







## Research Article

# Estimating Children Engagement Interacting with Robots in Special Education Using Machine Learning

George A. Papakostas <sup>1</sup>, George K. Sidiropoulos <sup>1</sup>, Chris Lytridis <sup>1</sup>,  
Christos Bazinas <sup>1</sup>, Vassilis G. Kaburlasos <sup>1</sup>, Efi Kourampa,<sup>2</sup> Elpida Karageorgiou,<sup>2</sup>  
Petros Kechayas,<sup>3</sup> and Maria T. Papadopoulou <sup>4</sup>

<sup>1</sup>Human-Machines Interaction Laboratory (HUMAIN-Lab), Department of Computer Science, International Hellenic University, 65404 Kavala, Greece

<sup>2</sup>Family Center KPG, 54352 Thessaloniki, Greece

<sup>3</sup>Department of Clinical Psychology, Papageorgiou General Hospital, Aristotle University of Thessaloniki, 56403 Thessaloniki, Greece

<sup>4</sup>Division of Child Neurology and Metabolic Disorders, 4th Department of Pediatrics, Papageorgiou General Hospital, Aristotle University of Thessaloniki, 56403 Thessaloniki, Greece

Correspondence should be addressed to George A. Papakostas; [gpapak@cs.ihu.gr](mailto:gpapak@cs.ihu.gr)

Received 10 March 2021; Accepted 30 May 2021; Published 19 June 2021

Academic Editor: Bhawani Shankar Chowdhry

Copyright © 2021 George A. Papakostas et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The task of child engagement estimation when interacting with a social robot during a special educational procedure is studied. A multimodal machine learning-based methodology for estimating the engagement of the children with learning difficulties, participating in appropriate designed educational scenarios, is proposed. For this purpose, visual and audio data are gathered during the child-robot interaction and processed towards deciding an engaged state of the child or not. Six single and three ensemble machine learning models are examined for their accuracy in providing confident decisions on in-house developed data. The conducted experiments revealed that, using multimodal data and the AdaBoost Decision Tree ensemble model, the children's engagement can be estimated with 93.33% accuracy. Moreover, an important outcome of this study is the need for explicitly defining the different engagement meanings for each scenario. The results are very promising and put ahead of the research for closed-loop human centric special education activities using social robots.

## 1. Introduction

Nowadays, we are witnessing the fourth industrial revolution commonly known in Europe as Industry 4.0 [1]. One of the most important parts of this revolution is the extension of the robots' usage beyond the industrial environments to social activities interacting directly with humans. This new kind of robot named social robots shows increased interaction capabilities, characterized by a certain degree of intelligence, and is very much safe to interact with children in any type of education.

Our interest here is the case of special education, which draws increased attention from modern societies aiming at

providing equal opportunities to children with special needs to develop their skills. Recent studies have demonstrated the positive role of social robots in delivering special education in person [2, 3] as well as in distance [4].

The ultimate goal of an advanced child-robot interaction is the establishment of a high level of an intelligence communication channel, in a closed-loop configuration with the child being at the center of the educational scenario. This goal can be achieved by developing efficient sensing mechanisms to the robot side, such as automatic engagement measuring, which will permit the robot to adapt its behavior or even the execution of the educational scenario [5], towards increasing the success—increased knowledge

transfer and achievement of the learning objectives—of the education delivery. Therefore, the development of a robust methodology for measuring the engagement state of the children in special education constitutes a challenging problem to tackle.

Children with learning disabilities (LD) are identified as having typical intelligence but manifest specific difficulties that interfere with their task performance and academic achievement [6]. This repeated failure and frustration experienced by the children with LD reduce their self-efficacy leading to a sense of helplessness, which is associated with lack of motivation and academic disengagement [7–9]. As academic engagement refers to active participation and attention and focuses on the task during the learning process, disengagement refers to apathy and lack of interest. The degree to which students are engaged is a critical precursor to learning, as without academic engagement, students are unlikely to benefit from instructions [10]. In other words, the more students are engaged, the more they learn [11]. Therefore, the development of a robust methodology for measuring the engagement state of children with LD constitutes a challenge.

Although several methods [12] for measuring the engagement level during child-robot interaction have been presented in the literature, all these attempts were focused on children with Autism Spectrum Disorder (ASD), and their experimental study was limited with a small number of children.

Taking into consideration the fact that children with learning disabilities are associated with maladaptive engagement compared with their typically developed peers [13, 14], it would be of great importance to have knowledge of each child's engagement level through social robots in order to use them in intervention programs that aim to promote child's learning by increasing their involvement in all kinds of learning tasks. Thus, as confirmed through researches, interventions using social robots as a tool to support the learning process have been demonstrated to enhance students' motivational skills, maintenance of engagement, and compliance during instructional interactions [15, 16]. The research of Pistoia et al. [17], which is one of the first attempts to investigate the use of a social robot in students with dyslexia, confirms that the presence of the robot to support the learning process showed high levels of response and engagement during child-robot interaction.

In this context, this work contributes along with the following directions:

- (1) A definition of the “Intelligent Interaction” based on psychology is provided
- (2) A machine learning-based methodology that allows a social robot to interact with intelligence with the child is proposed
- (3) The proposed methodology is evaluated with a large amount of in-house developed real data
- (4) For the first time the case of children with learning difficulties is considered for measuring their engagement state during interaction with the social robot
- (5) The need for customized engagement measuring methods based on the characteristics of the deployed scenario is touched for the first time

The rest of the paper is organized as follows: Section 2 provides a snapshot of the related work and Section 3 presents the definition of the “Intelligent Interaction,” the information of the designed educational scenarios, and the details of the proposed methodology. Section 4 provides the experimental study with the corresponding results. Section 5 discusses the results, concludes this study, and lays out the future work.

## 2. Related Work

Sidner et al. [18] proposed and Ahmad et al. [19] rephrased a general definition of the concept of engagement during human-robot interaction: “Engagement is the process by which interactors start, maintain, and end their perceived connection to each other during the interaction.” Measuring engagement of humans, executing a specific activity, constitutes a highly informative indication for analyzing the effectiveness of the activity design. This measurement can help the improvement of the design towards achieving the desired outcomes relative to the executed activity.

For this purpose, several methodologies have been proposed to measure the engagement of a user playing a video game [20], of a person when working [21], of students in a classroom [22], of TV viewers [22], of a consumer when purchasing products [23], and so on. Measuring engagement of a child with special needs during an educational process and/or intervention is very challenging due to the specially designed scenarios and interaction schemes, which must attract their attention and maintain engagement.

Early outstanding work for measuring the engagement of children with a game companion was proposed by Castellano et al. [24, 25] by using a multimodal processing scheme based on visual and contextual information. Moreover, Hernandez et al. [26] proposed a method to measure the engagement of children, which were difficult to engage during social interactions. In [26], wearable sensors were used to measure the electrodermal activity of the children and a Support Vector Machine (SVM) classifier was applied to classify the children being engaged or not. In [27], acoustic and linguistic data were utilized to detect the social engagement in conversational interactions of children with ASD and their parents, using an SVM classifier. The first in-depth study of measuring the engagement of children when interacting with social robots was proposed by Anzalone et al. [28]. In this work, the researchers analyzed visual information in a static and dynamic perspective, in several case studies of ASD child-robot interaction. Rudovic et al. [29] presented a very interesting study regarding the engagement measuring across cultures, which revealed that the engagement level of 30 ASD children can be increased by taking into account the cultural differences.

Recently, with the advent of deep learning technology, several attempts have been pointed out for measuring engagement during a child-robot interaction using advanced

intelligent models. Rudovic et al. [30] proposed the CultureNet model based on the typical ResNet-50 architecture for estimating the engaged or not engaged children of different cultures interacting with NAO robot in robot-assisted therapy for children with Autism Spectrum Condition (ASC). In [31], Hadfield et al. proposed a deep learning model consisting of three fully connected layers and a single LSTM layer, while the used features are computed using visual data relative to the position of the child's body parts. The reported results were of almost 80% accuracy, but the limited number (3) of Typical Developed (TD) children can justify the quite low accuracy. In a very recent work, Del Duchetto et al. [32] tried to measure the engagement level in human-robot interaction utilizing Convolutional Neural Networks (CNNs) and LSTM model. The novelty of the work in [33] is the tackling of the engagement estimation as a regression problem, aiming at providing a scalar value for the engagement level during human-robot interaction. The reported results were very promising with Mean Squared Error (MSE) 0.126.

Although the previous approaches have contributed significantly to the engagement estimation in human-robot interaction, they possess some limitations: (1) they were applied mostly on adults or children with TD or ASD, without examining other categories of children with special needs, such as children with learning difficulties; (2) they were experimented with a limited number of children; and (3) they did not study the engagement estimation in the framework of appropriately designed intervention scenarios or the designed scenarios were few and very simple.

It is important to realize the need to analyze and measure the engagement of children with learning difficulties. Considering dyslexia as the most frequent learning difficulty, Uta Frith [33] proposed a three-level theoretical framework for the interpretation of dyslexia, namely, behavioral, cognitive, and biological. In this context, Frith also distinguished the role of the environmental level that interacts with the abovementioned three levels. Therefore, dyslexia students interacting with a social robot can learn easier due to its interactive and fun performance, which also allows students to take their time during a learning task. In addition, a social robot engages pupils in mental information processing and captures their attention [34].

After reviewing the applications of social robots in special education from the international literature, we found that the usage of social robots in supporting the educational procedure of children with learning difficulties is limited. This observation contradicts the educational needs of a large percentage of the world's population, which accounts for 10–15% [35]. We believe that the high percentage of the population showing learning difficulties imposes the targeting of this part of the population as a potential application field for using social robots.

The current study aims to complement the previous works by investigating the engagement measuring when children with learning difficulties are interacting with the social robot NAO. The number of children that participated in the experiments was 10, while child psychologists carefully designed 10 scenarios, executed by each child.

### 3. Materials and Methods

**3.1. Intelligent Interaction: A Definition.** In order to understand the real needs for an engagement measuring methodology, it is crucial to provide a definition of what is the meaning of an “Intelligent Interaction.”

Considering the work of the psychologist Howard Gardner [36] regarding the type of intelligence, nine different types of intelligence can be considered. From these nine types of intelligence, the following five deal with the interaction of a human with the surrounding environment:

- (1) Linguistic intelligence: ability to find the right words to express what do you mean
- (2) Visual-spatial intelligence: having awareness of the surrounding environment
- (3) Bodily-kinesthetic intelligence: coordinating the mind with the body
- (4) Interpersonal intelligence: sensing children's feelings and motives
- (5) Logical-mathematical intelligence: quantifying things, making hypotheses, and proving them

From the engineering point of view though, the previous interaction-oriented intelligence can be summarized to the following two levels of intelligence:

- (1) 1st level of intelligence: ability to analyze the sensory data in order to understand the surrounding environment
- (2) 2nd level of intelligence: establishing a human-like closed-loop communication with the child

The above two levels of intelligence enclose the aforementioned five types of intelligence defined in terms of psychology and can be the ultimate goals of any research dealing with human-robot interaction.

An important part of the above two levels of intelligence is the measuring of the child's engagement state by processing the sensory data (1st level) for adapting the robot's behavior and/or the educational scenario towards establishing a closed-loop communication channel (2nd level).

**3.2. Educational Scenarios.** For the sake of this study, five child psychologists (three from the “Family Center KPG, Thessaloniki, Greece” and two from the “Department of Clinical Psychology, Papageorgiou General Hospital, Thessaloniki, Greece”) of our research team designed ten different educational scenarios for children with learning difficulties, as part of the national project titled “Social Robots as Tools in Special Education (SRTSE)” [37]. It is worth noting that each child executed each scenario on different days. More precisely, each child executed two scenarios per week and the average duration of each scenario was 35 minutes.

Table 1 shows what types of activities are included in each scenario.

The scenarios include the following types of activities:

- (i) Meet/greet
- (ii) Text decoding, comprehension, and reading
- (iii) Phonology composition, decomposition, discrimination, and addition
- (iv) Memory
- (v) Robot-child relaxation game
- (vi) Story listening and telling
- (vii) Sentence structuring
- (viii) Strategic visual representation

**3.3. Proposed Methodology.** Two are the main features of the proposed methodology: (1) the usage of multimodal data consisting of visual and audio modalities and (2) the usage of a machine learning model that provides the decision about the engagement state of the child. In the following subsections, the modules of the designed methodology depicted in Figure 1 are described in detail.

**3.3.1. Multimodal Sensing.** The sensing capabilities of the used social robot mainly control the type of sensory data to process in order to decide the engagement state of the child during the interaction. Our study considers the well-known NAO robot as the robot that is able to interact with the child, but other social robots [38] could also be used. This robot is equipped with two identical RGB video cameras located in the forehead and a microphone; thus, it can provide visual and audio sensing capabilities.

**(1) Visual Sensing.** The visual sensing capabilities of the NAO robot permit the acquisition of video frames that include the child's body and face. From each video frame, the body pose is extracted using the library [39], consisting of 25 key points (2 on the torso, 6 on the hands, 12 on the legs, and 5 on the head), as depicted in Figure 2(a). In addition, 68 key points called facial landmarks are extracted (see Figure 2(b)), from the child's face using the OpenFace library [40]. It is worth noting that the computed facial landmarks are used to define the child's emotional state in compliance with the Facial Action Coding System (FACS) [41]. Finally, the eye contact between the child and the robot is detected using the OpenGaze library [42] and following the methodology proposed by Xucong Zhang et al. [43].

**(2) Audio Sensing.** During the interaction with the child, the robot needs to keep facing the child at all times, in order for the robot to record and analyze the child's speech, by providing additional information related to the engagement state of the child.

**3.3.2. Feature Extraction.** The abovementioned multimodal sensing mechanism aims at collecting sensory raw data. This data, which has the form of 2D Cartesian points belonging to the child, is further processed to construct more informative descriptions named features. The feature extraction procedure is applied on the video frames ( $640 \times 480$  pixels

resolution) captured every 0.7 secs (1.4 fps) by using non-overlapping sliding windows of 60 secs. Although the camera of the NAO robot has 2.5 fps for  $640 \times 480$  video resolution, in a WiFi connection mode, in our case, the real-time performance of our system is 1.4 fps due to the execution of the algorithms. Moreover, it is decided to set the processing time window to 60 secs, in order to include enough event transitions and to help the manual annotation of the data. The features that are finally computed are the following:

- (1) Feature 1: number of blinks: the blinks count of the child on average
- (2) Feature 2: mean movement of the body in pixels
- (3) Feature 3: if the child's body was turned away from the robot (0 or 1)
- (4) Feature 4: percentage of the time window within which there was eye contact by the child
- (5) Feature 5: emotion (happy, sad, surprised, fear, anger, disgust, or contempt)
- (6) Feature 6: emotion intensity (0–5)
- (7) Feature 7: if the child's head was turned away from the robot (0 or 1)
- (8) Feature 8: mean response time (set to  $-1$  if the scenario did not require a response from the child)
- (9) Feature 9: mean voice level (in RMS)
- (10) Feature 10: percentage of the time window within which the child was silent
- (11) Feature 11: percentage of the time window within which the child was speaking

It should be noted that almost all the above visual features are computed by tracking and processing the extracted key points. For example, for a specific frame, the emotion is determined by combining the FACS corresponding to each feeling (Table 2), averaging their intensities, and choosing the emotion that has the highest intensity. In addition, features 3 and 7 are determined by counting the number of states (0 and 1) in the time window and choosing the one with the highest number of occurrences. Lastly, to determine if the child is speaking or not, we check if the voice volume is higher than 350 RMS and the mean voice level considers levels where the child is speaking.

To summarize, for each 60 secs video frame, a feature vector  $FV \in R^{11}$  is assigned, which is also manually annotated by three experienced child psychologists to an engaged time slot or not. The extracted features from the educational scenarios are used to train the machine learning model, so it will be able to detect the engagement state of the child.

**3.3.3. Machine Learning Models.** Herein, the detection of the child's engagement state (engaged or not engaged) is accomplished by solving a typical two-class classification problem by using a machine learning classifier. Machine learning has been proved to be an efficient technology in



TABLE 1: Educational scenarios' characteristics.

Scenario	Types of activities included
S1	Meet/greet, text decoding, phonology (de)composition, memory, and robot-child relaxation game
S2	Meet/greet, phonetic discrimination, text reading, decoding, and comprehension
S3	Meet/greet, story listening and telling, and sentence structuring
S4	Text comprehension and visual representation
S5	Phonemic addition, sentence playback from memory, and robot-child relaxation game
S6	Meet/greet, sentence playback from memory, and reading enhancement
S7	Meet/greet, phonetic awareness, and robot-child relaxation game
S8	Meet/greet, acoustic vocal discrimination, and acoustic syllable discrimination
S9	Memory enhancement and text decoding
S10	Text reading and robot-child relaxation game

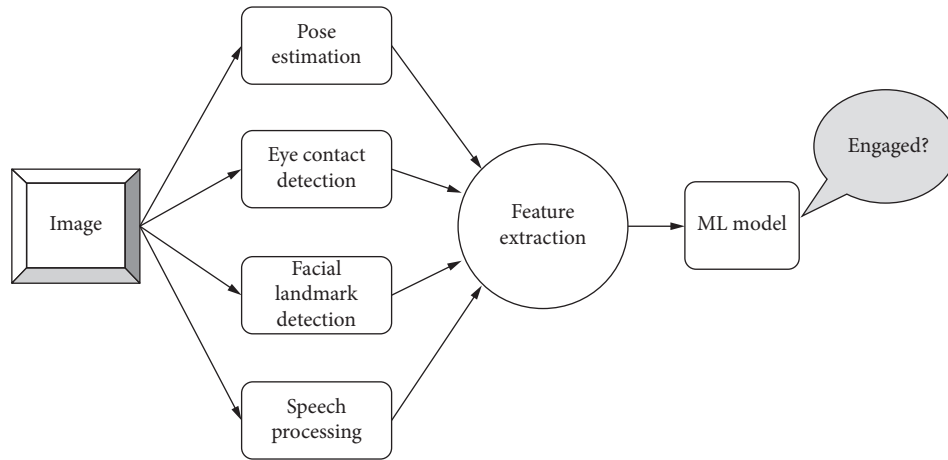


FIGURE 1: Block diagram of the proposed methodology.

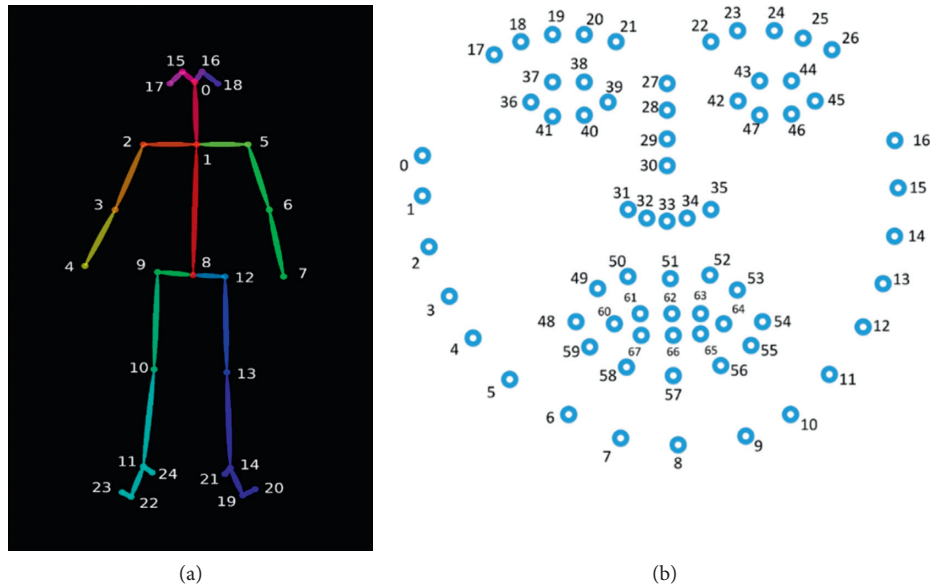


FIGURE 2: Visual sensing extracted data. (a) Body pose detection [39]. (b) Facial landmark (action units) [44] detection.

many disciplines such as signal processing [45] and computer vision [46]. More precisely, six traditional machine learning models, the Support Vector Machine (SVM) model with two different kernels (RBF and poly), the Decision Tree, the K-NN, the Naïve Bayes (NB), the Multilayer Perceptron

(MLP), and the Extreme Learning Machine (ELM) classifiers, are examined.

Additionally, three ensemble models are also considered, the Random Forest (RF), the AdaBoost Decision Tree, and AdaBoost Naïve Bayes ones. The advantage of the ensemble

TABLE 2: Facial action coding system (FACS).

Emotion	Action units (AU)	Description
Happiness/joy	6 + 12	Cheek raiser and lip corner puller
Sadness	1 + 4 + 15	Inner brow raiser, brow lowerer, and lip corner depressor
Surprise	1 + 2 + 5 + 26	Inner brow raiser, outer brow raiser, upper lid raiser, and jaw drop
Fear	1 + 2 + 4 + 5 + 7 + 20 + 26	Inner brow raiser, outer brow raiser, brow lowerer, upper lid raiser, lid tightener, lip stretcher, and jaw drop
Anger	4 + 5 + 7 + 23	Brow lowerer, upper lid raiser, lid tightener, and lip tightener
Disgust	9 + 15	Nose wrinkler, lip corner depressor, and lower lip depressor
Contempt	12 + 14	Lip corner puller and dimpler

classifiers is that they combine “weak learners” with strong ones, by reducing the bias and variance of the learner. The first ensemble uses the Bagging [47] and the last two use the AdaBoost [48] training techniques.

Most of the machine learning models owing to a set of configuration parameters that enables them to adjust their performance are subject to the considered problem and must be carefully selected.

#### 4. Experimental Study

In order to study the performance of the proposed methodology, a set of experiments was arranged. The experiments were carried out using the scikit-learn [49] Machine Learning Library for Python, on Python version 2.7. Moreover, the experiments were conducted on a laptop computer equipped with Intel i7-6700HQ CPU, 8 GB DDR4 RAM, and GTX 960M GPU.

**4.1. Dataset Design.** For the sake of the experiments, 10 children participated in the ten scenarios (see Table 1), 2 girls and 8 boys, aged from 9 to 10 years. Each scenario is executed in a classroom with the participation of a child, the NAO robot, and a child psychologist sitting behind the NAO robot. The robot also needs to keep facing the child at all times, in order for the speech recognition module of the robot to work more accurately, since in this position the microphones are oriented to the source of the sound [50]. From the recorded video files, a dataset of 819 samples, with 11 features for each sample, was designed. From these samples, 99 samples corresponded to children being engaged, while 720 samples corresponded to children being not engaged. Three experienced child psychologists derived the ground truth data after manual annotation. Since this dataset is imbalanced, an oversampling technique was employed, called Synthetic Minority Oversampling Technique (SMOTE) [51], in order to balance the dataset, by containing the same number of samples for each class. The final balanced dataset includes 1440 samples (720 per class).

**4.2. Settings of the Experiments.** A 10-fold cross-validation grid search technique [52] was applied in order to select the best parameters set for each model. The resulting parameters that optimize the accuracy of each model are presented in Table 3.

The performance of each model was evaluated using the Precision, Recall, Accuracy, and F-measure indices [53]. These measures are widely used in machine learning to

evaluate the performance of a model. They are taking into account the True Positive (TP) and True Negative (TN) cases, which correspond to those cases correctly identified as positive or negative, respectively, and False Positive (FP) and False Negative (FN) cases, which are falsely identified as positive or negative, respectively.

Accuracy is the proportion of the total number of correct predictions and is calculated from the equation

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (1)$$

Precision is the proportion of the correct predicted positive results and is calculated from the equation

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (2)$$

Recall is the proportion of correct positive results and is calculated from the equation

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (3)$$

F-measure combines both Precision and Recall and is the harmonic mean of those indices, calculated as follows:

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (4)$$

**4.3. Results.** A k-fold (with  $k = 10$ ) cross-validation technique is followed for the evaluation of each machine learning model in estimating the children’s engagement state. According to this training and testing protocol, the initial dataset of 1440 samples is divided into 10 equal and nonoverlapped subsets of 144 samples. Each one of these subsets is used to test the model trained with the remaining nine subsets. The process is repeated  $k$  times by using different subsets for testing only once. The results of the  $k$  experiments are averaged in order to conclude the generalization ability of each model. Table 4 summarizes the prediction performance of each model.

The results of Table 4 reveal two important conclusions. The first one is that the initial hypothesis that the children’s engagement can be measured by using multimodal data consisting of combined behavioral, pose, and emotional information is justified experimentally since the accuracy of the models is very high (up to 93.33%).

The second conclusion is that the AdaBoost Decision Tree model outperforms the other models, by a significant

TABLE 3: Settings of the ML models.

ML model	Best parameters
SVM (RBF)	$C = 10$ , $\text{tol} = 0.1$ , and $\text{gamma} = 0.001$
SVM (poly)	$C = 10$ , $\text{tol} = 0.0001$ , and $\text{gamma} = \text{"scale"}$
Decision Tree	Splitter = "best," min samples leaf = 10, criterion = "gini," max features = none, and max depth = 4
k-NN	n neighbors = 2, weights = "uniform," leaf size = 20, and algorithm = "ball tree"
Naïve Bayes	(Nothing to configure)
MLP	Solver = "Adam," learning rate = "constant," hidden layer sizes = (80, 40), tol = 10.0, and alpha = 0.01
ELM	Alpha = 100, n_hidden = 80, and rbf_width = 0.256
Random Forest	Max features = "sqrt," n_estimators = 4, criterion = "gini," max depth = 15, and min samples leaf = 15
AdaBoost Decision Tree	Criterion = "entropy," max depth = 15, max features = "auto," splitter = "best," and min samples leaf = 5
AdaBoost Naïve Bayes	(Nothing to configure)

TABLE 4: Predicted results.

ML model	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
SVM (RBF)	91.53	91.86	91.53	91.51
SVM (poly)	57.92	61.50	57.92	54.74
Decision Tree	87.01	88.59	87.01	86.83
k-NN	88.33	88.81	88.33	88.30
Naïve Bayes	78.12	81.41	78.12	77.49
MLP	78.75	79.73	78.75	78.50
ELM	89.72	90.16	89.16	89.66
Random Forest	88.54	89.61	88.54	88.42
AdaBoost Decision Tree	<b>93.33</b>	<b>94.09</b>	<b>93.33</b>	<b>93.28</b>
AdaBoost Naïve Bayes	82.29	83.48	82.29	82.09

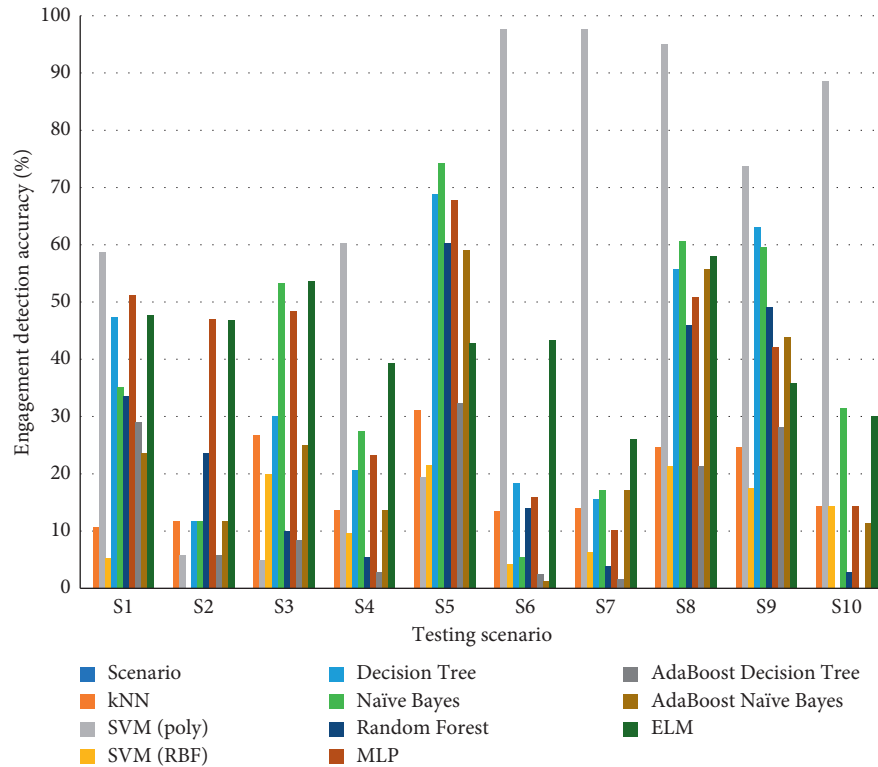


FIGURE 3: Engagement detection accuracy for the case of the modified leave-one-out training strategy.

factor in some cases, followed by the SVM (RBF). Despite the high accuracy of the AdaBoost Decision Tree model, the additional high Precision, Recall, and F-measure constitute the evidence that the model is able to estimate the

children's engagement of unseen data with minimum False Positive (FP) and False Negative (FN) decisions. The outperformance of the ensemble methods was expected since these models are more complex and they provide the

final decision considering the outcomes of multiple single classifiers working in a complementary way. Among the ensemble models, the AdaBoost shows the best performance, a result that reveals the ability of the sequential topology of the bootstrapping to improve the classification performance. On the other hand, the Bagging topology of the Random Forest model implies that the weak classifiers of the model operate with similar data, meaning that the dataset is quite homogenous, without including significant variations.

Moreover, in order to examine the bias of the machine learning models in the data of a specific scenario, a modified leave-one-out training strategy was applied with the samples that were left out in each case where all the samples corresponding to a specific scenario and the training were done with all other samples of the other scenarios. For each test case, the training data were again augmented to tackle the existing issue of the imbalanced number of samples per class. For example, in the first fold, the models were trained with the samples corresponding to all the scenarios except the first one and then tested with the samples corresponding to the first scenario and so on for each fold. Figure 3 depicts the performance of the machine learning models for each scenario when its data samples are used to test the models.

The results presented in Figure 3 reveal that the SVM (poly) model shows the lowest bias in the training data since it has the highest detection accuracy in 7 out of 10 training folds. Moreover, an interesting observation of this experiment is the different “definitions” of children’s engagement state in each scenario, since the performance of each model varies with the scenario type.

## 5. Discussion and Conclusion

The task of engagement detection of a child with learning difficulties interacting with a social robot for establishing a two-way intelligent interaction was studied in this work. The detection procedure was tackled as a two-class classification problem solved with high success by applying a machine learning model. The proposed methodology uses multimodal data (visual and audio) that describe the behavior of the child during the interaction. The initial hypothesis that an engaged child with learning difficulties can be identified by processing the body and head poses, the facial expressions, the eye contact, and the speech was accepted following the proposed method. However, this study brought to light the possible different “definitions” of engagement that apply in each educational scenario. This outcome is very important since it paves the way for more customized engagement measuring techniques oriented to the specific scenarios under deployment, towards providing an optimal interaction strategy.

In addition to the investigation of developing scenario-based engagement measuring methods, future work will consider the time parameter for each extracted feature and the handling of them as time series by deploying regression ML models such as Long Short Term Memory (LSTM) for predicting the engagement level at discrete time steps.

## Data Availability

The data used in this research will be provided upon request.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

This research had been cofinanced by the European Union and Greek National Funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (Project code T1EDK-00929).

## References

- [1] M. Hermann, T. Pentek, and B. Otto, “Design principles for industrie 4.0 scenarios,” in *Proceedings of the 2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 3928–3937, IEEE, Hawaii, HI, USA, January 2016.
- [2] L. I. Ismail, T. Verhoeven, J. Dambre, and F. Wyffels, “Leveraging robotics research for children with autism: a review,” *International Journal of Social Robotics*, vol. 11, no. 3, pp. 389–410, 2019.
- [3] V. Holeva, V.-A. Nikopoulou, M. Papadopoulou, E. Vrochidou, G. A. Papakostas, and V. G. Kaburlasos, “Toward robot-assisted psychosocial intervention for children with autism spectrum disorder (ASD),” in *Proceedings of the Social Robotics*, M. A. Salichs, S. S. Ge, E. I. Barakova et al., Eds., pp. 484–493, Springer International Publishing, Madrid, Spain, November 2019.
- [4] C. Lytridis, C. Bazinas, G. Sidiropoulos et al., “Distance special education delivery by social robots,” *Electronics*, vol. 9, no. 6, p. 1034, 2020.
- [5] M. Ahmad, O. Mubin, and J. Orlando, “A systematic review of adaptivity in human-robot interaction,” *Multimodal Technologies and Interaction*, vol. 1, no. 3, p. 14, 2017.
- [6] H. L. Swanson, “Learning disabilities: assessment, identification, and treatment,” in *The Oxford Handbook of School Psychology*, M. A. Bray and T. J. Kehle, Eds., pp. 334–350, Oxford University Press, Oxford, UK, 2011.
- [7] M. Boekaerts, E. De Koning, and P. Vedder, “Goal-directed behavior and contextual factors in the classroom: an innovative approach to the study of multiple goals,” *Educational Psychologist*, vol. 41, no. 1, pp. 33–51, 2006.
- [8] P. R. Pintrich, E. M. Anderman, and C. Klobucar, “Intra-individual differences in motivation and cognition in students with and without learning disabilities,” *Journal of Learning Disabilities*, vol. 27, no. 6, pp. 360–370, 1994.
- [9] G. D. Sideridis, “On the origins of helpless behavior of students with learning disabilities: avoidance motivation?” *International Journal of Educational Research*, vol. 39, no. 4–5, pp. 497–517, 2003.
- [10] K. Singh, M. Granville, and S. Dika, “Mathematics and science achievement: effects of motivation, interest, and academic engagement,” *The Journal of Educational Research*, vol. 95, no. 6, pp. 323–332, 2002.
- [11] R. Ben-ari and P. Kedem-Friedrich, “Restructuring heterogeneous classes for cognitive development: social interactive



- perspective," *Instructional Science*, vol. 28, no. 2, pp. 153–167, 2000.
- [12] C. Lytridis, C. Bazinas, G. A. Papakostas, and V. Kaburlasos, "On measuring engagement level during child-robot interaction in education," in *Robotics in Education*, M. Merdan, W. Lepuschitz, G. Koppensteiner, R. Balogh, and D. Obdržálek, Eds., Springer International Publishing, Cham, Switzerland, pp. 3–13, 2020.
  - [13] J. A. Baxter, J. Woodward, and D. Olson, "Effects of reform-based mathematics instruction on low achievers in five third-grade classrooms," *The Elementary School Journal*, vol. 101, no. 5, pp. 529–547, 2001.
  - [14] G. D. Sideridis, "Social, motivational, and emotional aspects of learning disabilities," *International Journal of Educational Research*, vol. 43, no. 4-5, pp. 209–214, 2005.
  - [15] A. Ramachandran, C.-M. Huang, E. Gartland, and B. Scassellati, "Thinking aloud with a tutoring robot to enhance learning," in *Proceedings of the 2018 ACM/IEEE International Conference*, Chicago IL, USA, March 2018.
  - [16] P. Baxter, E. Ashurst, R. Read, J. Kennedy, and T. Belpaeme, "Robot education peers in a situated primary school study: personalisation promotes child learning," *PLoS One*, vol. 12, no. 5, Article ID e0178126, 2017.
  - [17] M. Pistoia, S. Pinnelli, and G. Borrelli, "Use of a robotic platform in dyslexia-affected pupils: the ROBIN project experience," *International Journal of Education and Information Technologies*, vol. 9, pp. 45–49, 2015.
  - [18] C. L. Sidner, C. Lee, C. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," 2005, <https://arxiv.org/abs/cs/0507056>.
  - [19] M. I. Ahmad, O. Mubin, and J. Orlando, "Adaptive social robot for sustaining social engagement during long-term children-robot interaction," *International Journal of Human-Computer Interaction*, vol. 33, no. 12, pp. 943–962, 2017.
  - [20] E. N. Wiebe, A. Lamb, M. Hardy, and D. Sharek, "Measuring engagement in video game-based environments: investigation of the user engagement scale," *Computers in Human Behavior*, vol. 32, pp. 123–132, 2014.
  - [21] T. C.-t. Fong and S.-m. Ng, "Measuring engagement at work: validation of the Chinese version of the utrecht work engagement scale," *International Journal of Behavioral Medicine*, vol. 19, no. 3, pp. 391–397, 2012.
  - [22] Z. Wang, C. Bergin, and D. A. Bergin, "Measuring engagement in fourth to twelfth grade classrooms: the Classroom Engagement Inventory," *School Psychology Quarterly*, vol. 29, no. 4, pp. 517–535, 2014.
  - [23] B. J. Calder, M. S. Isaac, and E. C. Malthouse, "How to capture consumer experiences: a context-specific approach to measuring engagement," *Journal of Advertising Research*, vol. 56, no. 1, pp. 39–52, 2016.
  - [24] G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. W. McOwan, "Detecting user engagement with a robot companion using task and social interaction-based features," in *Proceedings of the 2009 International Conference on Multimodal Interfaces*, pp. 119–126, Cambridge, MA, USA, November 2009.
  - [25] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, "Detecting engagement in HRI: an exploration of social and task-based context," in *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, pp. 421–428, IEEE, Washington, DC, USA, September 2012.
  - [26] J. Hernandez, I. Riobo, A. Rozga, G. D. Abowd, and R. W. Picard, "Using electrodermal activity to recognize ease of engagement in children during social interactions," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 307–317, Seattle, WA, USA, September 2014.
  - [27] A. Chorianopoulou, E. Tzinis, E. Iosif, A. Papoulidi, C. Papailiou, and A. Potamianos, "Engagement detection for children with autism spectrum disorder," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5055–5059, IEEE, New Orleans, LA, USA, March 2017.
  - [28] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, "Evaluating the engagement with social robots," *International Journal of Social Robotics*, vol. 7, no. 4, pp. 465–478, 2015.
  - [29] O. Rudovic, J. Lee, L. Mascarell-Maricic, B. W. Schuller, and R. W. Picard, "Measuring engagement in robot-assisted autism therapy: a cross-cultural study," *Frontiers in Robotics and AI*, vol. 4, p. 36, 2017.
  - [30] O. Rudovic, Y. Utsumi, J. Lee et al., "A deep learning approach for engagement intensity estimation from face images of children with autism," in *Proceedings of the 2018 IEEE/RISJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 339–346, IEEE, Madrid, Spain, October 2018.
  - [31] J. Hadfield, G. Chalvatzaki, P. Koutras, M. Khamassi, C. S. Tzafestas, and P. Maragos, "A deep learning approach for multi-view engagement estimation of children in a child-robot joint attention task," 2018, <https://arxiv.org/abs/1812.00253>.
  - [32] F. Del Duchetto, P. Baxter, and M. Hanheide, "Are you still with me? continuous engagement assessment from a robot's point of view," 2020, <https://arxiv.org/abs/2001.03515>.
  - [33] U. Frith, "Paradoxes in the definition of dyslexia," *Dyslexia*, vol. 5, no. 4, pp. 192–214, 1999.
  - [34] K. Hamdan, A. Amorri, and F. Hamdan, "Robot technology impact on dyslexic students' English learning," *International Journal of Educational and Pedagogical Sciences*, vol. 11, pp. 1949–1954, 2017.
  - [35] S. Cramer, "Dyslexia and literacy, International: DI-Duke report," 2016, Available online: <https://www.dyslexia-and-literacy.international/wp-content/uploads/2016/04/DI-Duke-Report-final-4-29-14.pdf> Accessed on March 10, 2021.
  - [36] H. Gardner, *Multiple Intelligences: The Theory In Practice; Multiple Intelligences: The Theory In Practice*, vol. 304, Basic Books, New York, NY, USA, 1993, 978-0-465-01821-5.
  - [37] "Social robots as tools in special education (SRTSE)," 2018, Available online: <http://www.koiro3e.eu/en/homen/> Accessed on March 10, 2021.
  - [38] G. A. Papakostas, A. K. Strolis, F. Panagiotopoulos, and C. N. Aitsidis, "Social robot selection: a case study in education," in *Proceedings of the 2018 26th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pp. 1–4, Split, Croatia, September 2018.
  - [39] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2d pose estimation using part affinity fields," 2018, <https://arxiv.org/abs/1812.08008>.
  - [40] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: facial behavior analysis toolkit," in *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66, IEEE, Xian, China, May 2018.
  - [41] J. L. Tracy, R. W. Robins, and R. A. Schriber, "Development of a FACS-verified set of basic and self-conscious emotion expressions," *Emotion*, vol. 9, no. 4, pp. 554–559, 2009.
  - [42] X. Zhang, Y. Sugano, and A. Bulling, "Evaluation of appearance-based methods and implications for gaze-based

- applications,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, Scotland, UK, May 2019.
- [43] X. Zhang, Y. Sugano, and A. Bulling, “Everyday eye contact detection using unsupervised gaze target discovery,” in *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pp. 193–203, Québec City, Canada, October 2017.
  - [44] T. Baltrušaitis, M. Mahmoud, and P. Robinson, “Cross-dataset learning and person-specific normalisation for automatic action unit detection,” *IEEE in Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 6, pp. 1–6, Ljubljana, Slovenia, May 2015.
  - [45] G. A. Papakostas, K. I. Diamantaras, and F. A. Palmieri, “Emerging trends in machine learning for signal processing,” *Computational Intelligence and Neuroscience*, vol. 2017, Article ID 6521367, 2 pages, 2017.
  - [46] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, “Deep learning for computer vision: A brief review,” *Computational Intelligence and Neuroscience*, vol. 2018, Article ID 7068349, 13 pages, 2018.
  - [47] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
  - [48] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” in *Proceedings of the Computational Learning Theory*, P. Vitányi, Ed., pp. 23–37, Barcelona, Spain, March 1995.
  - [49] F. Pedregosa, G. Varoquaux, A. Gramfort et al., “Scikit-learn: machine learning in Python,” *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
  - [50] J. Kennedy, S. Lemaignan, C. Montassier et al., “Child speech recognition in human-robot interaction: evaluations and recommendations,” in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 82–90, Vienna, AT, USA, March 2017.
  - [51] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
  - [52] P. I. Kiratsa, G. K. Sidiropoulos, E. V. Badeka, C. I. Papadopoulou, A. P. Nikolaou, and G. A. Papakostas, “Gender identification through facebook data analysis using machine learning techniques,” in *Proceedings of the 22nd Pan-Hellenic Conference on Informatics*, pp. 117–120, Athens, Greece, November 2018.
  - [53] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.