

Facial Emotion Recognition-based Engagement Detection in Autism Spectrum Disorder

Noura Alhakbani

Information Technology Department, College of Computer and Information Sciences,
King Saud University, Riyadh 11543, Saudi Arabia

Abstract—Engagement is the state of alertness that a person experiences and the deliberate focus of their attention on a task-relevant stimulus. It positively correlates with many aspects such as learning, social support, and acceptance. Facial emotion recognition using artificial intelligence can be beneficial to automatically measure individual engagement especially when using automated learning and playing modalities such as using Robots. In this study, we proposed an automatic engagement detection model through facial emotional recognition, particularly in determining autistic children's engagement. The methodology employed a transfer learning approach at the dataset level, utilizing facial image datasets from typically developing (TD) children and children with ASD. The classification task was performed using convolutional neural network (CNN) methods. Comparative analysis revealed that the CNN method demonstrated superior accuracy compared to random forest (RF), support vector machine (SVM), and decision tree algorithms in both the TD and ASD datasets. The findings highlight the potential of CNN-based facial emotion recognition for accurately assessing engagement in children with ASD, with implications for enhancing learning, social support, and acceptance in this population. This research contributes to the field of engagement measurement in autism and underscores the importance of leveraging AI techniques for improving understanding and support for children with ASD.

Keywords—Engagement detection; facial emotion recognition; autistic children; convolutional neural networks

I. INTRODUCTION

A person's level of engagement in social activities is indicative of his or her socio emotional and cognitive well-being. It is usually possible to assess a person's engagement state by observing their behavior and physiological cues, such as the focus of their gaze, their smile, and their vocalizations [1].

Measuring engagement helps identify how to improve engagement in different settings, especially when targeting a particular health condition. There is an active field of research looking at measuring both engagement and attention. Measuring engagement is a particularly active subject of research, and advances in sensor technology and computer vision techniques have created a shift away from manual measurement to more automated approaches [2, 3]. Authors in study [3] aimed to detect emotions displayed by people viewing video commercials to understand how people respond to the media content. The researchers built a model using convolutional neural network (CNN) to measure the viewer's attention based on how the position of the head changed subtly over time.

Autism spectrum disorder (ASD) is a neurological and developmental disorder that typically begins in early childhood [4]. The recent increase in the utilization of assistive therapies during therapy sessions with autistic children has been driven, in part, by the societal demand for new technologies that can facilitate and enhance existing therapies for the growing number of children with ASD. One such assistive approach is robot-assisted autism therapy (RAAT), which is an emerging field. Although there are currently only a limited number of studies investigating the efficacy of RAAT, research has indicated that incorporating RAAT in treatment sessions can effectively motivate children with ASD to engage in activities. In line with this, one potential application of engagement measurement is its ability to significantly improve the learning experience with interactive robots. By accurately assessing and responding to a user's level of engagement, robots can adapt their interactions and instructional strategies accordingly, resulting in more personalized and effective learning outcomes [5].

An effort was made to estimate the visual attention of children with autism and other cognitive disabilities during robot-assisted autism therapy sessions by building a training engagement classifier. This study achieved 93% using a K-nearest neighbors (K-NN) classifier [6]. Moreover, in a study of 46 children, of whom 20 were diagnosed with autism, a face-based recognition model was used using CNN classification with an 89% accuracy result [2].

There are different automated approaches to measuring engagement, such as using face detection and neural networks. Previous studies used different parameters such as signal data from the brain, eye-tracking, galvanic skin conductance, face-tracking, blood flow, and heart rate to create their system. However, among these methods, face-tracking is the most promising approach because it is ubiquitous, cost-effective, and provides accurate results [6]. Moreover, face detection and localization through finding facial landmarks are widely used for determining visual attention from video cameras. Facial features can be used to estimate the head pose, eye gaze, and emotions, all of which are reliable indicators of engagement [7].

In this study, we used two datasets, one specifically collected from typically developing (TD) children and the other from children with Autism Spectrum Disorder (ASD). The motivation behind using these two distinct datasets was to account for the unique characteristics and expressions exhibited by children with ASD. By utilizing transfer learning at the dataset level, we aimed to leverage the knowledge learned from

the TD dataset to enhance the performance of engagement detection in the ASD dataset.

To measure the engagement state of children with ASD, we applied neural network-based deep learning in face detection and recognition. The selection of deep neural networks, specifically CNN, for measuring engagement in children with ASD is justified by its exceptional performance in image classification tasks. CNN demonstrates a high level of proficiency in extracting relevant features from complex visual data, making it highly suitable for capturing subtle facial cues related to engagement. Additionally, its robustness to noise and variability enhances its effectiveness in real-world scenarios. By exploring the application of CNN in measuring engagement, our objective is to leverage its feature extraction capabilities and capitalize on its resilience to variability, thereby advancing our understanding of engagement dynamics in children with ASD [8].

The remainder of this paper is arranged as follows: Section II introduces the main concepts of this study with background details. Section III presents the literature review presenting different machine learning models. Section IV presents an overview of datasets and the emotion model involved in the system design approach; Section V describes the proposed system starting with engagement model and ground-truth definition. Then, the implementation of the system which involves three main steps, namely pre-processing, feature extraction and classification; Section VI presents the evaluation results. Section VII discusses the results, and finally Section VIII presents the conclusion.

II. BACKGROUND

A. Autism Spectrum Disorder

Autism is a neurological and developmental disorder that begins early in childhood. Children with ASD face challenges in social interaction, communication skills, language development, and behavioral problems. They face many challenges in their lives, including persistent challenges in social communication, education, and many life skills [9].

Deficits in engagement\attention are one characteristic of ASD. Autistic children have difficulty with engagement, and it is challenging for them to pay attention to both an object and a person while interacting. At the same time, participation and engagement in a diverse range of social, play, educational, and therapeutic activities are essential for acquiring knowledge that is necessary for cognitive and social development [2].

B. Facial Emotion Recognition

Facial Emotion Recognition (FER) technology facilitates the recognition and interpretation of human emotions and affective states. It can analyze facial expressions from both static images and videos to reveal information on one's emotional state [10, 11].

Emotion detection is based on the analysis of facial land

mark positions (e.g. end of the nose, eyebrows). the basic steps of FER technology include (i) face detection, (ii) facial expression detection, and (iii) expression classification to an emotional state. FER has many applications including, but not limited to, human-computer interaction, human behavior understanding, computer vision, and gaming [10,11].

Facial symmetry refers to a complete alignment of the size, location, shape, and arrangement of each facial component about the sagittal plane whereas asymmetry refers to the bilateral difference between such components [12].

Previous studies in this field have revealed some facts about the connection between face symmetry and face recognition, including; symmetrical faces are judged as more emotion expressed than asymmetrical faces; the left face displays emotions more intensely than the right face; EFRs' decoding is modulated by complex interplays between the emotion and face asymmetry [13–15].

C. Neural Network-based Deep Learning

The brain is usually represented by neural networks, in which neurons connect to form a network. In computer science, an artificial neural network (ANN) is often called a neural network (NN), or a multi-layered perceptron (MLP), which is the most useful type of neural network.

ANNs are one of the best programming paradigms as they reflect the behaviour of the human brain, allowing computer programs to recognize patterns and resolve common problems. In contrast to conventional programming, in which a complex problem is broken down into a series of small, precisely defined tasks, ANNs do not require instructions to solve a particular problem; rather, they are instructed to use observational data to formulate a solution [16].

Deep learning is one of the machine learning methods that is based on ANN and uses larger numbers of hidden layers. NNs with more than two hidden layers are sometimes referred to as Deep Neural Networks (DNNs) [17].

Deep learning architectures are classified into two different learning algorithms, namely supervised and unsupervised learning. In supervised learning, the DNN is trained with the input of training data, all of which has a known label. Unsupervised learning is prepared by inferring the structures present in the unlabeled input data by a mathematical process [18].

A convolutional neural network (CNN) is a multi-layered neural network that draws biological inspiration from the visual cortex in animals [19]. CNN architecture is especially valuable in image processing applications, providing good results across many studies [20], and this is the architecture deployed in this study.

Fig. 1 shows CNN architecture relies on multiple layers that implement feature extraction and classification. The input data is broken down into receptive fields that feed into a convolutional layer and then extract the input data.

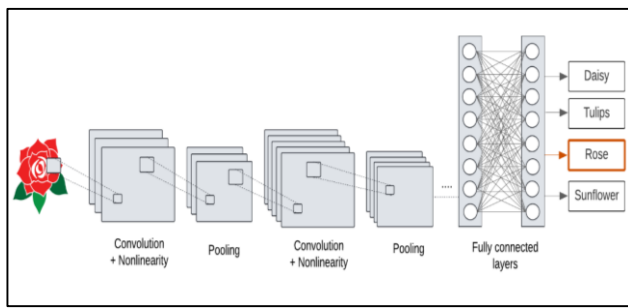


Fig. 1. Convolutional Neural Networks (CNN).

D. Transfer Learning

Transfer learning (TL) is a machine learning (ML) technique that focuses on applying knowledge gained while solving one task to a related task. By reusing and transferring previously learned information to new tasks, TL has demonstrated that it is possible to significantly improve learning efficiency [21].

There are several advantages of TL, the most important of which are reduced training time, improved neural network performance, and the absence of a large amount of data. For example, when training a neural model, a substantial amount of data is required, but access to that data is not always available. With TL, the model can be trained on an available labeled dataset, and then be applied to a similar task that may involve unlabeled data [22].

III. RELATED WORK

Engagement detection-based facial recognition technology has gained significant attention in recent years. Different machine learning techniques have been used for this purpose, including neural networks.

Trabelsi et al. in [23] proposed an automatic engagement detection system for classrooms using deep learning algorithms. A machine learning approach is employed to train behavior recognition models, including facial expression identification, to determine students' attention/non-attention in the classroom. They used AffectNet dataset. Various versions of the YOLOv5 model are evaluated for performance, showing promising results with an average accuracy of 76%.

Gupta et al. in [24] aimed to enhance the online learning environment by proposing a deep learning-based approach that utilizes facial emotions to detect real-time engagement of online learners. Facial expressions are analyzed to classify emotions and calculate the engagement index (EI), predicting "Engaged" and "Disengaged" states. Various deep learning models, including Inception-V3, VGG19, and ResNet-50, are evaluated and compared for the best predictive classification model. Benchmark datasets such as FER-2013, CK+, and RAF-DB are used for performance evaluation. Experimental results demonstrate that ResNet-50 achieves the highest accuracy of 92.3% for facial emotion classification in real-time learning scenarios, outperforming the other models.

Banire et al. in [25] proposed attention recognition for children with ASD using a face-based model. Two methods are proposed: geometric feature transformation with an SVM

classifier and transformation of time-domain spatial features to 2D spatial images using a CNN approach. The study involves 46 children (ASD $n=20$, typically developing children $n=26$) and examines participant and task differences. Results indicate that the geometric feature transformation with an SVM classifier outperforms the CNN approach. They reported that engagement detection is more generalizable for typically developing children and low-attention tasks.

Rathod et al. in [26] proposed a kids' facial emotion recognition system based on using deep-learning models. The aim was to improve interactive solutions in online platforms, particularly in the context of online education for children. They used LIRIS Children Spontaneous Facial Expression Video Database. The authors achieved the highest accuracy of 89.31%.

Overall, these studies demonstrate the effectiveness of using neural networks for engagement detection-based facial recognition. The use of deep learning techniques has enabled the extraction of high-level features from facial images and modeling of the temporal dynamics of facial expressions.

Furthermore, the integration of multimodal signals has shown to improve the accuracy of engagement detection. However, further research is needed to develop more robust and accurate engagement detection-based facial recognition systems that can be applied in real-world scenarios.

IV. MATERIALS AND METHODS

The following sub-sections present an overview of datasets and the emotion model involved in the system design approach.

A. Dataset

One of the challenges we ran into in this study was the lack of an open-access dataset of images of children with ASD, yet a benchmark dataset is necessary and important for researchers in machine learning-based image classification models [27].

Therefore, we found that transfer learning methods have been successfully applied in different domains where there is a lack of large datasets [21].

Moreover, it could be used to provide high-performance learners trained with more easily obtained data from different domains [22]. Therefore, in our study, we applied the transfer learning approach at the dataset level by using two different datasets to achieve good recognition results.

Our training dataset was for typically developing (TD) children, and the target dataset was images of autistic children. Our hypothesis assumes that knowledge gained while learning to recognize engagement in TD children could be applied when trying to recognize engagement in autistic children.

We initially implemented and trained our model denoted as TD_CNN on TD children. Next, the model was used to be trained on children with ASD and we named the resulting model ASD_CNN. Then, we compared and discussed the results of our implementation.

1) (LIRIS-CSE) dataset: This is a novel database of 189 video recordings for 12 TD children and is known as Children's Spontaneous Facial Expressions (LIRIS-CSE) [28].

It contains eight basic spontaneous facial expressions shown by 12 ethnically diverse children between the ages of 6 and 12 years, with a mean age of 7.3 years.

This unique database contains spontaneous/natural facial expressions of children in diverse settings with diverse recording scenarios showing eight universal or prototypical emotional expressions, namely happiness, sadness, anger, surprise, disgust, natural, fear, and confusion.

This dataset has been cited by 32 papers. For example, in this article [11] the author proposed a framework for automatic expression recognition based on CNN architecture and achieved an average classification accuracy of 75%. We extracted metadata from the video recordings and created our data frame as shown in Table I below.

TABLE I. DATASET (TD) DESCRIPTION

Id	Unique number for each row
Image name	Unique number for each image
Session number	The session number for each video recording
Iteration	The trial number in each session
Emotion	Label for one of the emotions (happiness, sadness, anger, surprise, disgust, natural, fear, and confusion)
Landmarks file	The name of the file that contains the extracted facial landmarks

2) *Autistic children dataset*: To create our second model, we used the autistic children dataset from the Kaggle repository [29], many versions of which are available online. The dataset includes images of ASD children aged 2 to 14 years, most of whom are two to eight years old. This dataset was used in recent studies, such as [27] [30].

For example, in [30], a deep learning model was built, using this dataset, that was designed to distinguish between healthy children and those potentially showing signs of autism, and it produced results with 94.6% accuracy.

The dataset contains 1,333 images of children with ASD categorized into five emotions, namely happy, angry, sad, fearful, and normal. Each image in the dataset presents the face of an ASD child experiencing one of the above types of emotion. We extracted the metadata and created the data frame as shown in Table II.

TABLE II. DATASET (ASD) DESCRIPTION

Id	Unique number for each row
Image name	Unique number for each image
Emotion	Label for one of the emotions (happy, angry, sad, fearful, and normal)
Landmarks file	The name of the file that contains the extracted facial landmarks

B. Emotion Model

We designed and developed an engagement detection system based on the most used emotion model, as presented in Fig. 2. Russell and Pratt (1980) suggested that all affective states originate from two fundamental neurophysiological systems,

embedded in a circumplex with two orthogonal dimensions, valence, and arousal [31].

One study into the structure of subjective learning experiences found that positive valence and high arousal were indicators of emotional engagement [1]. Another study that focused on measuring happiness found that people who experience positive valence and high arousal were more engaged and satisfied, and it is this emotion model that forms the foundation for our model [32] (see Fig. 3).

We used it for automatically estimating if the child is engaged and providing positive facial expressions, such as happiness or surprise, during the experiment, which are indicators of their engagement. In some multi-models, facial expressions are used as input to automatically estimate levels of valence and arousal, and they tend to have a high level of agreement with human coders [33].

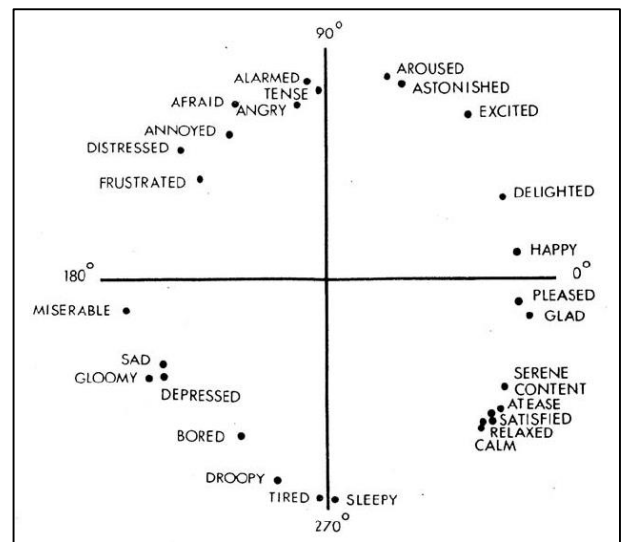


Fig. 2. Russell emotion model (1980).

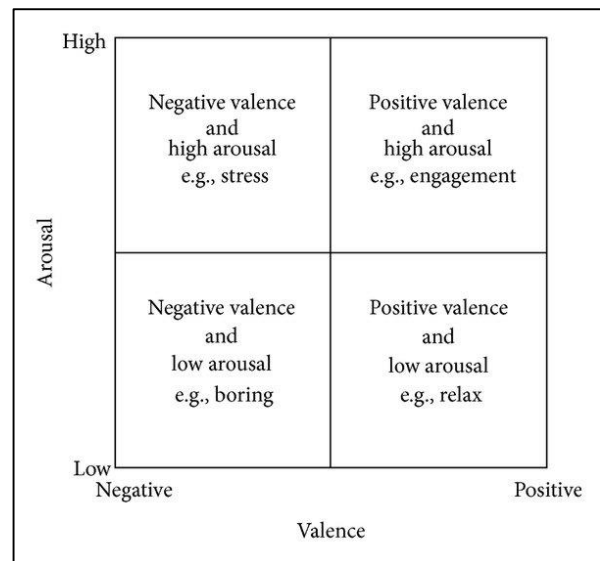


Fig. 3. Emotion model and classification labels.

V. PROPOSED SYSTEM

This section will cover our design and implementation steps in implementing the engagement detection system for autistic children.

First, we will describe how we built the data frame and ground truth. Afterward, we will present the steps to build our classifier based on the emotion model described above. Finally, we will present the evaluation of our classifier.

A. Engagement Model and Ground-Truth Definition

Based on the emotion model, we used the two-dimensional valence arousal model to classify only happy and surprised expressions as an engaged state, while other emotions we classified as non-engaged. After classification, we created data frames and saved them as a CSV file.

Table III below present examples of TD children collected from [29], and Table IV shows examples of children with ASD.

The following sub-sections present the implementation of engagement detection involving three main steps, namely pre-processing, feature extraction, and classification, which will now be discussed in turn.

TABLE III. EMOTIONS CLASSIFICATION OF (TD) CHILDREN











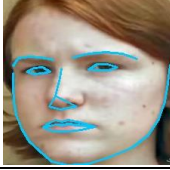
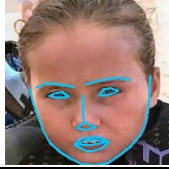
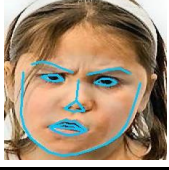
Happy (engagement)	Natural (non-engagement)
	
Sad (non- engagement)	Disgust (non- engagement)
	
Anger (non- engagement)	Surprise (engagement)
	
Confusing (non-engagement)	Fear (non- engagement)
	

TABLE IV. EMOTION CLASSIFICATION OF CHILDREN WITH (ASD)

Happy (engagement)	Natural (non-engagement)
	
Sad (non- engagement)	Disgust (non- engagement)
	
Anger (non- engagement)	
	

B. Pre-processing

Pre-processing is a necessary first step because the TD datasets are in video format, and we required these to be split into still images, known as frames. We pre-processed the input video files to extract the desired frames. First, we stored all the video files in one folder before reading each file individually and extracting the frames. To process the video files we set up OpenCV 1 (Open Source Computer Vision), which is an open-source library that includes several hundred computer vision algorithms and that is widely used in computer vision generally, and facial recognition more specifically [7].

We used the OpenCV library to read the video streams before creating a VideoCapture object using Python script, and then we split the videos into frames. Each video was split into approximately 125 frames, so from the 189 videos we extracted 23,132 frames. For each frame, we detected and cut out the face using the Dlib library [7].

C. Features Extraction

In facial recognition work, facial landmarks, defined as the detection and localization of certain characteristic points on the face, are considered very important features and are widely used in classification [20].

Features extraction is an important intermediate step in many subsequent facial processing processes that range from biometric recognition to understanding mental states. Facial landmarking is used to localize and represent salient regions of the face such as the eyes, eyebrows, nose, mouth, and jawline [34].

¹ Opencv: <https://opencv.org/>

Facial landmarks have been successfully applied to face alignment, head pose estimation, face swapping, blink detection, and much more.

Open-access software packages that automatically and efficiently detect facial landmarks include OpenCV, Imotion, MTCNN, and Open Face [20]. One of the main variants between them is the number of landmarks they detect; while Imotion detects 34, the MTCNN algorithm detects just five.

In our implementation, we used the Dlib library and OpenCV which can detect and extract 68 facial landmarks from each frame in both the TD and ASD datasets. Dlib, a facial landmark detector with pre-trained models, was used to estimate the location of 68 coordinates (x, y) that map the facial points on a person's face, as presented in Figure 4. These points are identified from the pre-trained model where the iBUG300-W dataset was used. Open CV read the video streams and splits it into frames [7].

We created a Python script that read all the frames for both the TD and ASD datasets and for each frame, we detected the face and extracted the 68 coordinates (x, y) which were then saved in a text file that contains all 68 coordinates (x, y) and then saved into a data frame.

D. Classification

In our work, the engagement recognition task is formulated as a binary classification problem in that we have two classes of engaged and non-engaged. We used a deep learning approach in our classification stage and, to compare our proposed methods, we implemented SVM, Decision Tree, and Random Forest.

To create the classification model, we divided our data into training and testing sets by using 80% of the data for training and 20% for testing and validation. We imported the Scikit-Learn library and used a train_test_split method to randomly split the data into training and testing sets. Additionally, to assess the robustness and generalization ability of our model, we employed cross-validation techniques.

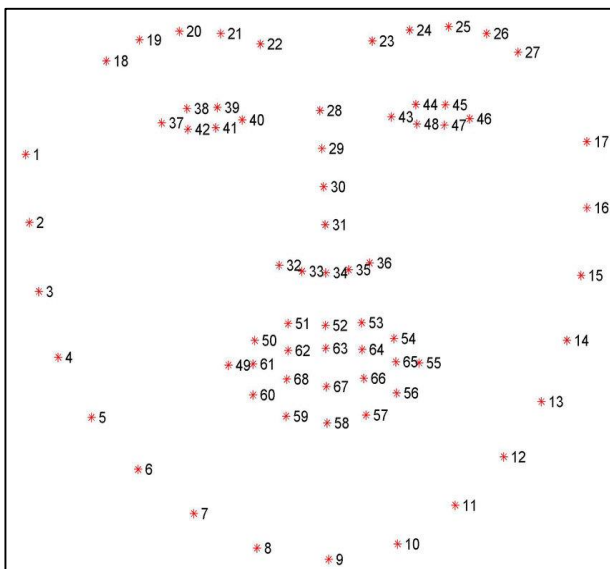


Fig. 4. Facial landmarks.

Specifically, we performed k-fold cross-validation, where we divided the training set into k subsets (folds) and iteratively trained and tested the model on different combinations of these subsets. This approach allowed us to obtain more reliable performance estimates by evaluating the model on multiple subsets of the data.

Firstly, we did our implementation using the TD dataset, the feature extraction part takes frames that contain the faces of the TD children as input, then we applied two convolution layers with 32 filters and (3,3) kernel size. Then we applied the third convolution layer which consists of 64 filters and (3,3) kernel size. After each convolution layer, there is an activation function called ReLU to set all the negative pixels to 0.

ReLU function introduces non-linearity to the network and generates an output-rectified feature map [2]. After the ReLU operation, there is a pooling layer for simple and salient elements. Finally, the flatten layer produces the output class (engagement/non-engagement).

To optimize the algorithms' hyperparameters, we employed a grid search approach. We defined a parameter grid containing a set of hyperparameters for each algorithm and exhaustively searched through the grid to find the combination that yielded the best performance. This hyperparameter tuning process was performed within each iteration of the cross-validation procedure.

We use the model implemented above TD_CNN as starting point for ASD_CNN model. We applied the transfer learning approach at the dataset level. The ASD_CNN model went through a similar implementation of TD_CNN but used the ASD dataset this round, with a few modifications to mitigate the differences in the data from images of autistic children. We removed one convolution layer with 32 filters and (3, 3) kernel size from the model. To evaluate the performance of the algorithms, we used multiple metrics, including precision, recall, and accuracy.

VI. RESULTS

After we trained our algorithms and made some predictions, we compared their accuracy using different algorithms.

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for classification or regression challenges and is one of the most popular classification algorithms used in machine learning. Its objective is to find the best hyperplane that correctly separates the points of different classes and provides the maximum margin among them [2]. SVM uses a set of mathematical functions that are defined as the kernel. The function of the kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions, such as linear, nonlinear, polynomial, and radial basis function (RBF) [35]. We used a linear kernel function in our implementation.

Decision tree is one of the most frequently and widely used supervised machine learning algorithms that can perform both regression and classification tasks. Decision trees build classification or regression models in the form of a tree structure, breaking down a dataset into subsets; the result is a tree with

decisional nodes and leaf nodes that represent the class [34] [10].

Random Forest (RF) is an ensemble learning method that is represented as a list of random trees. It works by creating a multitude of decision trees during training and outputting the class that is the mode of all classes. Basically, it functions by injecting randomness into the training of the trees and combining the output of multiple randomized trees into a single classifier [11].

To evaluate our algorithms, we used the `classification_report` and `confusion_matrix` methods to calculate these metrics. A confusion matrix is a table that is mostly used to explain the performance of a classification model on a set of test data for which the true values are known. Table V and Table VI show the confusion matrix for each classifier which contains information about actual and predicted classifications by different algorithms.

TABLE V. CONFUSION MATRIX FOR EACH CLASSIFIER USING (TD) DATASET

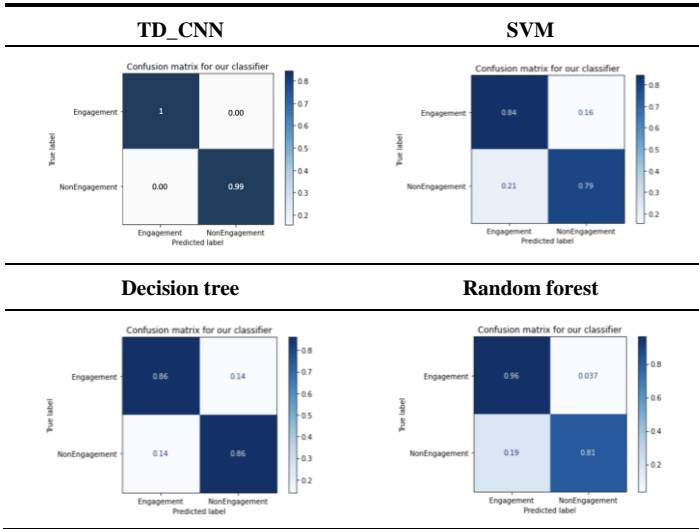


TABLE VI. CONFUSION MATRIX FOR EACH CLASSIFIER USING (ASD) DATASET

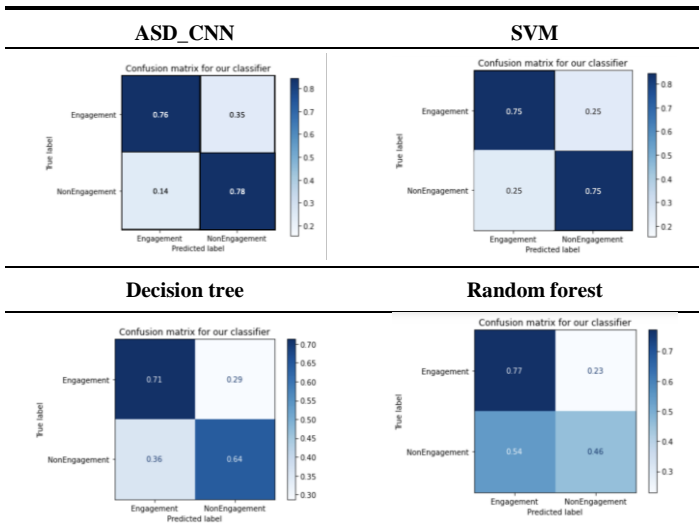


Table VII describes the evaluation of efficiency measures (accuracy, precision, recall) of our classifiers. Accuracy refers to the percentage of the total number of predictions that were correct; precision is the percentage of the predicted positive cases that were correct; recall is the percentage of positive cases that were correctly identified.

As shown in the table, for the TD dataset, the CNN model achieved the highest accuracy of 99%, indicating its ability to accurately detect engagement levels in typically developing children. In contrast, the random forest (RF), decision tree, and support vector machine (SVM) models achieved slightly lower accuracies of 89%, 86%, and 82% respectively. The precision and recall values for the CNN model were also consistently high at 99%.

TABLE VII. PERFORMANCE MEASUREMENTS

Dataset	Algorithm	Accuracy	Precision	Recall
TD	TD_CNN	99%	99%	99%
	Random Forest	89%	90%	89%
	Decision tree	86%	86%	86%
	SVM	82%	82%	82%
ASD	ASD_CNN	76%	76%	77%
	Random Forest	64%	63%	63%
	Decision tree	67%	67%	67%
	SVM	75%	75%	75%

On the ASD dataset, the CNN model again demonstrated superior performance with an accuracy of 75%. This suggests that the CNN model is effective in detecting engagement levels in children with Autism Spectrum Disorder (ASD). The RF, decision tree, and SVM models achieved lower accuracies of 64%, 67%, and 75% respectively. It is worth noting that the precision and recall values for the CNN model on the ASD dataset were 76% and 77% respectively.

VII. DISCUSSION

The results show that CNN was the most accurate of the four algorithms for the TD dataset (99%), and SVM was the least accurate. Also, on the ASD dataset, CNN achieved the highest level of accuracy (75%), while RF achieved the lowest accuracy compared to the other methods in the ASD dataset. It is evident that the CNN model consistently outperformed the other algorithms for both the TD and ASD datasets in terms of accuracy.

These findings suggest that the CNN model's ability to capture and learn complex patterns in facial images contributes to its superior performance. The CNN model's success in accurately detecting engagement levels can be attributed to its capacity to extract meaningful features from the facial images and effectively classify them.

We observe that the accuracy of the algorithms' using TD is higher than ASD this is due to the quality and quantity of data for TD children compared to the dataset for ASD children, and due to the limited scope of facial impressions in children with ASD. Although a person's emotional state can be detected from facial expressions, either consciously or subconsciously, there

are many theories about how children with ASD represent emotion.

A number of studies have found that recognizing facial emotions in children with ASD is challenging [14, 36, 37]. Studies have found that children with ASD are less responsive in the upper part of the face, with their eyes typically remaining emotionally neutral. Consequently, the lower half of the face, including the mouth, chin, jaw, and cheeks, is crucial in recognizing emotion in autistic children [36].

VIII. CONCLUSION

We proposed an engagement detection system using CNN for facial emotional recognition. AI and machine learning have been applied to automatically measure the engagement of children with ASD. This allows the therapist to track a child's engagement with ASD during therapy sessions without relying on traditional observation techniques.

The implications of this research are significant. The ability to accurately assess engagement levels has the potential to revolutionize various domains, including education, therapy, and human-computer interaction. By providing feedback and adaptive interventions, engagement detection technologies can enhance learning outcomes, facilitate personalized interventions for individuals with ASD, and create more immersive and interactive user experiences.

In this paper, we explained the implementation details to build an engagement detection model through facial emotion recognition. We presented the information of the datasets and the pre-processing steps for the videos and images to make them ready to be fed into the model. We used a transfer learning approach at the level of the dataset. Due to the small size of the datasets. Then, we described how we detected and extracted the 68 facial landmarks from each face in the frame and then created the data frame based on the two-dimensional emotion model to build the ground truth for detecting the engagement of the children. And due to the difficulty of recognizing facial emotions in children with ASD and their decreased response in the upper part the results we achieved were less accurate than with TD dataset. Then, we compared different machine learning algorithms and evaluated their performance, and our findings showed that the CNN outperformed other classifiers.

While we utilized two datasets in our study, it is important to acknowledge that these datasets have their own limitations, including sample size, demographic representation, and potential biases. Future work should aim to address these limitations by incorporating larger and more diverse datasets to ensure generalizability and robustness of the engagement detection models.

Our study focused on offline analysis of facial images to detect engagement. Future research should aim to develop real-time engagement detection systems that can operate in dynamic and interactive settings. This would enable the integration of engagement detection algorithms into interactive technologies, such as robots, to provide immediate feedback and adaptive interventions.

Facial emotion recognition is just one modality for measuring engagement. Future research should explore the integra-

tion of other modalities, such as speech analysis, body movement, and physiological signals, to create more comprehensive and accurate engagement detection models. Multimodal approaches can enhance the understanding of engagement dynamics and provide a more holistic assessment.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

- [1] Buckley, S.; Hasen, G.; Ainley, M. *Affective Engagement: A Person-Centered Approach to Understanding the Structure of Subjective Learning Experiences*. Melbourne, Australia: Australian Association for Research in Education 2004.
- [2] Banire, B.; Thani, D. Al; Qaraq, M.; Mansoor, B. Face-Based Attention Recognition Model for Children with Autism Spectrum Disorder. *J Healthc Inform Res* 2021, doi:10.1007/s41666-021-00101-y.
- [3] Schulc, A.; Cohn, J.F.; Shen, J.; Pantic, M. Automatic Measurement of Visual Attention to Video Content Using Deep Learning. In *Proceedings of the 2019 16th International Conference on Machine Vision Applications (MVA)*; IEEE; pp. 1–6, doi:10.23919/MVA.2019.8758046.
- [4] Sharma, S.R.; Gonda, X.; Tarazi, F.I. Autism Spectrum Disorder: Classification, Diagnosis and Therapy. *Pharmacol Ther* 2018, 190, 91–104, doi:10.1016/j.pharmthera.2018.05.007.
- [5] Rakhymbayeva, N.; Amirova, A.; Sandygulova, A. A Long-Term Engagement with a Social Robot for Autism Therapy. *Front Robot AI* 2021, 8, 14, doi:10.3389/frobt.2021.669972.
- [6] Di Nuovo, A.; Conti, D.; Trubia, G.; Buono, S.; Di Nuovo, S. Deep Learning Systems for Estimating Visual Attention in Robot-Assisted Therapy of Children with Autism and Intellectual Disability. *Robotics* 2018, 7, 21, doi:10.3390/robotics7020025.
- [7] Khan, M.; Chakraborty, S.; Astya, R.; Khepra, S. Face Detection and Recognition Using OpenCV. In *Proceedings of the 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*; IEEE, doi:10.1109/ICCCIS48478.2019.8974493.
- [8] Liu, W.B.; Wang, Z.D.; Liu, X.H.; Zeng, N.Y.; Liu, Y.R.; Alsaadi, F.E. A Survey of Deep Neural Network Architectures and Their Applications. *Neurocomputing* 2017, 234, 11–26, doi:10.1016/j.neucom.2016.12.038.
- [9] Foff, R.M. Applied Behavior Analysis Treatment of Autism: The State of the Art. *Child Adolesc Psychiatr Clin N Am* 2008, 17, 821–834, doi:10.1016/j.chc.2008.06.007.
- [10] Salmam, F.Z.; Madani, A.; Kissi, M. Facial Expression Recognition Using Decision Trees. In *Proceedings of the 2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*; IEEE; pp. 125–130, doi:10.1109/CGiV.2016.33.
- [11] Pu, X.; Fan, K.; Chen, X.; Ji, L.; Zhou, Z. Facial Expression Recognition from Image Sequences Using Twofold Random Forest Classifier. *Neurocomputing* 2015, 168, 1173–1180, doi:10.1016/j.neucom.2015.05.005.
- [12] Harguess, J.; Aggarwal, J.K. Is There a Connection between Face Symmetry and Face Recognition? In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*; IEEE Computer Society, 2011; pp. 66–73, doi:10.1109/CVPRW.2011.5981805.
- [13] Tan, D.W.; Gilani, S.Z.; Boutrus, M.; Alvares, G.A.; Whitehouse, A.J.O.; Mian, A.; Suter, D.; Maybery, M.T. Facial Asymmetry in Parents of Children on the Autism Spectrum. *Autism Research* 2021, 14, 2260–2269, doi:10.1002/aur.2612.
- [14] Briot, K.; Pizano, A.; Bouvard, M.; Amestoy, A. New Technologies as Promising Tools for Assessing Facial Emotion Expressions Impairments in ASD: A Systematic Review. *Front Psychiatry* 2021, 12, doi:10.3389/fpsy.2021.634756.
- [15] Dukić, D.; Sović Krzić, A. Real-Time Facial Expression Recognition Using Deep Learning with Application in the Active Classroom Environment. *Electronics (Basel)* 2022, 11, 1240, doi:10.3390/electronics11081240.
- [16] Nielsen, M. *Neural Networks and Deep Learning*; 2019.
- [17] Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Networks* 2015, 61, 85–117, doi:10.1016/j.neunet.2014.09.003.

- [18] Dalal, K.R.; Ieee Review on Application of Machine Learning Algorithm for Data Science. Proceedings of the 2018 3rd International Conference on Inventive Computation Technologies (Icict 2018) 2018, 270–273, doi:10.1109/ICICT43934.2018.9034256.
- [19] O’Shea, K.; Nash, R. An Introduction to Convolutional Neural Networks. ArXiv 2015.
- [20] Wang, Y.; Yuan, G.W.; Zheng, D.; Wu, H.; Pu, Y.Y.; Xu, D. Research on Face Detection Method Based on Improved MTCNN Network. In Proceedings of the 11th International Conference on Digital Image Processing (ICDIP); 2019; Vol. 11179, doi:10.1117/12.2539617.
- [21] Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A Survey of Transfer Learning. J Big Data 2016, 3, 1–40, doi:10.1109/SIBGRAP-T.2019.00010.
- [22] Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. Proceedings of the IEEE 2020, 109, 43–76.
- [23] Trabelsi, Z.; Alnajjar, F.; Parambil, M.M.A.; Gochoo, M.; Ali, L. Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student’s Behavior Recognition. Big Data and Cognitive Computing 2023, 7, 48, doi:10.3390/bdcc7010048.
- [24] Gupta, S.; Kumar, P.; Tekchandani, R.K. Facial Emotion Recognition Based Real-Time Learner Engagement Detection System in Online Learning Context Using Deep Learning Models. Multimed Tools Appl 2023, 82, 11365–11394, doi:10.1007/s11042-022-13558-9.
- [25] Banire, B.; Al Thani, D.; Qaraq, M.; Mansoor, B. Face-Based Attention Recognition Model for Children with Autism Spectrum Disorder. J Healthc Inform Res 2021, 5, 420–445, doi:10.1007/s41666-021-00101-y.
- [26] Rathod, M.; Dalvi, C.; Kaur, K.; Patil, S.; Gite, S.; Kamat, P.; Kotecha, K.; Abraham, A.; Gabralla, L.A. Kids’ Emotion Recognition Using Various Deep-Learning Models with Explainable AI. Sensors 2022, 22, 8066, doi:10.3390/s22208066.
- [27] Mujeeb Rahman, K.K.; Subashini, M.M. Identification of Autism in Children Using Static Facial Features and Deep Neural Networks. Brain Sci 2022, 12, 94, <https://doi.org/10.3390/brainsci12010094>.
- [28] Khan, R.A.; Crenn, A.; Meyer, A.; Bouakaz, S. A Novel Database of Children’s Spontaneous Facial Expressions (LIRIS-CSE). Image Vis Comput 2019, 83, 61–69.
- [29] Gerry Autistic Children Data Set 2020, 2022.
- [30] Hosseini, M.-P.; Beary, M.; Hadsell, A.; Messersmith, R.; Soltanian-Zadeh, H. Deep Learning for Autism Diagnosis and Facial Analysis in Children. <https://doi.org/10.3389/fncom.2021.789998>.
- [31] Russell, J.A. A Description of the Affective Quality Attributed to Environments. Journal of Personality and Social Psychology 38(2):311-322, doi:10.1037/0022-3514.38.2.311 1980.
- [32] Pietro, C.; Silvia, S.; Giuseppe, R. The Pursuit of Happiness Measurement: A Psychometric Model Based on Psychophysiological Correlates. Scientific World Journal 2014, doi:10.1155/2014/139128.
- [33] Nicolaou, M.A.; Gunes, H.; Pantic, M. Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space. IEEE Trans Affect Comput 2011, 2, 92–105, doi:10.1109/t-affc.2011.9.
- [34] Lytridis, C.; Kaburlasos, V.G.; Bazinas, C.; Papakostas, G.A.; Sidiropoulos, G.; Nikopoulou, V.-A.; Holeva, V.; Papadopoulou, M.; Evangelio, A. Behavioral Data Analysis of Robot-Assisted Autism Spectrum Disorder (ASD) Interventions Based on Lattice Computing Techniques. Sensors 2022, 22, 621, <https://doi.org/10.3390/s22020621>.
- [35] Hidalgo-Muñoz, A.R.; López, M.M.; Santos, I.M.; Pereira, A.T.; Vázquez-Marrufo, M.; Galvao-Carmona, A.; Tomé, A.M. Application of SVM-RFE on EEG Signals for Detecting the Most Relevant Scalp Regions Linked to Affective Valence Processing, <https://doi.org/10.3390/s22020621>.
- [36] Kuusikko, S.; Haapsamo, H.; Jansson-Verkasalo, E.; Hurtig, T.; Mattila, M.-L.; Ebeling, H.; Jussila, K.; Bölte, S.; Moilanen, I. Emotion Recognition in Children and Adolescents with Autism Spectrum Disorders. J Autism Dev Disord 2009, 39, 938–945, doi:10.1007/s10803-009-0700-0.
- [37] Weigelt, S.; Koldewyn, K.; Kanwisher, N. Face Identity Recognition in Autism Spectrum Disorders: A Review of Behavioral Studies. Neurosci Biobehav Rev 2012, 36, 1060–1084, doi:10.1016/j.neubiorev.2011.12.008.