Behavior Recognition for ASD Children through Audio & Video for Psychological Intervention using AI

Rubini.P[1], Midhunchakkaravarthy[2], Hemalatha[3]

[1] Lincoln University College, Malaysia; [2] Lincoln University College, Malaysia; [3]Panimalar Engineering College, India

Pdf.rubini@lincoln.edu.my, midhun@lincoln.edu.my, pithemalatha@gmail.com

Abstract

Autism Spectrum Disorder (ASD) is a neurological condition that impacts an individual's cognitive, emotional, physical, and social well-being. This research focus on a multimodal method approach that utilizes both video and audio data. By integrating analyses of facial expressions and speech-related emotional indicators, this approach seeks to enhance the accuracy and reliability of autism diagnostics. Traditional methods, often limited to observational techniques and behavioral assessments, may not fully capture the subtle nuances of autism spectrum disorders (ASD). However, by analyzing synchronized video and audio data, it becomes possible to detect intricate patterns and variations in facial and vocal expressions that are characteristic of ASD. This multimodal system not only provides a richer dataset for analysis but also enables a more comprehensive understanding of the emotional and communicative cues associated with autism. Recognizing the gestures of autistic children is crucial for preventing meltdowns and self-harm. We introduced a method to identify gestures by detecting poses through a person pose estimation technique. The features extracted from the pose estimation are then used to develop a gesture classification model using supervised learning algorithms. Our proposed model achieved the highest accuracy with the Random Forest technique, exhibiting evaluation metrics of 83% precision and 71% recall.

Keywords: Autistic Children; Gesture Identification; Person Pose Estimation; Supervised learning; Random Forest Technique

Introduction

Autism Spectrum Disorder (ASD) impacts social communication and interaction in both children and adults. Early detection and intervention can enhance the health, development, and mental well-being of affected children. Children with autism often respond differently to sensory inputs; for example, they might close their ears in response to loud noises or shut their eyes when disturbed by flickering fluorescent lights or fluttering curtains and posters. Research indicates that before a meltdown, children exhibit signs of distress, known as the rumble stage [1][2]. Observable involuntary cues of anxiety, such as head banging, self-scratching, wrist biting, kicking, and hand-flapping, typically precede a meltdown. Identifying these behaviors early can help parents and teachers manage the situation more effectively.

This research is organized as follows: Section II discusses related work on pose estimation and gesture recognition. Section III explains the proposed new framework for gesture detection.

Related work

Recently, artificial intelligence has made significant strides in various domains, including image classification, speech recognition, text classification, and more. Numerous studies have been conducted on autism detection utilizing machine learning and deep learning techniques. Authors in [1] propose a deep learning approach to identify autism meltdowns. A classifier was developed to predict gestures from images that can be used to identify meltdowns. The model was trained using a RCNN supported by the retrained GoogleNet model and achieved a validation accuracy of 93%. The training process included a loss classifier with a marginal loss of 0.4%. The authors in [3] propose a single-shot model using a box-free, downside-up approach for instance segmentation and pose estimation. A Convolutional Neural Network (CNN) is used to group key points for person pose estimation by detecting significant points and their displacements. The model was trained on the COCO dataset [4], achieving an average precision of 0.67 on the COCO test dataset using a single-scale inference model, and 0.69 using a multi-scale inference model, outperforming previous models. Specifically, it achieved an average precision of 0.42 for the person category. The classification of children as ASD based on body movement behaviors is discussed in [5].This study involved twenty-four autistic children and twenty-five children with neurodevelopmental disorders participating in a virtual reality event. Body movement variations were recorded using a depth-sensor camera, revealing that children with ASD exhibited more bodily movements compared to others. The head, body, and foot movements demonstrated the highest classification accuracy, achieving 83%.

Proposed Methodology

The proposed work introduces a novel method for identifying gestures in autistic children. The input image, captured via a webcam, is processed through the framework. The architecture is divided into two stages: the first stage focuses on person pose estimation, and the second stage handles gesture recognition, as illustrated in Fig. 1.

target classes namely “close ears” and “close eyes”.

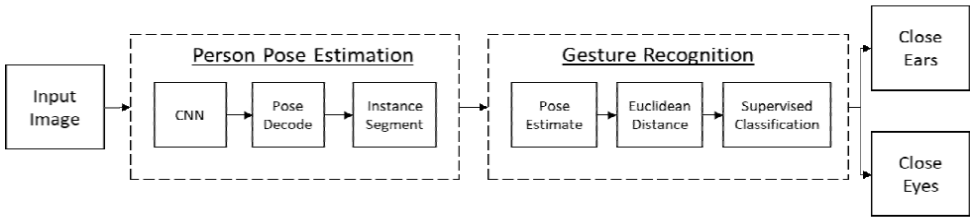

Fig. 1. Proposed Architecture

A. Person Pose Estimation

For this analysis, the COCO dataset [4] was utilized, specifically selecting images that depict actions such as closing ears and eyes. A CNN model with a ResNet-101 backbone was employed, using a

batch size of 8 images and trained over 30 epochs. The model predicts key points of a person in the image through a hybrid classification and regression approach, generating heat maps and displacements. Subsequent steps involve predicting short-range offsets and mid-range pairwise offsets for the pose estimation model. Person segmentation maps and long-range offsets are utilized in the instance segmentation module to produce pose estimates for 17 key points on the body, which are then used for the gesture recognition model.

B. Gesture Recognition

Seventeen pose key points coordinates are utilized to construct a feature vector for the gesture classification model based on Euclidean distance. Examples of pose key points include leftEye, rightEye, leftEar, rightEar, nose, leftWrist, rightWrist, among others. The distances between the wrist and the positions of the ears and eyes are computed using Euclidean Distance as shown in equation (1). These distances serve as features for building a binary classifier to detect gestures such as "close ears" and "close eyes." Various classifiers are tested, including Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Random Forest.

$$d_{1,2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

Learning Algorithms

Autism Detection Methodology

We introduce an innovative approach for detecting autism in children by leveraging both video and audio data to analyze facial and speech-related emotional indicators. The design of this autism detection system is depicted in Fig. 2.

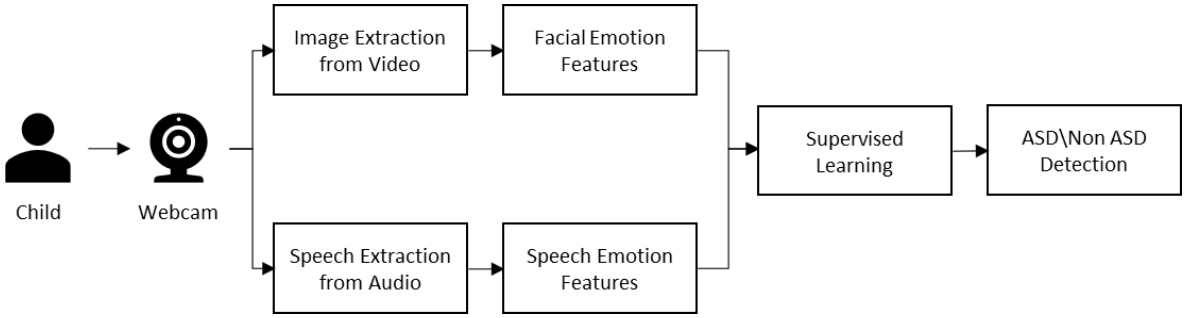

Fig 2. - Architecture for Autism Detection

Live video and audio data are collected using a webcam and inputted into the system. From the video stream, individual frames are processed to extract images, which are then analyzed by a facial emotion recognition module to obtain facial emotion features. Concurrently, audio is processed to isolate speech, which is analyzed by a speech emotion recognition module to extract features related to emotional expressions in speech. These extracted features from both facial and speech modules are then utilized in a supervised learning framework to classify and identify signs of autism.

The autism dataset, comprising facial images of autistic and non-autistic children, is sourced from a Kaggle competition organized by Gerry Piosenka [6]. It includes 1468 JPEG images. To train the facial emotion recognition model, we utilized the FER-2013 dataset [7]. As illustrated in Fig. 3, the pipeline consists of

several stages: image preprocessing, feature extraction, and model training using a Convolutional Neural Network (CNN) for facial emotion recognition, followed by predictions on images of autistic and non-autistic children to extract facial emotion features.

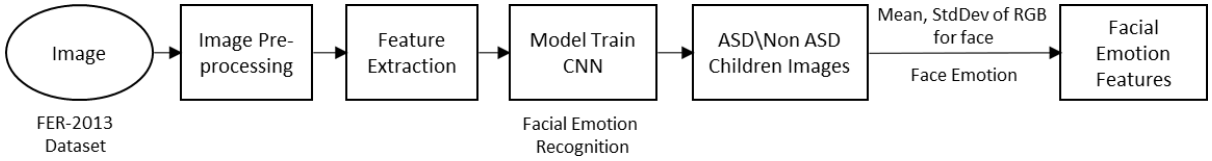

Fig. 3 - Model Architecture for Facial Emotion Recognition

In the preprocessing stage, images are enhanced to reduce noise and adjust exposure and brightness using OpenCV [8], which sets the stage for subsequent feature extraction. The architecture of the model, depicted in Fig. 4, is designed for Facial Emotion Recognition and categorizes emotions into seven distinct classes: angry, disgusted, fearful, happy, neutral, sad, and surprised. At the conclusion of the model, two fully connected convolutional layers are incorporated. Following this, a dense layer consisting of 128 neurons, employing an L2 regularization value of 0.015 and a ReLu activation function, is linked to the final prediction layer, which utilizes a Softmax activation function for classification.

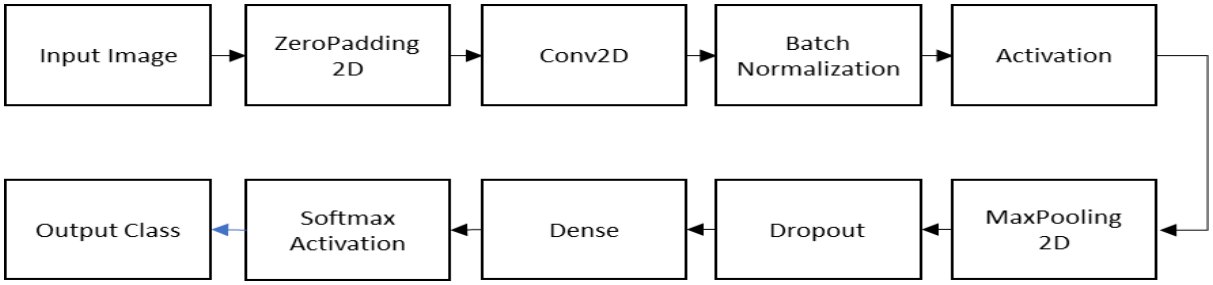

Fig. **Error! No text of specified style in document.** - CNN Model Architecture for Facial Emotion Recognition

Fig. 5 illustrates the Facial Emotion Recognition process applied to images of autistic children. Features from facial images are extracted within the designated boundary boxes for each face, involving the computation of mean and standard deviation values for the RGB channels of the images. These are then amalgamated with the detected facial emotions to form comprehensive facial emotion features, which are subsequently inputted into the supervised learning model. This provides as a robust approach to maintain the stability and performance on the model without comprising the accuracy. This model can be used in real-time as its light weight and can be used for diverse applications.

Fig. 5 - Facial Emotion Recognition

The model's accuracy and loss metrics are presented in Fig. 6. Throughout the training phase, which spanned 50 epochs, the model achieved an accuracy of 0.89 on the training set. On the test dataset, specifically for facial emotion recognition tasks, the model's accuracy stood at 0.62.

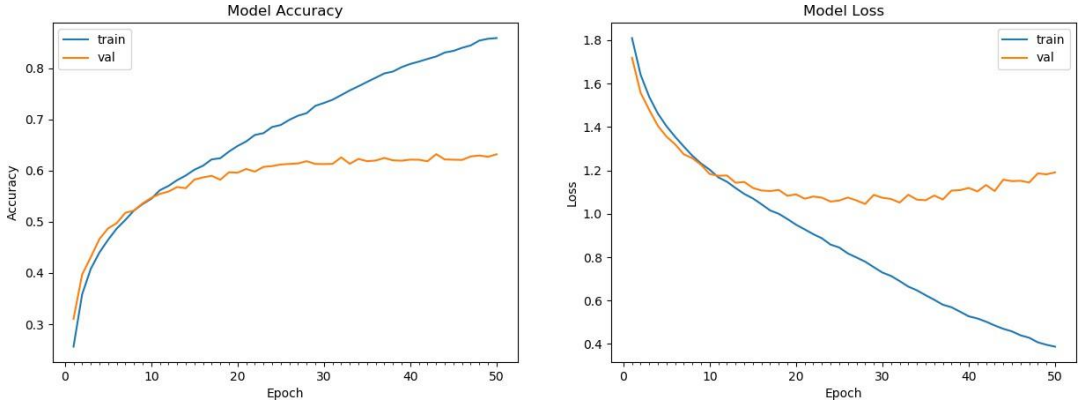

Fig. 6 - Model Accuracy and Loss for Facial Emotion Recognition

Speech Emotion Recognition (SER)

The dataset utilized for training the speech emotion recognition model is the Berlin Emotional Speech Database (Emo-DB) [9], which comprises recordings of emotional speech by actors, encapsulating emotions such as anger, happiness, sadness, fear, disgust, and neutrality. A separate speech dataset from [10] containing MP3 format recordings of voices from autistic children is employed. As depicted in Fig. 7, the architecture encompasses processes for extracting MFCC (Mel-Frequency Cepstral Coefficients) features, employing a CNN (Convolutional Neural Network) for Speech Emotion Recognition to analyze speech data from autistic and non-autistic sources. This aims to identify speech emotions, which then assist in extracting speech features for the subsequent supervised learning model.

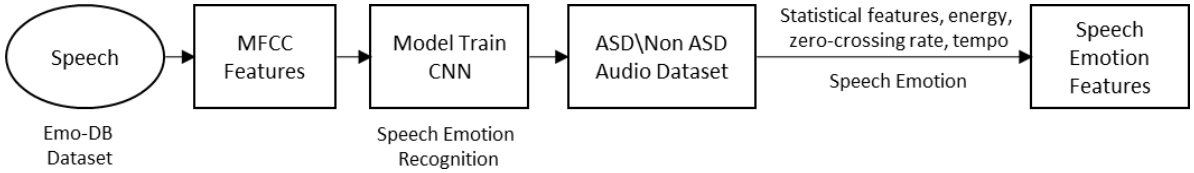

Fig. 7- Speech Emotion Recognition Architecture

The training architecture utilizes convolutional layers incorporating MFCC features extracted from the speech data. This model's architecture is illustrated in Fig. 8.

Fig. 8 - Model Architecture for Emotion Recognition from Speech

Mel Frequency Cepstral Coefficients (MFCC) features are extracted from the input speech signal, and the model's structure includes 1 to 6 convolutional layers, a Bi-LSTM (Bidirectional Long Short-Term Memory) network with 1 to 4 layers, and a dense layer employing a softmax activation function for categorization. Stochastic gradient descent is utilized for optimization, with a batch size set to 16 and L2 regularization applied. The model achieves an overall accuracy of 86% after training over 50 epochs. These attributes, combined with the detected speech emotions, are inputted into the supervised model to facilitate the prediction of ASD/non-ASD status.

Fig.9 displays a spectrogram for the sad emotion alongside the model's loss and accuracy metrics. Speech characteristics for the supervised model are extracted from the input speech data of both autistic and non-autistic individuals. This includes statistical attributes such as mean, median, and standard deviation, as well as the energy, zero-crossing rate, and tempo of the speech.

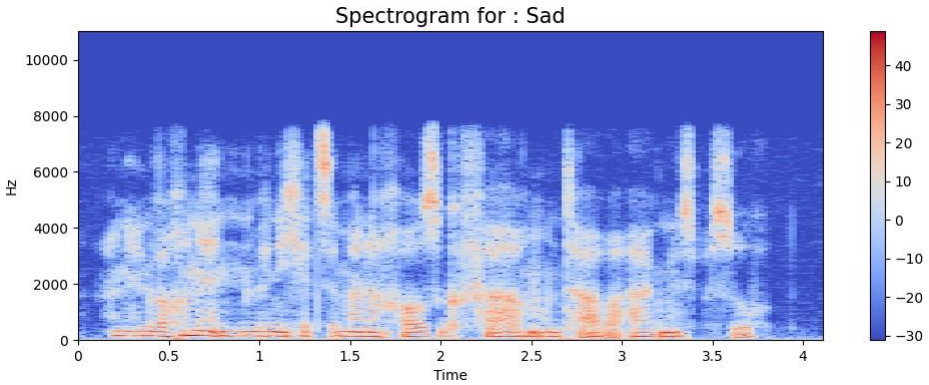

Fig.9 - Sample Spectrogram

Model loss and model accuracy is shown in Fig.10. Model achieved an overall test accuracy around 86% and a loss of 0.2.

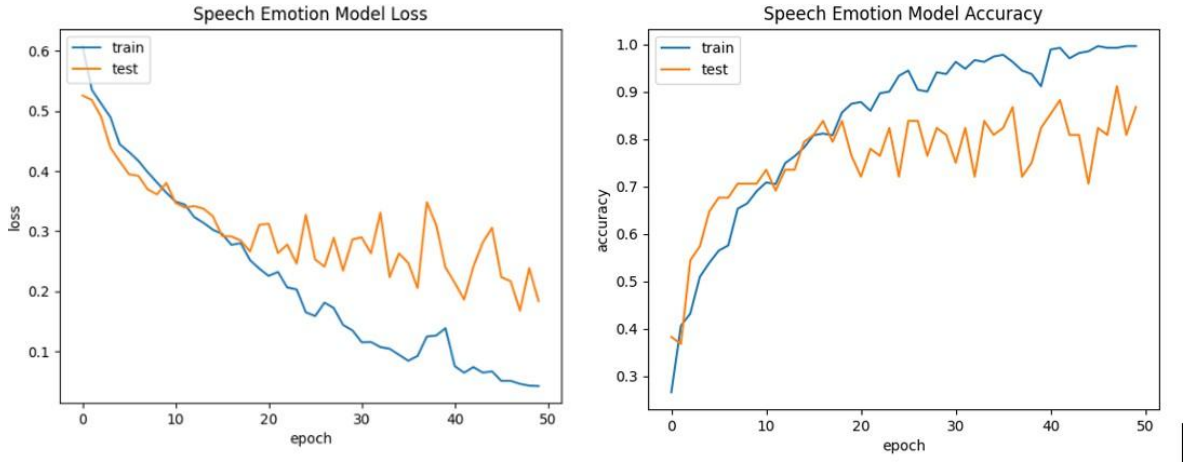

Fig.10 - Model loss and model accuracy

Supervised Models for ASD Detection

Facial and speech emotion features from the previous steps were used to build a binary classification model using supervised techniques like Logistic Regression, Random Forest, K-Neighbors Classifier, and Support Vector Machine. We had a total of 22 features for both facial and speech features combined. Classification reports for all the methods are shown. In Fig.11, we could see that Logistic Regression outperforms the other techniques with an F1 score of 0.70 for Non ASD and 0.6 for ASD.

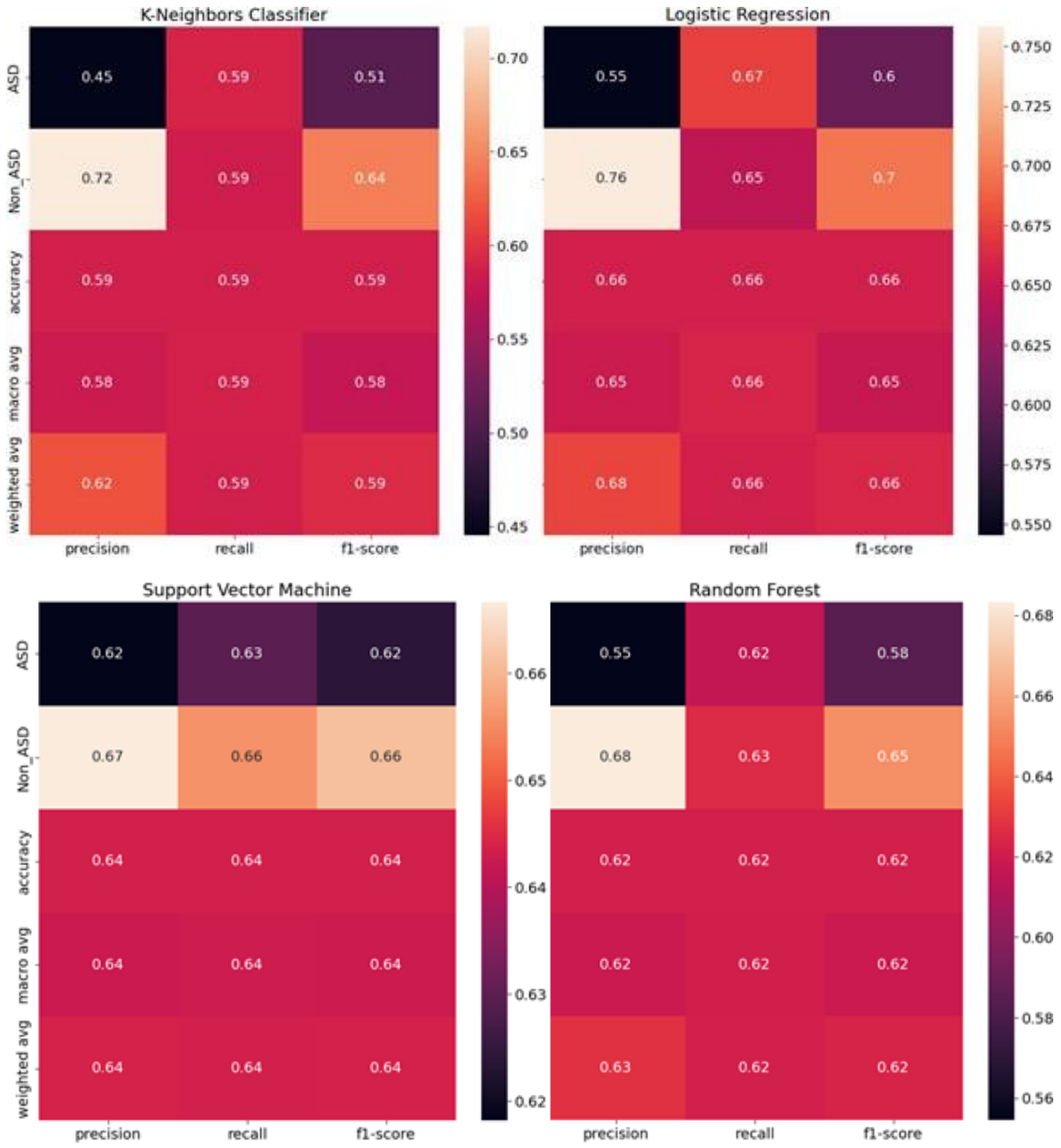

Fig.11 - Classification report for supervised techniques

Conclusion

This research explores an innovative multimodal approach to detecting autism in children, utilizing both video and audio data to analyze facial and speech-related emotional indicators. This method enhances the accuracy and reliability of autism diagnostics beyond traditional observational techniques, which may overlook the subtle nuances of autism spectrum disorders (ASD). By processing synchronized video and audio data, the system identifies complex patterns in facial and vocal expressions indicative of ASD.

References

[1] V. S. P. Patnam, F. T. George, K. George and A. Verma, "Deep Learning Based Recognition of Meltdown in Autistic Kids," 2017 IEEE International Conference on Healthcare Informatics (ICHI), 2017, pp. 391-396, doi: 10.1109/ICHI.2017.35.

[2] Erikk Linstead, and Renee, “An Application of Neural Networks to Predicting Mastery of Learning Outcomes in the Treatment of Autism Spectrum Disorder,” IEEE 14th International Conference on Machine Learning and Applications, California, 2015.

[3] George Papandreoua, Tyler Zhuu, Liang-Chieh Chenn, Spyros Gidaris, Jonathan Tompson, “PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model.” ECCV 2016.

[4] Lin, Tsung-Yii, M. Mairre, Serge J. Belongiee, James Hayes, P. Perona, D. Ramanan, Piotr Dollár and C. L. Zitnick. “Microsoft COCO: Common Objects in Context.” ECCV (2014).

[5] Alcañiz Raya and Mariano, “Machine Learning and Virtual Reality on Body Movements' Behaviors to Classify Children with Autism Spectrum Disorder.” Journal of clinical medicine vol. 9,5 1260. 26 Apr. 2020, doi:10.3390/jcm9051260.

[6] Gerry, "Detect Autism from a facial image" Retrieved May 03, 2021 from https://www.kaggle.com/gpiosenka/autistic-children-data-set-traintestvalidate.

[7] Gerry, "Learn facial expressions from an image" Retrieved December 15, 2021 from https://www.kaggle.com/deadskull7/fer2013.

[8] Bradski, G, “The OpenCV Library,” Dr. Dobb' Journal of Software Tools, 2000.

[9] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter Sendlmeier und Benjamin Weiss, “A Database of German Emotional Speech,” Proceedings Interspeech 2005, Lis-sabon, Portugal, 2005.

[10] Hendriks, P., Koster, C., & Hoeks, J. C. J. (2014). Referential choice across the lifespan: why children and elderly adults produce ambiguous pronouns. Language, Cognition and Neuroscience 29:4, 391-407.