



Experience with an Affective Robot Assistant for Children with Hearing Disabilities

Pinar Uluer^{1,2} · Hatice Kose² · Elif Gumuslu³ · Duygun Erol Barkana³

Accepted: 10 September 2021 / Published online: 16 November 2021
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract

This study presents an assistive robotic system enhanced with emotion recognition capabilities for children with hearing disabilities. The system is designed and developed for the audiometry tests and rehabilitation of children in a clinical setting and includes a social humanoid robot (Pepper), an interactive interface, gamified audiometry tests, sensory setup and a machine/deep learning based emotion recognition module. Three scenarios involving conventional setup, tablet setup and setup with the robot+tablet are evaluated with 16 children having cochlear implant or hearing aid. Several machine learning techniques and deep learning models are used for the classification of the three test setups and for the classification of the emotions (pleasant, neutral, unpleasant) of children using the recorded physiological signals by E4 wristband. The results show that the collected signals during the tests can be separated successfully and the positive and negative emotions of children can be better distinguished when they interact with the robot than in the other two setups. In addition, the children's objective and subjective evaluations as well as their impressions about the robot and its emotional behaviors are analyzed and discussed extensively.

Keywords Social robots · Human-robot interaction · Machine learning · Deep learning · Emotion recognition · Physiological signals

1 Introduction

Recent studies in socially assistive robotic systems enable the use of robots in many application areas such as health-care, education, and elderly care, as a part of our daily lives.

Research supported by The Scientific and Technological Research Council of Turkey (TÜBİTAK) under the grant number 118E214.

✉ Pinar Uluer
pulner@gsu.edu.tr

Hatice Kose
hatice.kose@itu.edu.tr

Elif Gumuslu
elif.gumuslu@std.yeditepe.edu.tr

Duygun Erol Barkana
duygunerol@yeditepe.edu.tr

¹ Department of Computer Engineering, Galatasaray University, Istanbul, Turkey

² Department of AI and Data Engineering, Istanbul Technical University, Istanbul, Turkey

³ Department of Electrical and Electronics Engineering, Yeditepe University, Istanbul, Turkey

When people see robots in human-robot interactive social environments, they expect a level of socially intelligent and socially aware behavior based on the robot's capabilities and human-robot interaction settings. Besides, most of the users are inexperienced in terms of robots, and might have special needs such as elderly, children with Autism spectrum disorder (ASD) or people with hearing disabilities [1–3]. Socially assistive robots can be personalized and adapted to these needs and requirements to increase the user performance and to improve their interaction with robots. Machine learning methods are beneficial in this personalization and adaptation phase to analyse the status of human in interaction, and to generate appropriate action or feedback accordingly.

Henschel et al. analyzed the gap between the general public's expectations about social robots and experience with them, they provided a review study to understand what makes a robot social in the eyes of general public and how they may contribute to the society [4]. The authors remark that even though the research domain present many challenges and there are still real limitations due to the present technology in social robots' capabilities, the early findings demonstrate that social robots have the potential to improve human life,

especially in educational settings, psychosocial support and rehabilitation contexts. In their review study, Robinson et al. investigated and discussed in details the evidence from randomized controlled trials on psychosocial health interventions by social robots [5]. In another study, Stower et al. presented a comprehensive review on the meta-analysis of the factors influencing children's trust in social robots [6].

In this sense, Cifuentes et al. [7] explored the implications of social robots in healthcare scenarios, and they conducted a review study on the applications of social robots, how they were perceived and accepted by children and adults. Their review revealed that adults and children who were exposed to an intervention with social robots, showed signs of improvements in social connection and communication and their current mood, and they displayed less signs of depression, anxiety and fear. Similarly, Kabacinska et al. [8] conducted a literature survey on how social robots were used as means to support mental health in children. The study outcomes suggested that interventions with social robots had positive impact on stress relief and improved positive affect level.

Ferrante et al. [9] explored how social robots can be used to increase therapeutic adherence in pediatric asthma based on the assumption that the robots increased the motivation of children and their engagement to treatment. Although they presented a conceptual study, the authors investigated the available robotic platforms, the nature of interaction and the intervention, and they discussed the feasibility, acceptability and efficacy of socially assistive robots in clinical practice.

Logan et al. [10] proposed to use social robots as engaging tools to address the emotional needs of hospitalized children. They introduced an interactive teddy bear robot, Huggable, into the inpatient setting of a pediatric clinic to improve engagement of hospitalized children and to reduce their stress during their intervention. They conducted a pilot study with 54 children (3 to 10 years old), the children were exposed to an intervention with a teddy bear social robot, a tablet-based avatar of it and a plush teddy bear accompanied by a human presence. The pilot study results pointed out that the children exposed to the social robot intervention showed more positive affect and they expressed greater levels of joyfulness and agreeableness than the other conditions.

In another study, Moerman and Jansens [11] used a baby dinosaur pet robot with children in a hospitalized setting to distract them and to improve their well-being through their interactions with the robot. They conducted a preliminary study with 9 children (aged 4–13 years old) who received personalized therapeutic sessions with the robot. The results showed that the children expressed less boredom, anxiety and stress and displayed more active and playful behavior based on the behavioral analysis of experts and the interviews with the parents.

In this paper, we present an assistive robotic system, RoboRehab, which is designed and developed to support the

auditory tests of children with hearing disabilities in clinical settings. Several tests are applied to diagnose the level of hearing and to adjust the hearing aids and cochlear implants of children accordingly. When the children get stressed or lose motivation, the performance of the tests and the cooperation of the children decrease significantly. In RoboRehab, we use a socially assistive humanoid robot Pepper, enhanced with emotion recognition, and a tablet interface, to support children in these tests. In the current study, we investigate the quantitative and qualitative effects of the test setups involving the robot+tablet, tablet and conventional methods. We employ traditional machine learning techniques and deep learning approaches to analyse and classify the affective data of children collected by E4 physiological wristband. Then this physiological data is used to evaluate the proposed assistive robotic system. We use blood volume pulse (BVP), skin temperature (ST), and skin conductance (SC) from the E4 wristband to classify three test setups (conventional, tablet, and robot+tablet). Furthermore, these BVP, ST and SC data are also used to classify the emotions (pleasant, neutral, unpleasant) of children under three test setups. We aim to detect the emotions or stress of children and personalize/adapt the feedback mechanism of the robot accordingly.

2 Emotion Recognition Studies

Emotion recognition is becoming an increasingly active field in research since it can contribute significantly in different applications such as video games, animations, psychiatry, education, and robotics, especially human-robot interaction (HRI).

One of the key concepts in the emotion recognition studies, is how to represent the emotions. Spezialetti et al. [12], in their review on the recent advances in emotion recognition studies focusing on the HRI context, report that currently adopted emotional models are the categorical models based on the discrete emotions [13–15] or the dimensional models [16], in which the emotions are defined by their characteristic features such as arousal and valence. Therefore, some of the studies focused on the classification of emotions such as joy, anger, disgust, etc. while others focused only on the classification of arousal and valence. The valence parameter is commonly used for the classification of pleasant (amusement, happiness, joy), unpleasant (anxiety, fear, disgust, anger, sadness) and neutral (neutral, surprise) emotions. Furthermore, some researchers have also classified three arousal classes: calm (neutral, disgust, calm), medium arousal (amusement, happiness, joy), and excited (anxiety, fear, surprise, anger). Six universal emotion classes (happiness, disgust, surprise, fear, sadness, and anger) have been classified in some studies. In this study, we are interested in understanding if the children feel positive (pleasant),

negative (unpleasant), or neutral, thus we use the valence dimension property of the two-dimensional emotion model.

On the other hand, several theorists argue that the emotions can not be categorized or discretized because they have evolved as an adaptation mechanism based on the individual's environment. Consequently, they claim that there is an appraisal system assessing the perceived environmental factors with respect to the individual's well-being, its plans and goals [17,18]. Scherer [19] suggests that this appraisal system requires a high level of cognition because people tend to appraise events with respect to a large variety of factors, such as novelty, pleasantness and goal significance as well as their coping mechanism. And, these appraisals evoke a particular emotional state that needs some sort of adjustment such as physiological changes, behavioural responses and expressions.

There are many ways to understand the emotional states and expressions of people such as facial expressions [20,21], speech [22], and electroencephalography (EEG) [23–26]. Physiological signals, such as heart rate (HR), heart rate variability [27,28], BVP [29–31], SC [32,33], and ST have been used to recognize emotions [34,35]. The physiological parameters are valuable because it is easy to measure them using wearable devices. A survey about the physiological sensors used to recognize emotions, has been presented in [36]. Although the facial expression, speech, EEG, and physiological signals have been widely employed in studies with adult subjects for emotion analysis and recognition, similar studies in children have rarely been addressed. Disgust, fear, happiness, sadness, and surprise have been recognized for typically developed children using facial emissivity changes detected by Infrared Thermal Imaging (IRTI) [37]. A stress detection framework has been previously proposed for children using heart rate data recorded with a wearable device [38]. A commercial device has also been used to detect the stress states of children such as iCalm [39]. Electrodermal activity signals, which are also known as galvanic skin response or skin conductance, have been used to classify emotions in children [40].

Empatica E4 is a compact, lightweight, and wireless wearable multi-sensor wristband that is used to collect physiological signals in real-time [41,42]. An emotion recognition system in the aspect of arousal, valence, and four emotions using E4 wristband has previously been developed in [43]. E4 wristband has also been used to detect the stress level of subjects in several studies [44–47]. The features from heart activity, skin conductance, and physical activity recorded by E4 wristband have also been used to recognize stress levels [48]. Electrocardiogram (ECG) signal from E4 wristband has been used to classify emotions based on Russell's four-class circumplex emotion model in a virtual reality environment [49,50]. However, to our knowledge E4 wristband has not been used to detect the emotions of children with hearing

disabilities. In this study, skin conductance (SC) or response of the skin as an electrodermal activity (EDA) via EDA sensor, blood volume pulse (BVP) with a photoplethysmogram (PPG) sensor, and skin temperature (ST) via infrared thermopile sensor are collected from the E4 wristband to recognize the emotions of the children with hearing disabilities.

Various machine learning methods such as Support Vector Machine (SVM), Multi-Layer Perceptron Back Propagation (MLP-BP) have been used to classify the emotions using physiological signals [36,49,51–56]. Several deep learning approaches have also been investigated for emotion classification because of the high-dimensional physiological features [57–65]. Furthermore, deep learning approaches have also been used to recognize the emotions of infants and children from their facial expressions [66,67].

We aim to develop an emotion recognition module using physiological signals that will be used to assist the rehabilitation of children with hearing disabilities via a socially assistive robot Pepper in this study. In our previous work, a classical machine learning approach, Gradient Boosting Machines (GBM) and a deep learning approach, Convolutional Neural Networks (CNN) were used to classify pleasant, unpleasant and neutral emotions from the recorded EEG and physiological signals of adults. We illustrated how an emotion recognition system can be trained on responses to standardized materials (IAPS images), and discussed how such a system might work as a feedback mechanism in human-robot interaction between robots and patients in a rehabilitation context [68].

In this study, we use machine learning and deep learning methods to classify the different test setups as well as the emotions of the children through their consultation. To the best of our knowledge, this is the first study which aims to understand the emotions of children with hearing impairment using machine/deep learning approaches using the physiological data recorded with E4 wristband.

3 RoboRehab: Affective Assistant Robot System and Auditory Perception Tests

3.1 RoboRehab

The main aim of the RoboRehab project is to support children with hearing impairments in audiometry tests and rehabilitation. The children take several auditory tests to reveal their hearing levels, and have their hearing aids or cochlear implants adjusted. Unfortunately audiologists and psychologists, who are working with these children, stated that children can not reveal their true performance during the conventional tests due to stress or boredom/motivation loss. Besides these parameters, the audiologist might be inexperienced or unfamiliar with the children in the test, which will

also effect the motivation and performances of the children. Therefore we design and develop a system called RoboRehab to overcome these limitations.

RoboRehab includes an affective module processing facial and physiological data, a tablet interface for the tests, and a social robot assistant Pepper to run the hearing tests and give feedback. Gamification is used in the tablet interface, and the scenarios with robot in order to increase the motivation and attention time of the children. The affective module for emotion recognition consists of three parts; facial data based emotion recognition, physiological data based emotion recognition, and a multimodal emotion recognition module which is based on the fusion of the facial and physiological data.

In this paper, we only focus on the physiological emotion recognition module, please refer to the study presented by Baglayici et al. [69] for further information on the facial emotion recognition and the challenges we faced during the recognition of children's emotion in-the-wild.

3.2 Motivation & Research Questions

In this study, our main motivation is to analyse the effect of test setup/media (conventional test, tablet and the robot+tablet setups). We want to observe not only the effect on the performance of children but also the change on children's emotions and attention during each test setup. We also want to recognize the emotions of children especially within the presence of robot. We plan to use the emotion recognition module to enhance the interaction skills of the robot in the tests with children and to increase their attention and lower negative emotions and stress in the long term. Therefore in this study we employ machine learning and deep learning approaches on the physiological data collected and labeled during the test studies with children with hearing disabilities. Based on these motivations we want to reveal the following research questions:

- RQ1 can we distinguish between the different setups (conventional/tablet/robot) using the physiological data with machine/deep learning approaches?
- RQ2 can we recognize the emotions of children when they interact with robot using machine/deep learning approaches?
- RQ3 is there any difference between the test performances based on the objective evaluation metrics (involving several parameters logged during the tests) of children in these three setups? If so, how can we interpret this difference?
- RQ4 can we reveal such difference between the test setups based on the subjective evaluation metrics (self-report survey answers and interview with children)?

Based on these research questions, we designed an experimental setting to show if the robot can stimulate children's emotions, and cause difference in their physiological signals. Furthermore, we investigate if machine learning approaches can be used to understand the emotions of children during their interaction with a socially assistive robot. Additionally, we explore if the proposed test setups have any effect on the performances of children and their impressions about the perceived character, intelligence and likability of the robot in the presented interaction context.

4 Experimental Setup

4.1 Auditory Perception Test

The auditory test used is a Turkish-translated version of the Developmental Test of Auditory Perception (DTAP). The test is designed for assessing the auditory perception of children without invoking higher order verbal reasoning [70]. The test is composed of different item sets to measure various aspects of children's auditory perception. Only the non-language related item sets are selected for this study:

- Environmental sounds: Non-language related sounds commonly found in environment such as crying babies, barking dogs, traffic noises, etc.
- Tonal patterns: Sequences of beeps with varying frequency and length.

The questions in the environmental sounds and tonal patterns item sets are adapted for the gamification process with the tablet and robot.

Each item set in the auditory test is composed of 30 questions and each question composed of a pair of auditory stimuli. The children are asked to indicate if the pairs are the same or different.

In the conventional test setup, an audiologist accompanies the child who is taking the test. The audiologists play the pair of auditory stimuli sequentially, and they ask the child to answer verbally if the pairs are the same or different, and finally they mark the child's answers on a paper-based questionnaire.

The auditory tests are digitized and two gamification-based setups are designed to motivate the children during their clinic consultation. In the proposed setups, the audiologists preserve their place along with the child but the child answers the test questions using a tablet. Both setups include a short tutorial session. The children are familiarized with the type of questions, the concepts "same" and "different" and their pictorial representations in this tutorial session. If the pairs of auditory stimuli are the same, then the child is advised to select the answer with the picture of two identical

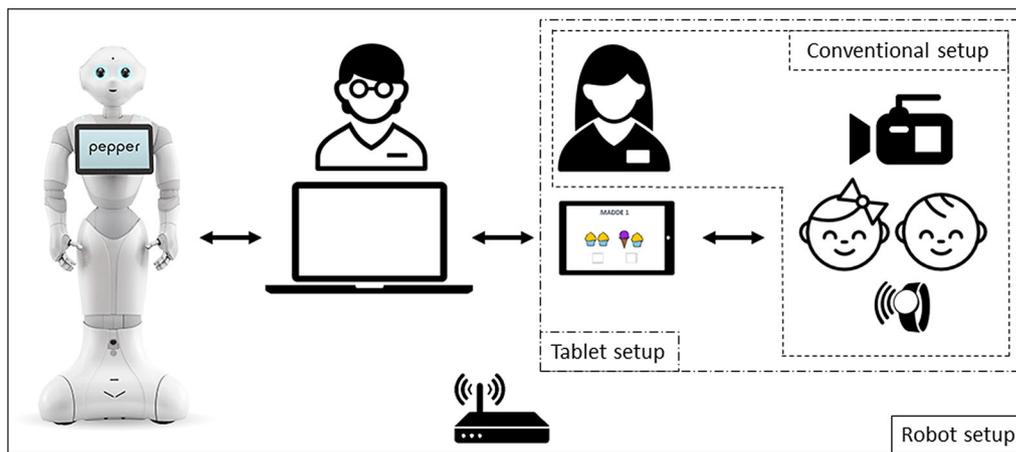


Fig. 1 Test setups for the auditory perception game

muffins, if not, then the answer with the picture of one muffin and one ice-cream cone should be selected.

The tablet setup consists of a tablet-based game, where the child plays the auditory perception game under an audiologist's supervision. If the child correctly answers the question, an audiovisual feedback is triggered, a verbal approval recording is played and a confetti animation is displayed on the tablet screen.

In the robotic setup, the same tablet-based game is played with Pepper robot and the feedback is given by the robot's affective behaviours. The robotic setup differs from the tablet setup mostly with the presence of the robot and the feedback mechanism. In the robot setup the children interact with the tablet accompanied by an audiologist, the tablet communicates with the robot via an intermediary device monitored by a researcher situated out of children's sight, as schematized in Fig. 1.

The robot's verbal and non-verbal behaviours are triggered by the correct answers of the children. The verbal and non-verbal feedback of the robot are designed and implemented under the guidance of the audiologists and a pedagogue. For the verbal feedback of the robot, a subset of recordings in Turkish with different voice profiles (male, female, machine-like voices with varying pitch and speed properties) were prepared using open source text-to-speech software and Pepper's embedded text-to-speech module. These recordings were voted by the audiologists and a small set of children, and the most liked voice was selected to generate the robot's verbal feedback. In brief, the robot uses a child-like and genderless voice that has a higher pitch and lower articulation speed with machine-like notes. The feedback sentences used to enhance the affective behaviour of the robot were also provided by the audiologists. The audiologists stated that in their conventional tests, no negative feedback are given to the children and they use only positive and neutral feedback. Therefore the verbal feedback of the robot are generated

accordingly. Thus, a positive utterance is randomly selected from a set of predefined sentences and voiced by the robot when a child answers the question correctly, and when the child's answer is incorrect, the robot tells the child to listen the next question carefully.

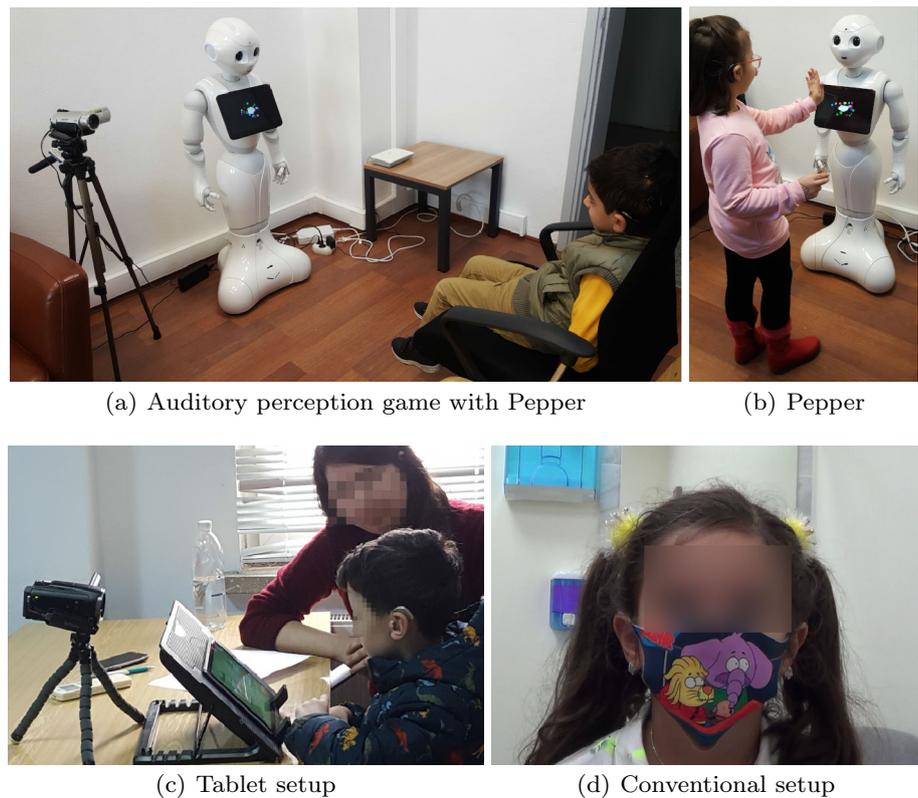
For the non-verbal feedback of the robot, all the predefined animations were investigated in detail and they were eliminated with respect to the cognitive developmental age and the cultural background of the children. In addition, the animations with brusque and high-powered movements were discarded in order not to intimidate children. The animation set for the study is composed of affirmative (e.g. head nod), exclamative (e.g. lean back with open arms) and affective animations (e.g. happy, excited). When the robot voices a positive utterance then it is accompanied randomly by a positive posture and a positive affective animation.

4.2 Interaction Data Collection

The children were equipped with the Empatica E4 smart wristband during the interaction session. The real-time physiological signals and facial expression data were collected with E4 wristband and a video camera, respectively for all the three test setups: conventional, tablet, and robot+tablet (Fig. 1). In addition, in the setup with the robot, the whole session was recorded with a supplementary video camera. The data captured by E4 was annotated by the audiologists and psychologists taking part in the study. First the video recordings of children were annotated (pleasant, neutral, unpleasant) and labeled using a video annotation tool designed and developed for this study. Then the labels were used to annotate the synchronous E4 data.

Two survey studies were conducted to explore the children's subjective evaluation of the robot when the interaction session with the robot was completed. The surveys were designed to measure and to evaluate the emotional responses

Fig. 2 Test setups for the experimental studies with children



(a) Auditory perception game with Pepper

(b) Pepper

(c) Tablet setup

(d) Conventional setup

of the children toward the robot and their impressions about its behaviours. Due to the age of children, the well-known questionnaires to measure attitudes towards robots such as NARS (Negative Attitude towards Robots Scale, [71]) or RAS (Robot Anxiety Scale, [72]) cannot be used since both questionnaires contains questions requiring a higher linguistic and cognitive understanding. Therefore, the first survey was designed as a simple questionnaire, mostly composed of yes-no questions.

The second survey was designed as an online survey to check the validity of the available animations of Pepper and to explore if the children in the target group could correctly identify which behaviour or emotion the robot tries to manifest. The online survey is composed of six questions, the first three questions are given as short narratives with pictures, and the children are shown short videos of the robot and select the most congruent one with the narrative. The remaining questions are composed of Pepper's video clips, each displaying a basic emotion. The children are asked to guess how the robot feels in the shown video clip. The structure of the survey studies, sample questions and narratives are given in details in [73].

4.3 Test Participants

The inclusion criteria in the study were determined by the audiologists, the children:

- with typical development based on AGTE [74], Denver [75], and WISC-R [76] scales for children,
- with typical language and speech development,
- with hearing loss, and who have been using hearing aids or cochlear implants for at least 1 year,
- who do not have any other known neurological or psychological diagnosis

were included in the study.

The parents of the children were informed about the experimental procedure for the test, data acquisition setup and they signed a written consent.

The tests for the gamified auditory perception studies with tablet and robot were conducted at the university laboratory (Fig. 2(a), (b), (c)). 25 children were invited for the study but only 19 of them completed the study with the Pepper robot. 3 of the children had multiple handicaps which affected their physical and cognitive performance in the tests therefore their data were not included in the results.

The gamified setups of the auditory perception test were performed with 16 children (10 female, 6 male) aged 5 to 8 years old ($M = 6.2$ and $SD = 0.9$). 10 of the children (6 female, 4 male) have cochlear implants, and 6 of them (4 female, 2 male) use hearing aid. 15 of them had no prior experience with real-world physical robots, except one of them, an 8-year-old girl who had priorly participated in our pilot study with a smartphone-powered education robot. All the children

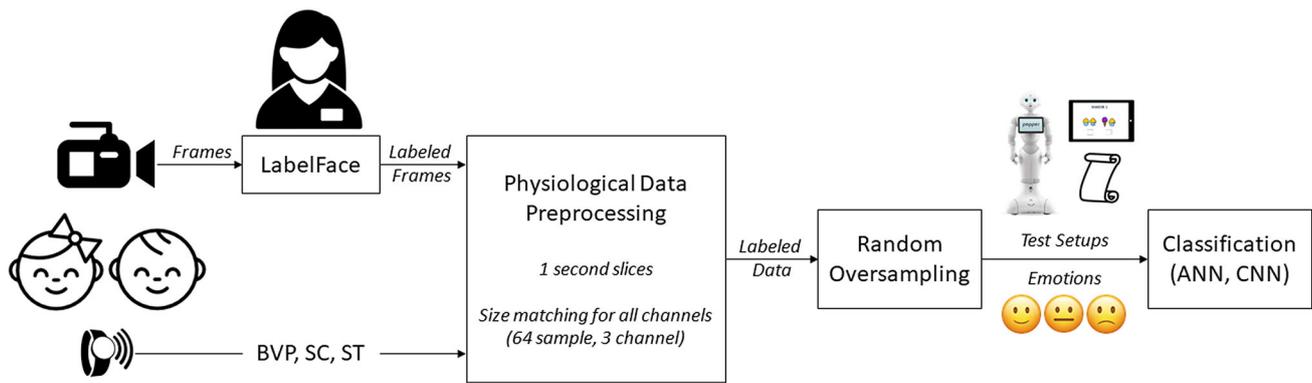


Fig. 3 Work flow of classification task using physiological signal

taking part in the study were tested in both tablet-based gamification with or without robot setups. They were tested in alternating order and with different item sets (environmental sounds or tonal patterns).

The control study with the conventional auditory perception test were conducted at the audiology clinic setup (Fig. 2(d)). The control study was planned to be performed with the same number of children but due to COVID-19 outbreak, only 6 of the families consented for their children to participate the study in an isolated and disinfected room. Therefore the conventional tests were conducted with 6 children (4 female, 2 male) aged 6 to 7 years old ($M = 6.6$ and $SD = 0.5$), 4 of them (3 female, 1 male) having cochlear implant and 2 (1 female, 1 male) of them having hearing aid. In addition to conventional test setup, the children were also tested with the gamified tablet setup. The tests were performed in alternating test order and with different item sets.

5 Classification Using Physiological Signals

5.1 Physiological Data Collection and Processing

Physiological data are collected using the E4 wristband. Skin conductivity (SC) data is acquired with an Electrodermal activity (EDA) sensor at 4 Hz. Blood volume pulse (BVP) data is acquired via the PPG sensor with a sampling rate of 64 Hz. Skin temperature (ST) data is acquired with infrared thermopile with a sampling rate of 4 Hz.

The data collected from the EDA and ST sensors were upsampled to 64 samples per second so that the number of samples of all sensors is equal. The data from all sensors were epoched into 1-second slices. Multisensor data fusion at the data-level was used to fuse the information from all sensors, then feature vectors (64x3) were obtained for each 1-second slice (Fig. 3).

5.2 Classifiers

The classifiers were trained for two classification tasks. The first one is the classification of test setups based on the captured physiological signals, and the second one is the classification of emotions based on the test setups.

We first selected commonly used machine learning methods for the classification task, which are Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Network (ANN). The implementation of the SVM classifier was based on LIBSVM [77] and used a Radial Basis Function (RBF) as kernel function. The Random Forest (RF) classifier used in the study was composed of 1000 estimators with Gini impurity criterion. The neural network architecture of the ANN classifier consisted of only 1 hidden layer with 64 units using a sigmoid function. A fully connected dense layer, which used the softmax function, was added. The loss function was selected as binary cross-entropy for binary classifications, and categorical cross-entropy was used for the multi-class classifications.

Later, we used two deep learning-based methods for the classification task: Long-Short Term Memory (LSTM) and Convolutional Neural Networks (CNN). We use physiological signals as input to our classifiers, which are time-series data. Thus, we use temporally dependent inputs, and this kind of dependency requires a memory unit, which provides capturing information from previously seen data. Therefore, we selected LSTM and adopted a classifier based on a vanilla LSTM model. The LSTM classifier had one LSTM layer with a time step of 64, connected to a dense layer for the classification task.

Besides the good performance of the CNN on image data, application of it on time-series signals such as physiological signals are also getting attention in the literature [78]. CNN yields spatially correlated outputs from the input, which means detecting salient spatial information in terms

Table 1 Classification Results of RT, RC, TC, RTC using ANN and CNN

Item set	Metric	ANN				CNN			
		RT	RC	TC	RTC	RT	RC	TC	RTC
env	Accuracy	0.990	0.980	0.747	0.809	0.988	0.976	0.753	0.817
	F1-score	0.990	0.980	0.743	0.734	0.988	0.976	0.744	0.740
ton	Accuracy	0.994	0.994	0.713	0.796	0.995	0.994	0.765	0.857
	F1-score	0.994	0.994	0.712	0.718	0.995	0.994	0.763	0.798
env+ton	Accuracy	0.989	0.984	0.645	0.747	0.991	0.985	0.619	0.750
	F1-score	0.989	0.984	0.642	0.656	0.991	0.985	0.615	0.655

of physiological signals. These spatially correlated connections contribute to a thoroughly spatial examination of the physiological signals. Therefore apart from the LSTM, we also used a CNN model as a deep learning-based classifier. The CNN model consisted of one convolutional layer with 8 filters which had kernel sizes of 8-by-2. A maximum pooling layer followed with the pool size of 64-by-1. The same arrangement as the classical neural network was set at the last layer because of the binary and multi-class classification tasks.

5.3 Classification Results

The classification performance of SVM, RF, and ANN were analyzed and compared based on their overall classification performance. Based on this analysis, the ANN model with the highest average accuracy and F1-score in all classification tasks was selected to be compared with the deep learning-based classifier with the highest performance metrics, which is the CNN model.

The selected ANN and CNN classifiers and their performance for the classification of test setups (conventional, tablet, and robot) and the classification of emotions (pleasant, unpleasant, and neutral) were given in detail in the following subsections. The performance of the classifiers (SVM, RF, and LSTM) with lower performance metrics were displayed in Appendix.

5.3.1 Classification of Test Setups

The effects of technology on children were evaluated by classifying the physiological signals captured during the three different test setup in this study (Fig. 3). The data samples were labeled considering the tools used in the experimental setups: the robot+tablet (R), the tablet (T), and no tool (conventional (C)) for different item sets (environmental sounds, tonal patterns, and both of them) The classifiers were trained and tested for both binary (robot - tablet (RT), robot - conventional (RC), tablet - conventional (TC)) and multi-class (robot - tablet - conventional (RTC)) classification. The training and test sets were prepared for three-fold cross-validation. Due to

the unbalanced numbers of samples for each class (R, T, C) of the classification tasks (RT, RC, TC, and RTC), the training dataset was augmented using the random oversampling method. The classification accuracies that were obtained for RT, RC, TC and RTC are presented in Table 1.

The results show that the classification accuracies of both ANN and CNN models were higher when robot was involved (Table 1). Both RT (environmental sounds, 99%-ANN and 98.8%-CNN and tonal patterns, 99.4%-ANN and 99.5%-CNN) and RC (environmental sounds, 98%-ANN and 97.6%-CNN and tonal patterns, 99.4%-ANN and 99.4%-CNN) were classified accurately. The results comply with the hypothesis that the robot acted as a stimulating agent and caused difference in the physiological signals of children. Since the test technology between two cases (R and T), test questions and tablet interface were identical in both cases; the interaction with the robot and the feedback from the robot, as well as the physical embodiment of the robot might cause such difference in the physiological signals compared to the other cases T and C (environmental sounds, 74.7%-ANN and 75.3%-CNN and tonal patterns, 71.3%-ANN and 76.5%-CNN). The multi-class classification (RTC) also resulted in high accuracy but with lower F1-scores. Increasing the number of children might be helpful to improve the performance of the multi-class classifiers for future studies.

5.3.2 Classification of Emotions

The results of the classification task for the test setups, presented in Table 1 in the previous subsection, comply with our assumption that the robot stimulated the children more than the other two test setups. Based on this, we decided to analyze the emotions of children during the auditory tests, especially with the robotic setup.

The emotions of the children when they were using a robot, tablet, or no tool at all (conventional setup) have been classified in this study (Fig. 3). Data samples were labeled as pleasant (P), unpleasant (U), and neutral (N) considering the emotions assigned by the psychologist in the experiment during the auditory perception tests. The classifiers were trained and tested for both binary (pleasant - unpleasant (PU), pleas-

Table 2 Classification Results of PU, PN, NU, PNU in Robot Setup using ANN and CNN

Item set	Metric	ANN				CNN			
		PU	PN	NU	PNU	PU	PN	NU	PNU
<i>R_{env}</i>	Accuracy	0.771	0.562	0.628	0.486	0.792	0.521	0.654	0.445
	Precision	0.843	0.565	0.646	0.507	0.809	0.544	0.657	0.476
	Recall	0.771	0.562	0.628	0.328	0.792	0.521	0.654	0.285
	F1-score	0.757	0.560	0.622	0.373	0.783	0.482	0.652	0.333
	Specificity	0.771	0.562	0.628	0.671	0.792	0.521	0.654	0.645
	GSP	0.782	0.562	0.630	0.394	0.792	0.506	0.654	0.355
	GSS	0.735	0.557	0.614	0.448	0.773	0.438	0.650	0.410
	MCC	0.610	0.127	0.273	0.001	0.600	0.062	0.311	-0.071
<i>R_{ton}</i>	Accuracy	0.521	0.542	0.682	0.472	0.458	0.604	0.636	0.389
	Precision	0.524	0.538	0.686	0.489	0.460	0.612	0.640	0.415
	Recall	0.521	0.542	0.682	0.314	0.458	0.604	0.636	0.249
	F1-score	0.514	0.523	0.680	0.368	0.455	0.593	0.631	0.287
	Specificity	0.521	0.542	0.682	0.650	0.458	0.604	0.636	0.584
	GSP	0.518	0.532	0.682	0.384	0.457	0.601	0.634	0.308
	GSS	0.507	0.500	0.677	0.440	0.450	0.580	0.624	0.354
	MCC	0.045	0.081	0.368	-0.035	-0.082	0.216	0.276	-0.179

Table 3 Classification Results of Emotions in Tablet and Conventional Setups using ANN and CNN

Item set	Metric	NU		Item set	Metric	PN	
		ANN	CNN			ANN	CNN
<i>T_{env}</i>	Accuracy	0.960	0.980	<i>C_{env}</i>	Accuracy	0.762	0.774
	Precision	0.963	0.979		Precision	0.785	0.846
	Recall	0.947	0.977		Recall	0.762	0.774
	F1-score	0.954	0.977		F1-score	0.758	0.760
	Specificity	0.947	0.977		Specificity	0.762	0.774
	GSP	0.954	0.978		GSP	0.766	0.784
	GSS	0.946	0.977		GSS	0.751	0.738
	MCC	0.910	0.956		MCC	0.546	0.614
<i>T_{ton}</i>	Accuracy	0.841	0.952	<i>C_{ton}</i>	Accuracy	0.795	0.487
	Precision	0.904	0.968		Precision	0.692	0.455
	Recall	0.762	0.929		Recall	0.690	0.444
	F1-score	0.789	0.943		F1-score	0.671	0.441
	Specificity	0.762	0.929		Specificity	0.690	0.444
	GSP	0.811	0.945		GSP	0.680	0.445
	GSS	0.722	0.924		GSS	0.508	0.365
	MCC	0.650	0.895		MCC	0.426	-0.100

ant - neutral (PN), neutral - unpleasant (NU)), and multi-class (pleasant - neutral - unpleasant (PNU)) classification. The training and test sets were again prepared for three-fold cross-validation. Because of the unbalanced numbers of samples for each class (P, N, U) for the classification tasks, the training dataset was again augmented using the random oversampling method.

The emotion classification accuracies using ANN and CNN were presented in Table 2 and Table 3. Since there was

few samples in emotion tasks, the number of samples belonging to different classes in the test sets remained unbalanced when trying to select as many samples as possible. Therefore, F1-score, recall, precision, specificity, Matthew’s correlation coefficient (MCC), the geometric mean of recall and precision (GSP), the geometric mean of recall and specificity (GSS) values besides classification accuracies for emotion classifications are also displayed in Table 2 and Table 3.

Table 4 The results of ANOVA for test metrics across test setups

	Mean (SD)			$F(2, 47)$	p
	Conventional	Robot	Tablet		
Test Score (%)	0.74 (0.22)	0.63 (0.18)	0.66 (0.19)	1.101	0.341
Tutorial (s)	150 (237)	190 (69)	258 (226)	1.319	0.277
Total Test (s)	461 (109)	596 (81)	502 (81)	9.169	< 0.001
Response (s)	15 (4)	20 (3)	17 (2)	9.541	< 0.001

The results showed that ANN and CNN classifiers performed best in the classification of PU in the environmental sounds with the robot (77.1%-ANN and 79.2%-CNN). However the highest classification accuracy results (68.2%-ANN and 63.6%-CNN) in the tonal patterns were obtained in NU classification for the robotic setup. Since the audiologists define tonal pattern tests as harder than the one with the environmental sounds, the children display less emotions during the tonal pattern test because they are more focused on the auditory task. Therefore, the highest accuracy in NU classification might be due to the higher number of samples in neutral and unpleasant labels than pleasant ones during the tests with the tonal patterns.

After the annotation task was finished by the experts, we also noticed that the number of samples in T and C test setups were very low for classification of emotions. Thus, we only run the NU classification for T (environmental sounds, 96%-ANN and 98%-CNN and tonal patterns, 84.1%-ANN and 95.2%-CNN) and PN classification for C setup (environmental sounds, 76.2%-ANN and 77.4%-CNN and tonal patterns, 79.5%-ANN and 48.7%-CNN) (Table 3). The reason behind the low number of labeled samples for the other emotional cases might be due to the fact that the children were less emotionally stimulated in these tests. When they used tablet, they showed only neutral or unpleasant emotions, which might be due to the fact that the children were not motivated. On the other hand, in C test setup, PN could be classified because of the high number of positive and neutral labeled samples and no unpleasant samples. This might be due to the individual motivation of the audiologist, who know the children for a long time (familiarity effect).

Note that the audiologist might be inexperienced or unfamiliar with the children in the test, which will effect the motivation and performances of the children in the conventional test setup. Thus, use of socially assistive robots and tablets may standardise the tests, and the feedback mechanism, and avoid the bias by the aforementioned causes.

6 System Evaluation of RoboRehab

The classification results for the test setup and emotional expressions of the children showed that the robotic system has a stimulating presence during the auditory perception tests. The system was assessed by objective and subjective evaluation metrics in order to explore if the gamified setups had any effect on the children performances and impressions.

6.1 Objective Evaluation: Test Metrics and Behavioral Analysis Study

The following test metrics were extracted from the interaction sessions for each child to explore if the gamified setups had any significant effect over the conventional setup:

- *Test score*: the percentage of the correctly answered questions,
- *Tutorial time*: the duration in seconds of the familiarization session with the gamified setups and the test questions,
- *Total test time*: the duration in seconds of the auditory perception test composed of 30 questions,
- *Response time*: the time difference in seconds between two answers.

A one-way ANOVA test was performed on the collected test metrics (Table 4), and Tukey and t-tests were computed to perform pairwise-comparison between the means of the 3 groups (Fig. 4). The results showed that the setup did not have any significant effect on the test scores of the children and the tutorial duration but it caused a significant difference on the total test time and response time of the children. As displayed in Fig. 4(c) and (d), the group means were significantly different between the conventional-robot setup and robot-tablet setup. The significance of the time difference is expected because the robot interacts with the children using

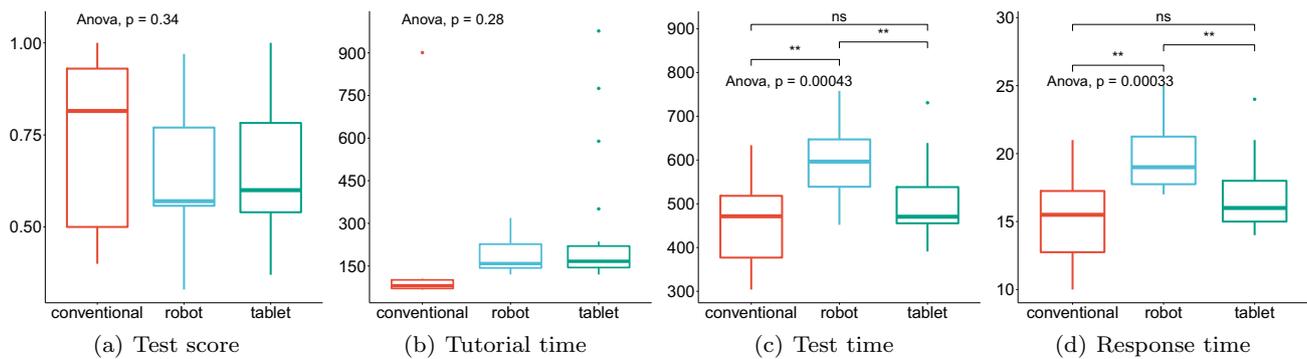


Fig. 4 The distributions of the test metrics for all the setups, and t-test significance between setup pairs: (ns) not significant, (**) are significant at $p \leq 0.01$

Table 5 The results of Welch two sample t-test for test setups based on item set

Setup	Item Set	N	Score (<i>M</i> , <i>SD</i>)	t	p
C	Env	6	0.85, 0.22	1.887	0.090
	Ton	6	0.63, 0.17		
R	Env	10	0.72, 0.16	3.311	0.002
	Ton	6	0.48, 0.10		
T	Env	14	0.75, 0.18	3.417	< 0.001
	Ton	8	0.51, 0.10		

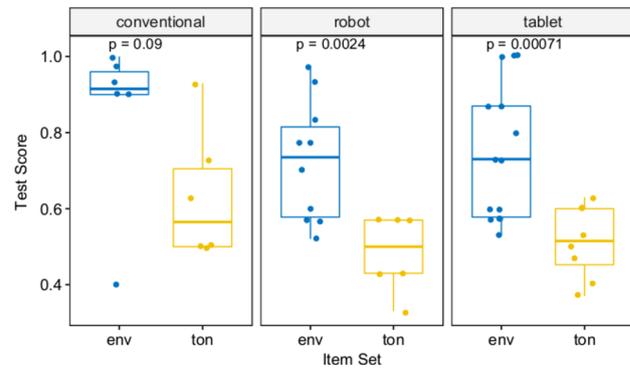


Fig. 5 The results of Welch two sample t-test for test setups based on item set

verbal and non-verbal cues to maintain their attention, therefore the interaction session takes longer.

The result of the one-way ANOVA demonstrated that test scores of the children are not affected by the test setups, in order to explore if the children’s profile (age, gender, hearing device) or auditory perception test-related factors (test order and item set) had any significant effect on the scores, further analysis were conducted. The results of ANOVA showed that only the item set (tonal pattern/environmental test) had a significant effect on the test scores in the gamified setups. The mean scores and number of test participants for both setup and item set are summarized in Table 5. One-tailed paired t-tests showed that the children’s scores in tests with environmental sounds item set were significantly higher than the test with tonal patterns for both test setup, as displayed in Fig. 5.

The audiologists comment that this difference in the scores is due to the nature of the auditory perception test, the auditory stimuli in the tonal pattern item set is harder to perceive and identify than the stimuli in the environmental sound set [79].

In order to eliminate the effect of the test-related factors on the test setup, a group of 6 children were tested with the same item set in the same test order for all the setups. They were first tested with the robotic setup, then the tablet and finally

in a conventional setup with a time difference designated by the audiologists between sessions.

One-way ANOVA test showed that the test setup did not have a significant effect on the test scores, total test duration and response time of the 6 children, it had only an effect on the tutorial duration (Table 6). The paired t-test computed for the pairwise comparison between the test setups showed that the tutorial time differed significantly between the conventional-robot and conventional-tablet setups (Fig. 6(b)). This difference is based on the fact that the tutorial of the conventional setup incorporates only instructions about the test questions and do not have any extra information about digital setup/media (i.e. robot and tablet). The t-test results also showed that, similar to the previous analysis on the all participants, the group means of total test duration (Fig. 6(c)) and response time (Fig. 6(d)) were significantly different between the conventional-robot setup and robot-tablet setup. As mentioned before, the significance was issued by the verbal and non-verbal feedback of the robot during the interaction session.

As a summary, the statistical analysis showed no significant difference on the test scores for the setups, but the timing characteristics differed significantly, as expected due to the

Table 6 The ANOVA results for test metrics of the group tested with the same item set in all the test setups

	Mean (SD)			$F(2, 16)$	p
	Conventional	Robot	Tablet		
Test Score (%)	0.75 (0.23)	0.67 (0.2)	0.72 (0.27)	0.177	0.84
Tutorial (s)	89 (15)	152 (38)	150 (23)	10.53	0.001
Total Test (s)	450 (113)	555 (47)	481 (66)	2.684	0.101
Response (s)	15 (4)	18 (2)	16 (2)	2.591	0.108

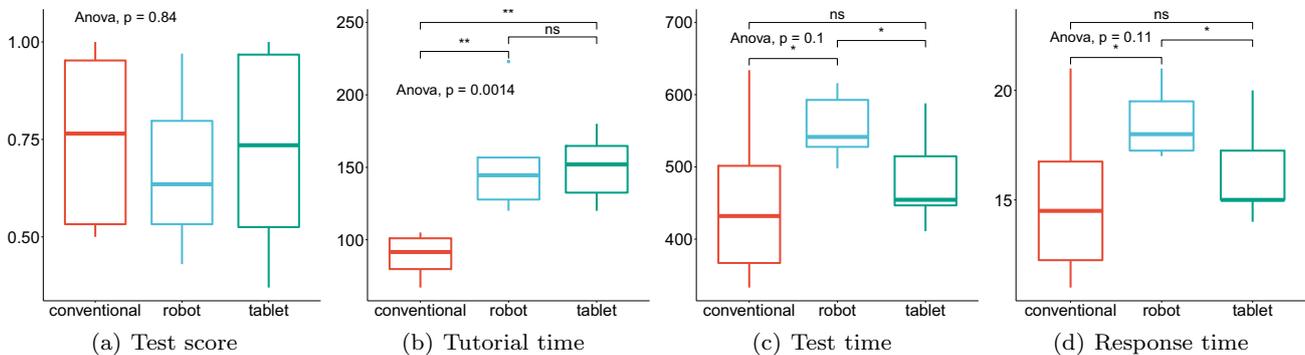


Fig. 6 The distributions of the test metrics of the group tested with the same item set in all the setups, and t-test significance between setup pairs: (ns) not significant, (*) and (**) are significant at $p \leq 0.05$ and $p \leq 0.01$, respectively

feedback mechanism of the robot. In order to investigate if the robot had any significant effect on the engagement of children despite the longer duration of the interaction session, a behavioral analysis study was conducted on the video recordings of the children tested with the same item set on all the setups. The videos were annotated by the start and end times of each child's speech, gaze direction, gestures (smiling, mimicking the robot) and boredom signs (sighing, yawning, fidgeting, etc.). Unfortunately, due to the fact that the conventional study were performed after the COVID-19 outbreak and all the children wear facial mask and protective clothing, the behavioral analysis results were inconclusive for the comparison of the conventional setup with the gamified setups. But a previous behavioral annotation study performed on the tablet and robot setups showed that there was no significant difference between both setup in terms of children's speech, their gaze direction and boredom signs. On the other hand, the behavioral coding revealed that children smiled and talked to the robot and mimicked it whereas there were no annotated occurrence of these behaviours in the tablet setup. The children's interactive engagement with the robot may be interpreted as a higher involvement with the game due to the presence of the robot, however the results are not significantly representative because of the small sample size.

6.2 Subjective Evaluation: Impressions of the Children About the Robot

Two survey studies were conducted: (1) to evaluate the children's perception of the robot, and, (2) to validate if the children correctly interpreted the feedback given by the robot as emotional behaviors during the auditory perception game. (For further details on both surveys, please refer to [73].)

The first survey about the perceived character, intelligence and likability of the robot revealed that the children were excited to see the robot (Q1) and they liked it very much. The results showed that the children liked playing with the robot (Q5), would like to play again with it (Q6), and would like to be friends with the robot (Q7). They also reported that they thought the robot was intelligent (Q10) and funny (Q12). Most of the children remarked that they would like to have a similar robot at home (Q8) but they were not in agreement about the presence of the robot in their school (Q9a).

The children were also asked open-ended "why" questions. When they were asked about what they did like the most about the robot (Q3), some of them mentioned only the physical properties of the robot saying "its hands" and "its face", on the other hand some of the children answered "I liked all of it", "I liked its being a robot" and "I liked the most

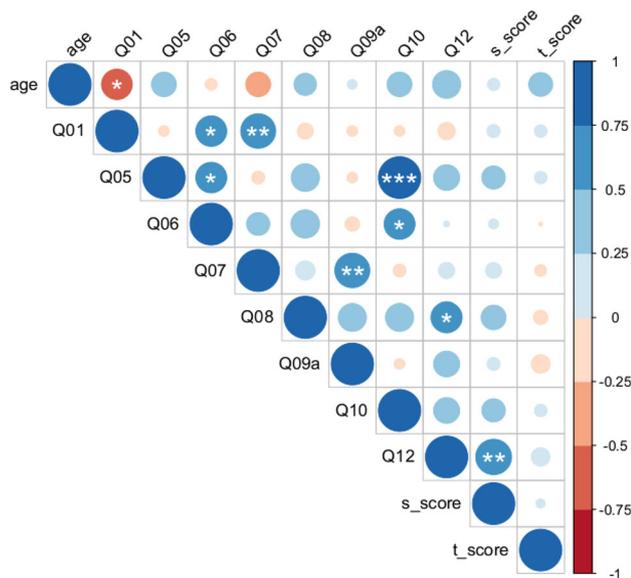


Fig. 7 Correlation among different factors on impressions of children: Correlations with (*), (**) and (***) are significant at $p \leq 0.05$, $p \leq 0.01$ and $p \leq 0.001$ respectively

its saying bravo to me when I answered correctly” . All the children were in agreement on that there was nothing about it they did not like (Q4). And lastly, when they are asked what they would teach the robot if they were the one “teaching it to do something” , they answered saying: “I would teach it to move like me” , “the alphabet” , “how to draw hearts” , “how to help me with my homework and studies” , and “to play my favorite games” .

The second survey on the interpretability of the robot’s affective behaviours showed that the children were able to match a happy, sad or surprised context with a congruent emotional behaviour and identify accurately the emotions displayed by the robot.

The answers of the 16 children taking the auditory test in robotic setup are encoded and quantified for both survey. The second survey were scored out of 100, the children’s scores varying from 0.33 to 1 ($M = 0.75$ and $SD = 0.21$) were integrated into a correlation analysis to examine the correlations among different factors influencing children’s impressions about the robot. The correlated factors are displayed in Fig. 7.

While the negatively correlated items showed that younger children were more excited to see the robot, the positively correlated items revealed that the children who described themselves as “excited to see the robot” would like to be friends with the robot and play again with it. The children who liked to play with the robot and would like to continue to play with it, found the robot intelligent and funny. There was also a positive correlation with the funny aspect of the robot and the children’s desire to have a similar robot in their home. Another positive correlation were found between the

answers of the children who would like to be friends with the robot and the children desiring to see the robot in their schools (Q9a). The presence of the robot in the school was the most disagreed item on the questionnaire, further analysis on the open-ended questions revealed that the answers of the children who do not want to see the robot in the school were dependent on the perceived character of the robot, i.e the children who perceived the robot as “a teacher” rather than “a friend” , said that they do not want the robot in their school [73].

The correlation study results also revealed that children who correctly interpreted robot’s behaviour in the second survey (having a higher s_score) found the robot more funny than the others.

Another aspect revealed by the correlation analysis was that the auditory perception test score (t_score) of the children did not have any significant correlation with any of the other factors. It is a promising outcome because it reveals that even though the children do not perform well during the auditory test, they do not relate it with their experience with the robot and their impressions about it.

7 Conclusion & Future Work

RoboRehab is an assistive robotic system enhanced with an affective module for children with hearing disabilities. The system is designed and developed for the audiometry tests and rehabilitation of children in clinic settings. The system involves a social robot Pepper, tablet, an interface specially designed for the verbal and nonverbal interaction of children with the robot, gamification of the tests, sensory setup and a machine learning based emotion recognition module to achieve this goal.

This paper presents a user study involving 16 children with hearing disabilities (using cochlear implant or hearing aid). Three scenarios involving conventional setup, setup with robot+tablet and setup with only tablet conditions are tested, and the results including detailed analysis of physiological signals captured by E4, performances of children during the auditory perception tests, and self-report questionnaire are presented.

An affective module based on facial and physiological data is being developed in the project. This study presents the results of the physiological data which is collected by E4 smart wristband and labeled by the psychologists involved in the project using the facial expressions of the children during the tests. Both traditional machine learning methods (ANN) and deep learning models (CNN) are used on the E4 data to classify the three test setups (conventional, tablet, robot). The results show that the robotic setup is well separated with high classification accuracy from tablet and conventional test setups for the auditory tests with both item set. The results

Table 7 Classification Results of RT, RC,TC, RTC using SVM, RF and LSTM

Item	Metric	SVM				RF				LSTM			
		RT	RC	TC	RTC	RT	RC	TC	RTC	RT	RC	TC	RTC
env	Accuracy	0.991	0.926	0.792	0.754	0.805	1.000	0.663	0.716	0.991	0.982	0.744	0.791
	F1-score	0.990	0.920	0.775	0.759	0.793	1.000	0.614	0.625	0.991	0.982	0.740	0.711
ton	Accuracy	0.945	0.999	0.675	0.838	0.717	0.926	0.588	0.599	0.991	0.995	0.719	0.794
	F1-score	0.948	0.999	0.536	0.829	0.691	0.926	0.541	0.564	0.991	0.995	0.717	0.714
env+ton	Accuracy	0.826	0.796	0.757	0.620	0.692	0.644	0.715	0.373	0.992	0.991	0.616	0.722
	F1-score	0.821	0.795	0.748	0.609	0.658	0.593	0.681	0.248	0.992	0.991	0.612	0.620

Table 8 Classification Results of PU, PN, NU, PNU in Robot Setup using SVM, RF and LSTM

Item	Metric	SVM				RF				LSTM			
		PU	PN	NU	PNU	PU	PN	NU	PNU	PU	PN	NU	PNU
R_{env}	Accuracy	0.521	0.524	0.397	0.384	0.690	0.963	0.678	0.629	0.750	0.521	0.564	0.319
	Precision	0.580	0.543	0.394	0.242	0.691	0.968	0.708	0.733	0.765	0.524	0.564	0.262
	Recall	0.521	0.524	0.397	0.287	0.690	0.963	0.678	0.537	0.750	0.521	0.564	0.205
	F1-score	0.499	0.477	0.392	0.221	0.688	0.962	0.668	0.516	0.744	0.479	0.564	0.215
	Specificity	0.521	0.524	0.397	0.619	0.690	0.963	0.678	0.782	0.750	0.521	0.564	0.504
	GSP	0.524	0.503	0.394	0.241	0.689	0.964	0.681	0.561	0.751	0.500	0.564	0.224
	GSS	0.440	0.428	0.387	0.149	0.683	0.961	0.655	0.548	0.736	0.435	0.563	0.234
	MCC	0.094	0.064	-0.208	-0.155	0.381	0.930	0.385	0.351	0.515	0.045	0.129	-0.338
R_{ton}	Accuracy	0.553	0.414	0.461	0.457	0.762	0.545	0.754	0.589	0.625	0.521	0.667	0.403
	Precision	0.568	0.407	0.456	0.575	0.772	0.592	0.828	0.619	0.697	0.441	0.702	0.378
	Recall	0.553	0.414	0.461	0.312	0.762	0.545	0.754	0.423	0.625	0.521	0.667	0.256
	F1-score	0.546	0.403	0.455	0.334	0.750	0.529	0.715	0.465	0.588	0.457	0.657	0.293
	Specificity	0.553	0.414	0.461	0.649	0.762	0.545	0.754	0.749	0.625	0.521	0.667	0.581
	GSP	0.553	0.407	0.457	0.361	0.758	0.548	0.745	0.492	0.618	0.469	0.670	0.304
	GSS	0.536	0.390	0.448	0.391	0.736	0.509	0.674	0.527	0.550	0.338	0.645	0.335
	MCC	0.121	-0.178	-0.083	-0.025	0.534	0.132	0.554	0.194	0.298	0.045	0.366	-0.190

also show that the tablet and conventional setup are distinguished from each other but the accuracy of the classifiers is not as high as in the robotic setup. These finding answers the first research question (RQ1) about the effect of the different test setups on the physiological signals of children.

To answer the second research question (RQ2), the physiological signals are used with ANN and CNN models to recognize the emotions of children (positive (pleasant), negative (unpleasant) or neutral) for the three test setups. The results reveal that the children's positive and negative emotions can be distinguished more accurately when they interact with the robot, than the other two cases in the environmental sounds test. However, negative and neutral emotions of children seem to be more distinguishable in the tonal pattern test. Since tonal pattern tests were harder than the environmental sounds tests, the children might feel less positive emotions during the tonal pattern test. Furthermore, negative and neutral emotions of children seem to be more distinguishable

in both environmental and tonal pattern test for tablet setup, which may infer that children get bored and are not much motivated. Pleasant and neutral emotions of children seem to be more distinguishable in both environmental sounds and tonal pattern tests for conventional setup. This is expected since children feel comfortable near audiologist whom they are familiar with and who are motivating them.

The system evaluation based on objective test metrics (RQ3) demonstrate that the test scores of children in tablet and robot setups were as good as the conventional tests with human audiologist. Moreover, the results show that the time spent with the robot is significantly higher than the conventional test but it is observed by the audiologists and experimenters that in the robotic setup, the children were more involved and showed attention to the test than the other setups. The subjective evaluation of the audiologist comply with the findings of previous studies: In robotic studies, even though the test performances of participants do not differ

Table 9 Classification Results of Emotions in Tablet and Conventional Setups using SVM, RF and LSTM

Item set	Metric	NU			Item set	Metric	PN		
		SVM	RF	LSTM			SVM	RF	LSTM
T_{env}	Accuracy	0.983	0.995	0.869	C_{env}	Accuracy	0.941	1.000	0.726
	Precision	0.988	0.996	0.869		Precision	0.947	1.000	0.827
	Recall	0.974	0.992	0.894		Recall	0.941	1.000	0.726
	F1-score	0.980	0.994	0.864		F1-score	0.940	1.000	0.695
	Specificity	0.974	0.992	0.894		Specificity	0.941	1.000	0.726
	GSP	0.981	0.994	0.873		GSP	0.942	1.000	0.733
	GSS	0.974	0.992	0.887		GSS	0.938	1.000	0.657
	MCC	0.962	0.988	0.763		MCC	0.888	1.000	0.537
T_{ton}	Accuracy	0.667	0.667	0.714	C_{ton}	Accuracy	0.385	0.701	0.769
	Precision	0.333	0.333	0.687		Precision	0.206	0.516	0.710
	Recall	0.500	0.500	0.738		Recall	0.308	0.514	0.625
	F1-score	0.400	0.400	0.679		F1-score	0.242	0.438	0.607
	Specificity	0.500	0.500	0.738		Specificity	0.308	0.514	0.625
	GSP	0.408	0.408	0.695		GSP	0.249	0.467	0.635
	GSS	0.000	0.000	0.585		GSS	0.029	0.098	0.402
	MCC	0.000	0.000	0.457		MCC	-0.409	0.082	0.358

significantly between the robotic and conventional setups and the test duration increases when the robot is present, the results report higher engagement and higher cooperation times by the participants [80,81]. It has also been reported in the literature that human - robot interaction increases participation in a given task [82,83]. Therefore, we expect that the proposed method will facilitate the children's experience during their auditory tests, increase the interaction and cooperation time and improve the interaction quality, especially for individuals with additional disabilities, who have short attention and cooperation time.

The self-report surveys conducted after auditory tests with the robotic setup show that the robot is accepted as an intelligent and funny social entity by the children. Although there was no significant difference of using a robot in terms of test scores, the subjective evaluation of the robotic setup reveal that the children were excited to see the robot and happy to play with it (RQ4). Additionally, the audiologists reported that the children were more motivated and engaged in the digitalized setups than the conventional setup. This is a motivating preliminary step to use these affective robots as a part of audiometry testing.

The overall results show that the robot can stimulate the children's emotions and cause difference in their physiological signals and machine learning approaches can be used to understand the emotions of children during their interaction with a socially assistive robot. To the best of our knowledge, this project is the first study towards an affective assistant robot for children with hearing disabilities as part of their rehabilitation and audiometry testing process. We expect this will encourage more research in this field, provide new ways

to address the emotional needs of children through their clinic interventions and improve their clinic experience. Especially in times of pandemics such technology assisted approaches supported by machine learning methods for rehabilitation will be very helpful for the vulnerable user groups.

Acknowledgements We would like to thank collaborating audiologists Dr. Selma Yilar, Talha Cogen and Busra Gokce from Istanbul University Cerrahpasa Medical Faculty for their contributions to this study. This study is supported by The Scientific and Technological Research Council of Turkey (TÜBİTAK) under the grant number 118E214. This work is supported by the Turkish Academy of Sciences in scheme of the Outstanding Young Scientist Award (TÜBA-GEBİP).

Appendix

The performance metrics of the classifiers for the classification of test setups, and the classification of pleasant, neutral and unpleasant emotions based on the test setup are displayed in Table 7, 8, and 9, respectively.

References

1. Pennisi P, Tonacci A, Tartarisco G, Billeci L, Ruta L, Gangemi S, Pioggia G (2016) Autism and social robotics: a systematic review. *Autism Res* 9(2):165–183
2. Martinez-Martin E, Escalona F, Cazorla M (2020) Socially assistive robots for older adults and people with autism an overview. *Electronics*. <https://doi.org/10.3390/electronics9020367>
3. Kose H, Akalin N, Uluer P (2014) Socially interactive robotic platforms as sign language tutors. *Int J Human Robot* 11(01):1450003. <https://doi.org/10.1142/S0219843614500030>

4. Henschel A, Laban G, Cross ES (2021) What makes a robot social? a review of social robots from science fiction to a home or hospital near you. *Social Robot Rep* 2:9–19
5. Robinson NL, Cottier TV, Kavanagh DJ (2019) Psychosocial health interventions by social robots: systematic review of randomized controlled trials. *J Med Internet Res* 21(5):e13203
6. Stower R, Calvo-Barajas N, Castellano G, Kappas A (2021) A meta-analysis on children's trust in social robots. *Int J Soc Robot*. 1–23
7. Cifuentes CA, Pinto MJ, Céspedes N, Múnera M (2020) Social robots in therapy and care. *Current Robot Rep* 1:57–74. <https://doi.org/10.1007/s43154-020-00009-2>
8. Kabacińska K, Prescott TJ, Robillard JM (2020) Socially assistive robots as mental health interventions for children: a scoping review. *Int J Soc Robot* 1–17
9. Ferrante G, Vitale G, Licari A, Montalbano L, Pilato G, Infantino I, Augello A, La Grutta S (2020) Social robots and therapeutic adherence a new challenge in pediatric asthma? *Paediat Resp Rev*. <https://doi.org/10.1016/j.prrv.2020.11.001>
10. Logan DE, Breazeal C, Goodwin MS, Jeong S, O'Connell B, Smith-Freedman D, Heathers J, Weinstock P (2019) Social robots for hospitalized children. *Pediatrics*. <https://doi.org/10.1542/peds.2018-1511>
11. Moerman CJ, Jansens RM (2020) Using social robot pleo to enhance the well-being of hospitalised children. *J Child Health Care* 1367493520947503. <https://doi.org/10.1177/1367493520947503>
12. Spezialetti M, Placidi G, Rossi S (2020) Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI* 7:145
13. Ekman P (1992) An argument for basic emotions. *Cognit Emot* 6(3–4):169–200
14. Plutchik R (2001) The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Am Sci* 89(4):344–350
15. Izard CE (2009) Emotion theory and research: highlights, unanswered questions, and emerging issues. *Ann Rev Psychol* 60:1–25
16. Lang PJ (1995) The emotion probe: studies of motivation and attention. *Am Psychol* 50(5):372
17. Levenson RW (1994) Human emotion: a functional view. In: Ekman P, Davidson R (eds) *The nature of emotion: Fundamental questions*. Oxford University Press, New York, NY, pp 123–126
18. Lazarus RS (1994) Universal antecedents of the emotions. In: Ekman P, Davidson R (eds) *The nature of emotion: Fundamental questions*. Oxford University Press, New York, NY, pp 163–171
19. Scherer KR (1994) Evidence for both universality and cultural specificity of emotion elicitation. In: Ekman P, Davidson R (eds) *The nature of emotion: fundamental questions*. Oxford University Press, New York, NY, pp 172–175
20. Grabowski K, Rynkiewicz A, Lassalle A, Baron-Cohen S, Schuller B, Cummins N, Baird A, Podgórska-Bednarz J, Pieniążek A, Łucka I (2019) Emotional expression in psychiatric conditions: new technology for clinicians. *Psyc Clin Neurosci* 73(2):50–62
21. Hassouneh, A., Mutawa, A., Murugappan, M.: Development of a real-time emotion recognition system using facial expressions and eeg based on machine learning and deep neural network methods. *Informatics in Medicine Unlocked* p. 100372 (2020)
22. Laban G, George JN, Morrison V, Cross ES (2021) Tell me more! assessing interactions with social robots from speech. *Paladyn J Behav Robot* 12(1):136–159. <https://doi.org/10.1515/pjbr-2021-0011>
23. Gonuguntla, V., Shafiq, G., Wang, Y., Veluvolu, K.C.: Eeg classification of emotions using emotion-specific brain functional network. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2896–2899. IEEE (2015)
24. Jatupaiboon N, Pan-ngum S, Israsena P (2013) Real-time eeg-based happiness detection system. *The Sci World J*
25. Khosrowabadi R, Quek C, Ang KK, Wahab A (2013) Ernn: a biologically inspired feedforward neural network to discriminate emotion from eeg signal. *IEEE Trans eural Netw Learn Syst* 25(3):609–620
26. López-Gil JM, Virgili-Gomá J, Gil R, Guilera T, Batalla I, Soler-González J, García R (2016) Method for improving eeg based emotion recognition by combining it with synchronized biometric and eye tracking technologies in a non-invasive and low cost way. *Front Comput Neurosci* 10:85
27. Agrafioti F, Hatzinakos D, Anderson AK (2011) Ecg pattern analysis for emotion detection. *IEEE Trans Affect Comput* 3(1):102–115
28. van Dooren M, Janssen JH et al (2012) Emotional sweating across the body: comparing 16 different skin conductance measurement locations. *Physiol ehav* 106(2):298–304
29. Goshvarpour A, Abbasi A (2017) An emotion recognition approach based on wavelet transform and second-order difference plot of ecg. *J AI Data Min* 5(2):211–221
30. Dissanayake T, Rajapaksha Y, Ragel R, Nawinne I (2019) An ensemble learning approach for electrocardiogram sensor based human emotion recognition. *Sensors* 19(20):4495
31. Marín-Morales J, Higuera-Trujillo JL, Greco A, Guixeres J, Llinares C, Scilingo EP, Alcañiz M, Valenza G (2018) Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. *Sci Rep* 8(1):1–15
32. Liapis A, Katsanos C, Sotiropoulos D, Xenos M, Karousos N (Springer (2015).) Recognizing emotions in human computer interaction: studying stress using skin conductance. pp 255–262
33. Greco A, Lanata A, Citi L, Vanello N, Valenza G, Scilingo EP (2016) Skin admittance measurement for emotion recognition: a study over frequency sweep. *Electronics* 5(3):46
34. Bruno P, Melnyk V, Völckner F (2017) Temperature and emotions: effects of physical temperature on responses to emotional advertising. *Int J Res Market* 34(1):302–320
35. Choi JS, Bang JW, Heo H, Park KR (2015) Evaluation of fear using nonintrusive measurement of multimodal sensors. *Sensors* 15(7):17507–17533
36. Dzedzickis A, Kaklauskas A, Bucinskas V (2020) Human emotion recognition: review of sensors and methods. *Sensors* 20(3):592
37. Goulart C, Valadão C, Delisle-Rodriguez D, Caldeira E, Bastos T (2019) Emotion analysis in children through facial emissivity of infrared thermal imaging. *PloS one* 14(3):e0212928
38. Choi Y, Jeon YM, Wang L, Kim K (2017) A biological signal-based stress monitoring framework for children using wearable devices. *Sensors* 17(9):1936
39. Fletcher RR, Dobson K, Goodwin MS, Eydgahi H, Wilder-Smith O, Fernholz D, Kuboyama Y, Hedman EB, Poh MZ, Picard RW (2010) icalm: wearable sensor and network architecture for wirelessly communicating and logging autonomic activity. *IEEE Trans Inf Technol Biomed* 14(2):215–223
40. Feng H, Golshan HM, Mahoor MH (2018) A wavelet-based approach to emotion classification using eda signals. *Exp Syst Appl* 112:77–86
41. Garbarino M, Lai M, Bender D, Picard R, Tognetti S (2014) Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. pp 39–42. <https://doi.org/10.1109/MOBIHEALTH.2014.7015904>
42. Menghini L, Gianfranchi E, Cellini N, Patron E, Tagliabue M, Sarlo M (2019) Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions. *Psychophysiol* 56(11):e13441
43. Zhao, B., Wang, Z., Yu, Z., Guo, B.: Emotionsense: emotion recognition based on wearable wristband. In: 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smart-

- World/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI), pp. 346–355. IEEE (2018)
44. Ollander S (2015) Wearable sensor data fusion for human stress estimation. Linköping University (**Ph.D. thesis**)
 45. Kikhia B, Stavropoulos TG, Andreadis S, Karvonen N, Kompatsiaris I, Sävenstedt S, Pijl M, Melander C (2016) Utilizing a wristband sensor to measure the stress level for people with dementia. *Sensors* 16(12):1989
 46. Gouverneur P, Jaworek-Korjakowska J, Köping L, Shirahama K, Kleczek P, Grzegorzec M (Springer (2017).) Classification of physiological data for emotion recognition. pp 619–627
 47. Ollander S, Godin C, Campagne A, Charbonnier S (2016). A comparison of wearable and stationary sensors for stress detection. *IEEE*, pp 004362–004366
 48. Can YS, Chalabianloo N, Ekiz D, Ersoy C (2019) Continuous stress detection using wearable sensors in real life: algorithmic programming contest case study. *Sensors* 19(8):1849
 49. Jalan U (2020) Four-class emotion classification using electrocardiography (ecg) in virtual reality (vr). *Int J Adv Sci Technol* 29(6):1523–1529
 50. Bulagang, A.F., Mountstephens, J., Wi, J.T.T.: Tuning support vector machines for improving four-class emotion classification in virtual reality (vr) using heart rate features. In: *Journal of Physics: Conference Series*, vol. 1529, p. 052069. IOP Publishing (2020)
 51. Jang EH, Park BJ, Park MS, Kim SH, Sohn JH (2015) Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. *J Physiol Anthropol* 34(1):25
 52. Ali M, Al Machot F, Haj Mosa A, Jdeed M, Al Machot E, Kyamakya K (2018) A globally generalized emotion recognition system involving different physiological signals. *Sensors* 18(6):1905
 53. Alarcao SM, Fonseca MJ (2017) Emotions recognition using eeg signals: A survey. *IEEE Transactions on Affective Computing*
 54. Mendoza-Palechor F, Menezes ML, Sant'Anna A, Ortiz-Barrios M, Samara A, Galway L (2019) Affective recognition from eeg signals: an integrated data-mining approach. *J Amb Intell Human Comput* 10(10):3955–3974
 55. Seo J, Laine TH, Sohn KA (2019) Machine learning approaches for boredom classification using eeg. *J Amb Intell Human Comput*. 1–16
 56. Domínguez-Jiménez J, Campo-Landines K, Martínez-Santos J, Delahoz E, Contreras-Ortiz S (2020) A machine learning model for emotion recognition from physiological signals. *Biomed Sig Process Cont*. 55:101646
 57. Supratak, A., Wu, C., Dong, H., Sun, K., Guo, Y.: Survey on feature extraction and applications of biosignals. In: *Machine Learning for Health Informatics*, pp. 161–182. Springer (2016)
 58. Ganapathy N, Swaminathan R, Deserno TM (2018) Deep learning on 1-d biosignals: a taxonomy-based survey. *Yearbook Med Inform* 27(01):098–109
 59. Rim B, Sung NJ, Min S, Hong M (2020) Deep learning in physiological signal data: a survey. *Sensors* 20(4):969
 60. Yin Z, Zhao M, Wang Y, Yang J, Zhang J (2017) Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Comput Meth Prog Biomed* 140:93–110
 61. Hammerla, N.Y., Halloran, S., Plötz, T.: Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880* (2016)
 62. Ordóñez F, Roggen D (2016) Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16(1):115
 63. Santamaria-Granados L, Munoz-Organero M, Ramirez-Gonzalez G, Abdulhay E, Arunkumar N (2018) Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos). *IEEE Access* 7:57–67
 64. Tripathi, S., Acharya, S., Sharma, R.D., Mittal, S., Bhattacharya, S.: Using deep and convolutional neural networks for accurate emotion classification on deap dataset. In: *Twenty-ninth IAAI conference* (2017)
 65. Kanjo E, Younis EM, Ang CS (2019) Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Inf Fus* 49:46–56
 66. Hammal Z, Chu WS, Cohn JF, Heike C, Speltz ML (2017). Automatic action unit detection in infants using convolutional neural network. *IEEE, IEEE*, San Antonio, TX, USA, pp 216–221
 67. Rao, A., Ajri, S., Guragol, A., Suresh, R., Tripathi, S.: Emotion recognition from facial expressions in children and adults using deep neural network. In: *Intelligent Systems, Technologies and Applications*, pp. 43–51. Springer (2020)
 68. Gümüşlü E, Erol Barkana D, Köse H (2020) Emotion recognition using eeg and physiological data for robot-assisted rehabilitation systems. pp 379–387
 69. Baglayici, E., Gurpinar, C., Uluer, P., Kose, H.: A new facial expression processing system for an affectively aware robot. In: *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, pp. 36–51. Springer International Publishing (2021)
 70. Reynolds CR, Voress JK, Pearson NA (2008) DTAP: Developmental Test of Auditory Perception. Pro-Ed
 71. Nomura, T., Kanda, T., Suzuki, T., Kato, K.: Psychology in human-robot communication: an attempt through investigation of negative attitudes and anxiety toward robots. In: *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, pp. 35–40 (2004). <https://doi.org/10.1109/ROMAN.2004.1374726>
 72. Nomura T, Suzuki T, Kanda T, Kato K (2006) Measurement of anxiety toward robots. pp 372–377. <https://doi.org/10.1109/ROMAN.2006.314462>
 73. Uluer, P., Kose, H., Oz, B.K., Can Aydinalev, T., Barkana, D.E.: Towards an affective robot companion for audiology rehabilitation: How does pepper feel today? In: *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 567–572 (2020). <https://doi.org/10.1109/RO-MAN47096.2020.9223534>
 74. Savaşır I, Sezgin N, Erol N (1995) Ankara gelişim tarama envanteri el kitabı (turkish). Ankara Üniversitesi Tıp Fakültesi, Ankara
 75. Frankenburg W, Dodds J, Archer P (1990) Denver I.I. Denver Developmental Materials. Inc
 76. Wechsler D (1974) Wechsler intelligence scale for children-revised. Psychological Corporation
 77. Chang CC, Lin CJ (2011) Libsvm: a library for support vector machines. *ACM Trans Intell Syst Technol (TIST)* 2(3):1–27
 78. Kanjo, E., Younis, E., Ang, C.S.: Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion* p. 33 (2019). <https://doi.org/10.1016/j.inffus.2018.09.001>
 79. Cinar M, Polat Z (2018) Normal işitmeye sahip bireylerde işitsel algı becerilerinin değerlendirilmesi. Master's thesis, İstanbul Üniversitesi ((in Turkish))
 80. Huijnen CA, Lexis MA, Jansens R, de Witte LP (2017) How to implement robots in interventions for children with autism? a co-creation study involving people with autism, parents and professionals. *J Autism Develop Disord* 47(10):3079–3096
 81. Yun SS, Choi J, Park SK, Bong GY, Yoo H (2017) Social skills training for children with autism spectrum disorder using a robotic behavioral intervention system. *Autism Res* 10(7):1306–1323
 82. Powers A, Kiesler S, Fussell S, Torrey C (2007) Comparing a computer agent with a humanoid robot. pp 145–152

83. Köse H, Uluç P, Akalın N, Yorgancı R, Özkul A, Ince G (2015) The effect of embodiment in sign language tutoring with assistive humanoid robots. *Int J Soc Robot* 7(4):537–548

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.