1 Research Description

This research is a combination of psychology and machine learning, and a part of the project *"An Affective Social Assistive Robot (SAR) for children with Generalized Anxiety Disorder (GAD)"* from HUMAN-AI Alliance Program. The project aims to build a SAR with Zenbo robot to assist children with GAD through several functions, including state anxiety detection, emotional communication and mental health monitoring. This thesis will introduce one part of this project, including the Bilingual Speech Dataset for Children's State Anxiety (BSDCSA) and state anxiety recognition algorithms that detect children's state anxiety through acoustic parameters. BSDCSA is an elicited datatset that was collected through experiments with students aged 10 to 12. The design of the data collection experiment, the method of dataset annotation and detailed information of BSDCSA will be exhibited in this thesis. A specific procedure of audio processing will be interpreted step by step. With BSDCSA, regression models and classification models will be trained. The performance of models with different training conditions will be evaluated and compared. In order to further explore the dataset and test the trained models, three research questions about the acoustic parameters related to children's state anxiety, the change of acoustic features with the increase of anxiety level and the influence of sociodemographic characteristics on anxiety recognition will be proposed and discussed. Limitations of this research will be discussed and provide future research direction for improvements. For one thing, this thesis will introduce a novel dataset for the field of children's anxiety analysis. For another thing, this thesis will provide constructive suggestions for further research regarding audios processing, acoustic parameters analysis and children's state anxiety detection algorithms.

Figure 1: The pictures of Zenbo

1.2 Thesis Structure

In the following subsections, the background and motivation for this research will be introduced. In the second section, the related work in this field will be summarized. In Research Question section, three research questions are proposed to further explore the data and models. In Methodology section, three aspects will be introduced. The first aspect is a detailed description of the data collection procedure

and BSDCSA. The second part is the standard data processing procedure for this dataset. The data processing will be further divided into two parts. The first part is raw audios processing and the second part is feature engineering. The models and evaluation metrics applied in this research will be introduced finally. In Results section, the performance of classifiers and regressors will be illustrated in the form of tables and figures. In Discussion section, the proposed research questions, the limitations and future works will be answered and discussed in specific. At last, a conclusion will be given to summarize this research.

1.3 Background and Motivation

Anxiety disorder in children is underexposed by researchers, psychologists and parents, although it is becoming common in children. The attention and experience on anxiety disorder of children are much less than that of adults. On the basis of recent research, the number of children suffering from anxiety disorders is increasing. Currently, about 10% - 15% of the children suffer from anxiety disorders globally [1]. According to the report of NSCH, in 2018 - 2019, 8.5% of American children aged 3-17 years (approximately 4.4 million) have been diagnosed with anxiety [2]. This number was 6.4% in 2011 - 2012. Especially, the isolation caused by the outbreak of COVID-19 has brought much more anxiety and panic to children than usual [3]. In contrast, due to the negligence or the lack of anxiety-related knowledge of parents, in most cases, the symptoms of anxiety disorder occurring among children are difficult to be noticed in time, and diagnosed children may not be treated properly. This may influence children's life and study and lead to more serious mental diseases in adolescence and adulthood. High levels of anxiety lead to a variety of negative effects, such as the decline of academic performance and social difficulties in a short term [4], and disrupting the developing architecture of the brain, which may cause lifelong consequences in the long term [5]. Moreover, anxiety attacks are unpredictable. It is possible that a child is unable to get professional treatment in time when she or he experiences anxiety attacks. In these cases, a method or system to detect children's anxiety levels in real-time and offer assistance to anxious children is needed.

Another reason to construct the anxiety detection system is that it is beneficial to improve the efficiency and accuracy of the diagnosis of anxiety disorders in children. Besides, the anxiety level monitoring function will assist therapists in tracking the psychological status of the patients. Right now, the diagnosis of anxiety disorders mainly depends on patients' self-reports or the experience of therapists [6], which is likely to be affected by subjective factors. An anxiety detection system is a more objective method. By summarizing the relevant literature in recent years, Daniel M. Low et al. [6] confirmed the role of speech-based machine learning methods or systems in the diagnosis and treatment of mental diseases. In conclusion, the method or system to recognize children's anxiety through speech is feasible and meaningful.

To realize the state anxiety detection and auxiliary therapy function, Socially Assistive Robots (SARs) are taken into account. On the one hand, it is proved that SARs can provide promising help in improving children's mental health and alleviating clinical symptoms [7]. On the other hand, children

show high acceptance of SARs in psychotherapy [8][9]. Alemi et al. applied a pre-programmed Nao in psychological interventions for children with cancer and found that after treatment with SAR, children's stress, depression and anger were significantly reduced [10]. Katarzyna Kabacińska et al. summarized the literature on the role of SARs in children's mental health up to 2020 and proposed a series of suggestions to guide further research in this area [11]. Therefore, employing a pre-programmed SAR to provide psychological help for children under anxiety attacks is possible. Besides, due to children's high acceptance of SAR, it is also a suitable executor for data collection experiments to decrease the influence of subjective factors of experimenters on collected data.

Anxiety to be analyzed in this research is state anxiety. Spielberger classified anxiety into state anxiety and trait anxiety [12]. Trait anxiety refers to a personality trait and describes the individual differences related to the tendency of present state anxiety [12]. Anxiety disorders in diagnosed patients are regarded as trait anxiety because the symptoms last for weeks or months. State anxiety is a transient emotional pattern caused by environmental stress, including physiological arousal and symptoms of anxiety, worry and tension [13]. It reflects the psychological and physiological instant reaction directly related to a particular moment and changes in a short time [14]. The short-term variability makes the level of state anxiety an appropriate indicator for anxiety detection. Because of the properties of state anxiety, it is not only suitable for monitoring the psychological state of the diagnosed children, but also it can reflect the potential anxiety of mentally healthy children.

As state anxiety can be regarded as an emotion, similar to happiness and sadness [15], the recognition method for basic emotions could be referred to. Right now, emotion recognition is mainly through videos, pictures, faces, speeches and physiological characteristics. However, it is difficult to capture effective static images that can be used to extract features from dynamic videos or faces when children are under anxiety. Besides, limited by equipment and time, physiological characteristics will not be considered. Therefore, speeches are ideal data. One advantage of acoustic features is that symptoms in speech are hard to hide and the emotions are directly expressed through speech [6].

Although the previous studies offer the feasibility for this research, realizing the algorithms specialized for detecting children's state anxiety is still challenging. For one thing, different from other basic emotions, anxiety is difficult to measure and evaluate, which leads to unreliable annotations of anxiety by unprofessional psychotherapists and a lack of baseline algorithms. However, fellow research is inseparable from reliable datasets. This is because that everyone has a different perception and cognition of anxiety. For another thing, the acoustic parameters specialized for children's anxiety detection are rarely analyzed and summarized. This leads to the need to try all the acoustic features when training the model. Therefore, it is necessary to collect a new dataset for the research in this field. Besides, a summarized feature set can help improve the efficiency of research. Baseline models are also required to provide a comparison for future studies.

2 Related Work

2.1 Data

As mentioned in the last section, there are few public datasets specialized for children's state anxiety. EmoReact published by Behnaz et al. contains 17 different emotions, including anxiety from children aged 4 to 14 [16]. It includes both videos and audios, which are suitable for visual and acoustic analysis. The data in it were collected from a Youtube channel and labeled by three annotators. However, due to the inappropriate annotation method, the labels for anxiety are not accurate in this dataset. McGinnis et al. used adapted TSST-C [17] to collect audio data from children aged 3 to 8. Each child was required to give a 3-minutes speech [18]. These speeches compose the experimental data. However, they didn't publish the data in their experiment as a public dataset.

2.2 Features

This research focuses on recognizing anxiety through speeches. Therefore, acoustic features related to anxiety disorders are paid more attention to. There are four types of acoustic features: Prosodic Features, Spectral Features, Voice Quality Features and Teager Energy Operator based Features [19]. Prosodic features include F0, energy, duration and so on. Spectral features include Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients(LPCC) and so on. Voice quality features include jitter, shimmer, HNR and so on. Teager energy operator based features are features depending on the Teager Energy Operator (TEO) [19]. Table 1 introduces the common features in detail [19]. Some of these features are related to anxiety. It has been studied how the values of acoustic features of people's voices change when they are under anxiety. According to the research of Banse et al. [20] and Weeks et al. [21], the mean of F0 increases. Murray et al. reported that the speech rate of an anxious person would increase, and the intensity and HNR will be irregular [22]. The values of jitter and shimmer will increase in anxious voices [23]. Turgut et al. analyzed and compared the voice of anxiety disorder diagnosed patients and mentally healthy people. acoustic features were considered. And 42 acoustic features were reported as anxiety-affected features [24]. Albuquerque et al. extracted 18 acoustic features from 112 individuals aged 35-97 to reveal the association between acoustic features and non-severe levels of anxiety [25]. Most studies focus on acoustic features of adults with anxiety. McGinnis et al. extracted 164 acoustic features that were used to identify anxiety in adults from speeches to analyze the voice of children with an internalizing disorder [18]. This research proves that some acoustic features which are effective for detecting adult anxiety can be a reference for children's anxiety recognizing.

2.3 Algorithms

As state anxiety is an emotion, the methods applied in general emotions recognition will also be references for this research. Mehmet et al. [19] summarized classifiers that have been used in emotion

Table 1: The description for common acoustic features

Prosodic Features	F0	Fundamental frequency. F0 is caused by the opening and closing of the glottis when a person utters sound. It reflects the time interval between two adjacent opening and closing of the glottis or the frequency of opening and closing.
	Energy	Also referred as volume or the intensity. It reflects amplitude variation of speech signals.
	Duration	Duration is the duration for words, silence and so on, such as Speech rate, duration of silence regions, rate of duration of voiced and unvoiced regions.
Spectral Features	MFCC	Mel-frequency cepstral coefficient. It is extracted frommel-frequency cepstrum and reflects the short term power spectrum of the speech signal.
	LPCC	Linear prediction cepstral coefficients. LPCC is derived from Linear PredictionCoefficient (LPC). LPC is the coefficients of all-pole filters.
Voice Quality Features	Jitter	Jitter reflects frequency instability
	Shimmer	Shimmer reflects amplitude instability
	HNR	Harmonics to noise ratio. It is the relative level of noise in the frequency spectrum of vowels.

classification by 2020 comprehensively. For anxiety detection, traditional machine learning methods have been used. Pintelas summarized the literature of supervised learning methods in the field of anxiety disorder detection up to 2018 [26]. According to this review, common machine learning methods such as logistic regression, SVM, random forest and so on, can be used to diagnose anxiety. Binary classifiers were applied to detect children's anxiety with acoustic features in McGinnis's research [18]. Salekin et al. proposed a weakly supervised learning framework for detecting social anxiety and depression [27]. Based on Occam's razor, in this research, traditional machine learning methods will be used firstly, then boosting algorithms will be developed.

3 Research Question

3.1 Which acoustic parameters are more relevant to children's state anxiety?

According to the literature review, some acoustic parameters, such as F0, MFCC, HNR, will be influenced by anxiety and also play a role in detecting anxiety. Right now, most research focuses on

acoustic features in adults' voices. The correlation between acoustic parameters and children's state anxiety is less researched. Because of the difference between the acoustic characteristics of children's voices and that of adults' voices, it is possible that the most anxiety-relevant features of adults' voices are less relevant or irrelevant to children's anxiety. Therefore, this research question will explore the acoustic parameters that are most relevant to children's state anxiety and provide a valuable acoustic feature set for future study in this domain.

3.2 How do acoustic parameters change with the increase of state anxiety level?

This question aims to inspect the correlation between acoustic parameters and the state anxiety levels of children. Similar studies in the field of adults' have been summarized in Related Work section. With these references, this research question will demonstrate the change of acoustic parameters that are common in previous research with the increase of children's state anxiety level.

3.3 Do sociodemographic characteristics have an impact on the state anxiety and acoustic parameters?

In BSDCSA, the sociodemographic information, including age, gender and education degree (grade) of participants, is recorded for each audio. There are two reasons for this research question. Firstly, it is uncertain whether the expression of anxiety is related to sociodemographic information. Secondly, it is possible that the acoustic parameters are influenced by the difference in sociodemographic characteristics other than anxiety. This research question will analyze the relationship between sociodemographic characteristics and anxiety as well as acoustic parameters.

4 Methodology

4.1 Data: BSDCSA

Bilingual Speech Dataset for Children State Anxiety (BSDCSA) is an elicited dataset. During data collection, participants are placed in a simulation situation which triggered their anxiety. BSDCSA contains Chinese and English audios from Chinese pupils, which is a novel dataset in the field of children's state anxiety detection. This property provides the possibility of applying BSDCSA in both Chinese and English contexts. Besides, in BSDCSA, there are two types of labels, the anxiety level which is a continuous variable and the anxiety label which is a discrete variable, which makes BSDCSA support for regression and classification tasks. In this section, how we collected the dataset will be described. And the methods of annotations will be explained.

4.1.1 Data Collection

Participants

The data was collected between 2021/03/26 and 2021/04/29 in Xingdong School located in Xiamen, China. Participants were 81 students aged 10 to 12 from the fourth and fifth grades. Before the beginning of the experiment, all students in fourth and fifth grades were required to take trait anxiety tests. Then we sent the consent forms to all students and their guardians. After this, we randomly selected 81 participants from students whose trait anxiety levels are lower than 39 and who signed the consent forms. 12 participants (6 males, 6 females) joined the pilot study and 69 participants (34 males, 35 females) joined the formal experiment.

Experiment

The data was collected in the form of experiments. All the experiments were approved by Utrecht University and Xingdong School. The whole experiment was divided into three sections and conducted in two classrooms. Before the experiment, participants would be informed of the precautions and confidentiality of the experiment, which aims to make participants more relax to express their authentic emotions.

The first section is the anxiety induction experiment. This experiment is an adapted version of TSST-C [17]. There are three tasks in the first section, including English Storytelling, Mental Arithmetic, and English Reading. To finish the English Storytelling task, participants had to continue a given story and tell the end of the story in English, lasting at least two minutes. To prevent participants from divulging experimental details, one of four stories was selected randomly for each participant. In the Mental Arithmetic task, participants need to subtract 7 from 758 continuously and stick to one minute's calculation. For the third task, participants were asked to read an English article on paper. Participants' voices in task 1 and task 3 will be recorded. All the materials of the three tasks, the contents of stories, and the numbers would not induce emotional changes.

The second section aimed to relieve participants' anxiety caused by the first section and assess the acceptance of participants to Zenbo. In this section, participants interacted with a SAR, which is a pre-programmed Zenbo. Based on ZenboLab, we designed three different programs corresponding to three groups of participants, including experimental groups are Therapeutic Autonomous (Group 1), Therapeutic Non-Autonomous (Group 2), and Non-Therapeutic (Group 3). Participants in Group 1 and Group 2 followed the instructions of SAR to relax with PMR therapy. Progressive Muscle Relaxation (PMR) therapy belongs to Cognitive Behavioral Therapy (CBT), which is a form of psychological treatment for anxiety problems [28]. The difference between Group 1 and Group 2 is that participants in group 1 had the autonomy to choose desired movements, while participants in group 2 could only follow the instructions of SAR. For the participants in group 3, SAR only had daily chats with them. It asked participants some questions that did not involve participants' preferences, such as "What did you have for breakfast?". Participants were randomly divided into three groups by the website

RANDOM.ORG. At the end of section 2, participants expressed their feelings about section 1 and section 2 in 2 minutes. Voices in this part were recorded.

In section 3, participants stayed alone in a classroom for 20 minutes and they could use the items in this room at will. However, they had to complete a story-writing task during this period. At the end of this section, they were required to read the finished story. The reading part was recorded.

Throughout the whole experiment, for each participant, a total of more than five minutes of speech was recorded. These recordings are the audio data of BSDCSA.

Annotations

In BSDCSA, both the anxiety level and the anxiety labels of participants for each audio were annotated. We modified "A short form of the Chinese version of the State Anxiety Scale for Children (CSAS-C)" [29] to acquire annotations. The CSAS-C is translated from Trait Anxiety Scale for Children, and it is proved that CSAS-C could be used to evaluate Chinese children's state anxiety through experiments [30]. After publishing CSAS-C, Ho et al. adapted CSAS-C into a short form containing ten items to make it more suitable for busy clinical settings [29]. They compared the evaluation results of the short form version and the original version and concluded that the original version could be replaced by the short form version in some situations.

In our questionnaire (the modified scale), there are ten items with four choices and one blank. One example of items is "Do you feel happy?" And the participants chose one option from "Hardly", "Appreciably", "Moderation", and "Extraordinary". The weighted scores of the ten items were calculated as the values of state anxiety (SA), which is anxiety levels. The blank is "What is your main emotion right now?", and the participants should fill in their main emotions at that moment. Any emotions were possible. These emotions filled in the blank are emotion labels. The anxiety labels were deduced from emotion labels. These emotions included negative emotions and positive emotions. Negative emotions corresponded to anxiety and positive emotions corresponded to no anxiety. The details will be introduced in the next section. Before the start of experiments and after each section, participants were required to completed the questionnaires.

In the pilot study, we noticed that participants tried to show their best, which means they were more likely to report positive emotions even though their real emotions were negative. For example, although he or she was nervous, the participant still filled in the scale with "hardly nervous". In self-reports, this is a normal phenomenon caused by social desirability bias [31]. In order to reduce the bias as much as possible, we adopted an "interview" form. Instead of distributing the scales to participants, we asked participants the questions (items in the questionnaire) and fill in their answers by ourselves. Each question would be repeated twice. In order to avoid the participants' impatience caused by repetitive questionnaires, the order of questions was randomly shuffled in four interviews.

4.1.2 BSDCSA Description

BSDCSA contains 294 raw audios (39 audios from the pilot study, 255 audios from the formal experiment) recorded by a recorder and a table containing the information of the audios. The duration of audios ranges from half a minute to eight minutes. The composition of the audio name is "ExperimentID0SectionNumber" and an example is "4033901". The participants' information, including ExperimentID, Record Date, Gender, Grade, and language, and annotations that include emotion labels, anxiety labels and anxiety levels, is demonstrated in the table. This information was collected from consent forms. The state anxiety level ranges from 10 to 40. A total of 14 emotions were labeled and were divided into 7 positive emotions and 7 negative emotions. 14 emotions are labeled from 0 to 13. The anxiety labels were deduced from emotion labels, positive emotions were regarded as anxiety-unrelated emotions and negative emotions were regarded as anxiety-related emotions. Anxiety-related emotions include sadness, fear, anger, tension, worry, annoyance, and uneasiness. Anxiety-unrelated emotions include delight, relaxation, calmness, expectation, happiness, excitement, and curiosity. Anxiety-related emotions are labeled as 1, representing anxiety, and anxiety-unrelated emotions are labeled as 0, representing no anxiety. Besides, we also marked the anxiety level of each participant based on our own observations. This level ranges from 1 to 7. The greater the number, the higher the degree of anxiety. Table 2 demonstrates the basic information of participants. Table 3 summarizes the labels of BSDCSA.

Table 2: The information of participants

Gender	Female 41
	Male 40
Education Level	Grade 4, Grade 5
Age	10,11,12
Languages	English, Chinese, mixed Chinese and English

Table 3: The annotations of BSDCSA

Emotion Labels		Anxiety Label		Anxiety Level	
Anxiety-Related	sadness, fear, anger, tension, worry, annoyance, uneasiness	Anxiety	1	Min	10
Anxiety-Unrelated	delight, relaxation, calmness, expectation, happiness, excitement, curiosity	No-Anxiety	0	Max	40
Degree of Interaction with Robot	1,2,3,4,5				
Label by experimenter	1,2,3,4,5,6,7 (Not complete)				

4.2 Data Process

For this research, only data collected from 69 participants (34 males, 35 females) in the formal experiments was used, because parts of the procedure and the contents of the formal experiment were adjusted according to the results of pilot experiments.

4.2.1 Audio Preprocess

Noise Reduction

Due to the poor sound insulation room and less advanced recording device, the noises were inevitable when recording. The noises maybe the sound of nature, the sounds made by students outside the classrooms, the voice of experimenters, and the electromagnetic sound of the recording equipment. These noises will affect the subsequent data processing and model training. Therefore, noise reduction is essential. Besides, the different periods of the experiment lead to different types of noise in recordings. For example, in the morning, most of the students are in class, and the surroundings are quiet. In the afternoon, there are many students in PE class, and the surroundings are noisy (The experimental classroom is close to the sports ground). Therefore, Adobe Premiere (PR) was used to reduce noises manually. On the one hand, noises were eliminated to the greatest extent without affecting the participants' voices. On the other hand, the overlapping parts of participants' voices and other voices were deleted as accurately as possible.

Audio Segmentation

For the convenience of audio data analysis and feature extraction, the noise-removed audios were segmented into smaller clips. Pydub module from python was used to automatically split the audios on silences. However, after splitting, there were still clips that were too long or too short for following analysis. For these clips, further segmentation or integration was conducted. At the end of segmentation, the duration of clips ranges from 1 second to 4 seconds. There are 3864 clips in total.

Pre-emphasis

Pre-emphasis is a signal processing method that compensates for the high frequency component of the input signal. The purpose of speech pre-emphasis is to enhance the high frequency part of speech, remove the influence of lip radiation, and increase the high frequency resolution of speech [32]. Pre-emphasis can be realized by time-domain technology and the frequency-domain technology. In this research, pre-emphasis was realized by adding a high pass filter in frequency domain, which is defined as Formula 1. Here, the value of α is 0.97.

$$S(n) = x(n) - \alpha \cdot x(n-1) \tag{1}$$

4.2.2 Features

Feature Extraction

OpenSMILE [33] (The following content about OpenSMILE are all quoted from this article) was applied to extract acoustic features. OpenSMILE is developed by audeering. It is an open-source toolkit for audio feature extraction and classification of speech and music signals. It also offers a python API which was used in this research. OpenSMILE can perform four kinds of feature extraction operations: Signal Processing, Data Processing, Audio Features Extraction, and Video Features Extraction. Seven standard dataset are supported by opensmile-python currently, including ComParE_2016, GeMAPSv01a, GeMAPSv01b, eGeMAPSv01a, eGeMAPSv01b, eGeMAPSv02 and emobase. For each dataset, two levels of features are available. The first is Low Level Descriptors (LLDs). LLDs refer to some low-level features designed by hand, which are generally calculated on a frame of speech and are used to represent the features of a frame of speech, such as F0, MFCC. The second is functional that maps variable series of LLDs to static values, such as average, mean, and so on. In general, the functional features are the results of functionals calculated on LLDs, and the number of functional features is the result of the number of LLDs times the number of functionals. Another level, LLD delta, is only available for ComParE 2016. Table 4 lists the number of features for each set and level.

For this research, ComParE 2016 feature set with Functionals level was applied and 6373 features were extracted as raw features. In ComParE 2016, there are 65 LLDs, including 4 energy related LLDs, 55 spectral LLDs, and 6 voicing related LLDs [34]. For each LLD, functionals are applied to get the features for this research [34].

Table 4: The datasets supported in opensmile-python and the number of features of each dataset[33][35]

Dataset Name	Number of Features LLDs/LLD delta/Functionals
ComParE_2016	65/65/6373
GeMAPSv01a	18/-/62
GeMAPSv01b	18/-/62
eGeMAPSv01a	23/-/88
eGeMAPSv01b	23/-/88
eGeMAPSv02	25/-/88

Outliers Remove

With raw features, the isolation forest was adopted to remove the outliers. Isolation forest is an anomaly detection method based on an ensemble algorithm [36]. There are two theoretical bases of isolation forests. Firstly, the proportion of abnormal data in the total sample size is very small;

Secondly, the values of outliers are quite different from those of normal points. The outliers finding process of isolation forest is continuously dividing the space containing all data points until a space contains only one data point [37]. This process is realized by isolation trees growing. The anomaly score of each data point is the synthesis of the anomaly calculation results of all trees [38]. Formula 2 is the calculation method of anomaly. In this formula, h(x) is the height of x in each tree, and c(ψ) is the average value of the path length for a given number of samples ψ. 387 clips were removed as outliers.

$$s(s, \psi) = 2^{\frac{E(h(x))}{c(\psi)}} \tag{2}$$

Feature Selection

Too many features may cause over-fitting. Feature selection can prevent this. Besides, it can also reduce the risk of computational cost and improve the performances and training speed of models [39]. Fewer features usually mean better interpretability [40]. Feature selection methods are divided into filtering, embedded, and wrapper. The filtering method uses statistical indicators to score each feature, which focuses on the characteristics of the data itself [41]. Its advantage is the fast calculation and does not depend on the specific model. Its disadvantage is that the selected statistical indicators are not customized for the specific model, so the final accuracy of models with filtering features may not be high [42]. The wrapper method uses models to filter features. By continuously adding or deleting features, the accuracy of the model is tested on the validation set to find the optimal feature subset [43]. Because of the direct participation of models, models with features selected by the wrapping method usually have high accuracy. However, the calculation cost is high, and it may cause over-fitting. The embedded method makes use of characteristics of models and embeds feature selection into the process of model construction. The accuracy of models with embedded features is high, and the computational complexity is between filter and wrapping [44].

For this research, two methods were applied. The first method is Mutual Information (MI). MI is a filtering method base on entropy. The entropy of a random variable is used to measure the uncertainty of the variable[45]. MI measures the degree of interdependence between two variables and evaluates the amount of information contributed by the appearance of one event to the appearance of another. Formula 3 is the definition of entropy, and formula 4 is the definition of MI based on entropy. H(X) is the entropy of X, and p(x) means the probability of x. In formula 4, I(X;Y) means the mutual information between X and Y. Because the feature selection process of MI does not involve models, the number of features can be set. For this research, 200 features were extracted through MI for anxiety level and anxiety label separately.

$$H(X) = -\sum_{x \subseteq R} p(x) log_2 p(x) \tag{3}$$

$$I(X;Y) = H(X) - H(X|Y) \tag{4}$$

The second feature selection method applied is an embedded method with Gradient Boosting Decision Tree (GBDT). GBDT is constructed based on boosting method. GBDT is trained through multiple iterations. In each iteration, GBDT generates a weak classifier, and each classifier is trained based on the residual of the previous one [46]. Finally, the predictions in each iteration are added together as the final prediction. M in formula 5 represents the number of cart trees. $T(x, \theta_i)$ represents the prediction result of the i-th regression tree. θ are parameters in each regression tree. f_M(X) means the importance of X. Generally, CART tree is chosen as the weak classifier. The training process of CART tree can be regarded as a feature selection process in a way. The importance of each feature is acquired by calculating the average value of the importance of the feature in a single tree. For this research, two more feature sets for anxiety level and anxiety label were extracted by GBDT.

$$f_M(X) = \sum_{i=1}^{M} T(x, \theta_i) \tag{5}$$

After feature selection, there are four subsets of features, including *MIC*, *MIR*, *GBDT_cla*, *GBDT_--reg*. *MIC* and *GBDT_cla* are the feature sets for the classification task. *MIR* and *GBDT_reg* are the features sets for the regression task. Each features subset contains 200 features.

4.3 Algorithms

In this section, the models which were trained in this research will be introduced. And the evaluation methods for models will also be illustrated.

4.3.1 Models

Linear Regression and Ridge Regression

Linear regression is one of the simplest and widely used regression models. It is a linear combination of model parameters with one or more regression coefficients. The definitions of linear models are exhibited below, in which θ is the weight for each variable.

$$Suppose\ Function : h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + ... + \theta_n x_n \tag{6}$$

$$Loss\ Function : J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_{(\theta)}(x^{(i)}) - y^{(i)})^2 \tag{7}$$

$$Objective : minJ(\theta_0, \theta_1, ..., \theta_n) \tag{8}$$

The training speed of linear regression models is fast, and its performance is better on data with obvious trends. However, linear regression is likely to over-fit. Therefore, an improved linear regression model, ridge regression model, was also used in this research.

Ridge regression is used to solve the over-fitting problem. It is more practical and reliable to obtain regression coefficient with the cost of losing part of information and reducing precision, and the fitting of bad data is better than linear regression. In ridge regression, a regularization term is added as loss function on the basis of linear regression, which is shown as follows.

$$Loss function of Ridge Regression : J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_{(\theta)}(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2 \tag{9}$$

In this formula, the definition of θ is the same as that of formula 6, and m is the number of samples. λ is called the regularization parameter. If λ is too large, all θ parameters will be minimized, resulting in under-fitting. If λ is too small, the overfitting problem will be solved improperly.

Logistic Regression

Logistic regression is a classification model mainly applied in the binary classification tasks. By nature, the principle of logical regression is to map the result of linear regression, which ranges from -∞ to ∞, to (0,1) by a sigmoid function. The formula 10 is the sigmoid function:

$$g(z) = \frac{1}{1+e^{-z}} \tag{10}$$

According to formula 10, the function of logistic regression is shown in formula 11. θ is the weight of each variable. If h(x) >0.5 and x >0, then y = 1. If h(x) >0.5 and x >0, then y = 0.

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \tag{11}$$

The loss function of logistic regression would be:

$$l(\theta) = logL(\theta) = \sum_{i=1}^{m}(y_i log h_\theta(x_i) + (1-y_i) log(1 - log h_\theta(x_i))) \tag{12}$$

Decision Tree

A decision tree is composed of nodes and directed edges. Generally, a decision tree contains a root node, several internal nodes, and several leaf nodes. Training a decision tree usually has three steps which are shown in Figure 2 [47]. And the decision making process of trees is demonstrated in Figure 3 [48].

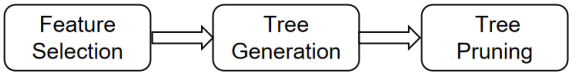


Figure 2: Training process of Decision Tree

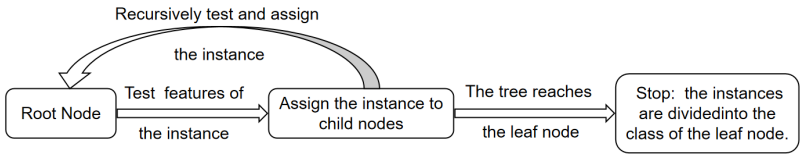


Figure 3: Decision making process of Decision Tree

There are three classical trees, namely ID3 [49], C4.5 [50], CART [51]. The core of the ID3 algorithm is to select the best feature of the current data set according to the principle of "maximum information entropy gain". In the ID3 algorithm, the feature with the largest entropy reduction is selected to divide the data. Formula 13 is the definition of information entropy, in which p_i means the proportion of i in the set. Formula 14 is the definition of information entropy gain. D represents the sets. E(D) is the information entropy of the original set, and the second part of the equation is the sum of information entropy of all subsets after dividing the original set into multiple subsets [49].

$$E(D) = -\sum_{i=1}^{n} p_i log_2 p_i \tag{13}$$

$$Gain(D, a) = E(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} E(D^v) \tag{14}$$

The process of the C4.5 algorithm is similar to ID3, but the information gain is changed as the information gain ratio to solve the problem of bias value more attributes. In addition, C4.5 can deal with continuous attributes. In formula 15 and formula 16, IV(a) is information entropy for of a [50].

$$Gain\_ratio(D, a) = \frac{Gain(D, a)}{IV(a)} \tag{15}$$

$$IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} log_2 \frac{|D^v|}{|D|} \tag{16}$$

CART tree is a binary tree, which is suitable for both regression and classification. It uses the Gini coefficient to replace the information gain ratio [52]. The Gini coefficient represents the impurity of models. The smaller the Gini coefficient is, the better the feature is. Gini is the opposite of information gain. Every iteration in CART trees aims to reduce the value of Gini index. The definition of Gini index of set D is illustrated in formula 17. K is one of the classes in D and C_k is the quantity of k [52].

$$Gini(D) = 1 - \sum_{k=1}^{K} (\frac{|C_k|}{|D|})^2 \tag{17}$$

Random Forest

Random Forest is an ensemble of decision trees and belongs to bagging algorithms. Its output is determined by the mode of the outputs of individual trees in it. There is no correlation between trees. Figure 4 is the training process of random forest.

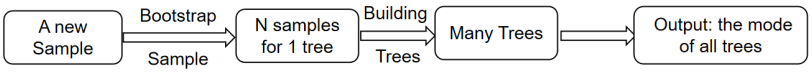


Figure 4: Training process of Random Forest

The performance of random forests is related to two factors. The first factor is the correlation between any two trees in the forest: the stronger the correlation, the higher the error rate. The second factor is the classification ability of each tree in the forest: the stronger the classification ability of each tree, the lower the error rate of the whole forest [53]. The random forests can deal with both discrete values and continuous values. Besides, it can also be used for unsupervised learning clustering [54] and outliers detection [55].

Gradient Boosting

Gradient boosting (GB) is a class of algorithms belonging to boosting methods. The basic principle of gradient boosting is to train new weak classifiers according to the negative gradient information of the loss function of the current model and then combine the trained weak classifiers into the existing model in the form of accumulation [56]. Gradient boosting is a machine learning algorithm for regression and classification problems [57]. It integrates weak learning models, typically decision trees, to produce a strong prediction model. In this research, GBDT was trained. The detailed information of GBDT has been introduced in Feature Selection section.

eXtreme Gradient Boosting (XGBoost)

XGBoost (XGB) is a machine learning algorithm implemented under the framework of gradient boosting. Therefore, the learning process of XGBoost is similar to GB. XGBoost has some improvements on the basis of GB. XGBoost adds a regularization term to the loss function to control the complexity of the model [58], which leads to better performance of XGBoost. Besides, the training of XGBoost is much faster due to the characteristics, including parallelization, distributed computing, and out-of-core computing, of XGBoost [58]. Formula 19 is the target function of XGB. There are two parts to this function. The first part is the training loss, and the second part is the complexity of the trees.

$$The\ target\ function\ of\ GBDT: L^{(t)} = \sum_{i=1}^{n} l(y_i, \widehat{y_i}^{t-1} + f_t(x_i)) + \Omega(f_t) \tag{18}$$

4.4 Unbalanced label

After data pre-processing, the ratio of clips with label 1 (anxiety) and clips with label 0 (no anxiety) is 1:3.6. With unbalanced data, models would tend to predict the majority, leading to high false accuracy. There are several methods to solve the problem of unbalanced data, such as over-sampling, under-sampling, threshold moving, and so on. Over-sampling repeats sampling data with a small proportion

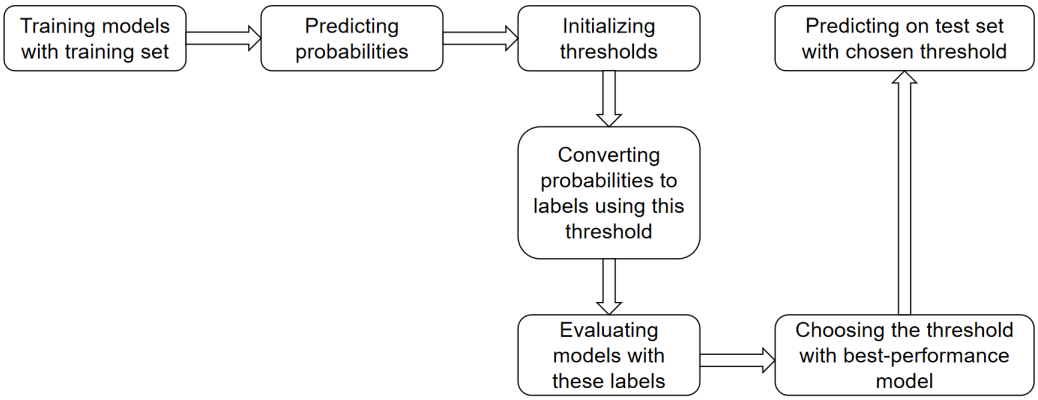


Figure 5: The procedure of threshold moving

to make their quantity be equal to that of data with a large proportion [59]. Under-sampling takes samples from data with a large proportion to make their quantity be equal to that of data with a small proportion [59]. Both of them operate on data, which leads to their shortcomings. Over-sampling may cause over-fitting on small-proportion data, as these data are simply repeated. Under-sampling may cause poor performances of models, as part of the data are removed. The threshold moving method moves the decision threshold of models. Normally, the threshold is 0.5, which means, when the probability of one case is higher than 0.5, then it will be mapped to one class [60]. The threshold moving method changes the value of decision threshold to reduce the impact of the unbalanced data on models when classifying. In this research, an improved under-sampling method and threshold moving method were applied.

4.4.1 Threshold-Moving

Threshold-moving is simpler than over-sampling and under-sampling, as it does not modify the data. Therefore, it will not lead to the problems caused by modifying data. Basically, there are four steps of threshold moving [60], which are shown in Figure 5.

4.4.2 Under-Sampling

To realize under-sampling, Easy Ensemble developed by Liu et al. was applied [61]. This method is a combination of bagging and under-sampling. The samples of majority class are randomly divided into n subsets. After division, the number of samples in each subset is equal to that of minority class samples. Then each subset of the majority class is combined with the minority class to train a model. Finally, n models are integrated, so that although the samples of each subset are less than the total samples, the total amount of information is not reduced after integration.

4.5 Evaluation Metrics

4.5.1 Classification Metrics

As the classification models in this research are binary models, three metrics were used. The first is TPR (True Positive Rate, or sensitivity) and TNR (Ture Negative Rate, or Specificity), which are shown in formulas 19 and 20. TP, TN, FP, FN means true positive, true negative, false positive, and false negative separately. TPR represents the ratio of correctly predicted positives to real positives. TFR represents the ratio of correctly predicted negatives to real negatives. Therefore, the values of TPR reflect the ability of the model to identify positives. Accordingly, TNR reflects the ability of the model to identify negatives [62]. In this research, the models should distinguish whether children are under anxiety as correctly as possible. Therefore, the higher the values of TPR and TFR, the better the models.

$$TPR = \frac{TP}{TP + FN} \tag{19}$$

$$TNR = \frac{TN}{TN + FP} \tag{20}$$

The second is F1 score. F1 score is an index used to measure the performance of classification models in statistics. It takes the accuracy and recall of the classification model into account. The F1 score can be regarded as a harmonic average of model accuracy and recall, with the maximum value of 1 and the minimum value of 0. The average value of precision and recall of all categories is calculated first, and then the F1 value is calculated according to precision and recall. F1 reflects the general performance of models directly. The following formula is the definition of recall, precision, and F1 [62].

$$Recall = \frac{TP}{TP + FN} \tag{21}$$

$$Precision = \frac{TP}{TP + FN} \tag{22}$$

$$F_1 = 2\frac{Recall \times Precision}{Recall + Precision} \tag{23}$$

The third evaluation metric is ROC and AUC. ROC (Receiver Operating Characteristic) curve is drawn according to a series of different binary classification methods, with true positive rate as ordinate and false positive rate as abscissa. AUC is defined as the area under the ROC curve. AUC is an effective metric for measuring the ability of a binary classifier to discriminate between positive and negative classes [63]. Models with larger AUC are better. The shapes of ROC curves and the values of AUC are evaluation metrics.

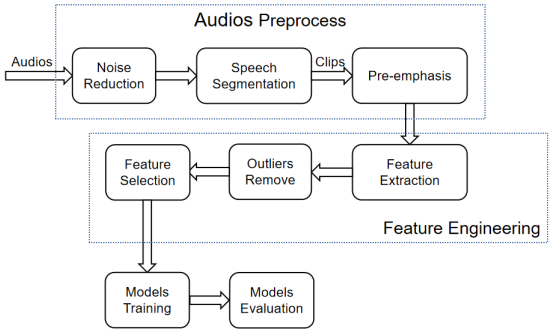


Figure 6: The procedure of audio preprocess, feature engineering, models training and models evaluation.

4.5.2 Regression Metrics

There are three evaluation metrics for regressors. The first one is Root Mean Squared Error (RMSE). It measures the average deviation, which is the square root of MSE (Mean Squared Error). MSE is the mean square sum of the difference between the real value and the predicted value. Formula 24 is the definition of RMSE [64].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{24}$$

The second metric is the Mean Absolute Error (MAE). It is the mean of absolute deviation, which is the more common form of deviation mean. It solves the problem of zero deviation caused by positive data and negative data [64].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{25}$$

The third metric is the Coefficient of determination (R^2 score). It reflects the proportion explained by the estimated regression equation in the variation of dependent variable y [64]. The closer R^2 is to 1, the greater the proportion of the sum of squares of regression to the total sum of squares, the closer the regression line to each observation point, the more part of the variation of y value explained by the change of X, and the better the fitting degree of regression.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y^2)^2} \tag{26}$$

Figure 6 demonstrates the whole procedure of audio preprocess, feature engineering, and models training.

5 Results

As mentioned in the last section, both regressors and classifiers were trained. To find the best parameters of models, grid search with 10-folds cross-validation was utilized [65]. In this section, the performances of models will be evaluated and compared.

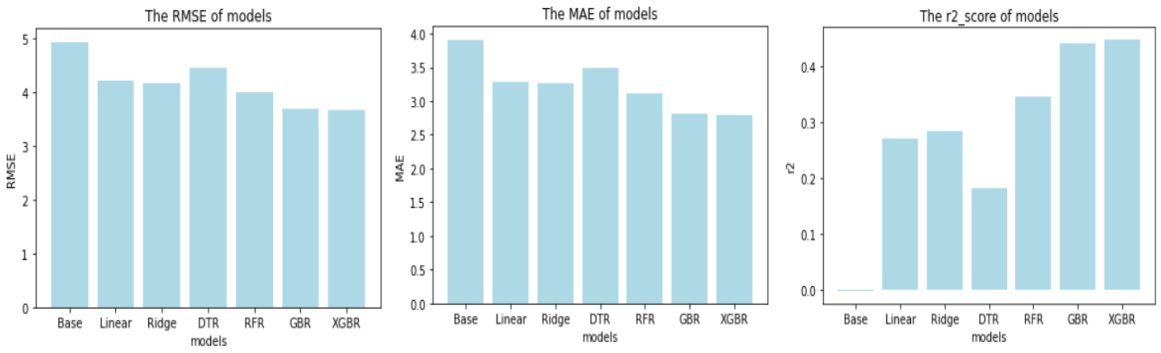


Figure 7: The performance of regressors with GBDT_reg feature set

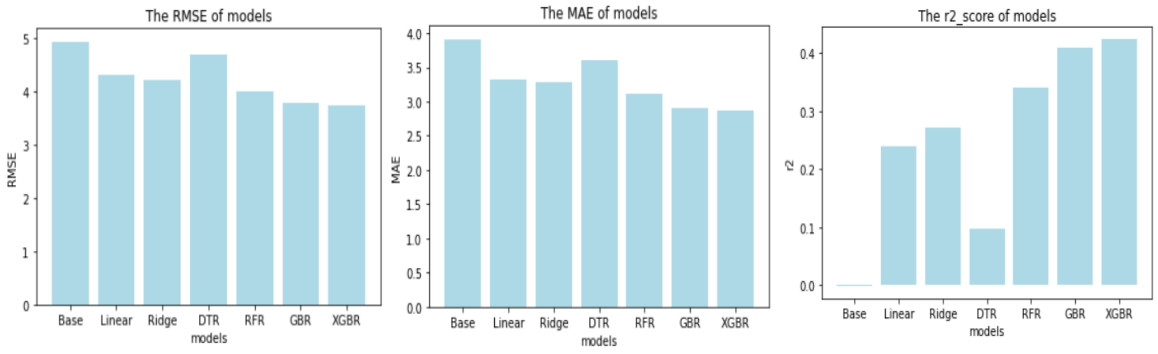


Figure 8: The performance of regressors with MIR feature set

5.1 Regression Models

The aim of regression tasks is to predict the levels of state anxiety based on acoustic features. Linear models with or without the regularization term (Linear or Ridge), Regression Tree (DTR), Random Forest (RFR), Gradient boosting (GBR), and eXtreme Gradient Boosting (XGBR) were trained with feature sets GBDT_reg and MIR separately. Besides, a baseline model which simply predicts anxiety levels as the mean of all anxiety levels was used as a reference. Figure 7 and figure 8 exhibit the performances of different regressors with the GBDT_reg or the MIR feature set. The performances of models on the two feature sets show the same trend. The models with the GBDT_reg feature set perform better than models with the MIC feature set. This is because that GBDT model was involved in the process of extracting the GBDT_reg feature set and most models used in this research are tree-based models. Therefore, the features in the GBDT_reg feature set are more suitable for models training. From the figures, XGBR with the GBDT_reg feature set acquired the best performance.

The RMSE of XGBR with GBDT_reg feature set is 3.5 and MAE is 2.7 approximately, which are the lowest. The R^2 score of thismodel is about 0.5 (50%), which is the highest. The performance of XBGR is similar to the performance of GBR. The decision tree performs worst. The models show reasonable performances compatible with their properties.

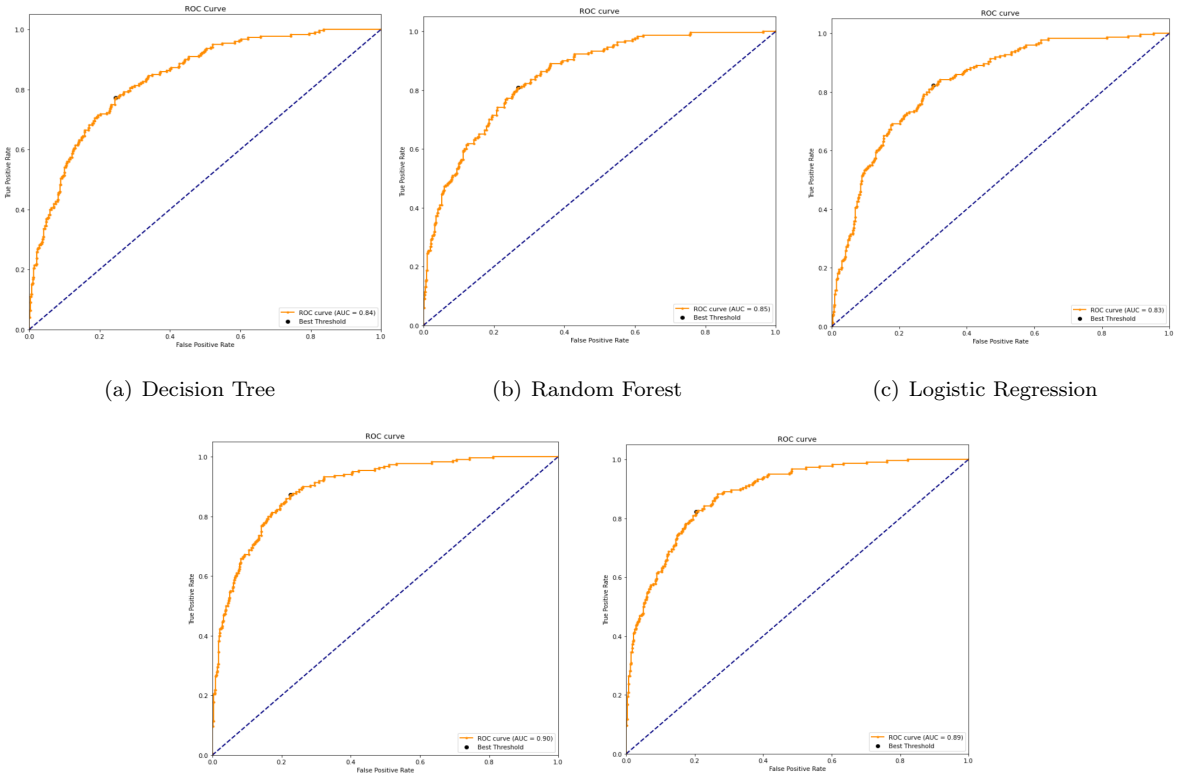
Table 5: The results of models with EasyEnsemble and Threshold Moving

Metrics	Bseline	Models with GBDT_cla					Models with MIC				
		DT	RF	LO	GBDT	XGB	DT	RF	LO	GBDT	XGB
TPR	0.5	0.77	0.81	0.82	0.87	0.82	0.64	0.65	0.68	0.83	0.80
TNR	0.5	0.75	0.73	0.70	0.77	0.80	0.81	0.82	0.81	0.71	0.74
F1	0.5	0.70	0.70	0.68	0.75	0.75	0.70	0.71	0.71	0.69	0.70
AUC	0.5	0.84	0.85	0.83	0.90	0.89	0.80	0.81	0.81	0.85	0.85

5.2 Classification Models

The aim of classifiers is to identify as accurately as possible whether a child is in a state of anxiety through acoustic features. As mentioned in the last section, to cope with the problem of the unbalanced labels in classification tasks, Threshold Moving and Easy Ensemble were used simultaneously. Decision Tree (DT), Random Forest (RF), Logistic Regression (LO), Gradient Boosting (GBDT), and eXtreme Gradient Boosting (XGB) were trained with the GBDT_cla and the MIC separately. Table 5 illustrates the performances of different classifiers with evaluation metrics. As the problem of data imbalance has been solved, the results of evaluation metrics can be compared directly. The baseline model simply predicts all labels as 0. Without the problem of unbalanced data, all values of TPR, TNR, F1, and AUC of the baseline model should be equal to 0.5. According to the results, GBDT with the GBDT_cla feature set gained the highest TPR (0.87), F1 score (0.75), AUC (0.90), and relatively high TNR (0.77). The original intention of the algorithm is to identify children's anxiety timely and accurately. Therefore, TNR is less concerned. The performances of partial models, including GBDT and XGB with the GBDT_cla feature set, DT, RF, and LO with MIC feature set, on predicting children's non-anxiety state are similar and better than the rest models. GBDT and XGB with the GBDT_cla feature set gain the highest values of F1 score and AUC. Figure 9 and Figure 10 demonstrate the shapes of ROC curves and the best thresholds after threshold moving of all models. The shapes are similar, while the ROC curve of GBDT with GBDT_cla is closer to the top-left of the graph. This means the performance of this classifier is better than that of other models based on ROC curves. With these results, GBDT with the GBDT_cla feature set gained the best performance. The performance of XGBT is similar to that of GBDT. On the one hand, GBDT and XGBT are effective and mature algorithms. On the other hand, the GBDT_cla feature set was selected with the GBDT model, which is more suitable for training GBDT and XGB models.

Combined the results of regression models and classification models, boosting algorithms have achieved relatively good performances. This provides baselines for future algorithms on BSDCSA.



(a) Decision Tree (b) Random Forest (c) Logistic Regression

(d) GBDT (e) XGB

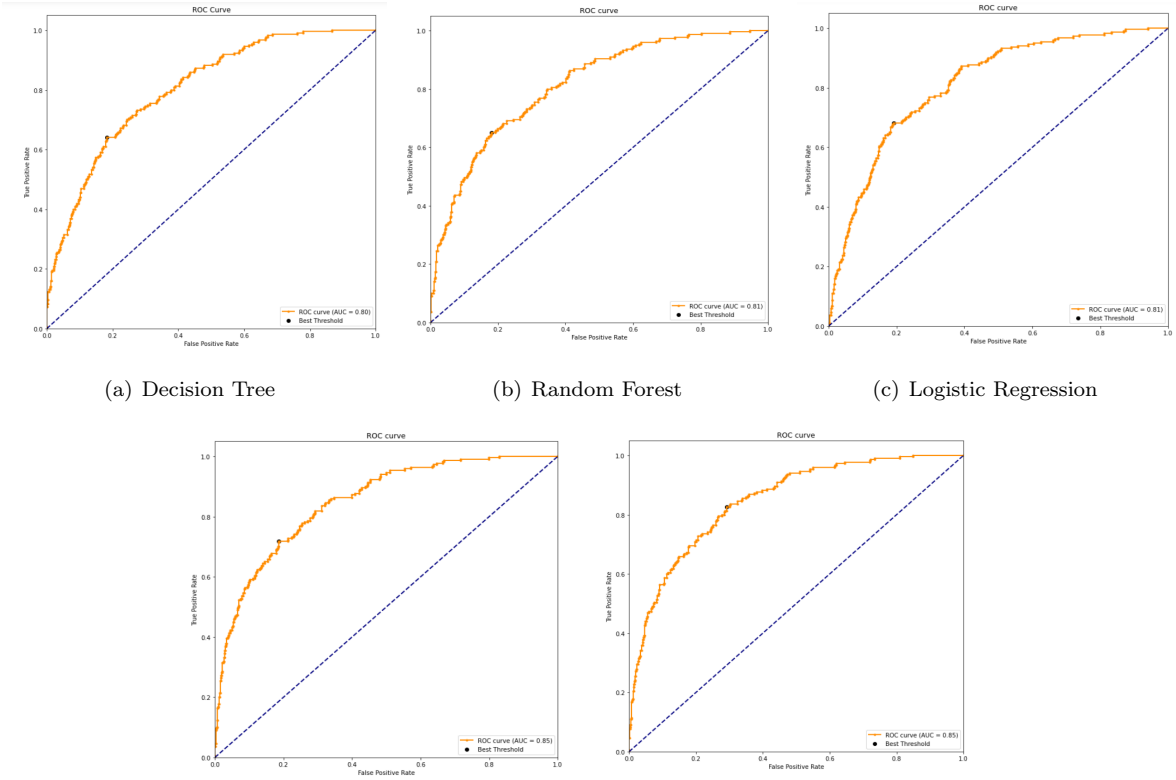
Figure 9: The ROC curves of models with GBDT_cla feature set

6 Discussion

6.1 The Effectiveness of Anxiety Labels

In BSDCSA, anxiety labels were not directly marked regarding whether children were under anxiety. Instead, they were labeled based on positive emotions and negative emotions. This is because, in the process of collecting data, we did not directly ask children whether they were anxious, but asked them about their main emotions. There are two reasons. The first reason is that the data was collected from children aged 10 to 12. Children of this age group may not have a clear understanding of anxiety. Directly asking whether they are anxious may bring incomprehension and ambiguity to children. The second reason is that we deliberately avoid asking questions with psychological implications, so as not to affect the experimental data. However, a problem caused is it is unclear whether anxiety can be represented by negative emotions. According to the performances of children in experiments, it is assumed that the anxiety labels are effective.

To verify the effectiveness of anxiety labels, the relationship between anxiety labels and anxiety levels was analyzed. If the anxiety levels corresponding to the two groups of labels are significantly different, it can be proved that the anxiety labels induced from negative emotions are effective. Table 6 summarizes the statistical features in two groups. Wilcoxon test is applied to verify the significance of



(a) Decision Tree (b) Random Forest (c) Logistic Regression

(d) GBDT (e) XGB

Figure 10: The ROC curves of models with MIC feature set

Table 6: The statistics of anxiety levels in two groups

Label	SA_min	SA_5%_qu	SA_median	SA_mean	SA_95%_qu	SA_max
0	10	11	17	16.94	22.8	29
1	15	16	24	24.18	32	33

the difference. According to the results in table 7, the anxiety levels in the two groups are significantly different, and the anxiety levels of no anxiety labels group are significantly lower than the anxiety levels in anxiety labeled group. Therefore, the anxiety labels induced from negative and positive emotions can represent the anxious state of children.

As there is a significant difference in anxiety levels (SA) between the two label groups, it is possible that SA can help improve the performance of models. To explore this, SA was added as a feature for anxiety label classification. Based on the results in table 8 and the results in table 8, the performances of models with SA as a feature are much better than performances of models without SA as a feature. The TPR, TNR, and F1 score of GBDT with added SA than original GBDT increase 11% 12%, the AUC increases 6%.

Table 7: The results of Wilcoxon test

Hypothesis	W	p_value
There is difference of anxiety levels between two groups	227698	2.2e-16
Anxiety levels of label 0 is lower than levels of label 1	227698	2.2e-16

Table 8: The results of models with SA as an additional feature

Metrics	Models with GBDT+SA					Models with MIC+SA				
	DT	RF	LO	GBDT	XGB	DT	RF	LO	GBDT	XGB
TPR	0.84	0.84	0.86	0.93	0.93	0.86	0.77	0.87	0.87	0.92
TNR	0.88	0.86	0.85	0.89	0.88	0.85	0.87	0.84	0.91	0.86
F1	0.82	0.81	0.80	0.87	0.86	0.80	0.80	0.80	0.86	0.84
AUC	0.92	0.93	0.92	0.96	0.96	0.92	0.90	0.92	0.95	0.96

6.2 Research Questions' Analysis

6.2.1 Research Question 1

The degree of relevance between features and children's state anxiety is evaluated through the importance of features during models training. And the importance of features is calculated through SHAP [66]. SHAP is a Python package used to explain the machine learning models. SHAP displays how each feature affects the prediction results. For each feature, the higher its SHAP value, the greater its contribution to the model. If the SHAP value of a feature is higher than 0, the effect of this feature on the model is positive. If the SHAP value of a feature is lower than 0, the effect of this feature on the model is negative.

This research question will explore the relevance between acoustic features and anxiety labels or anxiety levels separately. For this question, only features from best-performing models, including GBR and XBGR with GBDT_reg feature set and GBDT and XGB with GBDT_cla feature set, are taken into consideration. The SHAP values of all features used to train the models are calculated. The features used in this research are functional features in ComParE_2016 extracted by OpenSMILE, which are calculated from acoustic features (LLDs in ComParE_2016). This reseach question focuses on the relevance between acoustic features and anxiety instead of the relevance between functional features and anxiety. Therefore, to figure out the importance of acoustic features (LLDs in ComParE_2016), the absolute SHAP values of functional features that correspond to one acoustic feature (LLD) are added together as the importance of this acoustic feature for each model. Then 59 acoustic features from classifiers (GBDT and XGBT) and 56 acoustic features from regressors (GBR and XGBR) are ranked according to the calculated SHAP values. Higher SHAP values mean higher rankings. After this, four types of rankings are acquired from GBR, XGBR, GBDT, and XGBT respectively. For each acoustic feature, the average ranking from rankings of GBR and XGBR (or GBDT and XGBT) is its final ranking, indicating the importance of this feature to anxiety level (or anxiety label). The higher

the ranking, the more important the feature is to the model and the more relevant it is to anxiety. Figure 11 shows the whole procedure. Table 9 lists the top 20 acoustic features that are most relevant with anxiety labels and the top 20 acoustic features that are most relevant with anxiety levels. MFCCs, ZCR (Zero Cross Rate), audSpec_Rfilt (RASTA filt. aud. spect), F0, and spectral roll-off are related to both anxiety labels and anxiety levels. Spectral entropy is only related to anxiety labels, and the L1 norm is only relevant to anxiety levels. Combined with Turgut's research [24], it is concluded that MFCC4, MFCC5, MFCC7, MFCC9, MFCC11, and F0 are relevant to both children's and adults' anxiety. As there is little research on the relationship between acoustic features and children's state anxiety, the results of this research question also provide a reference for feature selection in this field. The features with higher importance should be considered first when training models in this field.

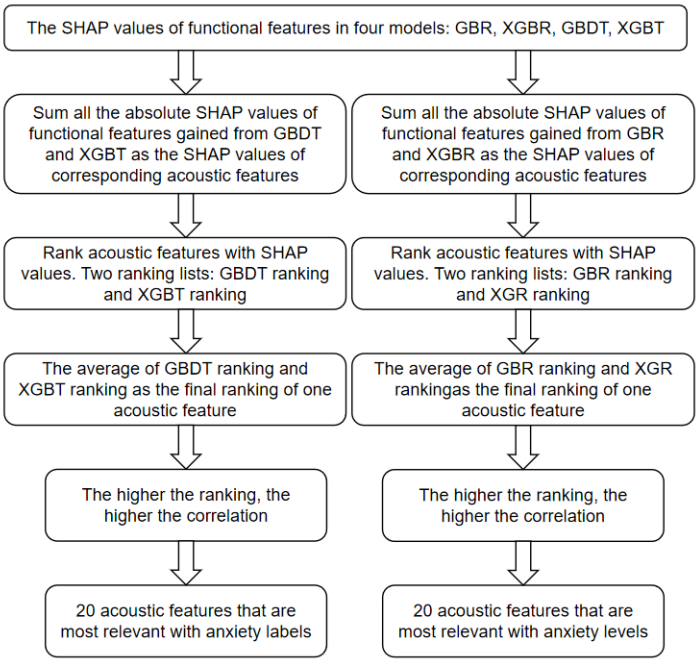


Figure 11: The calculation procedure of acoustic features' relevance to anxiety

Table 9: The top 20 features most relevant to anxiety

Features most relevant to anxiety labels	Features most relevant to anxiety levels
MFCC_[1]	MFCC_[1]
audSpec_Rfilt_[2]	audSpec_Rfilt_[1]
MFCC_[14]	audSpec_Rfilt_[4]
MFCC_[11]	F0final
audSpec_Rfilt_[4]	ZCR
MFCC_[6]	MFCC_[11]
MFCC_[10]	MFCC_[14]

Table 9 continued from previous page

Features most relevant to anxiety labels	Features most relevant to anxiety levels
audSpec_Rfilt_[3]	audspec_lengthL1norm
audSpec_Rfilt_[22]	MFCC_[6]
ZCR	spectralRollOff75.0
MFCC_[8]	spectralRollOff90.0
MFCC_[9]	MFCC_[5]
audSpec_Rfilt_[15]	MFCC_[10]
spectralEntropy	audSpec_Rfilt_[13]
audSpec_Rfilt_[21]	audSpec_Rfilt_[7]
MFCC_[5]	audSpec_Rfilt_[14]
spectralRollOff25.0	audSpec_Rfilt_[2]
F0final	audSpec_Rfilt_[0]
MFCC_[7]	spectralRollOff25.0
spectralCentroid	MFCC_[3]

6.2.2 Research Question 2

Spearman correlation analysis is applied to explore the trend of common acoustic parameters with the increase of state anxiety level. The common acoustic features are not only features involved in model training, but also features that are widely studied in previous research. To compare the results of this research question with previous research, the mean and standard deviation of F0, jitter, shimmer, HNR, and ZCR are calculated. Besides, the range and standard deviation of MFCC1-12 and RASTA (Relative Spectrum [67]) filter audSpec (auditory spectrum) 1-5, audSpec10-13, audSpec19, and audSpec21-22, which are part of the most relevant features, are also calculated. Table 10 summarizes the degree and direction of changes of these acoustic features with the increase of anxiety level. Compared with previous research, F0, jitter, and shimmer decreases with the increase of anxiety level, rather than the increase shown in previous studies (The previous studies in this aspect are summarized in Related Work section). The values of MFCC1, MFCC4, MFCC5, MFCC6, and MFCC7 increase with the increase of anxiety level, which shows opposite trends to Turgut's research [24]. The results are contrary to the previous research. The possible reason is that the acoustic features of children are quite different from that of adults. Most of RASTA filter auditory spectrum increase with the increase of anxiety level, which are rarely studied.

Table 10: The trends of acoustic parameters with the increase of state anxiety level. +++ : high increase, ++ : medium increase, + : low increase; — : high decrease, – : medium decrease, - : low decrease; ×

Features	Trends	Features	Trends
F0final_mean	- - -	mfcc_[12]_range	×
F0final_stddev	- - -	logHNR_mean	- - -
F0final_range	-	logHNR_stddev	+++
jitterLocal_mean	- - -	pcm_zcr_stddev	+++
jitterLocal_stddev	- - -	pcm_zcr_mean	×
shimmerLocal_mean	- - -	pcm_RMSenergy_range	- - -
shimmerLocal_stddev	×	audSpec_Rfilt_[1]_range	+++
MFCC_[1]_stddev	+++	audSpec_Rfilt_[1]_stddev	+++
MFCC_[1]_range	×	audSpec_Rfilt_[2]_range	+++
MFCC_[2]_stddev	×	audSpec_Rfilt_[2]_stddev	+++
MFCC_[2]_range	×	audSpec_Rfilt_[3]_range	+++
MFCC_[3]_stddev	+	audSpec_Rfilt_[3]_stddev	+++
MFCC_[3]_range	×	audSpec_Rfilt_[4]_range	+++
MFCC_[4]_stddev	+++	audSpec_Rfilt_[4]_stddev	+++
MFCC_[4]_range	+++	audSpec_Rfilt_[5]_range	+++
MFCC_[5]_stddev	++	audSpec_Rfilt_[5]_stddev	+++
MFCC_[5]_range	×	audSpec_Rfilt_[10]_range	+++
MFCC_[6]_stddev	+++	audSpec_Rfilt_[10]_stddev	+++
MFCC_[6]_range	+	audSpec_Rfilt_[11]_range	++
MFCC_[7]_stddev	+++	audSpec_Rfilt_[11]_stddev	++
MFCC_[7]_range	+++	audSpec_Rfilt_[12]_range	×
MFCC_[8]_stddev	×	audSpec_Rfilt_[12]_stddev	×
MFCC_[8]_range	×	audSpec_Rfilt_[13]_range	×
MFCC_[9]_stddev	×	audSpec_Rfilt_[13]_stddev	×
MFCC_[9]_range	×	audSpec_Rfilt_[19]_range	×
MFCC_[10]_stddev	-	audSpec_Rfilt_[19]_stddev	×
MFCC_[10]_range	×	audSpec_Rfilt_[21]_range	×
MFCC_[11]_stddev	×	audSpec_Rfilt_[21]_stddev	×
MFCC_[11]_range	×	audSpec_Rfilt_[22]_range	×
MFCC_[12]_stddev	×	audSpec_Rfilt_[22]_stddev	×

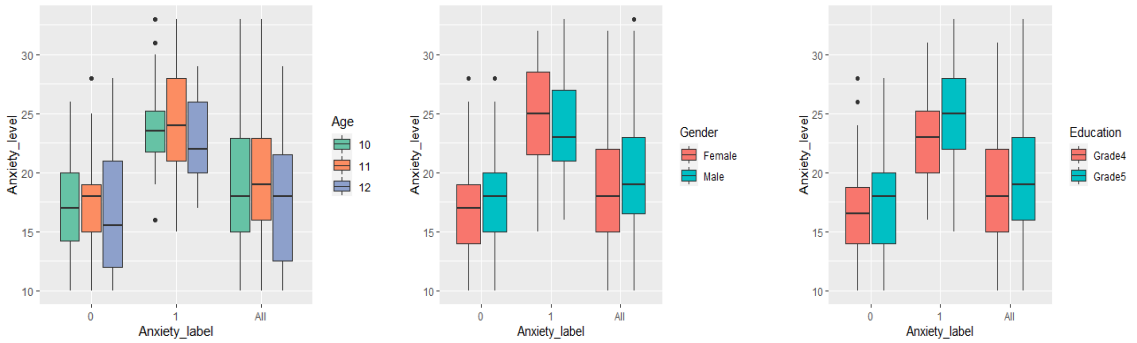


Figure 12: The boxplots of anxiety level in different groups

Table 11: The results of Spearman correlation analysis between acoustic parameters and anxiety level anxiety label. * p<0.05 ** p<0.01

	Anxiety_level	Anxiety_label
Age	-0.014	0.005
Gender	0.129*	0.159*
Education level	0.058	-0.069

6.2.3 Research Question 3

This research question will be answered in two aspects. Firstly, the influence of sociodemographic characteristics on anxiety will be evaluated through boxplots and Spearman correlation analysis. Secondly, the impact of sociodemographic characteristics on acoustic parameters will be assessed by Pearson correlation analysis and Linear regression analysis will be applied to explore how acoustic parameters are influenced. Figure 12 demonstrates the boxplots of anxiety levels. Tables 11 exhibits the results of Spearman correlation analysis for the influence of sociodemographic characteristics on anxiety levels and anxiety labels. Combining figure 12 and table 11, the impact of age, gender, and education level on anxiety are summarized as follows. Most of the participants aged 11 show higher anxiety levels than participants aged 10 and 12. However, age is not significantly related to anxiety. Females show higher anxiety levels in an anxious state, while males show higher anxiety levels in an un-anxious state and in general. Gender is significantly related to anxiety. In this analysis, females were labeled as 0 and males were labeled as 1. The results of table 11 mean that the anxiety levels of males are higher than that of females on the whole, which could also be proved by the boxplots, and males are more likely under anxiety state, which could be proved by table 12. A higher education level means a higher anxiety level for most participants. However, education level is not significantly related to anxiety. In a word, age and education level have little effect on anxiety, while gender is significantly related to anxiety.

To analyze the impact of sociodemographic characteristics on acoustic parameters, linear regression analysis. Therefore, pearson correlation analysis is applied firstly to find the acoustic parameters that are most relevant to sociodemographic characteristics. Table 13 analyzes the relationship between

Table 12: The number of audios corresponding to different genders and anxiety states

	Female	Male
Anxious	27	45
Un-anxious	101	82

common acoustic parameters and sociodemographic characteristics with Spearman correlation analysis. According to the results, MFCC[2]_range, MFCC_[6]_stddev, MFCC_[8]_stddev, MFCC_[8]_range, and MFCC_[10]_range are unrelated to age, gender and education level. There are 31 acoustic parameters significantly related to all three sociodemographic characteristics. RMSenergy_range is only related to age. F0_mean, MFCC_[6]_range, and zcr_mean are only related to Gender. MFCC_[7]_stddev, MFCC_[7]_range, MFCC_[10]_stddev logHNR_staddev, and zcr_stddev are only related to education level. The remaining acoustic parameters are related to two of the three sociodemographic characteristics. To further explore the level of sociodemographic characteristics' influence on acoustic features, linear regression analysis was applied.

Table 13: The results of Pearson correlation analysis beween acoutic features and sociodemographic characteristics. * $p<0.05$ ** $p<0.01$

	Age	Gender	Education Level
F0_mean	0.023	-0.159**	-0.022
F0_stddev	0.045**	-0.106**	-0.033
F0_range	-0.014	-0.079**	-0.071**
jitter_mean	0.106**	-0.162**	0.037*
jitter_stddev	0.036*	-0.094**	-0.010
shimmer_mean	0.087**	-0.082**	0.050**
shimmer_stddev	-0.012	0.012	-0.050**
MFCC_[1]_stddev	-0.075**	-0.014	-0.118**
MFCC_[1]_range	-0.055**	0.016	-0.104**
MFCC_[2]_stddev	-0.039*	0.003	-0.037*
MFCC_[2]_range	-0.025	-0.028	0.000
MFCC_[3]_stddev	0.034*	0.060**	-0.035*
MFCC_[3]_range	0.039*	0.062**	-0.028
MFCC_[4]_stddev	-0.034*	0.086**	-0.085**
MFCC_[4]_range	-0.030	0.066**	-0.086**
MFCC_[5]_stddev	0.043*	-0.032	-0.105**
MFCC_[5]_range	0.057**	-0.021	-0.083**
MFCC_[6]_stddev	0.005	0.021	0.022

Table 13 continued from previous page

	Age	Gender	Education Level
MFCC_[6]_range	0.015	0.046**	0.013
MFCC_[7]_stddev	-0.029	-0.014	-0.058**
MFCC_[7]_range	-0.026	0.022	-0.041*
MFCC_[8]_stddev	-0.010	-0.022	-0.023
MFCC_[8]_range	-0.004	-0.009	-0.010
MFCC_[9]_stddev	-0.079**	-0.047**	-0.102**
MFCC_[9]_range	-0.061**	-0.049**	-0.073**
MFCC_[10]_stddev	-0.027	-0.002	-0.048**
MFCC_[10]_range	-0.031	-0.010	-0.024
MFCC_[11]_stddev	-0.104**	0.059**	-0.060**
MFCC_[11]_range	-0.082**	0.071**	-0.038*
MFCC_[12]_stddev	-0.087**	0.064**	-0.005
MFCC_[12]_range	-0.073**	0.075**	0.026
logHNR_mean	-0.106**	0.023	-0.074**
logHNR_stddev	-0.029	0.009	-0.052**
zcr_stddev	-0.010	-0.007	0.035*
zcr_mean	-0.001	-0.045**	0.020
RMSenergy_range	-0.040*	-0.011	-0.026
audSpec_Rfilt_[1]_range	-0.070**	0.074**	-0.153**
audSpec_Rfilt_[1]_stddev	-0.090**	0.084**	-0.179**
audSpec_Rfilt_[2]_range	-0.075**	0.032	-0.128**
audSpec_Rfilt_[2]_stddev	-0.090**	0.040*	-0.157**
audSpec_Rfilt_[3]_range	-0.073**	0.049**	-0.065**
audSpec_Rfilt_[3]_stddev	-0.090**	0.063**	-0.083**
audSpec_Rfilt_[4]_range	-0.048**	0.093**	-0.125**
audSpec_Rfilt_[4]_stddev	-0.054**	0.105**	-0.141**
audSpec_Rfilt_[5]_range	-0.050**	0.081**	-0.102**
audSpec_Rfilt_[5]_stddev	-0.054**	0.096**	-0.112**
audSpec_Rfilt_[10]_range	-0.077**	0.146**	-0.071**
audSpec_Rfilt_[10]_stddev	-0.101**	0.166**	-0.094**
audSpec_Rfilt_[11]_range	-0.075**	0.137**	-0.061**
audSpec_Rfilt_[11]_stddev	-0.102**	0.152**	-0.081**
audSpec_Rfilt_[12]_range	-0.059**	0.078**	-0.083**
audSpec_Rfilt_[12]_stddev	-0.086**	0.086**	-0.099**
audSpec_Rfilt_[13]_range	-0.068**	0.050**	-0.123**

Table 13 continued from previous page

	Age	Gender	Education Level
audSpec_Rfilt_[13]_stddev	-0.093**	0.058**	-0.141**
audSpec_Rfilt_[19]_range	-0.080**	0.103**	-0.097**
audSpec_Rfilt_[19]_stddev	-0.100**	0.120**	-0.107**
audSpec_Rfilt_[21]_range	-0.071**	0.114**	-0.100**
audSpec_Rfilt_[21]_stddev	-0.103**	0.127**	-0.099**
audSpec_Rfilt_[22]_range	-0.051**	0.082**	-0.038*
audSpec_Rfilt_[22]_stddev	-0.075**	0.094**	-0.034*

Table 14 lists the results of linear regression analysis between sociodemographic characteristics and sociodemographic-characteristics-correlated acoustic parameters. In this table, B is the regression coefficient. It represents the change degree of model output when the feature changes by one unit. For example, if age increases 1, the value of F0_stddev will increase 2.473. When B is less than 0, the correlations between sociodemographic characteristics and sociodemographic-characteristics-correlated acoustic parameters are negative. When B is greater than 0, this kind of correlation is positive. The positive and negative of correlations demonstrated in table 14 are compile with table 13. In regression analysis, t and p indicate the significance of the influence of an independent variable on dependent variables. Generally speaking, The greater the absolute value of T, the smaller the value of P, and the higher the significance. If the value of p is smaller than 0.05, it could be proved that the result is significant. In table 14, 0.000 for p means the values of p are smaller than 0.001. The acoustic parameters analyzed in table 14 are those that have been proved to be significantly related to the sociodemographic characteristics. Therefore, most of the p values in table 14 are small.

Table 14: The results of linear regression analysis between sociodemographic characteristics and sociodemographic-characteristics-correlated acoustic parameters.

Acoustic parameters	Sociodemographic characteristics	B	t	p
F0_mean	Gender	-15.255	-9.499	0.000
F0_stddev	Age	2.473	2.682	0.007
	Gender	-7.516	-6.298	0.000
F0_range	Gender	-18.6	-4.644	0.000
	Education level	-18.047	-4.198	0.000
jitter_mean	Age	0.004	6.280	0.000
	Gender	-0.008	-9.690	0.000
	Education level	0.002	2.171	0.03

Table 14 continued from previous page

Acoustic parameters	Sociodemographic characteristics	B	t	p
jitter_stddev	Age	0.001	2.144	0.032
	Gender	-0.004	-5.537	0.000
shimmer_mean	Age	0.005	5.126	0.000
	Gender	-0.007	-4.874	0.000
	Education level	0.004	2.945	0.003
shimmer_stddev	Education level	-0.004	-2.949	0.003
MFCC_[1]_stddev	Age	-0.434	-4.456	0.000
	Education level	-0.944	-6.978	0.000
MFCC_[1]_range	Age	-1.233	-3.235	0.001
	Education level	-3.257	-6.151	0.000
MFCC_[2]_stddev	Age	-0.222	-2.282	0.023
	Education level	-0.295	-2.176	0.03
MFCC_[3]_stddev	Age	0.206	1.981	0.048
	Gender	0.480	3.560	0.000
	Education level	-0.301	-2.078	0.038
MFCC_[3]_range	Age	0.959	2.285	0.022
	Gender	2.004	3.679	0.000
MFCC_[4]_stddev	Age	-0.218	-1.998	0.046
	Gender	0.716	5.066	0.000
	Education level	-0.760	-5.014	0.000
MFCC_[4]_range	Gender	2.296	3.926	0.000
	Education level	-3.179	-5.075	0.000
MFCC_[5]_stddev	Age	0.275	2.533	0.011
	Education level	-0.931	-6.225	0.000
MFCC_[5]_range	Age	1.628	3.371	0.001
	Education level	-3.307	-4.920	0.000
MFCC_[6]_range	Gender	1.393	2.717	0.007
MFCC_[7]_stddev	Education level	-0.397	-3.416	0.001
MFCC_[7]_range	Education level	-1.323	-2.408	0.016
MFCC_[9]_stddev	Age	-0.355	-4.701	0.000
	Gender	-0.272	-2.763	0.006
	Education level	-0.634	-6.035	0.000
MFCC_[9]_range	Age	-1.373	-3.584	0.000
	Gender	-1.429	-2.866	0.004
	Education level	-2.319	-4.343	0.000

Table 14 continued from previous page

Acoustic parameters	Sociodemographic characteristics	B	t	p
MFCC_[10]_stddev	Education level	-0.288	-2.850	0.004
MFCC_[11]_stddev	Age	-0.393	-6.157	0.000
	Gender	0.289	3.475	0.001
	Education level	-0.318	-3.565	0.000
MFCC_[11]_range	Age	-1.602	-4.860	0.000
	Gender	1.801	4.199	0.000
	Education level	-1.020	-2.212	0.027
MFCC_[12]_stddev	Age	-0.265	-5.157	0.000
	Gender	0.252	3.760	0.000
MFCC_[12]_range	Age	-1.190	-4.286	0.000
	Gender	1.610	4.463	0.000
logHNR_mean	Age	-2.848	-6.288	0.000
	Education level	-2.788	-4.400	0.000
logHNR_stddev	Education level	-0.559	-3.098	0.002
zcr_stddev	Education level	0.002	2.094	0.036
zcr_mean	Gender	-0.004	-2.666	0.008
RMSenergy_range	Age	-0.001	-2.348	0.019
audSpec_Rfilt_[1]_range	Age	-0.181	-4.125	0.000
	Gender	0.248	4.346	0.000
	Education level	-0.555	-9.138	0.000
audSpec_Rfilt_[1]_stddev	Age	-0.031	-5.335	0.000
	Gender	0.038	4.987	0.000
	Education level	-0.087	-10.726	0.000
audSpec_Rfilt_[2]_range	Age	-0.410	-4.422	0.000
	Education level	-0.979	-7.606	0.000
audSpec_Rfilt_[2]_stddev	Age	-0.063	-5.303	0.000
	Gender	0.036	2.334	0.020
	Education level	-0.153	-9.400	0.000
audSpec_Rfilt_[3]_range	Age	-0.487	-4.333	0.000
	Gender	0.423	2.895	0.004
	Education level	-0.603	-3.848	0.000
audSpec_Rfilt_[3]_stddev	Age	-0.074	-5.309	0.000
	Gender	0.067	3.693	0.000
	Education level	-0.096	-4.919	0.000

Table 14 continued from previous page

Acoustic parameters	Sociodemographic characteristics	B	t	p
audSpec_Rfilt_[4]_range	Age	-0.529	-2.839	0.005
	Gender	1.333	5.526	0.000
	Education level	-1.913	-7.416	0.000
audSpec_Rfilt_[4]_stddev	Age	-0.074	-3.212	0.001
	Gender	0.186	6.242	0.000
	Education level	-0.266	-8.370	0.000
audSpec_Rfilt_[5]_range	Age	-0.656	-2.934	0.003
	Gender	1.384	4.766	0.000
	Education level	-1.875	-6.034	0.000
audSpec_Rfilt_[5]_stddev	Age	-0.086	-3.163	0.002
	Gender	0.200	5.665	0.000
	Education level	-0.251	-6.626	0.000
audSpec_Rfilt_[10]_range	Age	-0.659	-4.572	0.000
	Gender	1.617	8.693	0.000
	Education level	-0.849	-4.220	0.000
audSpec_Rfilt_[10]_stddev	Age	-0.118	-5.978	0.000
	Gender	0.252	9.925	0.000
	Education level	-0.153	-5.579	0.000
audSpec_Rfilt_[11]_range	Age	-0.653	-4.434	0.000
	Gender	1.549	8.145	0.000
	Education level	-0.735	-3.575	0.000
audSpec_Rfilt_[11]_stddev	Age	-0.120	-6.016	0.000
	Gender	0.233	9.082	0.000
	Education level	-0.134	-4.810	0.000
audSpec_Rfilt_[12]_range	Age	0.561	-3.508	0.000
	Gender	0.964	4.641	0.000
	Education level	-1.093	-4.909	0.000
audSpec_Rfilt_[12]_stddev	Age	-0.109	-5.091	0.000
	Gender	0.142	5.091	0.000
	Education level	-0.176	-5.865	0.000
audSpec_Rfilt_[13]_range	Age	-0.629	-4.004	0.000
	Gender	0.604	2.955	0.003
	Education level	-1.596	-7.322	0.000

Table 14 continued from previous page

Acoustic parameters	Sociodemographic characteristics	B	t	p
audSpec_Rfilt_[13]_stddev	Age	-0.117	-5.505	0.000
	Gender	0.095	3.415	0.001
	Education level	-0.247	-8.377	0.000
audSpec_Rfilt_[19]_range	Age	-0.659	-4.746	0.000
	Gender	1.099	6.098	0.000
	Education level	-1.111	-5.743	0.000
audSpec_Rfilt_[19]_stddev	Age	-0.111	-5.946	0.000
	Gender	0.173	7.113	0.000
	Education level	-0.166	-6.345	0.000
audSpec_Rfilt_[21]_range	Age	-0.496	-4.193	0.000
	Gender	1.309	6.781	0.000
	Education level	-0.977	-5.939	0.000
audSpec_Rfilt_[21]_stddev	Age	-0.099	-6.088	0.000
	Gender	0.159	7.539	0.000
	Education level	-0.133	-5.863	0.000
audSpec_Rfilt_[22]_range	Age	-0.326	-2.995	0.003
	Gender	0.688	4.864	0.000
	Education level	-0.340	-2.236	0.025
audSpec_Rfilt_[22]_stddev	Age	-0.067	-4.444	0.000
	Gender	0.109	5.570	0.000
	Education level	-0.042	-1.991	0.047

6.3 Limitation and Future Work

The limitations of this research are mainly caused by the data. Although this research proposed a new dataset for the field of children's state anxiety recognition, unbalanced label and less sufficient samples lead to the disadvantages of this research. The lack of data is due to time constraints. The data collection plan was postponed because of the delay of Zenbo's delivery. The imbalance of data is inevitable, as the participants in the data collection experiment are mentally healthy children. Therefore, in this research, Easy Ensemble and Threshold Moving methods are applied to solve this problem. Although the boosting algorithms perform well, the subsequent work may be affected by the disadvantages of data. Besides, although three types of languages are contained in this dataset, these audios are not suitable for analyzing languages-related questions. This is because all audios were collected from Chinese students who are native Chinese speakers. The differences shown by audios with different languages can not explain the influence of language on anxiety and anxiety recognition. Therefore, although BSDCSA contains three types of languages, there is little practical significance.

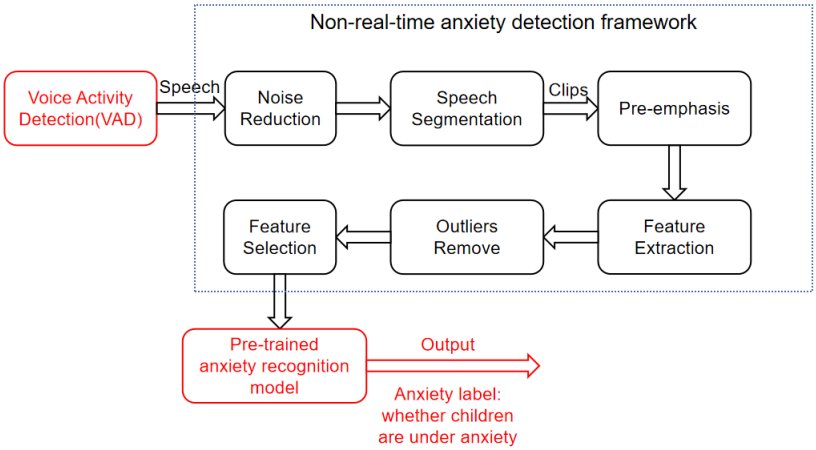


Figure 13: The real-time anxiety detection framework

For future work, third aspects will be devoted to. The first aspect focuses on data. Firstly, more data should be collected to expand BSDCSA. Secondly, if possible, GAD diagnosed children will be recruited to improve the reliability of BSDCSA in the clinical context. Thirdly, English native speakers will also be recruited to make BSDCSA applicable in English context. All data collection experiments should be consistent with the previous experimental conditions. With more data, three research questions will be re-analyzed to make sure the applicability of the results of these research questions in other research in this field. The third is the real-time anxiety detection algorithm for children. This research is a part of the HUMAN-AI project. The final target of this project is to develop an affective SAR for children with GAD. Therefore, a real-time anxiety recognition framework is needed. It can be developed from current anxiety detection process. Figure 13 is a basic framework for real-time children's anxiety detection. The framework will be inserted into Zenbo to build the SAR for children with GAD.

7 Conclusion

This paper mainly introduces three parts. First, a novel dataset for children's state anxiety detection, BSDCSA, is introduced for children' state anxiety detection field. The process and methods of data collection are also described in detail. Second, a standard data processing flow is provided, which provides a reference for future studies on BSDCSA. Third, traditional machine learning methods and boosting algorithms are applied to recognize children's state anxiety. Among the regressors that predict the anxiety level of children, GBR and XGBR with GBDT_reg best. Among the classifiers that classify whether children are under anxiety, GBDT and XGB with GBDT_cla feature set gained the best performance. Fourth, three research questions focusing on acoustic parameters that are most relevant to children's state anxiety, the change of common acoustic parameters with the increase of state anxiety level, the relationship between sociodemographic characteristics and children's state anxiety, and the

relationship between sociodemographic characteristics and common acoustic parameters were proposed and answered. The first research question listed 20 features that are most relevant to the anxiety labels and 20 features that are most relevant to the anxiety levels. Some of these features have been proved by previous studies that they are related to anxiety. The second research question analyzed the trends of common acoustic features with the increase of anxiety levels. The trends of acoustic parameters with the increase of anxiety level of children are quite different from that of adults, which is a novel discovery. The third research question explored the relationship between sociodemographic characteristics and children's state anxiety as well as acoustic parameters. The results show that only gender is related to state anxiety. Most of the common acoustic parameters are related to gender, age, and education level. As there is little research in the field of children's anxiety and children's state anxiety detection, this thesis provides a reference for the following research in the field of children's state anxiety recognition. From data collection to model training, each section of children's anxiety detection is introduced in detail. At last, the limitations and future work are concluded. BSDCSA needs more data from English native speakers and GAD diagnosed children, which is the biggest challenge in the case of a global COVID-19 epidemic.

References

[1] Sidney Kimmel Medical College of Thomas Jefferson University Josephine Elia, MD. An overview of anxiety disorders in children and adolescent, 2019.

[2] NSCH. Mental and behavioral health nsch data brief—october 2020. 2020.

[3] L. Araújo, Cássio Frederico Veloso, Matheus de Campos Souza, João Marcos Coelho de Azevedo, and G. Tarro. The potential impact of the covid-19 pandemic on child growth and development: a systematic review. *Jornal De Pediatria*, 2020.

[4] T.J. Huberty and A.C. Dick. Performance and test anxiety. *Children's needs III: Development, prevention and intervention*, pages 281–291, 01 2006.

[5] National Scientific Council on the Developing Child. Persistent fear and anxiety can affect young children's learning and development: Working paper no. 9. 2010.

[6] Ghosh SS. Low DM, Bentley KH. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investig Otolaryngol*, 2020.

[7] M. K. Crossman, A. Kazdin, and Elizabeth R Kitt. The influence of a socially assistive robot on mood, anxiety, and arousal in children. *Professional Psychology: Research and Practice*, 49:48–56, 2018.

[8] Marta Díaz, Neus Nuño, Joan Saez-Pons, Diego E. Pardo, and Cecilio Angulo. Building up child-robot relationship for therapeutic purposes: From initial attraction towards long-term social engagement. pages 927–932, 2011.

[9] José Carlos Pulido, José Carlos González Dorado, Cristina Suárez, Antonio Bandera, Pablo Bustos, and Fernando Fernández. Evaluating the child–robot interaction of the naotherapist platform in pediatric rehabilitation. *International Journal of Social Robotics*, 9:343–358, 06 2017.

[10] Clinical application of a humanoid robot in pediatric cancer interventions. *International Journal of Social Robotics*, 8(5):743–759, 3 2016.

[11] Katarzyna Kabacińska, Tony Prescott, and Julie Robillard. Socially assistive robots as mental health interventions for children: A scoping review. *International Journal of Social Robotics*, 07 2020.

[12] C. D. Spielberger. *Anxiety and Behavior*. New York: Academic Press., 1966.

[13] Gorsuch R. L. Lushene R. Vagg P. R. Jacobs G. A. Spielberger, C. D. Manual for the state-trait anxiety inventory. *Consulting Psychologists Press*, 1983.

[14] da Silva LCF Teixeira-Silva F. Leal PC, Goes TC. Trait vs. state anxiety in different threatening situations. *Trends Psychiatry Psychother*, Jul-Sep 2017.

[15] Charles Donald Spielberger and C Spielberger. Understanding stress and anxiety. 1979.

[16] Behnaz Nojavanasghari, Tadas Baltrusaitis, Charles E. Hughes, and Louis-Philippe Morency. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *ICMI*, pages 137–144. ACM, 2016.

[17] A Buske-Kirschbaum, S Jobst, A Wustmans, C Kirschbaum, W Rauh, and D Hellhammer. Attenuated free cortisol response to psychosocial stress in children with atopic dermatitis. *Psychosomatic medicine*, 59(4):419—426, 1997.

[18] Ellen W McGinnis, Steven P Anderau, Jessica Hruschak, Reed D Gurchiek, Nestor L Lopez-Duran, Kate Fitzgerald, Katherine L Rosenblum, Maria Muzik, and Ryan S McGinnis. Giving voice to vulnerable children: Machine learning analysis of speech detects anxiety and depression in early childhood. *IEEE journal of biomedical and health informatics*, 23(6):2294—2301, November 2019.

[19] Berkehan Akçay and Kaya Oguz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116, 01 2020.

[20] R Banse and KR Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614—636, March 1996.

[21] Justin W. Weeks, Chao-Yang Lee, Alison R. Reilly, Ashley N. Howell, Christopher France, Jennifer M. Kowalsky, and Ashley Bush. "the sound of fear": Assessing vocal fundamental frequency as a physiological indicator of social anxiety disorder. *Journal of Anxiety Disorders*, 26(8):811–822, 2012.

[22] Iain R. Murray and John L. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993. Medline is the source for the MeSH terms of this document.

[23] BF Fuller, Y Horii, and DA Conner. Validity and reliability of nonverbal voice measures as indicators of stressor-provoked anxiety. *Research in nursing amp; health*, 15(5):379—389, October 1992.

[24] Turgut Özseven, Muharrem Dügenci, A. Doruk, and Hilal İlkay Kahraman. Voice traces of anxiety: Acoustic parameters affected by anxiety disorder. *Archives of Acoustics*, 43:625–636, 2018.

[25] Teixeira A Figueiredo D-Sa-Couto P Oliveira C Albuquerque L, Valente ARS. Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan. *PLoS ONE*, 2021.

[26] Emmanuel Pintelas, Theodore Kotsilieris, Ioannis Livieris, and P. Pintelas. A review of machine learning prediction methods for anxiety disorders. 07 2018.

[27] A. Salekin, Jeremy W. Eberle, Jeffrey J. Glenn, B. Teachman, and J. Stankovic. A weakly supervised learning framework for detecting social anxiety and depression. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2:1 – 26, 2018.

[28] Beck JS. *Cognitive behavior therapy: Basics and beyond (2nd ed.)*. New York: The Guilford Press, 2011.

[29] Ho Li and Violeta Lopez. Development and validation of a short form of the chinese version of the state anxiety scale for children. *International journal of nursing studies*, 44:566–73, 06 2007.

[30] Ho Li and Violeta Lopez. The reliability and validity of the chinese version of the trait anxiety scale for children. *Research in nursing health*, 27:426–34, 12 2004.

[31] Ivar Krumpal. Determinants of social desirability bias in sensitive surveys: A literature review. *Quality Quantity*, 47, 06 2011.

[32] Yang Xinyi, Si Boyu, Meng Qingyun, and Huang Kailin. Design of the speech tone disorders intervention system based on speech synthesis. *Journal of Physics: Conference Series*, 1617:012078, aug 2020.

[33] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462. Association for Computing Machinery, 2010.

[34] Björn Schuller, Stefan Steidl, Anton Batliner, Jarek Krajewski, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Sebastian Schnieder. The interspeech 2014 computational paralinguistics challenge: Cognitive physical load. 09 2014.

[35] Audeering. opensmile python.

[36] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008.

[37] Goblin Guardian. Anomaly detection algorithm – analysis of isolation forest.

[38] Zhangyu Cheng, Chengming Zou, and Jianwei Dong. Outlier detection using isolation forest and local outlier factor. In *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, RACS '19, page 161–168, New York, NY, USA, 2019. Association for Computing Machinery.

[39] Mateusz Garbacz. Open-sourcing shaprfecv — improved feature selection powered by shap.

[40] Mirka Saarela and Susanne Jauhiainen. Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, 3, 02 2021.

[41] Jiliang Tang, Salem Alelyani, and Huan Liu. *Feature selection for classification: A review*, pages 37–64. CRC Press, January 2014.

[42] Liping Xie, Zilong Li, Yihan Zhou, Yiliu He, and Jiaxin Zhu. Computational diagnostic techniques for electrocardiogram signal analysis. *Sensors*, 20(21), 2020.

[43] Li-Yeh Chuang, Chao-Hsuan Ke, and Cheng-Hong Yang. A hybrid both filter and wrapper feature selection method for microarray classification, 2016.

[44] Huan Liu. *Feature Selection*, pages 402–406. Springer US, Boston, MA, 2010.

[45] Jorge R. Vergara and Pablo A. Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1):175–186, Mar 2013.

[46] Jerome Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 10 2001.

[47] Shuang Liu, Ziheng Yang, Yi Li, and Shuiqing Wang. Decision tree-based sensitive information identification and encrypted transmission system. *Entropy*, 22:192, 02 2020.

[48] Diego Lopez Yse. The complete guide to decision trees.

[49] J. R. Quinlan. Induction of decision trees. 1(1):81–106, March 1986.

[50] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[51] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone. Classification and regression trees. 1983.

[52] Rekha M. Entropy, information gain, and gini index; the crux of a decision tree.

[53] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[54] Tao Shi and Steve Horvath. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics - J COMPUT GRAPH STAT*, 15, 03 2005.

[55] Divya Pramasani Mohandoss, Yong Shi, and Kun Suo. Outlier prediction using random forest classifier. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0027–0033, 2021.

[56] Jason Brownlee. A gentle introduction to the gradient boosting algorithm for machine learning.

[57] T. Hastie, R. Tibshirani, and J. H. Friedman. 10. boosting and additive trees. In *The Elements of Statistical Learning (2nd ed.)*. New York: Springer, 2009.

[58] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.

[59] Kurtis Pykes. Oversampling and undersampling.

[60] Jason Brownlee. A gentle introduction to threshold-moving for imbalanced classification.

[61] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009.

[62] Namratesh Shrivastav. Confusion matric(tpr,fpr,fnr,tnr), precision, recall, f1-score.

[63] Sarang Narkhede. Understanding auc - roc curve.

[64] Akshita Chugh. Mae, mse, rmse, coefficient of determination, adjusted r squared — which metric is better?

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[66] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[67] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994.