Facial emotions are accurately encoded in the neural signal of those with Autism Spectrum Disorder: A deep learning approach

Juan Manuel Mayor Torres [a1*], Tessa Clarkson [b1*], Kathryn M. Hauschild [c], Christian C. Luhmann [c,d], Matthew D. Lerner [c,e], and Giuseppe Riccardi [a]

[1]Joint first authors

[a]Department of Information Engineering and Computer Science, University of Trento, Povo Trento, Italy 38123

[b]Department of Psychology, Temple University, Philadelphia, PA 19121

[c]Department of Psychology, Stony Brook University, Stony Brook, NY 11794

[d]Institute for Advanced Computational Science, Stony Brook University, Stony Brook, NY 11794

[e]Department of Psychology, University of Virginia, Charlottesville, VA 22904

Corresponding author:

Tessa Clarkson

1701 N 13th St, Philadelphia, PA 19122 Weiss Hall
Temple University
Philadelphia, PA 19122
Email: tessa.clarkson@temple.edu

https://orcid.org/0000-0001-8798-4652

Keywords

Deep Convolutional Neural Networks, Facial Emotion Recognition, Autism Spectrum Disorder, Electroencephalography

Abstract

Background: Individuals with autism spectrum disorder (ASD) exhibit frequent behavioral deficits in facial emotion recognition (FER). It remains unknown whether these deficits arise because facial emotion information is not encoded in their neural signal, or because it is encoded, but fails to translate to FER behavior (deployment). This distinction has functional implications, including constraining when differences in social information processing occur in ASD, and guiding interventions (i.e., developing prosthetic FER-vs.-reinforcing existing skills). **Methods:** We utilized a discriminative and contemporary machine learning approach - Deep Convolutional Neural Networks (CNN) – to classify facial emotions viewed by individuals with and without ASD(N= 88) from concurrently-recorded electroencephalography signals. **Results:** The CNN classified facial emotions with high accuracy for both ASD and non-ASD groups, even though individuals with ASD performed more poorly on the concurrent FER task. In fact, CNN accuracy was greater in the ASD group, and was not related to behavioral performance. This pattern of results replicated across three independent participant samples. Moreover, feature-importance analyses suggest that a late temporal window of neural activity (1000–1500ms) may be uniquely important in facial emotion classification for individuals for ASD. **Conclusions:** Our results reveal for the first time that facial emotion information is encoded in the neural signal of individuals with (and without) ASD. Thus, observed difficulties in behavioral FER associated with ASD likely arise from difficulties in decoding or deployment of facial emotion information within the neural signal. Interventions should focus on capitalizing on this intact encoding rather than promoting compensation or FER prosthetics.

Introduction

Deficits in facial emotion recognition (FER) are a core feature of the social-emotional information processing characteristic of autism spectrum disorder (ASD) (1). Variation in the ability to accurately identify and label the emotional valence of facial stimuli is associated with ASD symptom severity as well as measures of adaptive functioning (2). Meta-analyses support the conclusion that individuals with ASD perform poorer than controls on tasks of facial emotion identification and recognition across all basic emotions (3–5). Additionally, eye tracking, electrophysiological, and neuroimaging findings implicate atypical attentional and cognitive processing of facial emotions by individuals with ASD (6–8). These findings suggest that emotion-related information is either absent or insufficiently present, on a trial-by-trial basis, in the neural processing of individuals with ASD. However, it is instead possible that deficits reflect a failure to use emotion-related information to accurately guide behavior. This study aims to examine if facial emotion encoding occurred on a trial-by-trial basis for a particular individual, to understand the nature of observed facial emotion recognition deficits seen in some individuals with ASD.

Recent developments in machine learning techniques may provide a novel approach for addressing this question. One particularly promising technique is Deep Learning, more specifically a Deep Convolutional Neural Network (CNN) classifier. The CNN is a set of multifunctional layers that can be applied to discriminate multidimensional data (like most multilayer generative approaches). CNN can construct intermediate representations that are of specific use for a given classification problem (9). CNN can construct intermediate abstractions of neural data (or feature subspaces) that are of specific use for a given classification problem (10). Typically, CNN have been successfully applied to neuroimaging, and electroencephalography (EEG) data to identify patient population-level neural representations of psychological constructs in the brain, or biomarkers for disease (11,12). However, this is not always the case, with accuracy of representation varying considerably (13,14), and little work has applied these efforts to within-individual (rather than between-group) analyses (15,16). Thus, an important first step in

determining if neural representations of facial emotions are preserved in ASD is to assess whether this facial emotion information is even present within the neural signal of individuals with ASD, and whether that information can be successfully used to identify the facial emotion they are viewing at each trial. We can test if facial emotion information exists within the neural signal for an individual by inputting the CNN with EEG data on an intra-individual basis to predict emotion labeled by that individual. This can be applied in both control individuals as well as clinical populations (17,18).

In the current study, we used CNN to reanalyze EEG data collected as participants, both ASD and control, viewed images of facial emotions. Specifically, the CNN were trained using an individual's own neural data recorded during a FER task in order to predict which facial emotion was presented on each trial. Our analyses allow us to first determine (a) whether the neural activity (as represented in relatively sparse-array EEG channels) of individuals with and without ASD can be used to classify emotions, which tests if facial emotion information is present and extractable within the neural signal. Second, they allow us to determine (b) whether the relation between CNN predictions and behavioral performance is comparable between individuals with ASD and controls. This tests for suboptimal (but present) encoding compared to controls (CNN: ASD<Controls) or suboptimal deployment (ASD: CNN>Behavior) of encoded facial emotion in ASD. Lastly, these tests permit us to determine (c) whether differences in CNN performance (CNN accuracy) are related to differences in behavioral performance on the FER task for individuals with ASD and controls. If the classification performance of the CNN is not strongly related to behavior, it would suggest partially distinct neural processes for encoding and deployment of facial emotion information. In order to examine the robustness of findings, we conducted these analyses on data from three distinct study samples.

Though CNN have proven themselves to be useful in tackling difficult classification problems (19), interpretation of their behavior can be challenging. CNN, like many recent neural network models, are considered "black box" models, providing little insight into how, where, when and

why they perform the way they do. For this reason, we employ saliency methods that allow us to interrogate our trained CNN, shedding light on what portions of the input-set are most relevant to the CNN's performance and indicating what aspects of the EEG data may be relevant for classification (18); thus, allowing us to clarify the neural mechanisms underlying facial emotion processing. For example, there may be particular temporal (e.g., early or late) or spatial (e.g., frontal or lateral) components of the EEG signal that are particularly important when attempting to classify the facial emotional expressions. Moreover, associated group differences may shed light as to what stages of FER processing may be distinct or impaired in ASD.

Methods and Materials

Participants. The sample was comprised of 40 verbally-able adolescents with ASD and 48 adolescents without ASD ranging in age from 14 to 17 years[1*]. This age range was used as most impairments (neural and behavioral) seen in facial emotion recognition in ASD are relatively fixed by adolescence and are related to functional outcomes (2,5,7,20,21). ASD diagnoses were confirmed using the Autism Diagnostic Observation Schedule- 2 (ADOS-2; (22)) administered by research reliable study personnel. ADOS-2 comparison scores (CS) were then calculated as an index of core ASD symptom severity. Additionally, all study participants were determined to have a full-scale intelligence quotient (FSIQ) above 70 as measured by the Kaufman Brief Intelligence Test – Second Edition (KBIT-2; (23); see Table 1 for a summary of participant demographics). Replication sample participant information can be found in the Supplemental Material.

Facial Emotion Recognition Task (FER) and EEG Acquisition. Participants completed a standardized measure of facial emotion recognition that has been well validated in both typically developing and ASD populations, the Diagnostic Analysis of Nonverbal Behavior (DANVA-2;

[1*] The full sample demonstrated group differences in age and gender, such that the ASD group was younger and contained more male participants. Therefore, a propensity matching procedure (52) was used to optimize the sample matching on these variables (*p*'s>.183). Result replicated full sample analyses; thus, the entire sample was included in the manuscript.

(24,25)), while undergoing simultaneous EEG data collection. The facial expression subtests of the DANVA-2 include 48 color photographs of male (N=24) and female (N=24) faces that display depictions of four basic emotions (12 happy, 12 sad, 12 angry, and 12 fearful). During the FER task, participants were asked to view each face and make a behavioral determination of the emotion displayed via button press. Each face was presented on a computer screen for a minimum of 2000ms and a maximum of 4000ms, depending upon individual participant response time. If participants did not make a behavioral response within the maximum 4000ms time window, the face disappeared from the screen, but the response options remained requiring the selection of a response to advance to the next trial. Groups did not differ on reaction times (p=.182; Table 1), and no individual mean reactions times fell within the time window of EEG data used by the CNN (Range of subject mean reaction times = [2035.70-6881.17ms]; M=3933.76ms; SD =914.97ms). However, groups differed on FER accuracy (F(1,86)=4.369, p=0.040), such that the control group had more accurate performance (Table 1.0).

Electrophysiological responses to each stimulus presentation were recorded using Brain Products 32-channel BrainVision actiCHamp EEG recording system arranged in the international standard 10/20 system. Eye movements and eye blinks were recorded using four facial electrodes: 2 placed on the canthi of each eye to measure horizontal movements, and 2 placed at supra- and infra-orbital sites to measure vertical movements. The EEG signal was pre-amplified at the electrode to improve the signal-to-noise ratio by the BrainAmp system. The data were digitized at a 16-bit resolution with a sampling rate of 500 Hz using a band-pass filter of 0.016-1000 Hz, and notch filtered at 60Hz with a half-power cutoff of 12db/Oct. Each active electrode was measured online with respect to a common mode sense active electrode producing a monopolar (non-differential) channel. Data collection procedures adhered to best practices for EEG data collection in ASD (26). This EEG signal adequacy process has been executed before the automatic artifact and bad channels rejection methods explained below.

EEG Processing and Analyses. Our EEG analysis pipeline methodologies are divided in 1) EEG segmentation, 2) artifact removal, and 3) whitening normalization, 4) The CNN classifier architecture, and the detailed description of our training method, and 5) the feature importance analyses are outlined in Figure 4.

1) EEG data were processed using EEGlab (27) Matlab toolbox. First, data was segmented to between -200 and 1500ms relative to the emotional face onset for single-trial analyses (Figure 4A). Then, each single-trial EEG response was filtered using a 150 coefficient Blackman-Harris-Window non-linear phase band-pass filter 0.1-30 Hz and re-referenced to T9 - T10.

2) Automatic channel rejection/removal (Figure 4B) was conducted using the Prep pipeline (28), Koethe's cleanraw function, and the Artifact Subspace Removal (ASR) method to remove noise and artifactual channels. Next, the ADJUST EEGlab plugin (29) was used to remove additional spatio-temporal artifacts such as, temporal kurtosis, spatial average difference, maximum epoch variance, or generic discontinuities spatial feature a horizontal or vertical eye blink artifact, and reshape the single-trial EEG data through Independent Component Analysis (ICA) decomposition (30).

3) Data was then normalized using a Zero Component Analysis (ZCA) whitening normalization (Figure 4C; Mahanalobis Zero Phase Whitening; (31,32)) to create a 2D representation, which maximizes the average cross-covariance between each dimension of the whitened image and the EEG data. Further details are provided in the Supplemental Information.

4) After the whitened normalization procedures, the whitened normalized image was fed into the Deep ConvNet classifier (Figure 4D; See Supplementary Information). The Deep ConvNet was composed of three convolutional-pool (conv-pool) blocks. The first conv-pool block had a convolutional layer with a rectangular kernel-size of 100x10 units and 32 filters and a pooling layer with a size of 5x2 units connected to a local response normalization layer (31,33). The second conv-pool block was composed of a convolutional layer with a kernel-size of 20x5 units and a pooling layer with a size of 2x2 units also connected to a second normalization layer. The third conv-pool block was composed of a conv-layer with a size 10x2 and a max-pool layer

with a size of 2x2 and 128 filters. Each conv-pool layer has a stride factor of 2 and a non-zero padding, thus dividing the output size in half after each conv-pool layer. All the normalization/regularization layers used amplitude normalization and not batch normalization per block. The third max-pooling layer was attached to a dense, fully-connected (FC) layer with 1024 units in order to compute the final 4 class probabilities for each emotion (happy, sad, angry, and fear) using a *Softmax* function. A description of the CNN model, the training process, and the layer operators can be found in the Supplementary Information. A leave-one-trial-out (LOTO) per participant cross-validation was used for all three samples in order for each individual's data to be used to train and then predict their own data. Specifically, 47 of each subject's 48 trials were used to train a CNN and the single, remaining trial was used to test the performance of the trained CNN. This procedure was repeated 47 more times, swapping which trials were used for training and which trial was used for testing. This entire, 48-step cross-validation procedure was then repeated for each subject.

5) In order to examine which portions of the EEG signal are important and distinct in ASD for FER (32,34) we used the iNNvestigate package (35), which allows the comparison of multiple methods for extracting feature-importance information in the resulting CNN. In order to balance numerically the relevance maps across feature-space, we modified the LRP DeepTaylor baseline to balance the level of positive, and negative relevance propagated through the CNN (see Supplementary Information). This new balanced LRP relevance map is denoted as LRP A/B Flat Preset depending on the value of α and β relevance propagation adjustment parameters (35,36). Therefore, in order to statistically explore which temporal portions of the EEG signal are most relevant to the CNN's performance, we divided the segments into overlapped time-windows that may relate to various stages of facial emotion recognition. Specifically, these windows were 0-500ms, 250-750ms, 500-1000ms, 750-1250ms, and 1000-1500ms. For more background and analyses details please refer to the Supplementary Material.

Data Analyses: In order to determine (a) whether the neural activity of individuals with and without ASD can be used to classify emotions, and (b) whether the relation between CNN

predictions and behavioral performance are comparable between individuals with ASD and controls, we used a one-way ANOVA comparing the accuracies of the CNN and behavioral performance on the FER task for the control and ASD groups. Accuracies were calculated by determining the percent of correctly classified trails for each of the four emotions, meaning p erforming at chance levels would result in a .25 accuracy value. Next, to examine (c) whether differences in predictive accuracy are related to differences in behavioral performance on the FER task in control and ASD, we examined the associations between CNN accuracy and behavioral performance on FER and ADOS-CS within each group using Pearson correlations, then used Fisher r-to-z transformations to compare association strengths between groups, and used Pearson correlations collapsing across groups to examine performance's associations dimensionally. Lastly, in order to explore what temporal portions of the input are most relevant to the CNN's performance and indicate what aspects of the EEG data permit a successful classification we conducted a one-way ANOVA examining group difference in the importance of each pre-specified time window of EEG data in the CNN classification accuracy.

Data Availability. Data from the primary study can be found at the National Database for Autism Research (NDAR). Code for the CNN can be found here: https://github.com/meiyor/Deep-Learning-Emotion-Decoding-using-EEG-data-from-Autism-individuals. Data for replication samples can be made available upon request to the corresponding authors or Dr. Matthew D. Lerner.

Results

CNN successfully predicts viewed facial emotions. The CNN was able to successfully predict the viewed facial emotions on each test trial of the FER task (Figure 1). Indeed, the CNN was more accurate ($M = 0.88$, $SD = 0.17$) in predicting the facial emotions than participants themselves ($M = 0.82$, $SD = 0.08$, $F(1, 86) = 21.45$, $p < .001$). More importantly, we observed an interaction ($F(1, 86) = 7.56$, $p = .007$; Figure 2) such that the CNN achieved greater accuracy in

the ASD group (M = 0.932, SD = 0.13) than the control group (M = 0.863, SD = 0.213). This pattern of effects was replicated in two additional independent samples (see Supplemental Information).

Considering ASD symptoms dimensionally (N = 88; Figure 3), there was no significant relation between ADOS-2 CS and CNN classification accuracy (r = .186, p = .083). This was observed, despite a significant relationship between ADOS-2 CS and behavioral performance (r = -.306, p = .004), such that those with more severe ASD symptoms exhibited poorer behavioral accuracy. Furthermore, the relationship between ADOS-2 CS and behavioral performance was significantly stronger than the relationship between ADOS-2 CS and CNN accuracy (z = -3.29, p = .001). This effect replicated in the combined sample, including two other independent samples (see Supplementary Information)

CNN accuracy is unrelated to behavioral performance. There was no relation between behavioral performance and CNN accuracy. This was true for the control group (r = 0.19, p = .20), for the ASD group (r = -0.23, p = .16), and for the combined sample (r = 0.02, p = .82). Furthermore, these relationships did not differ between groups (z = 1.92, p = .055). These effects replicated across two other independent samples (see Supplementary Information).

CNN predictive value is temporally distinct across groups. In order to evaluate the LRP-B present saliency map, a group by time window ANOVA was evaluated. Findings revealed that the CNN predictive value differed across temporal windows (F(4, 83) = 3.42, p =.012). There was an interaction such that groups differed in the CNN's predictive value across temporal windows of the EEG during the FER task (F(4, 83) = 11.61, p < .001; Figure 3). Indeed, the CNN was more predictive during the early window for the control group (0-500ms; Controls: M = -0.06, SD = 0.13, ASD: M = -0.12, SD = 0.12; F(1, 86) = 5.56, p = .021), whereas the late time window was more predictive for the ASD group (1000-1500ms; Controls: M = -0.18, SD = 0.12, ASD: M = -0.06, SD

= 0.15; $F(1, 86) = 18.48$, $p < .001$). This pattern of effects is replicated in another independent sample (see Supplementary Information).

Discussion

This study is the first to utilize a CNN to identify facial emotions viewed by individuals with and without ASD from sparse-array raw EEG signal. These results indicate that (a) facial emotion information is encoded in and extractable from the neural signals of individuals with and without ASD, that (b) encoding is not detectably suboptimal, but is distinct, in ASD, and related impairments likely occur during the deployment of encoded information, and that (c) neural processes for encoding and deployment of facial emotion processing are distinct in both ASD and controls. These results are novel in that they show that, even in individuals with ASD who have impaired facial emotion recognition (worse behavioral performance), insufficient encoding is not related to or responsible for these impairments.

The CNN was able to use neural activity to accurately predict viewed emotional facial expressions in both groups, with accuracies (>86%) considerably greater than chance (25%). These results suggest that information required for predicting observed facial emotions is present within the neural signal in individuals with and without ASD. The success of the CNN is particularly impressive given that the CNN had an extraordinarily challenging task in that it was required to predict the emotional facial expression based on a single trial's-worth of sparse-array, intercorrelated EEG data – yet was still able to achieve very high accuracy.

Classification accuracy of the CNN was better in the ASD group, despite worse behavioral performance (3–5). Furthermore, CNN accuracy was not associated with ADOS-2 CS. These results suggest that emotional facial information is encoded in the neural activity of individuals with ASD regardless of the severity of ASD symptomatology. However, these findings alone need not suggest that neural encoding processes are the same for individuals across the ASD

spectrum – that is, such encoding of emotion as instantiated in the CNN may be achieving the same classification accuracy but interpreted differently by the CNN itself. However, successful translation of encoded facial emotion information into behavior is impaired in ASD, as indicated by the frequently observed impairments in FER associated with ASD. This is further substantiated by our results showing that the CNN performance was unrelated to behavior in both groups, which suggests partially distinct neural mechanisms involved in the encoding and deployment of facial emotions.

In order to elucidate what aspects of the neural signal encode facial emotion information, we explored which portions of the neural signal were most relevant to the CNN's performance, and if this varied by group using reliable saliency methods such as LRP. In fact, the early temporal window (250-500ms) was most important to successful CNN classification for the control group, whereas the later temporal window (1000-1500ms) was most important for the ASD group, suggesting key differences in the unfolding of neural processing of facial emotions between groups. These differences might reflect delayed latency of neural signal encoding of FER in ASD (37,38), though this seems unlikely given equivalent reaction times in the different groups. Alternatively, these differences could reflect compensatory mechanisms for FER (1,7,39) or suggest a greater reliance on later stages of facial emotion processing to better capture crucial nuances essential for encoding in ASD (40). As such, examination of these temporal windows represents a viable starting point for investigating portions of the neural signal that may lead to downstream differences in behavioral task performance. Notably, they provide ASD and cognitive neuroscience researchers new avenues to explore in attempting to understand how neural signals may encode facial emotions and subsequently be deployed to accurately perform FER tasks.

In addition to contributing to theoretical models of FER in ASD, these findings also have important implications for the development and implementation of intervention programs. Specifically, they suggest that intervention development should exploit the intact facial emotion information

encoding in ASD and aim to promote translation of this encoding into behavior. Currently, some interventions promote the use of FER prostheses (41), while others encourage individuals with ASD to adjust their attention when viewing faces (42), or integrate facial information (43) to more closely approximate that of their typically developing peers. Results of the CNN analyses suggest that such approaches – which either aim to compensate for poor encoding of facial emotions or to teach those with ASD to recognize emotions as typically developing peers do - may fail to capitalize on the most parsimonious *existing* route to facial emotion identification in ASD. That is, it may be more beneficial to build interventions that reinforce intact (but ASD-specific) neural facial emotion encoding and more closely target potential impairment in the deployment of accurately encoded facial emotion information (44).

Though promising, some limitations that can inform future studies bear note. The sample included only verbally-able individuals with ASD and the third replication sample did not have a control group. Additionally, all individuals performed fairly well on the task, limiting the variance with which we could examine the neural signals that may predict behavioral mistakes. Nevertheless, this work serves as an important step in evaluating the utility of applying CNN to better understand how typically developing individuals and individuals with ASD encode facial emotion information at the neural level. Future work should expand the current findings by examining emotion-specific effects (45–47), effects of correct versus incorrect trials, the effects of passive versus active viewing tasks, and the effects of facial emotion viewing duration (48–50) using different input representations, or more robust saliency methods. Studies should work to disentangle what aspects of the neural signal are being used to predict emotional faces, such as low-level visual information (but see Supplemental Information). Moreover, studies should examine developmental effects of encoding, particularly in childhood when face processing is maturing. Additional manipulations such as masking or occluding stimulus images during temporal windows of interest (e.g., 25-500ms and 1000-1100ms) may be particularly useful in leveraging both careful experimental design and innovative computational methods to elucidate portions of the neural stream that are essential for accurate processing of facial emotion

identification (51). In addition to future studies examining encoding, further work should be done to disentangle whether behavioral mistakes are made as a result of improper decoding or deployment of the encoded neural signal. Moreover, examining mistakes as on a continuous spectrum, rather than diagnostically could yield interesting results. Lastly, future studies could compare learning rates of facial emotion classification after receiving explicit feedback, rather than implicit feedback in this study, in individuals with ASD compared to the CNN to see if they learn in a similar manner.

Overall, these findings suggest that individuals with ASD indeed reliably encode facial emotional information in the neural signal, and that CNN can be used to accurately classify observed emotions from this signal. Therefore, individual performance deficits in FER associated with ASD are likely not accounted for by the absence of detectable neural encoding of facial emotion information, but rather, likely stem from aberrations later in the processing stream such as usage or deployment of facial emotion information. Findings reported here suggest that future studies should focus on identifying where and when the breakdown in translation from neural encoding to behavioral response may lie, which will be critical to further inform intervention development (Mayor-Torres et. al 2020 under review).

Acknowledgments

The authors would like to thank Stony Brook Research Computing and Cyberinfrastructure, and the Institute for Advanced Computational Science at Stony Brook University for access to the SeaWulf computing system, which was made possible by a $1.4M National Science Foundation grant (#1531492). This research was supported by NIMH grant R01MH110585, grants from the Alan Alda Fund for Communication, the American Psychological Association, and Association for Psychological Science, as well as Fellowships from the American Psychological Foundation, Jefferson Scholars Foundation, and International Max Planck Research to Matthew D. Lerner.

Portions of this article were presented at the 2018 International Society for Autism Research Annual Meeting.

Disclosures

The authors report no biomedical financial interests or potential conflicts of interest.

References

1. Harms MB, Martin A, Wallace GL (2010): Facial emotion recognition in autism spectrum disorders: A review of behavioral and neuroimaging studies. *Neuropsychology Review*, vol. 20. pp 290–322.
2. Trevisan DA, Birmingham E (2016, December 1): Are emotion recognition abilities related to everyday social functioning in ASD? A meta-analysis. *Research in Autism Spectrum Disorders*, vol. 32. Elsevier, pp 24–42.
3. Ekman P, Friesen W V, Ellsworth P (2013): *Emotion in the Human Face: Guidelines for Research and an Integration of Findings*, vol. 11. Elsevier.
4. Lozier LM, Vanmeter JW, Marsh AA (2014): Impairments in facial affect recognition associated with autism spectrum disorders: A meta-analysis. *Dev Psychopathol* 26: 933–945.
5. Uljarevic M, Hamilton A (2013): Recognition of emotions in autism: A formal meta-analysis. *J Autism Dev Disord* 43: 1517–1526.
6. Aoki Y, Cortese S, Tansella M (2015): Neural bases of atypical emotional face processing in autism: A meta-analysis of fMRI studies. *World J Biol Psychiatry* 16: 291–300.
7. Black MH, Chen NTM, Iyer KK, Lipp O V., Bölte S, Falkmer M, *et al.* (2017): Mechanisms of facial emotion recognition in autism spectrum disorders: Insights from eye tracking and electroencephalography. *Neurosci Biobehav Rev* 80: 488–515.
8. Kang E, Keifer CM, Levy EJ, Foss-Feig JH, McPartland JC, Lerner MD (2018): Atypicality of the N170 Event-Related Potential in Autism Spectrum Disorder: A Meta-analysis. *Biol Psychiatry Cogn Neurosci Neuroimaging* 3: 657–666.
9. Faust O, Hagiwara Y, Hong TJ, Lih OS, Acharya UR (2018, July 1): Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine*, vol. 161. Elsevier Ireland Ltd, pp 1–13.
10. Lecun Y, Bengio Y, Hinton G (2015): Deep learning. *Nature*, vol. 521. pp 436–444.
11. Sarraf S, Tofighi G (2016): Classification of Alzheimer's Disease Structural MRI Data by Deep Learning Convolutional Neural Networks. Retrieved April 2, 2019, from http://arxiv.org/abs/1607.06583

12. Woo CW, Chang LJ, Lindquist MA, Wager TD (2017): Building better biomarkers: Brain models in translational neuroimaging. *Nature Neuroscience*, vol. 20. pp 365–377.

13. Grossi E, Olivieri C, Buscema M (2017): Diagnosis of autism through EEG processed by advanced computational algorithms: A pilot study. *Comput Methods Programs Biomed* 142: 73–79.

14. Eslami T, Mirjalili V, Fong A, Laird AR, Saeed F (2019): ASD-DiagNet: A Hybrid Learning Approach for Detection of Autism Spectrum Disorder Using fMRI Data. *Front Neuroinform* 13. https://doi.org/10.3389/fninf.2019.00070

15. Knoth IS, Lajnef T, Rigoulot S, Lacourse K, Vannasing P, Michaud JL, *et al.* (2018): Auditory repetition suppression alterations in relation to cognitive functioning in fragile X syndrome: A combined EEG and machine learning approach. *J Neurodev Disord* 10: 1–13.

16. Müller KR, Tangermann M, Dornhege G, Krauledat M, Curio G, Blankertz B (2008): Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring. *J Neurosci Methods* 167: 82–90.

17. Schirrmeister RT, Springenberg JT, Fiederer LDJ, Glasstetter M, Eggensperger K, Tangermann M, *et al.* (2017): Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum Brain Mapp* 38: 5391–5420.

18. Vahid A, Mückschel M, Stober S, Stock A-K, Beste C (2020): Applying deep learning to single-trial EEG data provides evidence for complementary theories on action control. *Commun Biol* 3: 112.

19. Shen L, Yeung S, Hoffman J, Mori G, Fei-Fei L (2018): Scaling human-object interaction recognition through zero-shot learning. *Proc - 2018 IEEE Winter Conf Appl Comput Vision, WACV 2018* 2018-Janua: 1568–1576.

20. Rodger H, Vizioli L, Ouyang X, Caldara R (2015): Mapping the development of facial expression recognition. *Dev Sci* 18: 926–939.

21. Bayet L, Nelson CA (2019): Handbook of Emotional Development. *Handbook of Emotional Development*. https://doi.org/10.1007/978-3-030-17332-6

22. Lord C, Rutter M, DiLavore PC, Risi S, Gotham K, Bishop S, others (2012): *Autism Diagnostic*

Observation Schedule: ADOS-2. Western Psychological Services Los Angeles, CA.

23. Kaufman AS (2004): Kaufman Brief Intelligence Test--Second Edition (KBIT-2). *Circ Pines, MN Am Guid Serv*.
24. Booth AJ, Rodgers JD, Volker MA, Lopata C, Thomeer ML (2019): Psychometric Characteristics of the DANVA-2 in High-Functioning Children with ASD. *J Autism Dev Disord* 49: 4147–4158.
25. Nowicki S, Duke MP (2008): Manual for the receptive tests of the diagnostic analysis of nonverbal accuracy 2 (DANVA2). *Atlanta, GA Dep Psychol Emory Univ*.
26. Webb SJ an., Bernier R, Henderson HA, Johnson MH, Jones EJH, Lerner MD, *et al.* (2015): Guidelines and best practices for electrophysiological data collection, analysis and reporting in autism. *J Autism Dev Disord* 45: 425–443.
27. Delorme A, Makeig S (2004): EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods* 134: 9–21.
28. Bigdely-Shamlo N, Mullen T, Kothe C, Su KM, Robbins KA (2015): The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Front Neuroinform* 9: 1–19.
29. Mognon A, Jovicich J, Bruzzone L, Buiatti M (2011): ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology* 48: 229–240.
30. Hyvärinen A, Oja E (2000): Independent component analysis: Algorithms and applications. *Neural Networks* 13: 411–430.
31. Mayor Torres JM, Clarkson T, Stepanov EA, Luhmann CC, Lerner MD, Riccardi G (2018): Enhanced Error Decoding from Error-Related Potentials using Convolutional Neural Networks. *Proc Annu Int Conf IEEE Eng Med Biol Soc EMBS* 2018-July: 360–363.
32. Kindermans P-J, Hooker S, Adebayo J, Alber M, Schütt KT, Dähne S, *et al.* (2019): The (Un)reliability of saliency methods. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, Cham, pp 267–280.
33. Kingma DP, Lei Ba J (2014): *Adam: A Method for Stochastic Optimization*. Retrieved April 2, 2019, from https://arxiv.org/pdf/1412.6980.pdf %22 entire document

34. Kindermans P-J, Schütt KT, Alber M, Müller K-R, Erhan D, Kim B, Dähne S (2017): *Learning How to Explain Neural Networks: PatternNet and PatternAttribution*. Retrieved from http://arxiv.org/abs/1705.05598

35. Alber M, Lapuschkin S, Seegerer P, Hägele M (n.d.): How to iNNvestigate neural network ' s predictions !

36. Montavon G, Samek W, Müller KR (2018): Methods for interpreting and understanding deep neural networks. *Digit Signal Process A Rev J* 73: 1–15.

37. Rump KM, Giovannelli JL, Minshew NJ, Strauss MS (2009): The development of emotion recognition in individuals with autism. *Child Dev* 80: 1434–1447.

38. Clark TF, Winkielman P, McIntosh DN (2008): Autism and the Extraction of Emotion From Briefly Presented Facial Expressions: Stumbling at the First Step of Empathy. *Emotion* 8: 803–809.

39. Sasson N, Tsuchiya N, Hurley R, Couture SM, Penn DL, Adolphs R, Piven J (2007): Orienting to social stimuli differentiates social cognitive impairment in autism and schizophrenia. *Neuropsychologia* 45: 2580–2588.

40. Griffiths KR, Lagopoulos J, Hermens DF, Lee RSC, Guastella AJ, Hickie IB, Balleine BW (2016): Impaired causal awareness and associated cortical–basal ganglia structural changes in youth psychiatric disorders. *NeuroImage Clin* 12: 285–292.

41. Voss C, Schwartz J, Daniels J, Kline A, Haber N, Washington P, *et al.* (2019): Effect of Wearable Digital Intervention for Improving Socialization in Children With Autism Spectrum Disorder. *JAMA Pediatr* 173: 446.

42. Perlman SB, Pelphrey KA (2011): Developing connections for affective regulation: Age-related changes in emotional brain connectivity. *J Exp Child Psychol* 108: 607–620.

43. Tanaka JW, Wolf JM, Klaiman C, Koenig K, Cockburn J, Herlihy L, *et al.* (2010): Using computerized games to teach face recognition skills to children with autism spectrum disorder: The Let's Face It! program. *J Child Psychol Psychiatry Allied Discip* 51: 944–952.

44. Pineda JA, Juavinett A, Datko M (2012): Self-regulation of brain oscillations as a treatment for aberrant brain connections in children with autism. *Med Hypotheses* 79: 790–798.

45. Humphreys K, Minshew N, Leonard GL, Behrmann M (2007): A fine-grained analysis of facial expression processing in high-functioning adults with autism. *Neuropsychologia* 45: 685–695.

46. Enticott PG, Kennedy HA, Johnston PJ, Rinehart NJ, Tonge BJ, Taffe JR, Fitzgerald PB (2014): Emotion recognition of static and dynamic faces in autism spectrum disorder. *Cogn Emot* 28: 1110–1118.

47. Philip RCM, Whalley HC, Stanfield AC, Sprengelmeyer R, Santos IM, Young AW, *et al.* (2010): Deficits in facial, body movement and vocal emotional processing in autism spectrum disorders. *Psychol Med* 40: 1919–1929.

48. Baron-Cohen S, Jolliffe T, Mortimore C, Robertson M (1997): Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger Syndrome. Child Psychology & Psychiatry & Allied Disciplines, 38 (7), 813-822.

49. Rutherford MD, Towns AM (2008): Scan path differences and similarities during emotion perception in those with and without autism spectrum disorders. *J Autism Dev Disord* 38: 1371–1381.

50. Castelli F, Frith C, Happé F, Frith U (2002): Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain* 125: 1839–1849.

51. Zeiler MD, Taylor GW, Fergus R (2011): Adaptive deconvolutional networks for mid and high level feature learning. *Proc IEEE Int Conf Comput Vis* 2018–2025.

52. Jasjeet S. Sekhon (2011): Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. *J Stat Softw* 42: 1–52.

Table 1. Descriptive statistics for the participants. Demographics, ADOS-CSS, and FSIQ for all the participants in the main participant sample included in this study.

	Control (N = 48)		**ASD (N = 40)**	
Age (*M*, *SD*)	16.73*	3.41	14.89*	2.35
Male (*N*, *%*)	29*	60.42	32*	80.00
ADOS-2 CS (*M*, *SD*)	3.33*	2.71	8.15*	2.05
FSIQ (*M*, *SD*)	107.82	14.03	100.78	16.54
FER Accuracy (*%*; *M*, *SD*)	0.82*	0.08	0.78*	0.09
FER Reaction Times (ms; *M*, *SD*)	4081.46	957.16	3822.60	903.53

Note. Age = chronological age measured in years; ADOS-2 CS = Autism Diagnostic Observation Schedule- Second Edition Comparison Score; FSIQ = Full Scale Intelligence Quotient; FER = Facial Emotion Recognition behavioral performance accuracy. *p<.05 for group differences.

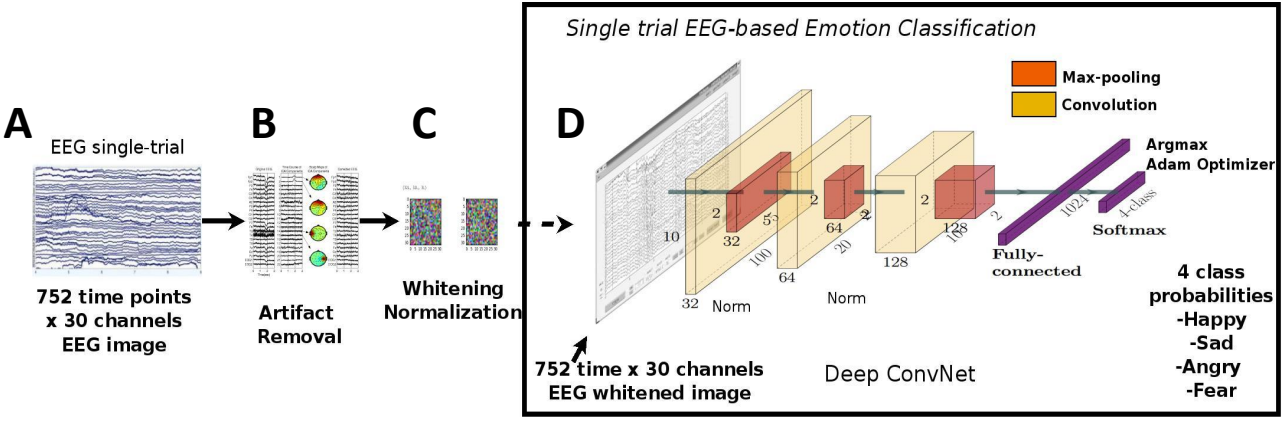
Figure Information:

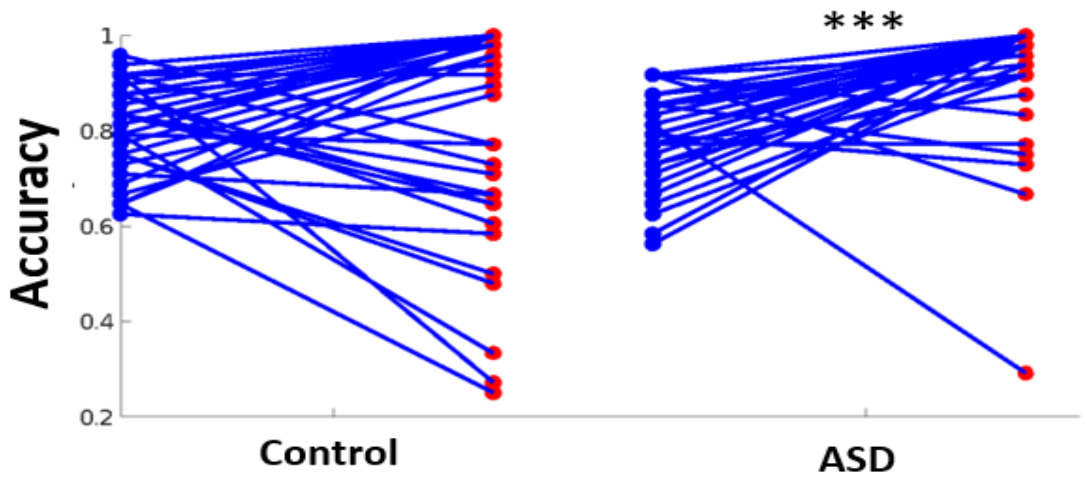
Figure 1. The pipeline used for emotion decoding in this study including (A) EEG data processing, (B) automatic channel and artifact removal using the Prep pipeline ASR EEGlab plugin and the ADJUST artifact removal plugin, (C) the ZCA whitening normalization/transformation process to increase the emotion class separability, and (D) the Convolutional Neural Networks (CNN) composed of 3 conv-pool blocks going from high to low in terms for conv-pool dimensionality, and low to high in terms of the number of filters. Two normalization /regularization layers were added, and a fully-connected (FC) layer with 1024 units with a final softmax layer was used to process the final class probabilities.

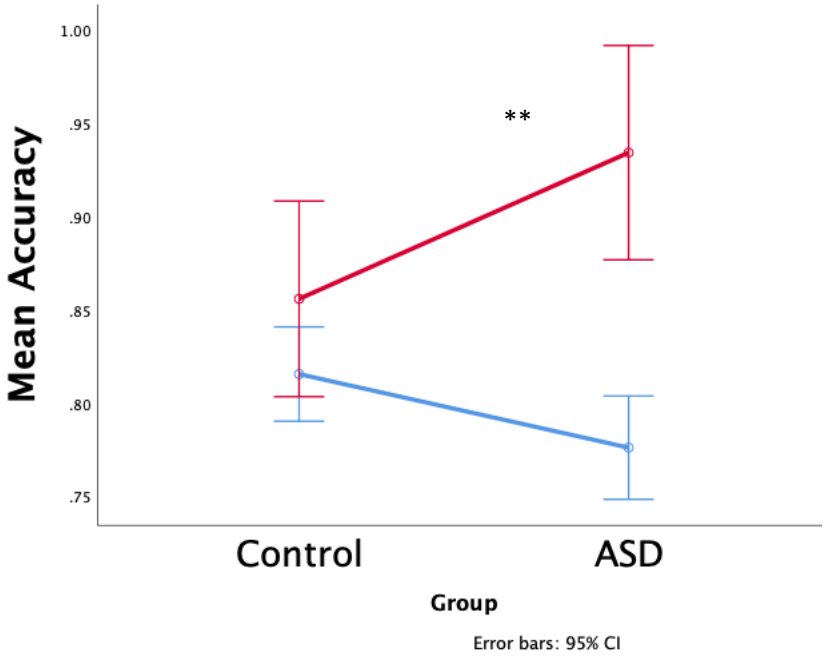
Figure 2. Spaghetti plots depicting the within-subject relationship between FER accuracies (in blue) and CNN accuracies (in red) for individuals in the control and ASD groups. Larger values on the y-axis indicate greater accuracies. ***p<.001.

Figure 3. Visual representation of the observed interaction between group (control, ASD) and accuracy type (behavioral, CNN) indicating greater CNN accuracy for the ASD group. Larger values on the y-axis indicate greater group mean accuracies. ** p<.01.

Figure 4. Relevance-maps created using the LRP B flat-rule preset saliency map (32). These maps depict the temporal windows of EEG signal most predictive of CNN classification accuracy, averaged across the four emotion classes: happy, sad, angry, and fearful, for the (A) control group, (B) ASD group, and (C) difference between control and ASD groups. The x-axis displays time in milliseconds from the onset of the stimulus and the y-axis shows the channels indexes on the scalp. Additionally, each relevance map shown here, we have included the 5 relevance topomaps associated with the time-windows utilized in this study: 0-500ms, 250-750ms, 500-1000ms, 750-1250ms, and 1000-1500ms. In (A) and (B), darker-red colors indicate higher relevance values while darker-blue colors indicate lower relevance values. In (C), red indicates a higher relevance value for the control group and blue indicates a higher relevance value for the ASD group. (A) and (B) relevance maps were normalized between [-1,1] plotted with the *jet* colormap, while (C) was normalized between [-0.1,0.1] and plotted with a redblue colormap.

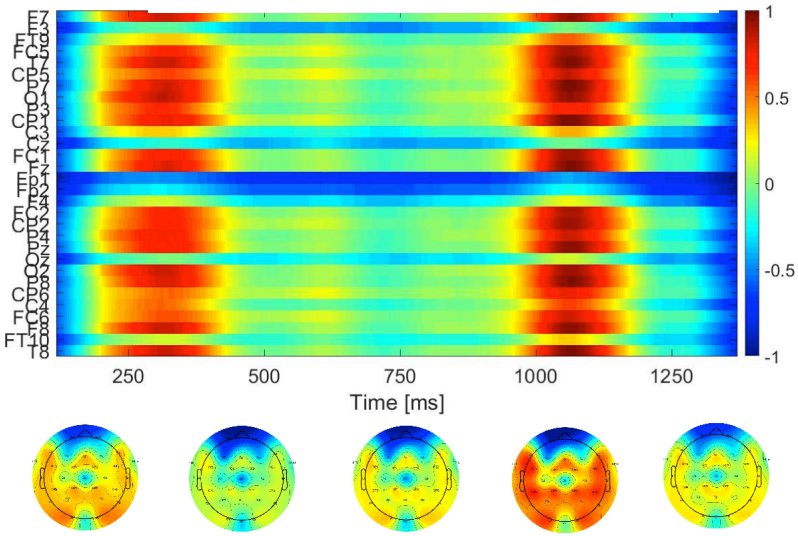

A
EEG single-trial
752 time points
x 30 channels
EEG image
B
Artifact
Removal
C
Whitening
Normalization
D
Single trial EEG-based Emotion Classification
Max-pooling
Convolution
Argmax
Adam Optimizer
Softmax
Fully-
connected
4-class
1024
10
32
2
55
100
64
20
128
Norm
Norm
752 time x 30 channels
EEG whitened image
Deep ConvNet
4 class
probabilities
-Happy
-Sad
-Angry
-Fear


***
Accuracy
1
0.8
0.6
0.4
0.2
Control
ASD


1.00
.95
.90
.85
.80
.75
Mean Accuracy
**
Control
ASD
Group
Error bars: 95% CI

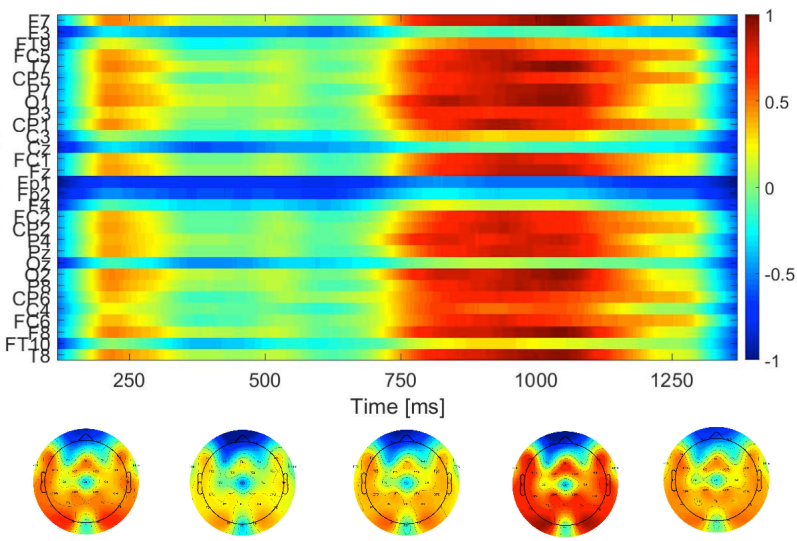
A

Control Average



B

ASD Average



C

Difference: Control-ASD