EyeXplain Autism: Interactive System for Eye Tracking Data Analysis and Deep Neural Network Interpretation for Autism Spectrum Disorder Diagnosis

Ryan Anthony J. de Belen
r.debelen@unsw.edu.au
University of New South Wales
Sydney, New South Wales, Australia

Tomasz Bednarz
t.bednarz@unsw.edu.au
University of New South Wales
Sydney, New South Wales, Australia

Arcot Sowmya
a.sowmya@unsw.edu.au
University of New South Wales
Sydney, New South Wales, Australia

ABSTRACT

Over the past decade, Deep Neural Networks (DNN) applied to eye tracking data have seen tremendous progress in their ability to perform Autism Spectrum Disorder (ASD) diagnosis. Despite their promising accuracy, DNNs are often seen as 'black boxes' by physicians unfamiliar with the technology. In this paper, we present EyeXplain Autism, an interactive system that enables physicians to analyse eye tracking data, perform automated diagnosis and interpret DNN predictions. Here we discuss the design, development and sample scenario to illustrate the potential of our system to aid in ASD diagnosis. Unlike existing eye tracking software, our system combines traditional eye tracking visualisation and analysis tools with a data-driven knowledge to enhance medical decision-making for physicians.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Explainable AI, Eye Tracking, Deep Neural Network, Visualization, Autism Diagnosis

ACM Reference Format:
Ryan Anthony J. de Belen, Tomasz Bednarz, and Arcot Sowmya. 2021. EyeXplain Autism: Interactive System for Eye Tracking Data Analysis and Deep Neural Network Interpretation for Autism Spectrum Disorder Diagnosis. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3411763.3451784

1 INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex life-long neurodevelopmental disorder often characterised by deficits in social, emotional and cognitive skills. While 'gold standard' assessment tools are available [22, 23], they are often costly, time-consuming and can be subjective. There is a need to develop new methods that augment existing clinical tools in order to reduce waiting periods for access to care. This is critical because early intervention can provide long-term improvements and greater effect on outcomes for the child [10].

There is a recent uptake on the use of data-driven approaches to automate medical diagnosis [37]. In the context of ASD diagnosis, computer vision and Deep Neural Networks (DNN) have been shown to be useful for the quantification of behavioural/biological markers that can further lead to a non-invasive, objective and automatic tool for ASD research [6]. A considerable amount of research has been applied to eye tracking data [1, 5, 7, 8, 11–13, 16, 17, 19–21, 25, 27, 31, 34–36].

Despite the increasing accuracy of DNN models, there is a need to open the 'black box' by providing physicians a better understanding on how models generate their final decisions [18]. This will help build trust and confidence in model's predictions, potentially augmenting existing clinical approaches for a faster and more efficient ASD screening.

In this paper, we present EyeXplain Autism, a interactive system for eye tracking data analysis and interpretation of DNN predictions for autism diagnosis. Unlike existing eye tracking software, our system augments statistical graphs and typical visualisations with a data-driven knowledge to enhance medical decision-making for physicians.

2 SYSTEM DESIGN AND IMPLEMENTATION

This section is split into three parts: (i) DNN for feature extraction and Support Vector Machine (SVM) for ASD diagnosis, (ii) EyeXplain Autism Interface and (iii) a sample scenario to illustrate the potential use of our system in a clinical setting. The system uses a Python back-end and a web-based interface built using HTML, CSS, Javascript and Plotly.

2.1 Deep Neural Network for Autism Diagnosis

Our system implements a DNN for feature extraction (Figure 1 A) and an SVM for autism diagnosis (Figure 1 B). This section introduces the dataset used, eye tracking procedure, model architecture, training protocols and results.

2.1.1 Dataset. The eye tracking study included 17 ASD children and 17 typically developing (TD) children. The ASD participants were recruited as part of the Australian Autism Biobank follow-up project by the Cooperative Research Centre for Living with Autism [9]. They were matched with TD children by age (4.6 ± 0.5 years old) at the time of the study. All participants in the ASD

CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8095-9/21/05…$15.00
https://doi.org/10.1145/3411763.3451784

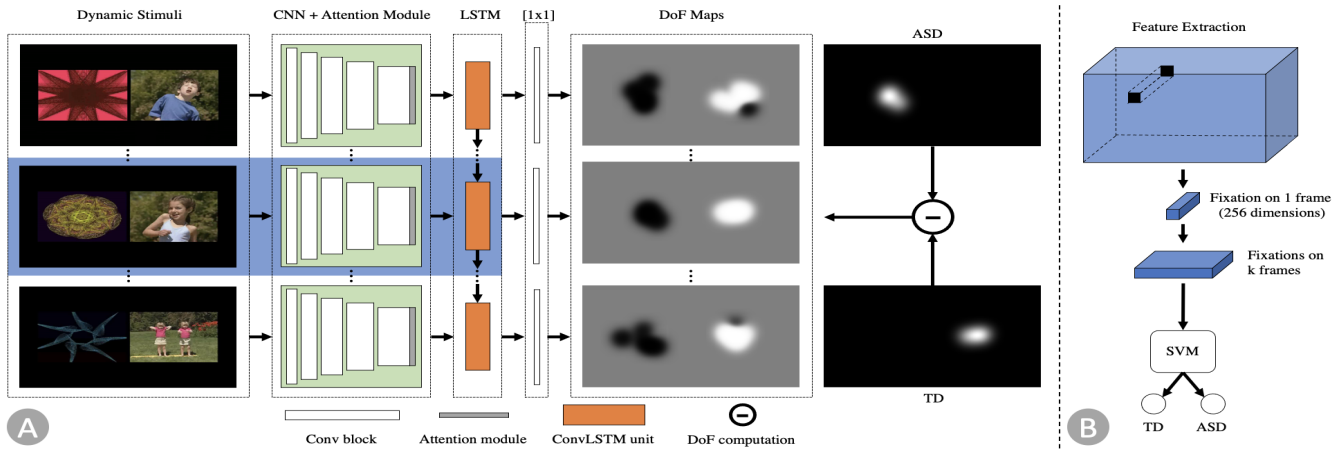

Figure 1: Overview of the proposed feature learning and autism diagnosis approach. A Given a video input, per-frame features are learned using an end-to-end approach to predict DoF maps; B Extracted features at fixated pixels from each fixation stage are cascaded and passed on to an SVM to identify individuals with ASD and TD.

Figure 2: Screenshots showing three frames of the dynamic stimuli (GeoPref test) used: It contains moving geometric images (left side) and moving social images (ride side)

group met criteria for ASD based on the DSM V [15] and diagnosis was confirmed using the Autism Diagnostic Observation Schedule (ADOS), Second Edition [22]. There is no specific exclusion criteria for the ASD group in this study. For the TD group, exclusion criteria included known neurodevelopmental disorders, significant developmental delays and known visual/hearing impairments.

2.1.2 Eye Tracking Procedure. Participants were tested using the Tobii X2-60 eye tracker. Eye movements were recorded at 60 Hz (with an accuracy of 0.5°) during the dynamic stimuli viewing. Each participant was seated approximately 60 cm in front of a 22" monitor with a video resolution of 1680 x 1050 pixels in a quiet room. A built-in five-point calibration in Tobii Studio was completed before administering the task for accurate tracking. Tobii Studio's I-VT filter [28] was used to process the raw eye-tracking data, exclude random noise and define variables (e.g. fixations and saccades) for further analysis. More specifically, short fixations (<100ms) were discarded and adjacent fixations (75ms, 0.5°) were merged.

2.1.3 Dynamic Stimuli. The GeoPref Test dynamic stimulus (Figure 2), which has been shown to be an effective stimulus for detecting ASD subgroups [26, 29, 30] was used. This stimulus consists of moving geometric images (MGIs) on the left side and moving social images (MSIs) on the right side. The MGIs were constructed from a collection of animated screen saver recordings. The MSIs were constructed from a series of short videos of children performing yoga exercises. It included video recordings of children performing a wide range of movements (e.g. waving arms and dancing). The stimulus contained a total of 28 different scenes.

2.1.4 Model Architecture. As shown in Figure 1 A, ACLNet [33], one of the best models available for dynamic saliency detection, is used for feature extraction. It consists of a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) network trained with an attention mechanism to enable rapid, end-to-end saliency prediction. Since ACLNet already contains an attention network trained on TD individuals, we trained our model with Difference of Fixation (DoF) maps that highlight more fixations of the TD group. This resulted in better training performance compared to a model trained on DoF maps that highlight more fixations of the ASD group. In particular, let I^+ and I^- be the fixation maps for the ASD and TD groups, respectively. The DoF map of an image is similar to [17]:

$$D = \frac{1}{1 + e^{-I/\sigma_I}} \tag{1}$$

where I = I^- - I^+ is a pixel-wise subtraction of ASD and TD fixation maps and σ_I represents the standard deviation of I. The resulting per-frame DoF maps encode the difference in visual attention of ASD and TD individuals. The white regions of the DoF map illustrate the visual attention of TD individuals while the black regions are for ASD individuals.

2.1.5 Training Protocol. Our model was optimised using the following loss function [14] that considers three different saliency evaluation metrics. We denote the predicted per-frame DoF map as $Y \in [0,1]^{28\times28}$ and the ground truth per-frame DoF map as $Q \in [0,1]^{28\times28}$. In particular, our loss function combines Kullblack-Leibler (KL) divergence, the Linear Correlation Coefficient (CC) and the Normalised Scanpath Saliency (NSS):

$$L = L_{KL} + 0.1L_{CC} + 0.1L_{NSS} \tag{2}$$

L_{KL} is widely used for training saliency models computed by:

$$L_{KL}(Y,Q) = \sum_x Q(x) log \frac{Q(x)}{Y(x)} \quad (3)$$

L_{CC} measures the linear relationship between Y and Q:

$$L_{CC}(Y,Q) = -\frac{cov(Y,Q)}{\sigma(Y)\sigma(Q)} \quad (4)$$

where $cov(Y,Q)$ is the covariance of Y and Q while σ is the standard deviation. L_{NSS} is defined as:

$$L_{NSS}(Y,Q) = -\frac{1}{N}\sum_x \bar{Y}(x) \times Q(x) \quad (5)$$

where $(\bar{Y}) = \frac{Y-\mu(Y)}{\sigma(Y)}$ and $N = \sum_x Q(x)$. It computes the mean of scores from the normalised per-frame DoF maps ($\bar{Y}$) at the predicted per-frame DoF maps Y.

We iteratively train our model with sequential DoF maps and video frames. We apply a loss defined over the predicted dynamic DoF maps from convLSTM. Let $\{Y_t^d\}_{t=1}^T$ and $\{Q_t^d\}_{t=1}^T$ denote the predicted dynamic DoF maps and continuous DoF maps. We minimise the following loss:

$$L^d = \Sigma_{t=1}^T L(Y_t^d, Q_t^d) \quad (6)$$

The parameters of ACLNet are initialized to the pre-trained parameters [33], then fine-tuned on the current dataset.

2.1.6 ASD Classification. Once the model has been trained to predict per-frame DoF maps of ASD and TD individuals from the given dynamic stimulus, feature extraction and classification are performed (Figure 1 B). Based on the eye tracking data, we determine the fixation positions and the corresponding frames in which they are recorded. Each saccade-fixation pair was considered as a fixation stage. For each fixation stage, features are extracted from the corresponding fixation position on the feature map obtained from the convLSTM output (note: convLSTM output is upsampled 4 times before extracting the feature map). More specifically, given a frame where a fixation has been identified, the feature map at the corresponding fixation is extracted, which results in a 256-dimensional feature vector. For a corresponding number of fixation stages, feature vectors for all fixations are concatenated in their temporal order starting from the first fixation to the last fixation stage. This serves as the feature space in which classification is performed. If there were fewer number of identified fixations, zeros are appended to the end. A linear decision boundary between ASD and TD individuals was determined by training an SVM on the extracted features.

2.1.7 Implementation Details. We implemented our model in Tensorflow with Keras and Scikit-learn libraries. For the training phase, we fine-tuned the network with Adam optimizer and a batch size of one image for a total of 20 epochs. The learning rate was set to 0.0001. There is no dropout and data augmentation performed. *L2* regularisation with the penalty parameter $C = 1$ was used for SVM classification.

2.1.8 Evaluation Metrics and Results. We report the performance of our model in terms of accuracy, sensitivity (i.e. true positive rate) and specificity (i.e. true negative rate). In the context of ASD diagnosis, these metrics have been widely used. To eliminate bias in the estimate of the probability of error [32], we used a leave-one-subject-out cross-validation. For our experiments, we fixed the weights of the fine-tuned DNN and perform SVM on top of the extracted features. In addition, we explored the optimal number of fixation stages, starting from 1 fixation to 30 fixations. Our model's performance on different fixation stages are as follows: accuracy: 68%-100%, sensitivity: 57%-100% and specificity: 65%-100%. For deployment, we set the number of fixation stages to 13, the value in which the highest accuracy has been reported. The SVM coefficients learned during the last fold are stored in a server for later processing.

2.1.9 Computational Load. The entire training procedure for feature learning takes about 1 hour with two NVIDIA RTX 2080 Super and a 3.5GHz Intel CPU (i7-7800X). Once the DNN model has been trained, feature extraction and SVM classification can be performed in less than 1 minute. The trained DNN model and SVM coefficients are then stored in a server and later accessed by the EyeXplain Autism Interface using a Python webframework, CherryPy.

2.2 EyeXplain Autism Interface

The EyeXplain Autism Interface consists of four components as shown in Figure 3: A control panel at the top A, a visualisation tab B with an interactive timeline at the bottom C and a statistics tab E.

2.2.1 Control Panel. The control panel A allows users to upload a new eye tracking data and dynamic stimulus used during the experiment. It also allows users to switch between the visualisation and statistics tabs for task-specific analysis. By clicking the 'Sort Fixations' button, users can sort fixations by feature importance from the SVM output. As discussed in the previous section, each recorded fixation is represented by 256 values. Since we set the value of fixation stages to be 13, there are 13*256 values as input to the SVM and the same number of coefficients are learned by the SVM. To sort the fixations by feature importance, we compute for the sum of the 256 coefficients from each frame, which results in 13 values. When these values are sorted in descending order, the frames with highest influence to diagnosis can be extracted. Finally, users can perform automated diagnosis shown in a popup box by clicking the 'Predict' button.

2.2.2 Visualisation Tab. The visualisation tab B enables users to play the dynamic stimulus, to scrub the timeline to jump to a different frame, to view recorded fixations **B2** and to turn on/off heatmap visualisation of ASD and/or TD samples that were seen by the DNN during training. The heatmaps are encoded using different colours, where ASD heatmap **B1** is red while TD heatmap **B3** is blue.

2.2.3 Interactive Timeline. The interactive timeline C shows the chronological sequence of fixations recorded in the uploaded eye tracking data. Each fixation is represented as a square thumbnail (100x100 pixels), centred at the fixation point (x,y) recorded in a given frame. When the mouse is hovered over a thumbnail, the interface shows a comparison popup **C1**, which displays randomly chosen ASD and TD fixations that the DNN has seen on that same

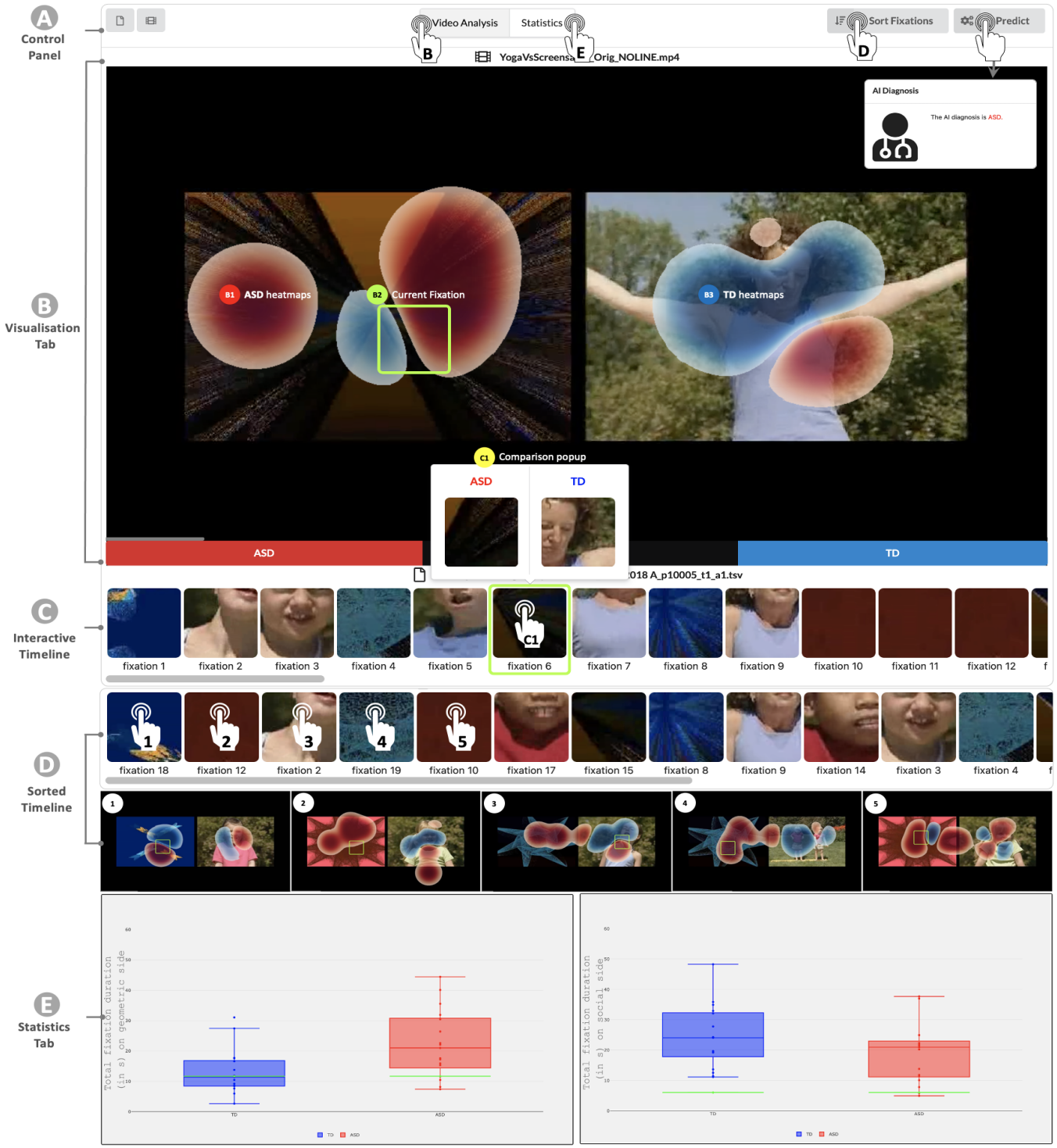

Figure 3: EyeXplain Autism Interface. The control panel (A) allows users to upload eye tracking data and dynamic stimulus, switch between visualisation and statistics tabs, sort fixations and perform diagnosis. The visualisation tab (B) shows a frame with recorded fixation (B2) and ASD (B1) and TD (B3) heatmaps. Each recorded fixation is represented as a thumbnail in the interactive timeline (C). Hovering over any fixation displays a comparison popup (C1). The 'Sort Fixations' button sorts feature importance (C) → (D). The 'Predict' button shows a popup with the predicted diagnosis. The statistics tab shows different variables for quantitative analysis (E).

frame during training. When a thumbnail is clicked, the visualisation tab will show the frame where the corresponding fixation was recorded.

2.2.4 Statistics Tab. The statistics tab E allows users to quantitatively analyse eye tracking data using box plots. For our dynamic stimulus, we defined two areas of interest (AOI): (i) geometric (left) side of the stimulus and (ii) social (right) side of the stimulus. Using our dataset, we pre-computed variables, such as fixation count, fixation duration and percentage of duration, around these two AOIs and saved them for later comparison. When a new eye tracking data is loaded, these variables are shown to help users compare them with variables computed from the new eye tracking data.

2.2.5 User Interaction Performance Test. In order to ensure interactivity and reduce latency, all computationally intensive processing (e.g. feature extraction, SVM classification, feature importance computation) are handled by the backend server. In exploratory user tests, the DNN model latency is at ~0.1 sec/prediction. On the other hand, the computation of SVM feature importance is performed within ~0.1 seconds.

2.3 Sample Scenario

We present a sample scenario showing how our system can help physicians better analyse eye tracking data and understand DNN predictions for ASD diagnosis (illustrated in the supplementary file). Our physician Hubble is studying a new eye tracking data collected from his patient. He decides to use the EyeXplain Autism interface to screen his patient for autism. Using the control panel A, he uploads a new eye tracking data and selects the dynamic stimulus used during the experiment. He sees the visualisation tab B where he plays the dynamic stimulus and observes fixation locations recorded during the experiment.

2.3.1 How different is the current fixation to the aggregate ASD and TD samples? Hubble starts by selecting a fixation of interest (fixation6 in the interactive timeline C) and activating the ASD and TD heatmap visualisation in the visualisation tab B. He finds that the majority of ASD samples B1 fixated on the geometric side of the screen, while the majority of the TD samples B3 fixated on the woman shown on the screen.

2.3.2 How different is the current fixation to an ASD and a TD sample? Hubble is interested to determine the similarity between the current fixation and individual ASD and TD fixations seen during training. In this case, he hovers to a fixation of interest (fixation6 in the interactive timeline C). The comparison popup C1 shows that the current fixation is similar to a fixation previously recorded from an ASD sample. Hubble observes that the current fixation is different to a representative TD sample where a fixation was recorded on the face of the woman.

2.3.3 What is the DNN's diagnosis? Why is the prediction like that? Hubble clicks the 'Predict' button in the control panel A and sees a popup appear showing that the DNN has determined that the current sample has ASD. Hubble now wants to know why the DNN prediction was ASD. He clicks the 'Sort Fixations' button in the control panel A to determine which frames are most important for diagnosis. The interactive timeline has now been updated and sorted by feature importance from the SVM output C → D. Upon interacting with the sorted timeline D and inspecting the top five fixations (labeled 1-5 in D) and their corresponding visualisation tab, Hubble notices that the top five recorded fixations fall within/near the ASD heatmaps, which helps explain why the DNN output was ASD.

2.3.4 How does the DNN output compare with quantitative analysis? After knowing and interpreting the DNN prediction, Hubble wants to know if quantitative analysis will reveal similar prediction. He switches to the statistics tab E to analyse the eye tracking data quantitatively. In particular, he is interested in two variables: (i) total fixation duration on the geometric side ($duration_{geo}$) and (ii) total fixation duration on the social side ($duration_{soc}$). The statistics tab displays these two variables with corresponding distribution of ASD (in red) and TD (in blue) samples seen during training. In addition, the current sample is highlighted as a green line. Upon analysis, he observes that the $duration_{geo}$ of the current sample is within the distribution of both ASD and TD samples. On the other hand, the $duration_{soc}$ of the current sample is lower than TD samples and within the distribution of ASD samples. He sequentially analyses the other variables (e.g. ratio of ($duration_{geo}$) to total duration) and finds out that the current sample has eye tracking variable values that are within the distribution of ASD samples seen during training.

2.3.5 Outcome. By using our interactive system to interpret DNN's output and quantitatively analyse eye tracking data, Hubble is convinced that the current sample has ASD. Standard clinical tests have revealed that the current sample indeed has ASD. Satisfied with the result, Hubble is now interested to use our system in his clinic for ASD screening.

3 LIMITATIONS AND ONGOING WORK

3.1 Small Dataset

Despite promising results, our DNN is currently trained on a relatively small dataset of 17 ASD and 17 TD children. We plan to collect more data to ensure reliability and robustness of autism diagnosis. Since multiple factors (e.g. age, content and dynamic nature of the stimuli) affect visual attention in children [24], we plan to build a collection of dynamic stimuli and determine the most discriminative videos that highlight differences in visual attention of ASD and TD participants. This would allow for a more efficient screening system.

3.2 DNN Architecture

Over the past decade, DNNs that predict visual attention have seen tremendous leap in performance [4]. While there exists a number of dynamic saliency models, we chose ACLNet[33] and did not explore other models for our system. The main rationale is that ACLNet has consistently held first place in all performance measures in several benchmark datasets. Furthermore, ACLNet has an attention module that enables rapid, end-to-end saliency prediction. We plan

to build a new saliency model, compare it with existing models and determine the suitable model for feature extraction.

3.3 Autism Severity and Treatment Response Prediction

We are working on extending the capability of our system to predict autism severity and treatment response. This would help determine which patients would have positive response to intervention, thereby allowing for more effective, efficient and even targeted treatments.

3.4 EyeXplain Autism Interface Design and Evaluation

We consulted the eye-tracking literature (see [2, 3] for review) and followed design recommendations for DNN-assisted medical systems [37] to establish the initial design and supported features of our interface. We will conduct co-design workshops with physicians to iterate the interface design, add more features and evaluate its effectiveness. Our goal is to determine if our system improves model's interpretability, user trust, confidence, and usability in a clinical setting.

4 CONCLUSION

In this paper, we presented EyeXplain Autism, an interactive system we are developing that visualises eye tracking data, enables an interpretable DNN for ASD diagnosis, and supports quantitative analysis commonly used in eye tracking studies. By presenting a sample scenario, we showed the potential of our system to help physicians analyse new eye tracking data, compare it to previously seen eye tracking data, perform automated ASD diagnosis, better interpret predictions of complex DNN models and build trust towards their predictions.

ACKNOWLEDGMENTS

We thank Valsamma Eapen, Anne Masi, Feroza Khan, Nisha Mathew and the Australian Biobank team for providing access to the eye tracking data.

REFERENCES

[1] Karan Ahuja, Abhishek Bose, Mohit Jain, Kuntal Dey, Anil Joshi, Krishnaveni Achary, Blessin Varkey, Chris Harrison, and Mayank Goel. 2020. Gaze-based Screening of Autistic Traits for Adolescents and Young Adults using Prosaic Videos. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*. 324–324.

[2] Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomas Ertl. 2014. State-of-the-Art of Visualization for Eye Tracking Data.. In *EuroVis (STARs)*.

[3] Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomas Ertl. 2017. Visualization of eye tracking data: A taxonomy and survey. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 260–284.

[4] Ali Borji. 2019. Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[5] Shi Chen and Qi Zhao. 2019. Attention-based autism spectrum disorder screening with privileged modality. In *Proceedings of the IEEE International Conference on Computer Vision*. 1181–1190.

[6] Ryan Anthony J de Belen, Tomasz Bednarz, Arcot Sowmya, and Dennis Del Favero. 2020. Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. *Translational psychiatry* 10, 1 (2020), 1–20.

[7] Huiyu Duan, Xiongkuo Min, Yi Fang, Lei Fan, Xiaokang Yang, and Guangtao Zhai. 2019. Visual Attention Analysis and Prediction on Human Faces for Children with Autism Spectrum Disorder. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 3s (2019), 1–23.

[8] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Yi Fang, Zhaohui Che, Xiaokang Yang, Cheng Zhi, Hua Yang, and Ning Liu. 2018. Learning to predict where the children with asd look. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 704–708.

[9] V Eapan, A Masi, and F Kahn. 2020. Australian Autism Biobank follow-up cohort pilot study: Final Report. *Cooperative Research Centre for Living with Autism, Brisbane. Copies of this report can be downloaded from the Autism CRC website autismcrc. com. au* (2020).

[10] Annette Estes, Jeffrey Munson, Sally J Rogers, Jessica Greenson, Jamie Winter, and Geraldine Dawson. 2015. Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry* 54, 7 (2015), 580–587.

[11] Yi Fang, Huiyu Duan, Fangyu Shi, Xiongkuo Min, and Guangtao Zhai. 2020. Identifying Children with Autism Spectrum Disorder Based on Gaze-Following. In *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 423–427.

[12] Yuming Fang, Hanqin Huang, Boyang Wan, and Yifan Zuo. 2019. Visual Attention Modeling for Autism Spectrum Disorder by Semantic Features. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 625–628.

[13] Jesús Gutiérrez, Zhaohui Che, Guangtao Zhai, and Patrick Le Callet. 2021. Saliency4ASD: Challenge, dataset and tools for visual attention modeling for autism spectrum disorder. *Signal Processing: Image Communication* 92 (2021), 116092.

[14] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 262–270.

[15] Marisela Huerta, Somer L Bishop, Amie Duncan, Vanessa Hus, and Catherine Lord. 2012. Application of DSM-5 criteria for autism spectrum disorder to three samples of children with DSM-IV diagnoses of pervasive developmental disorders. *American Journal of Psychiatry* 169, 10 (2012), 1056–1064.

[16] Ming Jiang, Sunday M Francis, Diksha Srishyla, Christine Conelea, Qi Zhao, and Suma Jacob. 2019. Classifying individuals with ASD through facial emotion recognition and eye-tracking. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 6063–6068.

[17] Ming Jiang and Qi Zhao. 2017. Learning visual attention to identify people with autism spectrum disorder. In *Proceedings of the IEEE International Conference on Computer Vision*. 3267–3276.

[18] Ayush Kumar, Prantik Howlader, Rafael Garcia, Daniel Weiskopf, and Klaus Mueller. 2020. Challenges in Interpretability of Neural Networks for Eye Movement Data.. In *ETRA Short Papers*. 12–1.

[19] Olivier Le Meur, Alexis Nebout, Myriam Cherel, and Elise Etchamendy. 2020. From Kanner Austim to Asperger Syndromes, the Difficult Task to Predict Where ASD People Look at. *IEEE Access* 8 (2020), 162132–162140.

[20] Sidrah Liaqat, Chongruo Wu, Prashanth Reddy Duggirala, Sen-ching Samson Cheung, Chen-Nee Chuah, Sally Ozonoff, and Gregory Young. 2021. Predicting ASD diagnosis in children with synthetic and image-based eye gaze data. *Signal Processing: Image Communication* (2021), 116198.

[21] Wenbo Liu, Ming Li, and Li Yi. 2016. Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research* 9, 8 (2016), 888–898.

[22] Catherine Lord, Michael Rutter, Susan Goode, Jacquelyn Heemsbergen, Heather Jordan, Lynn Mawhood, and Eric Schopler. 1989. Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of autism and developmental disorders* 19, 2 (1989), 185–212.

[23] Catherine Lord, Michael Rutter, and Ann Le Couteur. 1994. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders* 24, 5 (1994), 659–685.

[24] Ann M Mastergeorge, Chanaka Kahathuduwa, and Jessica Blume. 2020. Eye-tracking in infants and young children at risk for autism spectrum disorder: A systematic review of visual stimuli in experimental paradigms. *Journal of Autism and Developmental Disorders* (2020), 1–22.

[25] Pramit Mazumdar, Giuliano Arru, and Federica Battisti. 2021. Early detection of children with autism spectrum disorder based on visual exploration of images. *Signal Processing: Image Communication* (2021), 116184.

[26] Adrienne Moore, Madeline Wozniak, Andrew Yousef, Cindy Carter Barnes, Debra Cha, Eric Courchesne, and Karen Pierce. 2018. The geometric preference subtype in ASD: identifying a consistent, early-emerging phenomenon through eye tracking. *Molecular autism* 9, 1 (2018), 19.

[27] Alexis Nebout, Weijie Wei, Zhi Liu, Lijin Huang, and Olivier Le Meur. 2019. Predicting Saliency Maps for ASD People. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 629–632.

[28] Anneli Olsen. 2012. The Tobii I-VT fixation filter. *Tobii Technology* (2012), 1–21.

[29] Karen Pierce, David Conant, Roxana Hazin, Richard Stoner, and Jamie Desmond. 2011. Preference for geometric patterns early in life as a risk factor for autism. *Archives of general psychiatry* 68, 1 (2011), 101–109.

[30] Karen Pierce, Steven Marinero, Roxana Hazin, Benjamin McKenna, Cynthia Carter Barnes, and Ajith Malige. 2016. Eye tracking reveals abnormal visual preference for geometric images as an early biomarker of an autism spectrum disorder subtype associated with increased symptom severity. *Biological psychiatry* 79, 8 (2016), 657–666.

[31] Yudong Tao and Mei-Ling Shyu. 2019. SP-ASDNet: CNN-LSTM based ASD classification model using observer scanpaths. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 641–646.

[32] Vladimir N Vapnik. 1999. An overview of statistical learning theory. *IEEE transactions on neural networks* 10, 5 (1999), 988–999.

[33] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. 2018. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4894–4903.

[34] Weijie Wei, Zhi Liu, Lijin Huang, Alexis Nebout, and Olivier Le Meur. 2019. Saliency prediction via multi-level features and deep supervision for children with autism spectrum disorder. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 621–624.

[35] Weijie Wei, Zhi Liu, Lijin Huang, Alexis Nebout, Olivier Le Meur, Tianhong Zhang, Jijun Wang, and Lihua Xu. 2020. Predicting atypical visual saliency for autism spectrum disorder via scale-adaptive inception module and discriminative region enhancement loss. *Neurocomputing* (2020).

[36] Weijie Wei, Zhi Liu, Lijin Huang, Ziqiang Wang, Weiyu Chen, Tianhong Zhang, Jijun Wang, and Lihua Xu. 2021. Identify autism spectrum disorder via dynamic filter and deep spatiotemporal feature extraction. *Signal Processing: Image Communication* (2021), 116195.

[37] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.