Affective Computational Model to Extract Natural Affective States of Students with Asperger Syndrome (AS) in Computer-based Learning Environment.

Amina Dawood[1], Scott Turner[2], Prithvi Perepa[3]

University of Northampton, Northampton, United Kingdom

amina.dawood@northampton.ac.uk

Abstract

This study was inspired by looking at the central role of emotion in the learning process, its impact on students' performance; as well as the lack of affective computing models to detect and infer affective-cognitive states in real time for students with and without Asperger Syndrome (AS). This model overcomes gaps in other models that were designed for people with autism, which needed the use of sensors or physiological instrumentations to collect data. The model uses a webcam to capture students' affective-cognitive states of confidence, uncertainty, engagement, anxiety, and boredom. These states have a dominant effect on the learning process. The model was trained and tested on a natural-spontaneous affective dataset for students with and without AS, which was collected for this purpose. The dataset was collected in an uncontrolled environment and included variations in culture, ethnicity, gender, facial and hairstyle, head movement, talking, glasses, illumination changes and background variation. The model structure used deep learning (DL) techniques like convolutional neural network (CNN) and long short-term memory (LSTM). DL is the-state-of-art tool that used to reduce data dimensionality and capturing non-linear complex features from simpler representations. The affective model provide reliable results with accuracy 90.06%. This model is the first model to detected affective states for adult students with AS without physiological or wearable instruments. For the first time, the occlusions in this model, like hand over face or head were considered an important indicator for affective states like boredom, anxiety, and uncertainty. These occlusions have been ignored in most other affective models. The essential information channels in this model are facial expressions, head movement, and eye gaze. The model can serve as an aided-technology for tutors to monitor and detect the behaviors of all students at the same time and help in predicting negative affective states during learning process.

INDEX TERMS Affective Model, Affective-Cognitive States, Autism, Asperger Syndrome, AS, CNN, Deep Learning, LSTM.

I. Introduction

Autism Spectrum Disorder (ASD) is a development disorder which can cause three main impairments: social communication, social interaction and social imagination [1].There are three common types of autism spectrum disorders: Autism, Asperger Syndrome, and Pervasive Developmental Disorder - Not Otherwise Specified (PDD-NOS). All these types share most symptoms; therefore, it is difficult to distinguish between ASD types. But they can be recognised based on severity and impact [1]. There is no cure or treatment for ASD [2]. But, on the other hand, there is agreement that educational intervention programs and intensive behavioural training, can lead to improve deficits in social communication for ASD [3]; [4]. Despite this, these resources are expensive and not easy to use for ASD and their families [5]. To fulfil these needs different methods and technical tools have been developed to help people with autism to improve communication in their daily life. Researchers have focused on interactive intervention tools to teach those with ASD how to interact with others socially [4], [5], [6], [6], [7] [8]. Even with all of this attention from researchers, there is still a gap in models concerning the detection of affective and emotional states of autism in a real environment [5]. Normally, standard systems and models that detect emotions or affective states of students with TD, can't detect the affective state of students with autism, due to the lack of facial expressions and eye gaze of people with autism [9]; [10].

Recently, attention has been shifted towards recognising and analysing facial expressions of those with ASD [11] [12] based on physiological signals and wearable instruments like electromyography (EMG). Unfortunately, using these tools come with restrictions, most importantly; they often need to be implemented in controlled environments instead of real-world environments. Using wearable sensors can lead to negative reactions because people can become annoyed and uncomfortable when wearing these instruments, this may obscure natural-spontaneous emotions. There can also be high cost of these instruments and the applications needed to control them [13]. As a result, it is undesirable to use these tools in a natural interactive environment [14]. The relation between emotion and learning continues to attract the attention of researchers from different areas in education, psychology, and social sciences [15] [16]. This relationship influences learning quality and performance domains and concludes that learning experience implicitly involves a range of emotions, which motivate students during the learning process [17]. Considering the inadequacy of existing works in detecting and analysing facial expressions of AS in the learning environment, this research aims to develop an affective computational model to capture, and infer affective states for facial expressions, eye gaze, head gestures of students with AS and TD in computer-based learning environments, through a low-cost device, a webcam. The study proposed a novel model to detect affective-cognitive states for AS students of confidence, uncertainty, engagement, anxiety, and boredom. These behaviours are the most dominant on particpant's learning interaction [18]; [19]; [20]. In addition, anxiety is strongly coupled with autism in both children and adults, also it has a strong impact on students' communication skills [21] and can obstruct students learning, performance, and cognition [22]; [23]. Most of the existing automated affective models depend on posed or acted emotional datasets, collected in a controlled environment with specific scenarios which differ from that in a real environment [24]. Therefore, this model used a dataset of natural and spontaneous examples collected from students with AS and TD in an uncontrolled environment.

Recently, Deep learning has popularity in problem classification and recognition. This popularity due its high performance and less need to features engineering. Deep learning has two main models are convolutional neural network (CNN) and Recurrent neural network (RNN) [25].

CNN has the power to learn from images (features-map). It has ability to abstract high level of features from raw data. RNN is the best model for sequential data. RNN has two types are gated recurrent unit (GRU) and long short-term memory (LSTM). The model in this paper was built by choosing CNN and LSTM to learn from a series of temporal-data (response-maps series). To date, there are a few studies, which have adapted these techniques in facial expressions recognition methods [26]. This is believed to be the first computational model developed to extract complex emotions of students with AS using hybrid deep learning techniques.

The rest of paper is organised as follows. Next section presents the literature review. Section III the important methodology that used to build the proposed model and the results. Finally, the conclusion of the paper.

II. Literature review

Previous studies focused on classifying students' emotions in the learning environment, and detected affective states of autism or AS.

Generally, there have been efforts made towards modelling students (typical development) affective states in a computer-based learning environment. The nature and concept of affective and emotional states depends on different factors such as context environment, stimuli, and source of information (verbal, non-verbal cues) [27]. For example, emotions or affects that are generated from facial expressions in Human-Computer Interaction (HCI), differ from those extracted from verbal channels and may also differ from those generated in Human-Human Interaction (HHI) or through interacting with a game, or intelligent tutoring system (ITS).

Modelling affective states of students has witnessed growing development in the past decades and with work targeted on students' affective state in the learning environment. For example, [19]; [28]; [29] refined user affects through educational games to analyse students' attitudes towards an agent. The researchers used OCC cognitive theory of emotions [30], which contains 22 of emotions as valanced reaction and dynamic decision networks in modelling students affect. However, these studies did not achieve empirical model accuracy, as the model parameters were estimated by hand. In [14]the students' motivation was diagnosed, they used screen capture of students during interaction with ITS to infer rules of motivation. The study relied on visible material such as mouse and keyboard movement, and students' performance; but these rules have not been validated by the researchers and remained as theoretical assumptions. Another work carried out in the education domain by [31]; [32]; [33]; and [34] to detect frustration and stress in automated learning companions, and intelligent tutoring systems. In a study of [35] the focus was on action units of facial expressions (AU). The researchers used a commercially available Computer Expression Recognition Toolbox (CERT), to detect action units (AUs) for facial expressions of students. This study emphasises the relation between AU and learning outcome. Another commercial tool, FaceReader (5.0) was used in the research of [36] [37] to recognise learners' emotions in interacting with pedagogical agent in MetaTutor.

The relationship between affective state and learning with another tool Auto-Tutor was investigated by [15] [38]. In these studies, the observers were trained on the facial action coding system (FACS) to recognise affective states like neutral, confusion, flow, frustration, and boredom; experienced by the students. Another focus on AU can be seen in [39], where they investigated the relationship between engagement and frustration via tutoring system outcomes through AU of students.

It can be seen that, all these works focused on detecting affective states or AU for the typical development of students. There is no explicit automated model for students with autism to infer their affect in a computer-based learning environment compared with those works used to analyze students with TD.

In the past decades, many works [40]; [41]; [42]; [43]; [44] succeeded in capturing facial expressions of those with autism to compare the quality of expressions between ASD and TD. These works used mimic or posed expression methods for specific basic emotions, then trained-observers analysed these expressions. For instance, in [40], a group of children with autism and another group with typical development (TD) were instructed to act out happy and sad emotions. Images of the emotions portrayed were captured and analysed by observer. The researcher concluded that, the facial expressions of children with autism were lower in quality when compared with the typical group and tried in another study to extract more from non-verbal cues. The experiment was carried out again by instructing those with ASDs to mimic emotions with a visual feedback (mirror) and without it. The researcher found that with the mirror, facial expressions of children ASD were on par with those of typical development. Their findings were that the lower quality in emotions production was in students with ASD rather than in TD, in addition to the inability to observe and capture all facial expression components by the observer. [41] extended the study of [40] to conclude another facial expression comparison between those with high-functioning ASD and adult TDs during in emotion production. The researchers photographed those with ASD and adult TDs during the production of facial expressions for emotions including happy, sad, anger, and neutral. The photographs were analysed and labelled by observers. They found errors in labelling negative emotions, while there were no differences in labelling positive emotions like happy. Comparisons continued between those with ASD and TD in the quality of production of expressions by mimicry (based on an external actor) and acted emotion (based on emotion concepts) [45], [46]. These experiments showed that ASD on par with TD in mimicking expressions.

Moreover, another mimicry method used to produce facial expressions in those with ASD can be found in [47], the subject was instructed to mimic the emotion that was acted out by and actor in a Mind Reading DVD [48]. The purpose was to understand the dynamics and mechanisms of facial expressions, and the deviation between those shown by participants with ASD and TD. Another example following the mimicry method can be seen in [49], in this study, a robot presented basic emotions and a child tried to imitate these emotions, a camera was prepared to capture the child's face during the imitation process. Another study [50], tried to improve vocal communication for children with autism, the study was based on imitation methods. However, mimic methods are the same as acted methods; they do not produce real feelings. Also, a manual observation of facial expressions in these methods is a difficult process and may miss micro interpretations; in addition, it is resource intensive, and time consuming [44].

Nowadays, there is an increase in the number of people with autism, which leads to an increase their numbers in mainstream schools, colleges, and universities [51]. Also, due to the growing development in technology-based intervention of ASD, researchers have shown that individuals with autism display a different range of affective and emotional states when interacting with computers [52]. Therefore, there is a need to modulate, perceive, and predict ASD affective states during interaction with a computer, automatically and without any manual intervention.

Recently, studies have often focused their attention on improving autism capabilities in enhancing social communication [4]; [5]; [6]; [7]; [8]; [53]; [54]. Few works are concerned in detecting affective states of autism [5]. These studies have considered the physiological signals as the important source to collect the data needed to recognise facial expressions for ASD or detect emotions. Researchers have commonly focused their attention on physiological signals-based affective state detection [11], [12] and wearable instruments. The best example of these studies can be found in [55]; [56]; [57]; [58]; [59]; [60]. The physiological signals are a rich source of data collection for studies concerning people with autism. However, the use of physiological instruments in affective state detection comes with restrictions as stated in early section.

However, with the growing development of computer vision and machine learning techniques, the question whether there is a continued need for special equipment to capture emotions of those with Autism is raised. These techniques have the potential to help us to overcome the limitation in physiological instruments; as well as developing a computational model to explore affective states of individuals with AS in the education domain. [61] stated that computers and advanced technologies have become very helpful and convenient to educate students with autism and TD, they are valuable resources to predict their emotions during interaction with a computer. Evidence approved that individuals with autism prefer to interact with computers instead of interacting with others [52]. Based on [62] there is no affective models were developed to analyse and investigate the interactions between AS or ASD with computers. As consequent of the important role of computers in education environment and the needs to affective model in natural settings, this study adapted one-to-one computer interaction to develop a new affective model to track and detect affective-cognitive states of adult students with AS, without any physiological tools, and without any intervention or manual observations. The proposed model based on natural-spontaneous affective-cognitive dataset collected in an uncontrolled environment for adult students with AS and TD.

III. Methodology

Initial stage of model data preperation process starts with face detection and tracking. From the captured face data, the model extracted a set of features; Action Units (AUs), eye gaze and head movements then fed this to the classification algorithm.

Instead of handcrafting the desired features, recent algorithms use deep learning (DL) for feature detection. DL used to reduce data dimensionality and capture non-linear complex features from simpler representations, in contrast to other methods that cannot handle complex representations.

This study used CNN to extract features from raw data and LSTM for time-series data prediction [63]. CNN extracted features in spatial space domain while LSTM deal with time-series (temporal) data.

CNN consists of input layer and output layer as well as number of hidden layers. The hidden layers in CNN consist of Convolutional layers, Pooling layers, Dropout layers, Fully Connected layers, and Normalization layers. Convolution layer is the core of CNN, it is a mathematic operation to merge two functions. These operations work as filters for certain information from raw input data. Convolutional network terminology has two fundamental keys Input data and Kernel filters. The output of these operations called feature map [25]; [26].

RNN have internal memory to store temporal information of input series, this memory helps RNN to predict desired classes based on previously stored information [63]. However, storing information for a long time makes RNN failed to capture long length sequence dependencies. LSTM introduced by Hochreiter & Schmidhuber as an efficient and gradient-based method to overcome RNN problems in the way it can deal with long length input sequences. LSTM structure commonly consists of number of gate units (an input gate, forget gate, and an output gate). The input and output gates designed to prevent irrelevant information to access the memory. In addition, it contains memory cells; each cell has fixed self-hidden units with recurrent connections. Then, a forget gate to discard far history data. More details about the layers and parameters of model structure in next sections

A. *Video pre-processing*

This research used collected natural-spontaneous data from facial expressions, head movements, and eye gaze. The dataset consists of 862 videos of students with AS and 545 videos for TD students. Each video has 30 frames/second and a duration (1s-1.5min) with resolution 640x480. Frame sequences of a fixed length were used as the basic structure to be reviewed in model training and testing. Data augmentation processes conducted on each frame included colour-space transformation (RGBA to Grayscale) and unified frame size through frame resizing (32x32 pixels). The steps below state the features and cues, which were extracted from the video.

B. *Features extraction*

The tracker took each video and processed it frame by frame and designated output, spatial and orientation values for the face models (2D and 3D) and 18 AUs. Features and dimensions used in this research include:

1. Head rotation axes, rotations about 3D space [Rx, Ry, Rz] with the origin as the centre of the head 3D model.
2. Left-eye gaze vector, 3D rotation angles [$G0_x$, $G0_y$, $G0_z$].
3. Right-eye gaze vector, 3D rotation angles [$G1_x$, $G1_y$, $G1_z$].
4. Eyes gaze angle, 2D angles [Ga_x, Ga_y].
5. AUs intensity (17 action-unit), the intensity is a value range between [0, 5] with 0 for minimum AU intensity and 5 for maximum intensity. Fig. 1 stated the intensity of A12, for intensity of AU12 the output would range from 0 (not present), 1 (present at minimum intensity), 5 (present at maximum intensity), with continuous values in between.

   The AUs intensities taken are; AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU 26, and AU45.
6. AUs occurrence (18 AUs), will take the value of zero (not detected), or one (detected). The tracker can detect the presence of the following AUs; AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU28, and AU45.

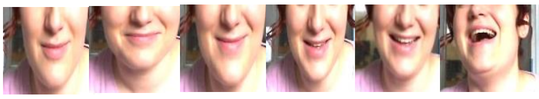

Fig. 1. **Stated the intensity range of AU12, 0 for absence of AU12 and 5 for maximum intensity**.

These features contain the main characteristics of head and face that fill the research interests. From these features, we get head pitch, yaw and roll from Rx, Ry and Rz respectively. Eye features like looking inside or outside the screen and their direction (up, down, center, left and right) from G0x, G0y, G0z, G1x, G1y, G1z, Gax and Gay. In addition to AUs added intensities and occurrences to get richer descriptors in the spatio-temporal domain of the video. The total number of features of interest to the research were 46 features per frame (listed above).

C. *Prediction cues*

Extracting complex affective states from each single frame is unfeasible due to the time component related to human capability to build a recognizable complex emotion, which cannot be captured in a single frame, on the other hand these affective states can be captured through a meaningful segment of a video. Time oriented video analysis approaches are adequate for complex emotion recognition because predictions of this type of affective states are governed abstractly by three stages (onset, apex and offset). These stages cover the first signs of the affective state (onset), the highest intensity of the shown state (apex), and the vanishing of this emotion (offset). Time can vary for the existence of each of these stages depending on various factors including physiological and behavioral arousals for the affective state.

To make the predicted period feasible an empirical time segment is taken to be one second. Each one-second segment of the video is a cue and labelled using the video's label.

The methodology for extracting emotion and prediction cues was done by taking the videos from the selected dataset first and then feeding them to the tracker and extracting the desired features. Each cue was encoded by five actions each action is a six-frame sequence used for predicting a sub domain from the whole cue. The cue consists of 30 frames each with 46 features, which yielded 1380 features, and each action has 230 features each.

IV. Building of Prediction model

The prediction model is an empirical design to examine the benefits of using the strengths of two most used architectures in deep learning. The model incorporates a sandwich model of CNN-LSTM-CNN. This structure takes the power of CNN to learn from images (frame's response-map) and LSTM to learn from series of temporal-data (sequence of response-maps). The parameters of the proposed model are outlined in tables 1-3.

The model is a sequential system, takes input data by reading frames from the input videos. A response map constructed using the features taken from the tracker by a transformation function T_f, $T_f(F) = F'$ where T_f is a local scaling function, F is the raw frame response-map, and F' is the transformed frame response-map. The sampling rate of this model is 30 frames per second; these frames are then sliced and stacked into 6 five-frame actions. The stacking function transforms the frames dimensionality from 1D to 2D (image) (Fig.2) as follows:

$$\{F'_0, F'_1, \dots, F'_m\} \overset{S}{\Rightarrow} \begin{bmatrix} F'_0(f_0) & F'_0(f_1) \dots & F'_0(f_n) \\ F'_1(f_0) & F'_1(f_1) & F'_1(f_n) \\ & \vdots & \\ F'_m(f_0) & F'_m(f_1) \cdots & F'_m(f_n) \end{bmatrix}$$

Where f_n is a transformed feature value; m = 0 to number of frames per action -1; n = 0 to number of features per frame -1; and S is the stacking function.

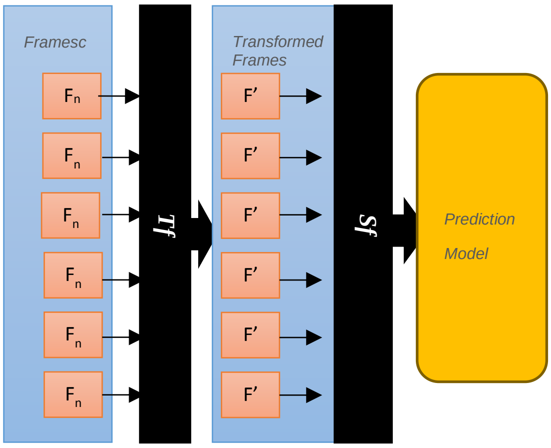

Fig. 2. **Model input process (Tf is transformation function and Sf is stacking function)**

The model network structured as four sub-networks namely First Convolutional Network, LSTM Network, Second Convolutional Network, and Prediction Network. The first convolutional layer with 256 nodes and *Tanh* activation function. Followed by a dropout layer with drop ratio of 20% and finally a fully connected layer serves as an output layer for first convolutional network Table-1.

TABLE 1. FIRST CNN NETWORK STRUCTURE

LAYER	SHAPE	PARAMS
INPUT LAYER	(6, 27)	0
CONV. 1D	(6, 64)	1792
DROPOUT	(6, 64)	0
FULLY CONNECTED	(6, 27)	1755
SUM		3547

The first convolutional net connected to the mid-subnetwork (i.e., LSTM stack) to extract the temporal features of the input sequences. The upper LSTM layer input has 162 weights from the first convolutional net. Both LSTM layers have 972 unit each, which indicate the number of states reside inside that layer. The second LSTM layer take a sequence of states provided from the first LSTM layer and yield the final temporal-prediction from the LSTM network, Table-2.

TABLE 2 LSTM NETWORK STRUCTURE

LAYER	SHAPE	PARAMS
LSTM	(6, 162)	123120
LSTM	(6, 162)	210600
SUM		333720

To complete the sandwich structure, a second CNN network added to gain more skills from the preceding networks. This sub-net structured as a block of four 1D convolutional nets with 768 nodes per net and *Tanh* activation function. A pooling layer reduces the dimensionality with a factor of 50% added as the last layer here Table-3.

TABLE 3 SECOND CNN NETWORK STRUCTURE

LAYER	SHAPE	PARAMS
CONV1D	(6, 128)	20864
CONV 1D	(6, 128)	16512
CONV 1D	(6, 128)	16512

CONV 1D	(6, 128)	16512
POOLING 1D	(128)	0
SUM		70400

At the end, the Prediction Network used to provide the final categorical prediction. This network will concatenate the output from all the sub-models then feed this vector to a fully connected layer. This layer has five units (the number of the dataset classes) and a *SoftMax* activation function.

The optimizer used is Adam Stochastic Gradient Descent optimizer, and the loss function is categorical cross-entropy.

A. *Training Data*

The natural dataset that was collected to perform this model contains different occlusions like hand over face or head, face turned left or right, and head up or head down, as shown in Fig. 3.

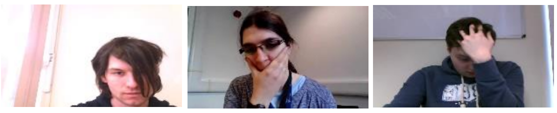

Fig. 3. Some of occlusion features that recognized as affective-cognitive states in this model

All works that are interested in automated emotions ignored these occlusions and the face was considered pure without these limitations. But, in this model for example hand over face or head refers to anxiety, or the affect339 of boredom, or may represent the thinking affect. For this purpose, this is study based on CNN and LSTM to detect features like these occlusions.

Number of video clips for autistic students provided by this dataset was 862 clips. From these clips, the researcher extracted a total number of 684364 cues, which were split using a ratio (80%, 20% (validation, test)) for training, validation, and test sets respectively, into the following:

SPLIT	NUMBER OF CUES
TRAINING	554334
VALIDATION	61593
TEST	68437

The model trained with 100 epochs reaches 5,543,340 training steps and 615,930 validation steps to achieve the following results:

	ACCURACY	LOSS
TRAIN	98.30%	0.0505
TEST	89.00%	0.5095
VALIDATION	89.05%	0.4999

Fig. 4 (a) and (b) stated plots of training and validation accuracy and loss error respectively. We noticed that the performance of the model improved by increasing number of epochs. It can be noticed that there is overfitting in the model results

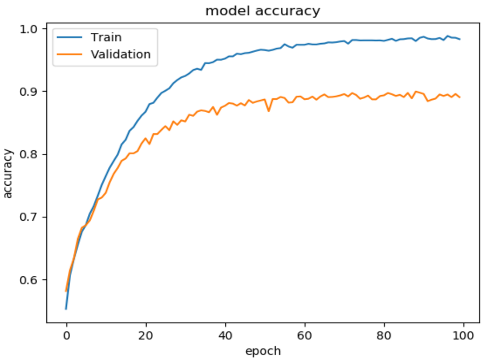

Fig. 4(a). Stated the training and validation accuracy

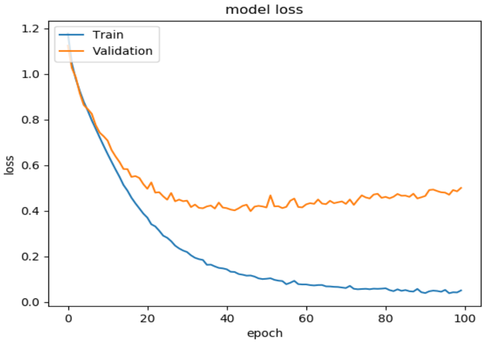

Fig. 4*(b)*. stated the loos error of training and validation

The overfitting in this model was solved by changing in the model structure by adding batch-normalization layer after the convolution layer. The normalized output divided into six action groups each group will be the input of a sub-LSTM network. These sub-nets inherit robustness from the regularizer (batch-normalization) and increase generality from LSTM architecture. The resulted model eliminate overfitting. The results of this step are:

	ACCURACY	LOSS
TRAINING	90.06	0.28
TEST	91.49	0.24
VALIDATION	91.38	0.25

B. *Model Generality*

The best performance of the proposed model when it applies in natural environment to test its generality. To do this step, labelled videos of AS should be publicly available to test the model generality. Unfortunately, there is no available data for this purpose. Therefore, we take a sample of the data, excluding them from the dataset before training phase of the model and considered them as the validation set. The unseen data (validation set) are used to challenge the model architecture in producing the correct affective-cognitive states. The trained model produced a vector of probabilities of prediction to each emotion class, prediction result is the maximum argument from the prediction vector.

Fig. 5 presented the confusion matrix of unseen data that was not used in the training and validation

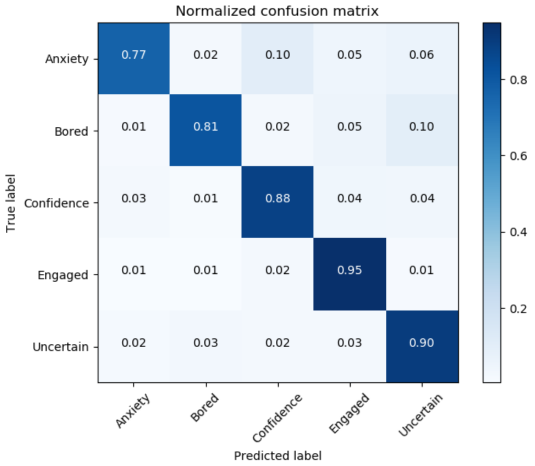

Fig. 5. Confusion matrix of unseen data of students with AS to test the model generality

process, to test the generality and efficiency of this model of classifying-affect per second.

Then, the model was tested on data for TD students, below the confusion matrix stated the results of this test (Fig. 6).

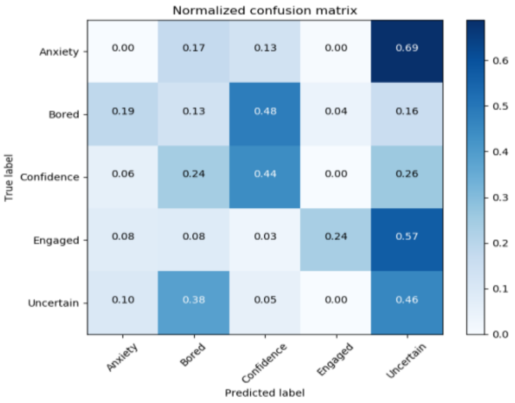

Fig. 6. Confusion matrix of unseen data of TD students

We can show how the model accuracy dropped to 26% because of the differences in AU, eye gaze, and head movement between AS and TD [64]; [65]; [66]; [67].

V. Conclusion

This study adopted deep learning techniques to capture and infer affective-cognitive states of AS students in real time. Within the study, a new affective model extended was produced without the need for wearable instruments. The prominent role of facial expressions in emotion recognition integrated with eye gaze and head motion, was a rich source of information to capture affect and cognition states especially in subjects with verbal impairment. It was difficult challenge to infer natural affective-cognitive states in an uncontrolled environment, especially with different variations and occlusions. In the learning environment, students experience a rich diversity of positive and negative of emotions. Therefore, this study focused on confidence, engagement, anxiety, uncertainty, and bored as dominant factors during the learning process. The model trained and tested on natural-spontaneous data collected for this purpose from students with AS. Because there is no available dataset for students with autism to model affective-cognitive states detection. The data collected in an uncontrolled environment with different variations in head movement, hand over face and head, facial and hairstyle, talking, glasses, face-semi-out plane, background, and illumination changes. The model structure was (CNN-LSTM-CNN), which achieved best results to detect complex natural affective-cognitive states comparing with other methods that used CNN or LSTM to extract basic emotions. CNN architecture produced a high quality of feature extraction, especially the occlusion features. The affective-cognitive states in this research were detected moment by moment by reading cues off videos instead of frame-by-frame. The model training, test, and validation is (90.06%), (91.49%), (91.38%) respectively. The proposed architecture model can compete with other models in the field of affective recognition. Also, the study explored the difference in model performance when applied on unseen data of TD, these data were collected parallel with AS under the same conditions. The model accuracy dropped to (26.2%) because of the divergence in AU, eye gaze, and head movement between AS and TD. For future work, the dataset should include all three types of ASD and they do not limit to AS type. This will help to explore more results about all atypical development of students and the differences in their facial expression, and eye gaze. Another direction for future work, using verbal and non-verbal communication channels for students with autism. This will lead to explore how verbal channel will increase in model performance. Also, we will investigate the impact of increasing number of features for both channel on network efficiency, time of training, and model generalization.

References

[1] V. S. Ramachandran and L. M. Oberman, "Broken mirrors: a theory of autism," *Scientific American*, vol. 295, no. 5, pp. 62--69, 2006.

[2] M. R. Sherer and L. Schreibman, "Individual behavioral profiles and predictors of treatment effectiveness for children with autism," *Journal of consulting and clinical psychology*, vol. 73, no. 3, p. 525, 2005.

[3] H. Cohen, M. Amerine-Dickens and T. Smith, "Early intensive behavioral treatment: Replication of the UCLA model in a community setting," *Journal of Developmental \& Behavioral Pediatrics*, vol. 27, no. 2, pp. S145--S155, 2006.

[4] S. J. Rogers, "Empirically supported comprehensive treatments for young children with autism," *Journal of clinical child psychology*, vol. 27, no. 2, pp. 168--179, 1998.

[5] C. Liu, K. Conn, N. Sarkar and W. Stone, "Physiology-based affect recognition for computer-assisted intervention of children with Autism Spectrum Disorder," *International journal of human-computer studies*, vol. 66, no. 9, pp. 662--677, 2008.

[6] D. Moore, P. McGrath and J. Thorpe, "Computer-aided learning for people with autism--a framework for research and development," *Innovations in Education and Training International*, vol. 37, no. 3, pp. 218--228, 2000.

[7] H. Kozima, C. Nakagawa and Y. Yasuda, "Interactive robots for communication-care: A case-study in autism therapy," *IEEE International Workshop on Robot and human interactive communication*, pp. 341--346, 2005.

[8] G. Pioggia, R. Igliozzi, M. Ferro and A. Ahluwalia, "An android for enhancing social skills and emotion recognition in people with autism," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 13, no. 4, pp. 507--515, 2005.

[9] P. A. Prelock, Autism spectrum disorders: Issues in assessment and intervention, 1 ed., USA: Pro-Ed Austin, TX, 2006.

[10] E. Pellicano, G. Rhodes and A. J. Calder, "Reduced gaze aftereffects are related to difficulties categorising gaze direction in children with autism," Neuropsychologia, vol. 51, no. 8, pp. 1504--1509, 2013.

[11] J. Groden, M. S. Goodwin, M. G. Baron and G. Groden, "Assessing cardiovascular responses to stressors in individuals with autism spectrum disorders," *Focus on Autism and Other Developmental Disabilities,* vol. 20, no. 4, pp. 244--252, 2005.

[12] D. B. Shalom, S. Mostofsky, R. Hazlett, M. Goldberg and R. Landa, "Normal physiological emotions but differences in expression of conscious feelings in children with high-functioning autism," *Journal of autism and developmental disorders,* vol. 36, no. 3, pp. 395--400, 2006.

[13] M. D. Samad, J. L. Bobzien, J. W. Harrington and K. M. Iftekharuddin, "Non-intrusive optical imaging of face to probe physiological traits in Autism Spectrum Disorder," *Optics \& Laser Technology*, vol. 77, pp. 221--228, 2016.

[14] A. De Vicente and H. Pain, "Motivation diagnosis in intelligent tutoring systems," *International Conference on Intelligent Tutoring Systems,* pp. 86-95, 1998.

[15] S. Craig, A. Graesser, J. Sullins and B. Gholson, "Affect and learning: an exploratory look into the role of affect in learning with AutoTutor," *Journal of educational media*, vol. 29, no. 3, pp. 241--250, 2004.

[16] A. Graesser, P. Chipman and B. King, "Emotions and learning with auto tutor," *Frontiers in Artificial Intelligence and Applications*, vol. 158, p. 569, 2007.

[17] T. W. Malone and M. R. Lepper, "Making learning fun: A taxonomy of intrinsic motivations for learning," *Aptitude, learning, and instruction*, vol. 3, pp. 223-253, 1987.

[18] F. A. Khan, E. R. Weippl and A. M. Tjoa, "Integrated approach for the detection of learning styles and affective states," *EdMedia: World Conference on Educational Media and Technology*, pp. 753--761, 2009.

[19] C. Conati, "Probabilistic Assessment of User's Emotions in Educational Games," *Applied Artificial Intelligence*, vol. 16, no. 7-8, pp. 555--575, 2002.

[20] N. a. J. W. L. Wang and R. E. Mayer, "The politeness effect: Pedagogical agents and learning outcomes," *International Journal of Human-Computer Studies*, vol. 66, no. 2, pp. 98-112, 2008.

[21] J. M. Olson, "Misattribution, preparatory information, and speech anxiety,*" Journal of personality and social psychology,* vol. 54, no. 5, p. 758, 1988.

[22] J. A. Harrigan and D. M. O'Connell, "How do you look when feeling anxious? Facial displays of anxiety,*" Personality and Individual Differences*, vol. 21, no. 2, pp. 205--212, 1996.

[23] J. Sanghvi, G. Castellano and I. Leite, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," *Proceedings of the 6th international conference on Human-robot interaction*, pp. 305--312, 2011.

[24] Z. Zeng, M. Pantic, G. I. Roisman and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39--58, 2009.

[25] W. Yin, K. Kann, M. Yu and H. Scutze, "Comparative study of cnn and rnn for natural language processing," *arXiv preprint arXiv*:1702.01923, 2017.

[26] B. C. Ko, "A Brief Review of Facial Emotion Recognition Based on Visual Information," *sensors*, vol. 18, no. 2, p. 401, 2018.

[27] S. Afzal and P. Robinson, "Modelling Affect in Learning Environments-Motivation and Methods," *IEEE 10th International Conference on Advanced Learning Technologies (ICALT),* pp. 438--442, 2010.

[28] C. Conati and X. Zhou, "Modeling students' emotions from cognitive appraisal in educational games,"

International Conference on Intelligent Tutoring Systems, pp. 944--954, 2002.

[29] C. Conati and H. Maclaren, "Data-driven refinement of a probabilistic model of user affect," *International Conference on User Modeling,* pp. 40--49, 2005.

[30] A. Ortony, G. L. Clore and A. Collins, The cognitive structure of emotions, Cambridge : Cambridge university press, 1990.

[31] W. Burleson, *Affective learning companions: strategies for empathetic agents with real-time multimodal affective sensing to foster meta-cognitive and meta-affective approaches to learning, motivation, and perseverance,* Doctoral dissertation,Massachusetts Institute of Technology, 2006.

[32] W. Burleson and R. W. Picard, "Affective agents: Sustaining motivation to learn through failure and a state of stuck," *Workshop on Social and Emotional Intelligence in Learning Environments,* 2004.

[33] H. Prendinger and M. Ishizuka, "The empathic companion: A character-based interface that addresses users'affective states," *Applied Artificial Intelligence,* vol. 19, no. 3-4, pp. 267--285, 2005.

[34] A. Kapoor, W. Burleson and R. W. Picard, "Automatic prediction of frustration," *International journal of human-computer studies,* vol. 65, no. 8, pp. 724--736, 2007.

[35] G. C. Littlewort, M. S. Bartlett, L. P. Salamanca and J. Reilly, "Automated measurement of children's facial expressions during problem solving tasks," *Automatic Face \& Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on,* pp. 30--35, 2011.

[36] J. M. Harley, F. Bouchet and R. Azevedo, "Measuring learners’ co-occurring emotional responses during their interaction with a pedagogical agent in MetaTutor," *International Conference on Intelligent Tutoring Systems,* pp. 40--45, 2012.

[37] J. M. Harley, F. Bouchet and R. Azevedo, "Aligning and comparing data on emotions experienced during learning with MetaTutor," *International Conference on Artificial Intelligence in Education,* pp. 61--70, 2013.

[38] S. D. Craig, S. D'Mello, A. Witherspoon and A. Graesser, "Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive--affective states during learning," *Cognition and Emotion,* vol. 22, no. 5, pp. 777--788, 2008.

[39] J. Grafsgaard, J. B. Wiggins, K. E. Boyer and E. N. Wiebe, "Automatically recognizing facial expression: Predicting engagement and frustration," *Educational Data Mining* , 2013.

[40] T. Langdell, *Face perception: An approach to the study of autism,* PhD Thesis, University of London, 1981.

[41] H. Macdonald, M. Rutter and P. Howlin, "Recognition and expression of emotional cues by autistic and normal adults," *Journal of Child Psychology and Psychiatry,* vol. 30, no. 6, pp. 865--877, 1989.

[42] K. A. Loveland, B. Tunali-Kotoski and D. A. Pearson, "Imitation and expression of facial affect in autism," *Development and Psychopathology,* vol. 6, no. 3, pp. 433--444, 1994.

[43] D. N. McIntosh, A. Reichmann-Decker and P. Winkielman, "When the social mirror breaks: deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism," *Developmental science,* vol. 9, no. 3, pp. 295--302, 2006.

[44] N. Yirmiya, C. Kasari, M. Sigman and P. Mundy, "Facial expressions of affect in autistic, mentally retarded and normal children," *Journal of Child Psychology and Psychiatry,* vol. 30, no. 5, pp. 725--735, 1989.

[45] D. N. McIntosh, A. Reichmann-Decker and P. Winkielman, "When the social mirror breaks: deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism," *Developmental science,* vol. 9, no. 3, pp. 295--302, 2006.

[46] J. H. Williams, A. Whiten, T. Suddendorf and D. I. Perrett, "Imitation, mirror neurons and autism," *Neuroscience \& Biobehavioral Reviews,* vol. 25, no. 4, pp. 287--295, 2001.

[47] T. Guha, Z. Yang, A. Ramakrishna and R. B. Grossman, "On quantifying facial expression-related atypicality of children with autism spectrum disorder," *International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 803--807, 2015.

[48] S. Baron-Cohen, "Mind Reading: The Interactive Guide to Emotions," London, 2004.

[49] S. Mavadati, *Spontaneous Facial Behavior Computing in Human Machine Interaction with Applications in Autism Treatment,* University of Denver, 2015.

[50] V. Bernard-Opitz, N. Sriram and S. Sapuan, "Enhancing vocal imitations in children with autism using the IBM speech viewer," *Autism,* vol. 3, no. 2, pp. 131--147, 1999.

[51] K. Gobbo and S. Shmulsky, "Faculty experience with college students with autism spectrum disorders: A qualitative study of challenges and solutions," *Focus on Autism and Other Developmental Disabilities,* vol. 29, no. 1, pp. 13--22, 2014.

[52] D. J. Ricks and M. B. Colton, "Trends and considerations in robot-assisted autism therapy," *International Conference on Robotics and Automation (ICRA),* pp. 4354--4359, 2010.

[53] R. W. Schlosser and O. Wendt, "Effects of augmentative and alternative communication intervention on speech production in children with autism: A systematic review," *American Journal of Speech-Language Pathology,* vol. 17, no. 3, pp. 212--230, 2008.

[54] W. Ahmed, S. Mitra, K. Chanda and D. Mazumdar, "Assisting the autistic with improved facial expression

recognition from mixed expressions," *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013 Fourth National Conference on,* pp. 1-4, 2013.

[55] D. B. Shalom, S. Mostofsky, R. Hazlett and M. Goldberg, "Normal physiological emotions but differences in expression of conscious feelings in children with high-functioning autism," *Journal of autism and developmental disorders,* vol. 36, no. 3, pp. 395--400, 2006.

[56] M. Toichi and Y. Kamio, "Paradoxical autonomic response to mental tasks in autism," *Journal of autism and developmental disorders,* vol. 33, no. 4, pp. 417--426, 2003.

[57] O. Rudovic, J. Lee, M. Dai, B. Schuller and R. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *arXiv preprint arXiv:1802.01186,* 2018.

[58] A. Rozga, T. Z. King, R. W. Vuduc and D. L. Robins, "Undifferentiated facial electromyography responses to dynamic, audio-visual emotion displays in individuals with autism spectrum disorders," *Developmental science,* vol. 16, no. 4, pp. 499--514, 2013.

[59] O. Golan, E. Ashwin, Y. Granader and S. McClintock, "Enhancing emotion recognition in children with autism spectrum conditions: An intervention using animated vehicles with real emotional faces," *Journal of autism and developmental disorders,* vol. 40, no. 3, pp. 269--279, 2010.

[60] M. Benedek and C. Kaernbach, "Decomposition of skin conductance data by means of nonnegative deconvolution," *Psychophysiology,* vol. 47, no. 4, pp. 647--658, 2010.

[61] M. Shoaib, I. Hussain and H. T. Mirza, "The role of information and innovative technology for rehabilitation of children with autism: a systematic literature review," *International Conference on Computational Science and Its Applications (ICCSA),* pp. 1--10, 2017.

[62] J. Hailpern, K. Karahalios, J. Halle and L. Dethorne, "A3: Hci coding guideline for research using video annotation to assess behavior of nonverbal subjects with computer-based intervention," *ACM Transactions on Accessible Computing (TACCESS),* vol. 2, no. 2, p. 8, 2009.

[63] R. Pascanu, C. Gulcehre, K. Cho and Y. Bengio, "How to construct deep recurrent neural networks," *arXiv preprint arXiv:1312.6026,* 2013.

[64] T. Guha, Z. Yang, A. Ramakrishna and R. B. Grossman, "On quantifying facial expression-related atypicality of children with autism spectrum disorder," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* p. 803, 2015.

[65] A. Metallinou, R. B. Grossman and S. Narayanan, "Quantifying atypicality in affective facial expressions of children with autism spectrum disorders," *IEEE International Conference on Multimedia and Expo (ICME),* pp. 1--6, 2013 .

[66] D. J. Faso, N. J. Sasson and A. E. Pinkham, "Evaluating posed and evoked facial expressions of emotion from adults with autism spectrum disorder," *Journal of Autism and Developmental Disorders,* vol. 45, no. 1, pp. 75--89, 2015.

[67] D. Mathersul, S. McDonald and J. A. Rushby, "Automatic facial responses to affective stimuli in high-functioning adults with autism spectrum disorder," *Physiology \& behavior,* vol. 109, pp. 14--22, 2013.

[68] P. Carcagni, M. Coco, M. Leo and C. Distante, "Facial expression recognition and histograms of oriented gradients: a comprehensive study," *SpringerPlus,* vol. 4, no. 1, p. 645, 2015.

[69] T. Baltruvsaitis, P. Robinson and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," *IEEE Winter Conference on Applications of Computer Vision (WACV),* pp. 1-10, 2016.

[70] R. A. Khan, *Detection of emotions from video in non-controlled environment,* University Claude Bernard-Lyon, 2013.