\documentclass[10pt,twocolumn,letterpaper]{article}

\usepackage{cvpr}

\usepackage{times}

\usepackage{epsfig}

\usepackage{graphicx}

\usepackage{amsmath}

\usepackage{amssymb}

\usepackage{graphicx}

\usepackage{float}

\usepackage{gensymb}

\usepackage{placeins}

% Include other packages here, before hyperref.

% If you comment hyperref and then uncomment it, you should delete

% egpaper.aux before re-running latex. (Or just hit 'q' on the first latex

% run, let it finish, and you should be clear).

\usepackage[breaklinks=true,bookmarks=false]{hyperref}

\cvprfinalcopy % \*\*\* Uncomment this line for the final submission

\def\cvprPaperID{\*\*\*\*} % \*\*\* Enter the CVPR Paper ID here

\def\httilde{\mbox{\tt\raisebox{-.5ex}{\symbol{126}}}}

% Pages are numbered in submission mode, and unnumbered in camera-ready

%\ifcvprfinal\pagestyle{empty}\fi

\setcounter{page}{1}

\begin{document}

%%%%%%%%% TITLE

\title{ Predicting Autism Spectrum Disorder (ASD) using Eye Tracking Data and Deep Learning

}

\author{Arya Sunil Gangatkar\\

{\tt\small agangadk@umd.edu}

\and

FNU Koustubh\\

{\tt\small koustubh@umd.edu}

\and

Suhas Nagaraj\\

{\tt\small suhas99@umd.edu}

\and

Swaraj Mundruppady Rao\\

{\tt\small swarajmr@umd.edu}

}

\maketitle

\begin{abstract}

Detecting Autism Spectrum Disorder (ASD) is a complex process due to the lack of definitive medical tests and the reliance on subjective behavioral assessments. ASD is characterized by persistent difficulties in social communication and repetitive or restricted behaviors. Among key diagnostic markers, atypical eye movement patterns have been identified as a distinguishing characteristic of people with ASD. In this work, we propose a novel system that leverages eye tracking data as an early test for ASD detection. Eye movement data, collected over time, were compressed into RGBA images, creating a dataset that encapsulates both spatial and temporal gaze dynamics [1]. Using this dataset, we employ advanced Deep Learning techniques to classify the images and diagnose ASD. This automated approach offers an efficient and scalable diagnostic tool that minimizes the need for extensive behavioral observation in clinical environments. Furthermore, our method can serve as an early screening mechanism, paving the way for more comprehensive and targeted evaluations of suspected individuals, thus improving early detection and timely intervention for ASD.

\noindent\textbf{Keywords: }Autism Spectrum Disorder (ASD), Eye Tracking Data, Deep Learning, CNN, LSTM, Transformers, Attention Mechanism

\end{abstract}

\maketitle

\section{Introduction}

Autism Spectrum Disorder (ASD) is a lifelong neurodevelopmental condition characterized by challenges in social interaction, communication, and behavior, including restrictive interests and repetitive actions. The World Health Organization estimates the global prevalence of ASD at approximately 0.76\%, equating to about 16\% of the global child population [3]. Despite its widespread occurrence, the diagnosis of ASD is often delayed for years after symptoms emerge, largely due to the limited availability of trained professionals and effective diagnostic tools. Early identification is critical, particularly in childhood, as timely interventions can significantly improve long-term outcomes for affected individuals [4].

Recent developments in artificial intelligence (AI) and machine learning (ML) have introduced innovative approaches to ASD diagnosis. Among these, eye-tracking technologies have shown considerable promise in capturing subtle and atypical visual attention patterns associated with ASD. These behavioral biomarkers, when integrated into machine learning models, hold potential for enhancing the accuracy and efficiency of early diagnosis [4], [5].

However, several challenges remain. Implementing these technologies in real-world clinical settings necessitates their integration with existing diagnostic frameworks, rigorous validation across diverse datasets, and addressing the ethical concerns related to collecting data from young children [2], [5]. This study seeks to enhance diagnostic capabilities by employing deep networks to analyze eye-tracking data, optimizing pre-processing and model performance to improve diagnostic accuracy [6], [7].

In this work, we propose a novel approach to ASD diagnosis by analyzing multiple eye-tracking images from individuals under evaluation. Unlike traditional methods that typically rely on a single image to classify individuals as autistic or non-autistic, our method aggregates data from a series of eye-tracking images to create a more comprehensive and dynamic profile for each individual. This approach mirrors clinical practices, where doctors observe behavioral patterns over time to inform their diagnosis. By using multiple images, we aim to capture a more complete range of visual attention behaviors, potentially improving diagnostic accuracy.

However, the picture changes if the models are trained on larger datasets (14M-300M images). We

find that large scale training trumps inductive bias. Our Vision Transformer (ViT) attains excellent

results when pre-trained at sufficient scale and transferred to tasks with fewer datapoints. When

pre-trained on the public ImageNet-21k dataset or the in-house JFT-300M dataset, ViT approaches

or beats state of the art on multiple image recognition benchmarks. In particular, the best model

reaches the accuracy of 88.55\% on ImageNet, 90.72\% on ImageNet-ReaL, 94.55\% on CIFAR-100,

and 77.63\% on the VTAB suite of 19 tasks.

\section{Related Work}

{Related Work: Discuss published work that relates to your project. How is your approach similar or different from others? Include 2-3 sentences summary of closest papers.}

\section{Dataset}

The dataset utilized in this study is designed to facilitate the analysis of eye-tracking patterns in children diagnosed with Autism Spectrum Disorder (ASD) and their neurotypical counterparts. The data collection involved 59 children recruited from schools in the Hauts-de-France region of France. Participants were divided into two balanced groups of children diagnosed with ASD and neurotypical children (Non-ASD), with ages ranging from approximately 3 to 13 years. The Childhood Autism Rating Scale (CARS), a widely recognized tool for assessing autism severity, was used by psychologists to classify and evaluate participants within the ASD cohort.

A total of 547 images were generated from eye-tracking visualizations, with 219 images representing children diagnosed with ASD and 328 images corresponding to neurotypical participants. These visualizations are derived from scanpaths, which provide a spatiotemporal mapping of the sequence of fixations and saccades made by the participants during the experimental sessions. Each participant was assigned a unique identifier to maintain anonymity, and metadata files documenting participant demographics (age, gender) and diagnostic scores (CARS) were compiled for analysis.

\subsection{Data Acquisition and Equipment}

Eye movement data were collected using the SMI RED-m remote eye-tracker, a high-precision device capable of recording at a sampling rate of 250 Hz. The eye-tracker was mounted below the display screen, ensuring unobtrusive and accurate tracking of participants' gaze as they engaged with visual stimuli. Participants were seated approximately 60 cm from the screen to optimize the accuracy of the recordings, with environmental conditions controlled to minimize visual distractions.

The experimental protocol involved presenting participants with a series of carefully curated videos designed to stimulate eye movements. These videos included dynamic and engaging content such as animated objects and human presenters, tailored to capture the attention of young children. The stimuli were presented in French, the native language of the participants. The objective was to elicit naturalistic eye movements, allowing researchers to observe and record patterns of gaze and attention under conditions approximating real-world interactions.

\subsection{Visualization Methodology}

The core innovation of this study lies in the transformation of raw eye-tracking data into visual representations, enabling detailed analysis of gaze dynamics. Each scanpath visualization encodes the trajectory of eye movements as a sequence of lines connecting consecutive fixations. The dynamics of these movements—velocity, acceleration, and jerk—are captured using RGB color gradients. Velocity is represented by a gradient from black (low) to red (high). Acceleration is encoded using a gradient from black (low) to green (high). Jerk is visualized with a gradient from black (low) to blue (high). The resulting images offer a compact yet comprehensive representation of the temporal and spatial characteristics of eye movements. The vertical mirroring of images ensures that the y-axis aligns correctly with the screen layout, enhancing interpretability. To standardize the amount of information contained within each image, a threshold of 200 fixation points was applied. This threshold strikes a balance between capturing sufficient detail and avoiding excessive visual clutter.

\subsection{Dataset Structure and Accessibility}

The dataset is publicly available via the Figshare Data Repository (\url{https://figshare.com/s/5d4f93395cc49d01e2bd}) and includes two primary components. Images are stored in two subfolders, one for ASD participants and another for neurotypical participants. Metadata includes CSV and JSON files documenting participant attributes (e.g., age, gender, CARS scores) and mapping images to unique participant IDs. The images are formatted as 640×480 pixels, providing a resolution suitable for both human interpretation and computational analysis. The dataset is structured to support diverse research applications, including machine learning, statistical analysis, and clinical studies.

\begin{figure}[htbp]

\centering

\includegraphics[width=0.9\linewidth]{pictures/asd-non-asd.png}

\caption{Comparison of eye-tracking scanpaths between a non-ASD diagnosed participant (left) and an ASD-diagnosed participant (right). The images illustrate the difference in eye movement patterns, with the ASD-diagnosed participant showing more dispersed and irregular gaze paths, indicative of atypical visual attention and fixation behavior associated with Autism Spectrum Disorder.}

\label{fig:asd\_comparison}

\end{figure}

\subsection{Exploratory Data Analysis}

We conducted an exploratory data analysis to examine the structure of the dataset and identify potential biases. This analysis will help us in uncovering trends that may be relevant for diagnosing Autism Spectrum Disorder.

As mentioned previously, the dataset consists of data collected from 59 children who participated, each labeled as ASD or neurotypical (non-ASD). The class labels were balanced, with 30 participants clinically diagnosed with ASD and 29 classified as neurotypical. This balance reduces the risk of model bias due to class imbalance, ensuring a fair assessment of the performance of the classifier.

Key demographic observations include age and sex. The participants ranged from approximately 3 to 13 years of age, with a mean age of 7.8 years. The age distribution was normal for both the classes with a slight skew towards younger participants in the neurotypical group. The dataset comprised 38 male and 21 female participants. Among ASD-labeled participants, 76\% were male, reflecting the higher prevalence of ASD in boys.

\begin{figure}[htbp]

\centering

\includegraphics[width=0.6\linewidth]{pictures/EDA/age-distribution.png}

\caption{Age distribution for dataset}

\label{fig:age\_distrivution}

\end{figure}

\begin{figure}[htbp]

\centering

\includegraphics[width=0.8\linewidth]{pictures/EDA/age distribution class wise.png}

\caption{Class-wise age distribution for ASD and non-ASD participants.}

\label{fig:class\_wise\_age\_distribution}

\end{figure}

\begin{figure}[htbp]

\centering

\includegraphics[width=0.7\linewidth]{pictures/EDA/Gender vs class distribution.png}

\caption{Gender and class distribution for ASD and Neurotypical(Non-ASD) participants.}

\label{fig:gender\_class\_distribution}

\end{figure}

For participants diagnosed with ASD, the Childhood Autism Rating Scale (CARS) score was also recorded. It's a 15-item observational rating scale completed by a clinician after observing a child's behavior and interviewing caregivers. A score below 30 indicates a non-autistic range, 30 to 36 indicates moderate autism and 36 to 60 indicates severe autism. Although CARS is a valuable tool, it has limitations. It may overdiagnose or underdiagnose autism in young children with intellectual disabilities. For our dataset, the minimum reported CARS score is 17, and the maximum score is 45 with an average CARS score of approximately 33.

\begin{figure}[htbp]

\centering

\includegraphics[width=0.7\linewidth]{pictures/EDA/CARS score distribution.png}

\caption{CARS score distribution for ASD diagnosed participants.}

\label{fig:gender\_class\_distribution}

\end{figure}

\begin{figure}[htbp]

\centering

\includegraphics[width=0.7\linewidth]{pictures/EDA/correlation heatmap.png}

\caption{Correlation Heatmap for Eye Tracking dataset}

\label{fig:gender\_class\_distribution}

\end{figure}

EDA revealed significant differences in eye-tracking behaviors and demographic characteristics between ASD and neurotypical participants. These findings informed the preprocessing steps and feature engineering techniques used in the models' development, ensuring that the models leveraged the most informative aspects of the dataset.

\section{Method}

The primary objective of this study is to leverage Deep Learning techniques to classify eye tracking images and detect Autism Spectrum Disorder (ASD). As the task involves binary classification (positive: ASD, and negative: Non-ASD), a sigmoid activation function is employed in the final layer, and Binary Cross-Entropy (BCE) is used as the loss function.

\subsection{Initial Approach}

We began by adopting standard methodologies from prior works in this field. A custom Convolutional Neural Network (CNN) was implemented as the initial model, consisting of three convolutional layers, each followed by max-pooling layers, and a three-layer fully connected head for classification. This served as a baseline approach for image-level classification. The Custom CNN model, while effective for training, exhibited poor validation performance with an AUC of 0.74, highlighting insufficient generalization and low recall, which are critical for medical applications. Furthermore, the model suffered from overfitting, as evidenced by the disparity between training and validation metrics.

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/CNN/loss.png}

\caption{Training and validation loss curves for the custom Convolutional model}

\label{fig:train\_val\_loss\_cnnLSTM}

\end{minipage}

\hfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/CNN/accuracy.png}

\caption{Training and validation accuracy curves for the custom Convolutional model}

\label{fig:train\_val\_accuracy\_cnnLSTM}

\end{minipage}

\end{figure}

To address this, we implemented transfer learning using pre-trained models such as ResNet50, DenseNet121, and VGG19, replacing their classification heads and fine-tuning select layers. This approach leveraged the feature extraction capabilities of these architectures, improving classification accuracy. Among these models, ResNet50 achieved the highest AUC (0.85) with stable validation accuracy and minimal over-fitting due to its residual connections that prevent gradient vanishing. DenseNet121 followed with an AUC of 0.83, leveraging its dense connectivity to enhance feature reuse and reduce redundancy. VGG19, while achieving an AUC of 0.80, was computationally heavier and showed slightly inferior generalization.

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.32\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Resnet/accuracy.png}

\label{fig:accuracy\_resnet}

\end{minipage}

\hfill

\begin{minipage}[t]{0.32\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Densenet/accuracy.png}

\label{fig:accuracy\_densenet}

\end{minipage}

\hfill

\begin{minipage}[t]{0.32\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/VGG19/accuracy.png}

\label{fig:accuracy\_vgg}

\end{minipage}

\caption{Training and validation accuracy curves for ResNet50 (left), DenseNet121 (center), and VGG19 (right) models.}

\label{fig:accuracy\_all}

\end{figure}

The ROC and PR curves for each model indicate that, while the models performed well, their performance may not yet meet the requirements for medical applications. ResNet50 achieves the highest Area Under the Curve (AUC) of 0.85, demonstrating superior discriminative ability. DenseNet and CNN follow closely with an AUC of 0.83, while VGG19 achieves an AUC of 0.80.

Analyzing the Precision-Recall (PR) curves, ResNet50 again leads with the highest Average Precision (AP) of 0.81. VGG19 and CNN both achieve an AP of 0.80, with DenseNet slightly trailing at 0.79.

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Results/BASE/download.png}

\label{fig:ROC}

\end{minipage}

\hfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Results/BASE/download (1).png}

\label{fig:PR }

\end{minipage}

\caption{ROC and PR curves of ResNet50

, DenseNet121, and VGG19 }

\end{figure}

The confusion matrices highlight the performance of different models in classifying positive and negative cases. ResNet50 shows balanced performance with 23 true positives (TP) and 56 true negatives (TN) but struggles with 16 false positives (FP). DenseNet improves slightly with 27 TP and 53 TN but has 12 FP. VGG19 achieves 28 TP but records 17 FN, indicating challenges in correctly identifying negatives. CNN demonstrates strong performance in identifying negatives with 61 TN but under performs in correctly classifying positives, with only 13 TP and 26 FP. These results emphasize the trade-offs in sensitivity and specificity across models, with ResNet50 and DenseNet showing better overall balance.

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Results/BASE/download (2).png}

\label{fig:roc\_resnet}

\end{minipage}

\hfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Results/BASE/download (3).png}

\label{fig:pr\_resnet}

\end{minipage}

\vfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Results/BASE/download (4).png}

\label{fig:roc\_vgg}

\end{minipage}

\hfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Results/BASE/download (5).png}

\label{fig:pr\_vgg}

\end{minipage}

\caption{Confusion Matrix for ResNet50, VGG19, CNN and DenseNet121}

\label{fig:roc\_pr\_curves}

\end{figure}

These approaches significantly improved validation accuracy, loss stabilization, and AUC, with ResNet50 and DenseNet121 emerging as the most robust options due to their superior feature extraction. However, the noisy and variable nature of the data limited their effectiveness in achieving our broader objective of individual-level classification. Variability in the number and quality of images per individual, combined with inconsistencies in eye-tracking data, led to suboptimal results.

While these approaches were effective for image-level classification, they did not align with our broader objective of classifying individuals rather than individual images. Due to the noisy nature of the data, the image-level accuracy was not very high, and the outputs did not fully reflect the goal of classifying individuals. Specifically, the variability in the number of images per individual and the quality of the eye tracking data led to suboptimal performance. However, we did implement a voting mechanism in the next section, which helped mitigate the impact of noisy data. By aggregating predictions from multiple images for each individual, incorrect classifications from individual images were effectively ignored, ensuring that the final classification reflected the majority view. This approach helped maintain accuracy to a reasonable level. Despite this, we still found that the voting-based method did not fully address the challenge of classifying individuals accurately. As a result, we shifted our focus towards sequence models, which are better suited to handle variable-length inputs and account for the relationships between multiple images belonging to the same individual. This decision allowed us to develop a more accurate method for individual-level classification, better aligned with our ultimate objective of diagnosing autism spectrum disorder based on eye tracking data.

\subsection{Voting-Based Approach}

Our first approach focused on classifying individuals through a voting-based algorithm. Individual images were classified using a custom CNN model, and the predictions were aggregated by counting the number of positive (ASD) and negative (Non-ASD) classifications. This method employed a majority voting mechanism to determine the final classification for each individual. Specifically, the number of positive and negative predictions for each individual was compared, and the final class was assigned based on the majority vote. This approach allowed us to account for the variability in the number of images associated with each participant and provided a simple yet effective method for individual-level classification.

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Vote/Va.png}

\label{fig:accuracy\_resnet}

\end{minipage}

\hfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Vote/roc.png}

\label{fig:accuracy\_densenet}

\end{minipage}

\caption{ROC Curve and Validation Accuracy Curve for the Combined CNN and Voting Model}

\hfill

\end{figure}

The training accuracy curve shows a steady improvement, surpassing 95\% by the final epoch, while the validation accuracy peaks near 100\% at epoch 8 before a slight decline, indicating potential minor overfitting. Training and validation loss exhibit consistent downward trends, with validation loss stabilizing at a low value, demonstrating effective learning. The ROC curve, with an AUC of 0.90, highlights strong discrimination capability between classes, a significant improvement over the baseline model.

\subsection{LSTM-Based Approach}

The second approach incorporated a Long Short-Term Memory (LSTM) network to address the challenge of handling variable-length inputs. In this approach, we first extracted features from each image using a ResNet50 model, with the final classification layer removed. These features, representing the image embeddings, were treated as input sequences to the LSTM. By leveraging the LSTM’s ability to process sequences, we were able to capture temporal dependencies across images from the same individual. The LSTM processed the sequence of features and passed the result through a fully connected classification head to generate a probability score for the individual being autistic. A threshold was applied to the probability score to classify the individual as autistic or non-autistic. This method effectively incorporated the temporal aspect of the eye movement data, providing a more nuanced classification of individuals based on their complete image sequence.The training and validation loss curves showed consistent convergence, reinforcing the robustness of this method in leveraging temporal dependencies for individual-level classification.

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/CNN+LSTM/loss.png}

\caption{Training and validation loss curves for the combination of Convolutional and LSTM model}

\label{fig:train\_val\_loss\_cnnLSTM}

\end{minipage}

\hfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/CNN+LSTM/accuracy.png}

\caption{Training and validation accuracies for the combination of Convolutional and LSTM model}

\label{fig:train\_val\_accuracy\_cnnLSTM}

\end{minipage}

\end{figure}

\subsection{Attention-Augmented LSTM}

To enhance the LSTM-based approach, we integrated an attention mechanism, enabling the model to focus on the most informative parts of the input sequence. This addition allowed the model to assign varying importance to different features, improving its ability to identify the most relevant images in the sequence for accurate classification. The attention mechanism not only enhanced the precision of predictions but also improved the interpretability of the model's decisions.

The performance improvement is evident in the training and validation curves. The training and validation loss curves demonstrate consistent convergence, with validation loss declining steadily alongside the training loss, indicating minimal overfitting and strong generalization to unseen data. Similarly, the training and validation accuracy curves show steady progress, with validation accuracy reaching over 90\%, reflecting the model's ability to effectively learn from variable input sequences.

Furthermore, the ROC curve with an AUC of 1.0 highlights the model’s high discriminative power in differentiating between classes. These results validate the effectiveness of the attention mechanism in augmenting LSTM performance for individual-level diagnosis, particularly in scenarios requiring the analysis of sequential data.

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.32\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/LSTM+ Attention/loss.png}

\end{minipage}

\hfill

\begin{minipage}[t]{0.32\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/LSTM+ Attention/accuracy.png}

\end{minipage}

\hfill

\begin{minipage}[t]{0.32\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/LSTM+ Attention/roc.png}

\end{minipage}

\caption{Performance of the LSTM model with attention mechanism: (Left) Training and validation loss curves showing consistent convergence, (Center) Training and validation accuracy curves demonstrating steady improvement with validation accuracy exceeding 90\%, and (Right) ROC curve with an AUC of 1.0 highlighting the model's high discriminative power.}

\label{fig:lstm\_attention\_performance}

\end{figure}

\subsection{Transformer-Based Approach}

In our final approach, we replaced the LSTM architecture with a Transformer-based model. The Transformer offers several advantages over LSTM, including parallel sequence processing and the ability to capture long-range dependencies through its self-attention mechanism. This scalability and capability to model complex relationships in the data make it highly effective for classifying individuals based on their eye movement data. The Transformer-based approach proved more efficient and powerful than the LSTM model, while retaining the flexibility to handle variable-length inputs.

The training and validation loss curves showed consistent convergence, with validation loss declining steadily alongside the training loss, indicating strong generalization to unseen data. Similarly, the accuracy curves demonstrated substantial improvement over successive epochs, achieving over 90\% validation accuracy. These results highlight the model's ability to learn meaningful patterns from eye movement data effectively.

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/transformers/train\_val\_loss\_transformer.png}

\caption{Training and validation loss curves for the Transformer-based model.}

\label{fig:train\_val\_loss\_transformer}

\end{minipage}

\hfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/transformers/accuracy.png}

\caption{Training and validation accuracies for the Transformer-based model.}

\label{fig:train\_val\_accuracy}

\end{minipage}

\end{figure}

The ROC curve with an AUC of 0.95 further demonstrated the model's strong discriminative power in differentiating between autistic and non-autistic individuals, supported by a low false positive rate and high true positive rate. These results validate the robustness of the Transformer-based approach in individual-level ASD classification.

\begin{figure}[htbp]

\centering

\includegraphics[width=0.9\linewidth]{pictures/transformers/roc.png}

\caption{Receiver Operating Characteristic (ROC) curve for the Transformer-based model, with an Area Under the Curve (AUC) of 0.95, indicating high discriminative performance in differentiating between autistic and non-autistic individuals.}

\label{fig:roc\_curve\_transformer}

\end{figure}

Through these iterative improvements, we developed a robust methodology for accurate individual-level classification. Notably, increasing the number of input images per participant significantly improved prediction accuracy. Aggregating multiple images for individual-level classification provided substantially better accuracy than image-level classification, as it utilized richer information to capture participant-specific features effectively.

As shown in the prediction probability versus the number of input images, the model achieved perfect classification when provided with 13 or more images of the individual, demonstrating the importance of aggregating data for robust classification.

\begin{figure}[htbp]

\centering

\includegraphics[width=0.9\linewidth]{pictures/Results/ours/download (6).png}

\caption{Incorporating more images provides a more comprehensive representation of the individual's patterns, enabling the model to make more accurate and reliable predictions.}

\label{fig:Increase in accuracy with increase in number of inputs of the individuals image}

\end{figure}

However, certain misclassifications, such as Participant 48 in the test set (misclassified as autistic despite being non-autistic), shows the presence of noise or outliers in the data. Further investigation showed that the image data for this participant exhibited patterns resembling those of autistic individuals, which likely contributed to the misclassification.\\

\section{Experiments}

Discuss the experiments that you performed to demonstrate that your approach solves the problem. The exact experiments will vary depending on the project, but you might compare with previously published methods, perform an ablation study to determine the impact of various components of your system, explain your cross validation setups, experiment with different hyper parameters or architectural choices, use visualization techniques to gain insight into how your model works, use various evaluation metrics, discuss common failure modes of your model, etc. You should include graphs, tables, or other figures to illustrate your experimental results.

\section{Limitations}

\textit{Limitations (5\%): Summarize the limitation of using your method. Under what conditions does it work and when does it not work? }

\\

The limited amount of data available after augmentation impacts the model's ability to fully utilize its potential, even though it performs well overall. Additionally, noise and outliers in the dataset present challenges. For instance, the image data of a patient diagnosed as ASD-positive may visually appear consistent with non-ASD characteristics, and vice versa for a patient diagnosed as ASD-negative. These outliers highlight variability within the dataset.

Moreover, the data collection process involves capturing eye-tracking data while children watch videos. However, as children are naturally prone to distraction, this behavior can introduce additional noise and variability into the dataset. Such factors may influence the reliability of the data and limit the model's ability to achieve even greater accuracy. Addressing these issues through larger datasets, better handling of outliers, and controlled data collection protocols could further enhance the model's performance.

\section{Conclusion}

\begin{thebibliography}{99}

\bibitem{figshare-ASD}

Figshare, ``Dataset for Autism Spectrum Disorder Detection Using Eye Tracking Data,'' [Online]. Available: \url{https://figshare.com/s/5d4f93395cc49d01e2bd}.

\bibitem{Hussaini-ML-ASD}

A. Hussaini, et al., ``A Systematic Literature Review on the Application of Machine-Learning Models in Behavioral Assessment of Autism Spectrum Disorder,'' \textit{Journal of Personalized Medicine}, vol. 11, no. 4, Apr. 2021.

\bibitem{Hodges2020}

Hodges, H., Fealko, C., Soares, N., ``Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation,'' \*Transl Pediatr.\*, vol. 9, Suppl 1, pp. S55–65, 2020.

\bibitem{Raj2020}

Raj, S., Masood, S., ``Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques,'' \*Procedia Computer Science\*, 2020. DOI: 10.1016/j.procs.2020.03.399.

\bibitem{Crippa2015}

Crippa, A., Salvatore, C., et al., ``Use of Machine Learning to Identify Children with Autism and Their Motor Abnormalities,'' \*Journal of Autism and Developmental Disorders\*, 2015. DOI: 10.1007/s10803-015-2379-8.

\bibitem{Tao2019}

Tao, Y., Shyu, M.L., ``SP-ASDNet: CNN-LSTM Based ASD Classification Model Using Observer Scanpaths,'' \*2019 IEEE ICME Workshops\*. DOI: 10.1109/ICMEW.2019.00062.

\bibitem{Alam2018}

Alam, M.E., et al., ``An IoT-Belief Rule Base Smart System to Assess Autism,'' \*Proceedings of iCEEiCT 2018\*. DOI: 10.1109/ICEEICT.2018.8646020.

\bibitem{Loth2017}

Loth, E., Charman, T., Mason, L., Tillmann, J., Jones, E. J. H., Wooldridge, C., Ahmad, J., Auyeung, B., Brogna, C., Ambrosino, S., et al., ``The EU-AIMS Longitudinal European Autism Project (LEAP): design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders,'' \*Molecular Autism\*, vol. 8, pp. 1–19, 2017.

\bibitem{Guillon2014}

Guillon, Q.; Hadjikhani, N.; Baduel, S.; Roge, B. ``Visual social attention in autism spectrum disorder: Insights from eye tracking studies,'' \*Neurosci. Biobehav. Rev.\*, 2014, 42, 279–297. [Google Scholar] [CrossRef] [PubMed].

\bibitem{Lord2006}

Lord, C.; Risi, S.; DiLavore, P.S.; Shulman, C.; Thurm, A.; Pickles, A. ``Autism from 2 to 9 years of age,'' \*Arch. Gen. Psychiatry\*, 2006, 63, 694–701. [Google Scholar] [CrossRef] [PubMed].

\bibitem{CDC-ASD-Diagnosis}

Centers for Disease Control and Prevention, ``Diagnosing Autism Spectrum Disorder (ASD),'' [Online]. Available: \url{https://www.cdc.gov/autism/diagnosis/index.html}

.

\end{thebibliography}

\section{Novelty and Key Contributions}

\section{Method}

The primary objective of this study is to leverage Deep Learning techniques to classify eye tracking images and detect Autism Spectrum Disorder (ASD). The task involves binary classification (positive: ASD, and negative: Non-ASD), and we employ a sigmoid activation in the final layer with Binary Cross-Entropy (BCE) as the loss function.

\subsection{Baseline and Initial Approach: Custom CNN}

Our baseline for image-level classification originates from a prior analysis on the same dataset, where Carette et al. employed classical machine learning algorithms and achieved 74.2\% accuracy. We aimed to improve upon this established baseline by exploring Convolutional Neural Network (CNN)-based deep learning techniques. As a first step, we implemented a custom CNN model consisting of three convolutional layers, each followed by max-pooling, and a three-layer fully connected head for classification. This model served as our initial deep learning approach for image-level classification.

Through iterative tuning of hyperparameters, regularization strategies (e.g., dropout, weight decay), and data augmentation, the custom CNN eventually reached a test accuracy of 76.24\% with a ROC AUC of 0.83. This exceeded the classical machine learning baseline (74.2\%), demonstrating the potential of deep learning techniques for this task, though there remained room for further improvement.

\subsection{Exploring Transfer Learning}

To potentially enhance performance beyond the custom CNN, we explored transfer learning using pre-trained models such as ResNet50, DenseNet121, and VGG19. The rationale was to leverage the robust feature extraction capabilities of models trained on large image datasets. Residual connections in ResNet50, dense connectivity in DenseNet121, and the deep architecture of VGG19 offered distinct advantages. However, as shown later, the overall improvements in AUC and accuracy over the tuned custom CNN were marginal, indicating that while transfer learning stabilized validation performance, it did not dramatically surpass the carefully tuned baseline CNN model.

\subsection{Individual-Level Classification: Voting-Based Baseline}

While we established a baseline for image-level classification (Carette et al.'s 74.2\% and our improved CNN at 76.24\%), there was no previously established baseline for individual-level classification. To address this gap, we implemented a voting-based method as our baseline approach for individual-level classification. After classifying each image from an individual using a transfer learning model, we aggregated the predictions. The final individual-level prediction was determined by majority vote. This simple yet effective method served as the baseline for evaluating more advanced sequence-based models introduced subsequently.

\subsection{Sequence Modeling with LSTM}

We extended the modeling capabilities by incorporating a Long Short-Term Memory (LSTM) network to handle sequences of images. First, features from each image were extracted using a pre-trained CNN backbone (e.g., ResNet50 with the final classification layer removed). These features formed input sequences to the LSTM, allowing the model to capture temporal or relational patterns across multiple images of the same individual. Our aim was to surpass the voting-based baseline for individual-level classification by modeling the sequential nature of the data.

\subsection{Attention-Augmented LSTM}

To further enhance sequence modeling, we integrated an attention mechanism. This allowed the model to focus selectively on the most informative parts of the input sequence. By assigning varying importance to different images, the attention-augmented LSTM could highlight critical patterns, potentially improving both accuracy and interpretability compared to the baseline methods.

\subsection{Transformer-Based Approach}

Finally, we explored a Transformer-based approach to replace the recurrent architecture. The Transformer leverages self-attention to efficiently capture long-range dependencies and relationships within the sequence of image embeddings. This architecture scales well and can handle variable-length inputs in parallel, potentially providing a more powerful and faster alternative to the LSTM-based models.

\section{Experiments}

\subsection{Experimental Setup and Evaluation Metrics}

The models described in the Method section were trained and evaluated using a dataset of eye-tracking images. We measured performance using metrics such as accuracy, Area Under the Receiver Operating Characteristic Curve (AUC), Average Precision (AP), and confusion matrices. Given the medical context, generalization and recall were emphasized. Experiments were conducted at both the image level (to improve upon the Carette et al. baseline of 74.2\%) and the individual level. For individual-level classification, our voting-based method served as the baseline, against which we compared sequence-based models (LSTM, attention-augmented LSTM, and Transformers).

\subsection{Results: Custom CNN and Transfer Learning}

The custom CNN initially faced issues with generalization and overfitting. However, after careful tuning, it achieved a test accuracy of 76.24\% (improving upon the 74.2\% baseline of Carette et al.) and a ROC AUC of 0.83. This demonstrated the feasibility of surpassing the classical machine learning baseline using a deep learning approach.

We then explored transfer learning. Among the pre-trained models, ResNet50 achieved an AUC of 0.85, DenseNet121 achieved 0.83, and VGG19 achieved 0.80. While these performances were in a similar range as the tuned custom CNN, the margin of improvement was not substantial. The tuned custom CNN’s competitive results indicate that a domain-specific, carefully optimized model can perform comparably to more complex pre-trained architectures.

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/CNN/loss.png}

\caption{Training and validation loss curves for the custom Convolutional model}

\label{fig:train\_val\_loss\_cnnLSTM}

\end{minipage}

\hfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/CNN/accuracy.png}

\caption{Training and validation accuracy curves for the custom Convolutional model}

\label{fig:train\_val\_accuracy\_cnnLSTM}

\end{minipage}

\end{figure}

Figure~\ref{fig:accuracy\_all} shows the training and validation accuracy curves for ResNet50, DenseNet121, and VGG19. Although these models provide stable validation metrics, the gains over the custom CNN are modest.

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.32\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Resnet/accuracy.png}

\label{fig:accuracy\_resnet}

\end{minipage}

\hfill

\begin{minipage}[t]{0.32\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Densenet/accuracy.png}

\label{fig:accuracy\_densenet}

\end{minipage}

\hfill

\begin{minipage}[t]{0.32\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/VGG19/accuracy.png}

\label{fig:accuracy\_vgg}

\end{minipage}

\caption{Training and validation accuracy curves for ResNet50 (left), DenseNet121 (center), and VGG19 (right) models.}

\label{fig:accuracy\_all}

\end{figure}

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Results/BASE/download.png}

\label{fig:ROC}

\end{minipage}

\hfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Results/BASE/download (1).png}

\label{fig:PR }

\end{minipage}

\caption{ROC and PR curves of ResNet50, DenseNet121, and VGG19. While pre-trained models show strong feature extraction, their improvements over the tuned custom CNN remain limited.}

\end{figure}

\subsection{Confusion Matrices and Performance Trade-Offs}

Confusion matrices for the various models highlight different trade-offs in sensitivity and specificity. Despite their architectural complexities, transfer learning models do not consistently outperform the tuned custom CNN across all metrics. These results suggest that the complexity of pre-trained architectures must be balanced with domain adaptation and careful tuning.

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Results/BASE/download (2).png}

\label{fig:roc\_resnet}

\end{minipage}

\hfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Results/BASE/download (3).png}

\label{fig:pr\_resnet}

\end{minipage}

\vfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Results/BASE/download (4).png}

\label{fig:roc\_vgg}

\end{minipage}

\hfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Results/BASE/download (5).png}

\label{fig:pr\_vgg}

\end{minipage}

\caption{Confusion matrices for ResNet50, VGG19, CNN, and DenseNet121. The tuned custom CNN is competitive with transfer learning models, illustrating that careful optimization can yield strong baseline performance.}

\label{fig:roc\_pr\_curves}

\end{figure}

\subsection{Individual-Level Classification: Voting-Based Baseline Results}

For individual-level classification, no established baseline existed. By adopting a voting-based approach (aggregating predictions from multiple images of the same individual), we established a baseline to evaluate sequence-based models. The voting-based method improved stability at the individual level and reached an AUC of 0.90, indicating the value of combining multiple samples per individual.

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Vote/Va.png}

\label{fig:accuracy\_resnet}

\end{minipage}

\hfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/Vote/roc.png}

\label{fig:accuracy\_densenet}

\end{minipage}

\caption{ROC Curve and Validation Accuracy Curve for the Combined CNN and Voting Model. The voting-based method serves as the baseline for individual-level classification, providing a reference point for evaluating sequence-based approaches.}

\hfill

\end{figure}

\subsection{LSTM and Attention-Augmented LSTM Results}

By introducing LSTM-based models, we aimed to surpass the voting-based baseline. The LSTM incorporated sequence modeling, capturing temporal dependencies and improving classification beyond simple aggregation. Adding an attention mechanism further refined this approach, allowing the model to emphasize crucial images. This resulted in near-perfect performance (AUC = 1.0) on validation data, underscoring the advantages of sequence modeling and informed attention in boosting individual-level classification accuracy.

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/CNN+LSTM/loss.png}

\caption{Training and validation loss curves for the Convolutional+LSTM model}

\label{fig:train\_val\_loss\_cnnLSTM}

\end{minipage}

\hfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/CNN+LSTM/accuracy.png}

\caption{Training and validation accuracies for the Convolutional+LSTM model}

\label{fig:train\_val\_accuracy\_cnnLSTM}

\end{minipage}

\end{figure}

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.32\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/LSTM+ Attention/loss.png}

\end{minipage}

\hfill

\begin{minipage}[t]{0.32\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/LSTM+ Attention/accuracy.png}

\end{minipage}

\hfill

\begin{minipage}[t]{0.32\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/LSTM+ Attention/roc.png}

\end{minipage}

\caption{Performance of the LSTM model with attention: (Left) Training and validation loss curves showing consistent convergence, (Center) Training and validation accuracy curves demonstrating steady improvement with validation accuracy exceeding 90\%, and (Right) ROC curve with an AUC of 1.0 highlighting the model's high discriminative power.}

\label{fig:lstm\_attention\_performance}

\end{figure}

\subsection{Transformer-Based Model Results}

Transitioning to a Transformer architecture facilitated more efficient and potentially more powerful sequence modeling. The Transformer-based model achieved over 90\% validation accuracy and a ROC AUC of 0.95, illustrating strong discriminative capability while handling variable-length inputs in parallel. These results confirm that more advanced sequence modeling architectures can significantly outperform the voting-based baseline for individual-level classification.

\begin{figure}[htbp]

\centering

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/transformers/train\_val\_loss\_transformer.png}

\caption{Training and validation loss curves for the Transformer-based model.}

\label{fig:train\_val\_loss\_transformer}

\end{minipage}

\hfill

\begin{minipage}[t]{0.48\linewidth}

\centering

\includegraphics[width=\linewidth]{pictures/transformers/accuracy.png}

\caption{Training and validation accuracies for the Transformer-based model.}

\label{fig:train\_val\_accuracy}

\end{minipage}

\end{figure}

\begin{figure}[htbp]

\centering

\includegraphics[width=0.9\linewidth]{pictures/transformers/roc.png}

\caption{Receiver Operating Characteristic (ROC) curve for the Transformer-based model, with an AUC of 0.95, indicating high discriminative performance.}

\label{fig:roc\_curve\_transformer}

\end{figure}

\subsection{Impact of the Number of Input Images and Data Quality}

Providing more images per participant enhanced individual-level classification. With 13 or more images per individual, the model achieved perfect classification accuracy, illustrating the importance of comprehensive data aggregation for robust diagnosis.

\begin{figure}[htbp]

\centering

\includegraphics[width=0.9\linewidth]{pictures/Results/ours/download (6).png}

\caption{Increasing the number of images per individual leads to more accurate and reliable predictions, emphasizing the value of richer participant-specific data.}

\label{fig:Increase in accuracy with increase in number of inputs of the individuals image}

\end{figure}

Nevertheless, certain misclassifications (e.g., Participant 48 being misclassified as autistic) suggest that data noise or outliers can mislead even advanced models. Such cases underscore the need for careful data quality checks, potential domain-specific preprocessing, or incorporating clinical insights.

---

This final version integrates the baseline from Carette et al., clarifies the baseline comparisons for both image-level and individual-level classification, and retains the full detail and structure of the methods and experimental results.

\end{document}