Engineered Science
DOI: https://dx.doi.org/10.30919/es1396

Utilizing Large Language Models for the Analysis of Video Data in Early Attention Deficit Hyperactivity Disorder Detection in Children

Abhilasha Kulkarni[1,*] and Jayashree Rajesh Prasad[1]

Abstract

This study proposes a novel framework leveraging large language models (LLMs) for early detection of attention deficit hyperactivity disorder (ADHD) in children based on video data analysis. The pipeline begins with behavioral sessions where video recordings are collected and stored systematically. These videos undergo preprocessing involving frame extraction, image processing, and feature extraction to identify key behavioral patterns. The extracted features are analyzed using LLMs to gain insights into attention patterns, movements, and behavioral trends. LLMs are particularly suited for this task as they can process multimodal data and contextualize subtle behavioral cues within a broader diagnostic framework, offering a robust tool for nuanced analysis. Subsequently, an ADHD detection model is trained to classify the likelihood of ADHD, providing high- and low-risk predictions. Results are visualized through a dashboard for clinicians or researchers, enabling further refinement of the detection process. This framework aims to support early ADHD diagnosis with data-driven insights, potentially improving the intervention strategies.

Keywords: ADHD detection; Video analysis; Large language models; Behavioral insights; Feature extraction; Early diagnosis.
Received: 11 December 2024; Revised: 16 January 2025; Accepted: 22 January 2025.
Article type: Research article.

1. Introduction

Attention deficit hyperactivity disorder (ADHD) is one of the most common neurodevelopmental disorders, diagnosed primarily in children but often persisting into adulthood. Characterized by significant issues with hyperactivity, impulsivity, and inattention, ADHD can severely impact a child's ability to function and succeed in academic, social, and home environments. Despite its prevalence, estimated to affect approximately 5% of children globally, ADHD remains underdiagnosed in many populations, leading to children not receiving the interventions that could significantly improve their developmental trajectories and quality of life. ADHD is a prevalent neurodevelopmental disorder affecting millions ofchildren worldwide. It shows itself as impulsivity, hyperactivity, and inattention, which can seriously hinder a child's socialization, scholastic achievement, and general development.[1] Since untreated ADHD frequently results in secondary issues, including low self-esteem, poor academic achievement, and increased incidence of comorbid illnesses like anxiety and depression, early detection is essential for effective management of ADHD.[2] By treating symptoms before they worsen, early therapies such as behavioral therapy, educational assistance, and medication can enhance results.[3] Research into cutting-edge early diagnostic methods is crucial because early detection of ADHD can help kids develop coping mechanisms, social skills, and improved outcomes throughout their lives.

Traditional ADHD diagnosis relies heavily on subjective assessments from parents, teachers, and clinicians, who observe a child's behavior in various settings and complete standardized questionnaires. Due to observers' biases, inconsistent results across settings, and the child's mood or surroundings at the time of observation, this method has several drawbacks. Furthermore, some symptoms might be concealed in specific settings or confused with typical behavioral variances, especially in younger children.[4] These elements may result in an underdiagnosis, incorrect diagnosis, or delayed diagnosis, delaying prompt interventions. The complex dynamic nature of ADHD symptoms may be difficult for standardized tests to measure, and there are not enough

[1] *Department of Computer Engineering, MIT School of Computing, MIT Art Design and Technology University, Loni Kalbhor, Pune, 412201, India*
**Email:* abhilasha1101@gmail.com (A. Kulkarni)

impartial diagnostic resources to back up these findings. Investigating data-driven, technologically based methods that can reliably and impartially examine behavior is necessary to meet these problems. Video data analysis offers a novel approach to studying ADHD-related behaviors, providing objective and continuous observation of a child's actions, movements, and interactions. Video data allows for the long-term monitoring of tiny behavioral indicators by capturing youngsters in controlled or natural environments.[5] Certain patterns that are hard to reliably perceive in real-time, like frequent changes in gaze, bodily restlessness, or trouble focusing, can be detected by sophisticated video processing algorithms. Researchers and physicians can measure and examine these patterns with automated video analysis, providing a more uniform evaluation that lessens subjective biases.[6] Additionally, it is possible to analyze video data longitudinally, which is useful for continuous monitoring and early detection since it tracks behavioral changes over time. By offering reliable impartial insights into a child's behavioral profile, this method may be used in addition to conventional ADHD tests.

Originally designed for tasks in natural language processing, large language models (LLMs) are evolving to support broader applications, including multimodal data analysis that combines text, image, and video inputs. The ability of LLMs to decipher intricate contextual information can be used to analyze nonverbal behaviors in video footage.[7] LLMs can be trained to identify trends in children's movements, gaze changes, and interactions to detect ADHD by intelligently and systematically analyzing contextual data. For instance, by contextually examining video frames and sequences, LLMs can detect frequent movements or attentional shifts, both of which may be signs of ADHD. This feature gives LLMs the capacity to automatically extract behavioral signals linked to ADHD, offering a scalable, automated approach for early identification.[8] Thus, LLMs provide a novel way to handle massive amounts of video data by turning behavioral cues into insightful information that might help doctors make diagnoses.

Numerous approaches, such as behavioral evaluations, psychometric testing, and neuroimaging techniques, are being used in the current study on ADHD detection. The conners rating scale and the ADHD rating scale are two examples of parent and teacher questionnaires used in traditional behavioral approaches to assess a child's hyperactivity, impulsivity, and attentiveness.[9] Although neuroimaging techniques like electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) offer insights into brain activity patterns linked to ADHD, they are frequently expensive, require specialist equipment, and may not be available for routine evaluations. Computational methods, like as eye-tracking data or speech patterns, are also gaining popularity as more objective indicators of ADHD. These approaches still have drawbacks, though, in terms of affordability, usefulness in routine clinical settings, and accessibility.

In psychological and behavioral evaluations, video data analysis is being utilized more and more to record and measure non-verbal indicators including body language, gaze direction, and facial expressions. Using machine learning (ML) algorithms to find behavioral markers, researchers have successfully used video analysis to detect a range of disorders, such as anxiety, sadness, and autism. Continuous observation of a child's behavior in controlled or natural environments is made possible by video analysis, which yields objective information to support subjective reports.[10] Frequent fidgeting, gaze aversion, or trouble sitting are behaviors frequently linked to ADHD that can be captured in video data for ADHD identification.[11] Nevertheless, there is still a dearth of studies particularly relating video data to ADHD, and existing systems frequently lack automated, scalable solutions that can objectively and consistently examine these behaviors. ML techniques, particularly deep learning, have been widely used in the medical field for tasks like image and text analysis, disease prediction, and diagnostic assistance. When modified to incorporate data formats such as video and image sequences, LLMs, which were first created for text-based natural language processing, have demonstrated promise in multimodal data interpretation. LLMs are useful for studying complex behavioral data because they can identify correlations and contextual patterns in huge datasets.[12,13] ML models have been used in ADHD research to analyze structured behavioral observations, eye-tracking, and EEG data, providing unbiased insights into impulsive and attentional behaviors. Despite being relatively new, LLMs in video-based ADHD detection have the potential to improve diagnostic model accuracy by offering deeper behavioral insights through contextual and sequential data interpretation. ML applications are beyond pandemic-related research, such as using supervised learning algorithms to explore hidden patterns in medical datasets. ML is widely applicable for predicting health outcomes, diagnosing diseases, and optimizing treatment protocols through knowledge discovery and feature selection.[14] ML has wide-ranging applications beyond traditional areas. It is employed in healthcare for disease diagnosis and treatment optimization, in industries for predictive maintenance and automation, and in scientific research to analyze complex datasets for pattern recognition and predictions. Additionally, it supports advancements in areas like personalized medicine, environmental monitoring, and smart city management.[15] ML finds applications across diverse domains beyond traditional settings. It is instrumental in healthcare for early disease detection and treatment, in finance for fraud detection, and in industry for automation and predictive maintenance. These innovations improve efficiency, decision-making, and outcomes in numerous fields.[16] ML applications extend beyond healthcare and are transformative in various fields. For example, it is extensively used in financial services for fraud detection and credit scoring, in agriculture for crop yield prediction and disease detection, and

in transportation for autonomous vehicles and traffic optimization systems.[17]

The study surveys how Generative AI and Large Language Models (LLMs) are revolutionizing video generation, understanding, and streaming.[18] It investigates how these technologies can be used to produce realistic video content, extract useful information from videos, and improve streaming by delivering content in an efficient and customized manner. Important developments include LLM-based methods for video captioning, segmentation, and question-answering, as well as GANs, VAEs, autoregressive models, and diffusion models for video production. LLMs enhance video compression, forecast user perspectives, and maximize bandwidth utilization in streaming. The poll points up issues include the requirement for large-scale datasets, significant computational costs, and temporal consistency in video creation. It draws attention to challenges with modern technology, such as false information, privacy concerns, and moral dilemmas. The study emphasizes the necessity for responsible growth and regulation while highlighting the revolutionary potential of generative AI and LLMs in the fields of networking, multimedia, and artificial intelligence.

The study introduces LLM-grounded video diffusion (LVD), a framework addressing challenges in text-to-video generation by leveraging dynamic scene layouts (DSLs) generated using LLMs to guide video diffusion models.[19] First, LVD leverages LLMs to generate spatiotemporal layouts based on text prompts, capturing object dynamics. Second, these layouts use attention-based modifications to direct the video generating process without the need for fine-tuning. As shown across benchmarks covering tasks such as spatial dynamics, visibility, and sequential actions, this method greatly enhances text-video alignment and video fidelity. Although there are still issues with managing ambiguities and enhancing visual quality, LVD beats baseline models, demonstrating the potential of LLMs to improve video creation. The framework emphasizes how LLM-grounded approaches are revolutionizing video synthesis.

LLoVi, a language-based long-range video question-answering (LVQA) framework that leverages LLMs, was introduced for analyzing long-form videos.[20] Short-term video clips are first processed by a pre-trained visual captioner to produce textual descriptions, and then an LLM aggregates these captions to reason over extended temporal extents. This is how LLoVi divides the LVQA problem. The system includes a multi-round summarizing prompt to tackle noisy captions, improving LVQA accuracy by removing extraneous information and concentrating on important facts. Empirical testing demonstrates LLoVi's resilience on datasets such as EgoSchema, NExT-QA, and IntentQA, surpassing previous techniques with notable improvements in accuracy. Key findings emphasize the significance of optimal prompting strategies and the selection of powerful visual captioners and LLMs (*e.g.*, GPT-4). The framework is model-agnostic and offers potential for improvements with advancements in visual captioning and LLM designs, aiming to simplify and enhance long-range video understanding tasks.

Memory-augmented large multimodal model (MA-LMM), a framework for efficient and effective long-term video understanding, was introduced to eliminates the drawbacks of current multimodal models, which are limited by the high GPU memory utilization and context length of LLMs.[21] By processing video frames in a sequential manner and preserving past data in a long-term memory bank, MA-LMM takes an online method as opposed to processing whole films at once. The model can effectively reference temporal context thanks to this memory bank, which records and aggregates previous video information. In order to minimize redundancies and preserve memory size, the framework also incorporates a compression technique. Existing multimodal models can be seamlessly integrated with MA-LMM without the need for retraining. The state-of-the-art performance of MA-LMM is demonstrated by extensive tests in a variety of tasks, such as long-term video understanding, video question answering, and video captioning. The model improves temporal modelling capabilities while drastically lowering the computing costs. The design of MA-LMM facilitates real-time applications as well, demonstrating an enhanced accuracy in action prediction and online video reasoning tasks. Its versatility and promise to further video analysis across domains are highlighted by its modular architecture.

Reframe any video agent (RAVA), a LLM-based framework for automated video reframing, was introduced to address the growing demand for adapting videos to diverse screen aspect ratios on social media platforms.[22] Three phases make up RAVA's operation: perception, where it deciphers user commands and extracts video content by identifying objects and scenes; planning, where it establishes aspect ratios, ranks objects, and creates layout and visual effects strategies; and execution, where it carries out the planned modifications and applies effects in real time. The system improves its ability to comprehend and edit video information by utilizing sophisticated technologies such as CLIP, segment anything model (SAM), and multimodal LLMs. Results from experiments show that RAVA performs competitively when compared to professional editing tools in both real-world reframing tasks and video salient object recognition. The study emphasizes RAVA's promise to make video editing more accessible, but it also points out its limits in terms of foundational model dependence and its room for improvement in terms of object recognition and timeline editing features.

A Socratic video understanding system was proposed for unmanned aerial vehicles (UAVs) using LLMs and vision-language models (VLMs) to enhance scene understanding in real-world environments.[23] The system creates thorough world-state logs that record objects, events, and risks by combining the BLIP-2 framework with OpenAI's da-vinci-003 and GPT-3.5-turbo. This allows for zero-shot reasoning and decision-making. The system uses video footage from lightweight drones, such as RYZE Tello, to process scenes

using pre-programmed prompts for hazard assessment, object detection, action prediction, and captioning. By recommending the best commands for drone operations, the framework highlights human-in-the-loop involvement and cost-effective deployment. Readability measures show that the model produces high-quality, easily readable textual outputs and experimental findings show that it can accurately anticipate activities and identify threats in a variety of situations. While addressing latency and scalability issues for future development, the study emphasizes the promise of LLM-enhanced UAV systems for uses such as safety monitoring, surveillance, and real-time navigation.

Although current techniques for detecting ADHD offer insightful information, there are still significant drawbacks, particularly about the impartiality, usability, and scalability of tests. Conventional approaches depend on subjective observations, which may not always be reliable, while sophisticated neuroimaging or eye-tracking techniques are frequently unavailable for everyday use. Despite the potential of video data analysis, present systems frequently lack standardized markers for behaviors specific to ADHD and automated, real-time analysis capabilities. By using LLMs to automate the interpretation of video data, this study seeks to close these gaps and offer a more scalable and objective method of identifying behaviors associated with ADHD. The goal of this work is to provide a more affordable and easily accessible technology for early ADHD detection that can be used in clinical and educational settings by concentrating on video analysis reinforced by LLM-based contextual interpretation. This study aims to develop a consistent, objective method for recognizing behavioral patterns suggestive of ADHD by automating the analysis of recorded actions. This research is important because it has the potential to improve diagnostic accuracy and lessen the need for subjective observations, which are two drawbacks of conventional ADHD assessment techniques. This study presents a scalable, technology-driven method for incorporating LLMs into the ADHD screening process, which might be used in clinical and educational contexts to promote early identification. For children with ADHD, early and accurate diagnosis can increase access to timely therapies and enhance long-term results. In the end, this study might help create more objective and approachable instruments for evaluating ADHD, expanding the use of AI in mental health diagnosis.

2. Methodology

Fig. 1 depicts system architecture for early ADHD detection in children using LLMs. It is divided into five main modules: data collection, preprocessing, analysis, ADHD detection model, and results & feedback.

Data Collection: In this module, behavioral sessions of children are recorded on video, capturing real-time interactions, movements, and attention shifts. This raw video data is then stored in a secure, structured video data storage system for further processing.

Preprocessing: The preprocessing module prepares the video data for analysis. First, frames are extracted from the video to create a sequence of still images, breaking down continuous behavior into manageable units. These frames are then processed to enhance image quality and remove irrelevant visual noise. The system identifies and extracts key behavioral features, such as gaze direction, head movements, and other physical gestures, which are relevant to ADHD detection.

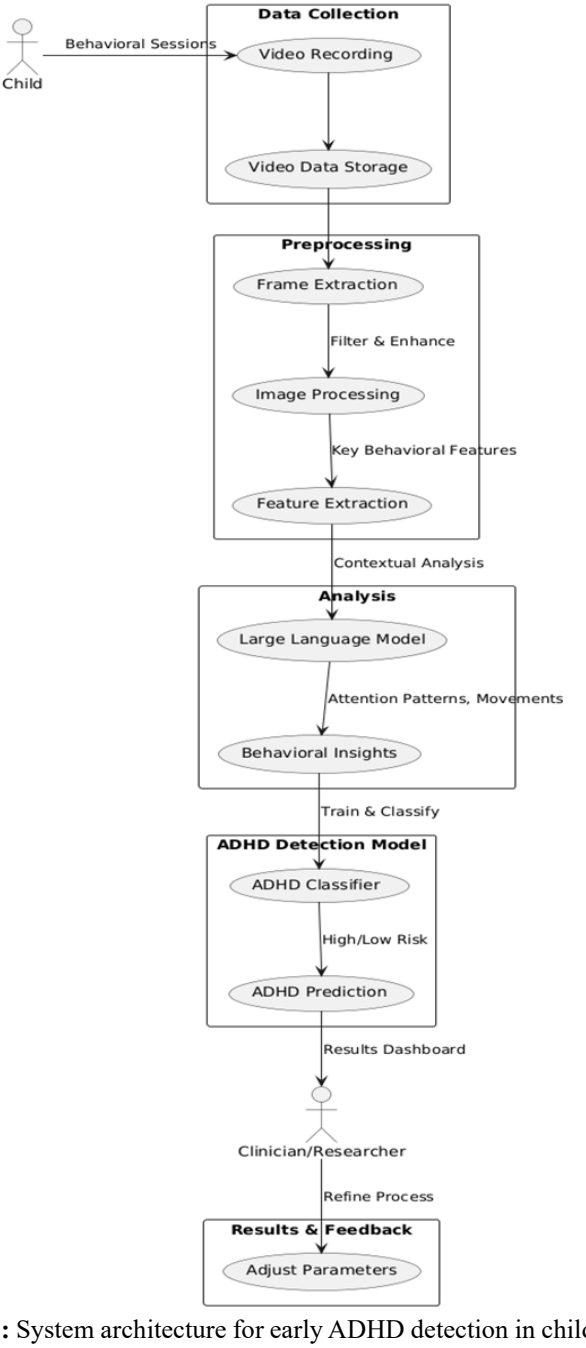

Fig. 1: System architecture for early ADHD detection in children using large language models.

Analysis: This module employs a pre-trained LLM to analyze contextual features extracted from video data. The LLM leverages its transformer-based architecture, which uses attention mechanisms to process and interpret temporal and

spatial patterns in the data. By identifying correlations between these patterns and ADHD-related behaviors, such as attention lapses, impulsive movements, and difficulty maintaining focus, the model generates detailed behavioral insights. The analysis begins with feature embeddings derived from video frames, representing attention and movement characteristics. The LLM uses these embeddings to establish relationships between sequential features, enabling it to capture subtle temporal dynamics indicative of ADHD-related symptoms. Through multiple layers of processing, the model contextualizes these behavioral cues within a diagnostic framework, refining the representation of key patterns associated with attention and movement. The final output of this module includes classifications and behavioral insights that are synthesized into actionable results. These results, delivered via a clinician-friendly dashboard, provide an objective and scalable basis for identifying children at high or low risk for ADHD. By automating the interpretation of complex behavioral data, this analysis module minimizes subjectivity and improves the precision and efficiency of ADHD detection.

ADHD Detection Model: This module employs a ML classifier to distinguish ADHD-related behaviors from typical behaviors by analyzing the contextual insights generated by the LLM. The classifier is a supervised learning model trained on labeled data, where each instance corresponds to a child's observed behaviors and their associated ADHD risk level. Examples of classifiers commonly used for such tasks include decision trees, support vector machines (SVMs), or neural networks, depending on the complexity of the problem and the dataset size.

The classifier is designed to process behavioral features such as attention patterns, impulsive movements, and sustained focus levels, which are extracted and contextualized by the LLM. These features are numerical representations derived from video data, summarizing temporal and spatial behaviors over a defined period. The classifier distinguishes ADHD-related behaviors based on patterns in these features that align with known diagnostic criteria. The training process involves the following steps:

Feature Engineering: Behavioral features identified by the LLM are aggregated and encoded into a structured dataset. For example, features might include frequency and duration of attention lapses or variability in movement.

Training on Labeled Data: The classifier is trained using a labeled dataset that includes examples of both ADHD-related behaviors and typical behaviors. Labels are assigned based on clinical diagnoses or expert annotations. The classifier learns to map the input features to the corresponding labels (e.g., "high risk" or "low risk").

Decision-Making Process: Once trained, the classifier evaluates new behavioral data. It applies learned patterns and decision boundaries to classify the behaviors into categories, for instance, High Risk: Frequent, prolonged attention lapses combined with erratic movement patterns. Low Risk: Consistent focus and controlled movements within normal ranges.

Output and Interpretation: The classifier generates a probability or confidence score for each prediction, indicating the likelihood of ADHD-related behaviors. These predictions are synthesized into an overall risk level (*e.g.*, "high risk" or "low risk"). The ADHD detection model provides an objective data-driven assessment of each child's risk level, reducing reliance on subjective human observation. By learning from diverse datasets, the classifier can generalize across various contexts, improving the accuracy and reliability of early ADHD detection. The results are then visualized in a user-friendly dashboard for clinicians, aiding in informed decision-making and early intervention.

Results & Feedback: The results and feedback module present the ADHD risk assessment to clinicians or researchers through a results dashboard. This dashboard allows the user to review the ADHD prediction along with the analysed behavioral insights. Additionally, clinicians or researchers have the option to adjust system parameters, allowing for further refinement of the model based on user feedback and enhancing the accuracy of future assessments.

This modular architecture provides a comprehensive approach to ADHD detection, from data collection to actionable insights, supporting clinicians in early diagnosis by leveraging the analytical power of LLMs.

3. Implementation

To analyze behaviors in children potentially indicative of ADHD, we utilized video data capturing naturalistic interactions of children in diverse settings. A critical step in our methodology involves the extraction and preparation of video data, which ensures that the input to our LLM-based tool is optimized for high-fidelity analysis.

3.1 Video frame extraction

The initial step involves the systematic breakdown of video into individual frames, enabling detailed examination of each moment captured on video. This process is crucial as it allows the LLM to analyze discrete instances of behavior, which could indicate signs of ADHD.

3.1.1 Python code implementation

This Python script (Fig. 2) uses the OpenCV library to capture video frames and save them into a designated output folder, which our model then uses for further analysis. This script is efficient and automated, ensuring that data handling remains consistent and accurate across different datasets. Fig. 2 shows the sample python code.

3.1.2 Opening the video file

The function cv2.VideoCapture (video_path) is used to read the video file. The video file consists of frames captured at a

```python
def extract_frames(video_path, output_folder):
    cap = cv2.VideoCapture(video_path)
    if not os.path.exists(output_folder):
        os.makedirs(output_folder)

    fps = cap.get(cv2.CAP_PROP_FPS)
    total_frames = int(cap.get(cv2.CAP_PROP_FRAME_COUNT))

    print(f"Video FPS: {fps}")
    print(f"Total Frames: {total_frames}")

    frames_folder = []

    for frame_num in range(total_frames):
        ret, frame = cap.read()
        if not ret:
            break

        frame_path = os.path.join(output_folder, f"frame_{frame_num:04d}.png")
        cv2.imwrite(frame_path, frame)
        frames_folder.append(frame_path)

    cap.release()
    return frames_folder
```

Fig. 2: Sample python code.

specific frame rate (FPS). Let the video duration be T seconds. The total number of frames N in the video is given by Eq. (1):

$$N = FPS \times T \quad (1)$$

where FPS is obtained using cv2.CAP_PROP_FPS, and N is obtained using cv2.CAP_PROP_FRAME_COUNT.

3.1.3 Extracting frames

The code iterates over all N frames of the video. For each frame ii, the frame is read using cap. read() and is saved as an image file. This can be represented mathematically by Eq. (2):

$$Fi = V(i), \text{ for } i=0, 1, \ldots, N-1 \quad (2)$$

where Fi is the i-th extracted frame; V(i) is the frame returned by the video at index i. Each frame is saved as an image file with a name following the pattern frame_{frame_num:04d}.png.

3.1.4 Saving frames

The frames are saved to a specified output folder. The saved path for each frame is given by Eq. (3):

$$P_i = \text{output\_folder}+\text{"/frame\_"}+i+\text{".png"} \quad (3)$$

where P_i represents the path to the i-th saved frame.

3.1.5 Output

The function returns a list of paths to all the extracted frames in Eq. (4):

$$\text{Frames\_folder}=[P_0, P_1, \ldots, P_{N-1}] \quad (4)$$

3.2 Video processing

Once the frames are extracted, each is analyzed using our LLM Gemini, which has been specifically trained to recognize and interpret behavioral patterns that might be symptomatic of ADHD. The process includes:

Frame Selection: From the vast array of extracted frames, a subset is carefully chosen based on predetermined criteria such as clarity, relevance to behavioral cues, and representation of diverse interactions.

Text Generation: For each selected frame, descriptive text is generated using LLM Gemini, which provides a detailed narrative of the observed behaviors. This text is crucial for subsequent analysis stages as it converts visual data into textual data that can be more deeply analyzed for patterns and anomalies.

Conclusion Summarization: The generated texts are then compiled and analyzed to summarize findings and draw conclusions about the presence of behavioral indicators of ADHD.

This workflow, supported by robust technological tools and advanced AI models like LLM Gemini, sets a new standard in the field of behavioral analysis, particularly in the context of diagnosing developmental disorders such as ADHD. The use of such advanced technology allows for a more nuanced understanding and potentially earlier detection of ADHD, offering significant improvements over traditional diagnostic methods.

3.3 Text generation and analysis

3.3.1 Overview

The methodology for analyzing video data involves a multi-step process where video frames are first extracted, reviewed, and then processed using a LLM to generate textual interpretations of the observed behaviors. This text-based

analysis aims to capture nuanced behavioral signals that might be indicative of ADHD or other syndromes, thereby supporting more objective and detailed assessments.

3.3.2 Video review and prompt formulation

The initial phase involves a comprehensive manual review of the video frames that have been extracted from longer recordings. Each frame is analyzed to understand the context and the specific behaviors displayed by the children. This meticulous examination is crucial for crafting accurate and descriptive prompts that will guide the LLM in its analysis. These prompts are tailored to each frame, ensuring they encapsulate key information about the child's behavior, actions, and any visible symptoms or interactions that are relevant to the study.

3.3.3 Python code implementation for text generation

The following Python script shown in Fig. 3 is integral to processing the frames with the LLM. It is designed to automate the interaction with the LLM, manage the flow of data, and handle the output to ensure all generated responses are systematically recorded and analyzed:

3.3.4 Code explanation

The extract_frames function is designed to extract frames from a video file specified by video_path and save them to a designated output_folder. The script begins by importing necessary libraries: cv2 for video and image processing and os for file and directory operations. Using cap = cv2.VideoCapture(video_path), the video file is opened, creating a cap object for interacting with the video. Before proceeding, the script checks if the output_folder exists using if not os.path.exists(output_folder): os.makedirs(output_folder) and creates the folder if it does not, ensuring a ready location for storing extracted frames. To handle frame extraction correctly, the total number of frames in the video is determined using total_frames = int(cap.get(cv2.CAP_PROP_FRAME_COUNT)). Fig. 3 shows the sample python code.

The loop for frame_num in range(total_frames): iterates through each frame in the video, processing them sequentially. For each iteration, ret, frame = cap.read() attempts to read the next frame; ret is a boolean indicating success. If reading fails (if not ret:), likely signaling the end of the video, the loop is exited. Each frame's file path is constructed using frame_path = os.path.join(output_folder, f"frame_{frame_num:04d}.png"), ensuring sequentially numbered file names with zero padding for consistency. The frame is then saved as a PNG image using cv2.imwrite(frame_path, frame). After all frames are processed, cap.release() is called to release the video file and free up system resources. Finally, the function returns output_folder, which contains the saved frames and can be used for further processing.

3.3.5 Significance in research

The extraction and individual analysis of video frames form a critical foundation for behavioral studies, enabling precise frame-by-frame scrutiny of subtle actions and patterns that may be overlooked in real-time observation. By converting

```python
def process_frames_with_model(model, frames_folder, output_file, num_frames=12):
    frames_per_iteration = len(frames_folder) // num_frames
    delay_between_iterations = 30

    with open(output_file, "wb") as text_file:
        for i in range(0, len(frames_folder), frames_per_iteration):
            frame_path = frames_folder[i]
            frame = Image.open(frame_path)
            response = model.generate_content([prompt, frame])
            var = response.prompt_feedback
            var_str = str(var).strip()

            if var_str == "block_reason: OTHER":
                print(f"Skipping frame {frame_path} due to model block reason.")
            elif response.parts:

                print("Done")
                text_file.write(f"# Response {i + 1}\n".encode('utf-8'))
                text_file.write(response.parts[0].text.encode('utf-8'))
                text_file.write(b"\n\n")
            else:
                print(f"No response parts found for frame {frame_path}.")


            time.sleep(delay_between_iterations)
```

Fig. 3: Sample python code.

dynamic video data into a sequence of static frames, researchers can apply advanced image processing techniques and AI models to extract meaningful insights with greater accuracy and consistency. This approach ensures that even fleeting or nuanced behavioral cues are captured and analyzed, providing a robust basis for data-driven assessments and reducing the risk of missing critical moments during analysis.

3.3.6 Examples of generated responses

Response 1: The child in the video appears to have difficulty pronouncing words and holding the microphone. It is possible that the child has a speech disorder, such as apraxia of speech, which can make it difficult to produce sounds and words correctly. The child may also have difficulty holding the microphone due to a lack of coordination or fine motor skills. It is important to note that the child may have a variety of other conditions, such as ADHD or down syndrome, which can also affect their speech and motor skills.

Response 2: The child in the video appears to have difficulty pronouncing words and holding the microphone. The child may have a speech disorder, such as apraxia of speech, which is a motor speech disorder that affects a person's ability to produce sounds and words correctly. The child may also have a physical disability, such as cerebral palsy, which can affect a person's muscle tone and coordination. It is difficult to say for sure what condition the child has based on the video alone. Further evaluation by a qualified professional would be necessary to determine a diagnosis.

Response 3: The child in the video appears to have difficulty pronouncing words and holding the microphone. The child may have a speech disorder, such as apraxia of speech, which is a motor speech disorder that affects a person's ability to produce speech sounds correctly. The child may also have a physical disability, such as cerebral palsy, which can affect a person's muscle control over muscles and movements. The child may also have a cognitive disability, such as down syndrome, which can affect a person's intellectual and developmental abilities. The child may also have a behavioral disorder, such as ADHD, which can affect a person's ability to focus and control their behavior. It is important to note that the child in the video may not have any of the conditions mentioned above. The child may simply have a difficult time singing the song. It is also important to note that the child may be nervous or uncomfortable being on camera. If you are concerned about the child in the video, it is important to talk to a qualified professional, such as a doctor or a speech therapist. The professional can help you determine if the child has a speech disorder, a physical disability, a cognitive disability, or a behavioral disorder. The professional can also help you develop a plan to help the child.

Response 4: The child in the video appears to have difficulty pronouncing words and holding the microphone. It is possible that the child has a speech disorder, such as apraxia of speech, which can affect a person's motor planning for speech. The child may also have a cognitive disability, such as down syndrome, which can cause difficulty with language and communication. The child's behavior may also be related to a behavioral disorder, such as ADHD. ADHD can cause difficulty with attention, impulsivity, and hyperactivity.

These responses illustrate the LLM's capability to provide detailed and varied interpretations of observed behaviors, aiding in the comprehensive analysis of ADHD indicators.

4. Results

4.1 Distribution of syndromes

Fig. 4 illustrates the distribution of different syndromes observed in the dataset. The syndromes include ADHD, Autism, down syndrome, and Apraxia of Speech.

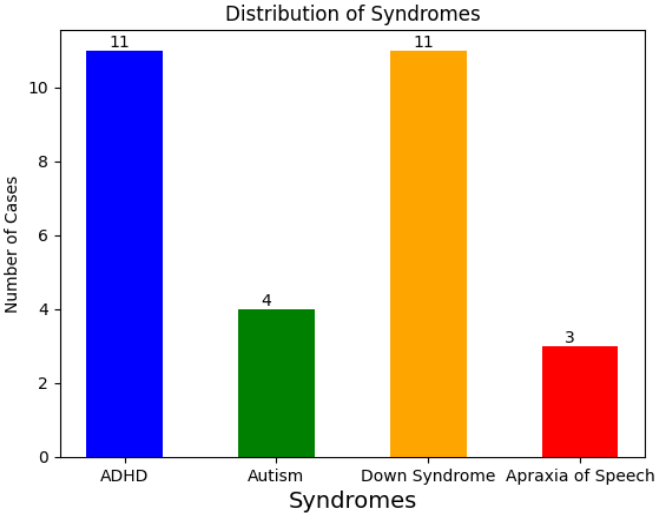

Fig. 4: Distribution of these syndromes in the dataset.

4.2 Observations

ADHD and Down Syndrome: The most prevalent syndromes in the dataset are ADHD and down syndrome, each with 11 cases. This indicates that these two conditions are commonly observed and might require more focused attention in early detection and intervention strategies.

Autism: There are 4 cases of Autism in the dataset. While not as prevalent as ADHD and down syndrome, the presence of Autism still represents a significant portion of the dataset, highlighting the need for detection tools that can accurately identify Autism alongside other neurodevelopmental disorders.

Apraxia of Speech: The dataset includes 3 cases of Apraxia of Speech. This relatively lower number might suggest that Apraxia of Speech is less common compared to the other syndromes studied. However, it is essential to have reliable diagnostic methods for identifying even the less prevalent conditions to ensure that all affected children receive appropriate support and intervention.

The distribution of these syndromes in the dataset underscores the diversity of neurodevelopmental disorders and the importance of developing robust, versatile diagnostic tools. The use of LLMs in analyzing video data provides a promising approach to understanding and identifying these conditions with greater accuracy and efficiency.

5. Discussions

5.1 Interpretation of results and their implications for ADHD early detection

When LLMs are used to analyze video data, the results show promise in detecting behavioral indicators of attention deficit hyperactivity disorder. Through the analysis of text descriptions produced from videos, such as verbal replies that have been transcribed or interactions that have been watched, LLMs are able to identify patterns that point to characteristics associated with ADHD, such as impulsivity, hyperactivity, and inattention.[24] These results imply that by offering unbiased, data-driven insights, LLMs can supplement conventional diagnostic techniques. LLMs' capacity to examine subtle characteristics, like speech patterns, nonverbal clues, or movement inclinations, provides an early identification method that may identify kids for more clinical assessment. By reducing diagnostic delays, this strategy may facilitate earlier interventions, which are essential for enhancing long-term results.[25] Additionally, LLMs can provide a more comprehensive view of a child's behavioral profile by combining multimodal data sources (such as textual observations, video, and audio), which may improve diagnosis precision and assist clinicians in prioritizing cases.

5.2 Advantages and limitations of using LLMs for video data analysis

5.2.1 Advantages

Scalability: LLMs can process large volumes of video data efficiently, allowing widespread screenings in schools, clinics, or other settings. This scalability is critical for addressing resource constraints in traditional ADHD diagnostic processes. Objectivity: By reducing reliance on subjective human observation, LLMs minimize biases and ensure consistent analysis. This uniformity is essential for standardizing ADHD detection across diverse populations.

Multimodal Capabilities: Advanced LLMs can integrate textual transcripts, audio cues, and descriptive annotations, offering comprehensive insights into behaviors that may be missed by human evaluators.

Accessibility: The use of video data enables remote assessments, which is particularly advantageous for children in underserved or geographically isolated areas where in-person evaluations may not be feasible.

5.2.2 Limitations

Data Quality Dependency: The performance of LLMs heavily depends on the quality of input data. Poor video resolution, incomplete transcriptions, or missing contextual information can lead to inaccurate predictions.

Interpretability: The "black-box" nature of LLMs poses challenges in explaining how specific behavioral markers are identified, potentially reducing trust among clinicians and caregivers.

Ethical Concerns: Privacy issues arise when analyzing sensitive video data, particularly for children. Ensuring compliance with data protection laws and obtaining informed consent are critical.

Contextual Nuances: LLMs may struggle to differentiate ADHD behaviors from those caused by other conditions, such as anxiety or sleep disorders, without additional contextual information.

Resource Intensive: Running LLMs on video data can require significant computational resources, which may be a barrier for widespread deployment.

5.3 Potential for real-world applications in clinical settings

ADHD diagnosis in actual clinical settings could be significantly improved by integrating LLMs for video data processing. These systems can be used in conjunction with conventional assessments as pre-screening techniques. For LLM analysis, for instance, caregivers could send in video recordings of kids performing structured activities like storytelling or playing.[26] The findings might direct medical professionals toward areas of emphasis for upcoming assessments, increasing the effectiveness of diagnostic procedures. Furthermore, by offering real-time insights during virtual consultations, LLMs could assist with telemedicine services. This device could serve as a diagnostic tool for medical professionals operating in underdeveloped areas, filling in knowledge or experience shortages. Healthcare professionals could identify children who are at risk early, track their progress over time, and evaluate the effectiveness of therapies by integrating LLMs into clinical procedures.

Furthermore, LLMs' scalability and automation make them ideal for extensive screening campaigns in community programs or educational institutions.[27] Proactive measures like behavioral therapy or educational assistance may be made possible by early identification, which would lessen the long-term effects of ADHD on kids and their families. To guarantee dependability and inclusion, however, strong ethical protections, clinical training, and continual assessment will be necessary for successful deployment. By addressing these considerations, LLMs for video data analysis hold promise as an innovative tool for enhancing ADHD detection and management in diverse real-world scenarios.

6. Future work

6.1 Suggestions for improving model accuracy and scalability

To enhance the accuracy of ADHD detection models, it is essential to refine preprocessing techniques for video data, ensuring high-quality feature extraction while minimizing noise. The ability of LLMs to recognize subtle behavioral signs can be improved by applying sophisticated fine-tuning strategies, such as domain-specific pretraining with datasets relating to ADHD or multi-task learning.[28] Additionally, the identification of intricate patterns in video data may be enhanced by hybrid models that combine LLMs with specialized computer vision frameworks like convolutional neural networks (CNNs) or transformers like vision

transformers (ViTs). By using cloud-based platforms for model deployment and distributing training across several GPUs, scalability can be attained. Using effective model designs, like quantized or distilled LLMs, can help lower computational overhead and increase the viability of real-time processing.

6.2 Exploration of different data sources or modalities for ADHD detection

Accurately detecting ADHD can be greatly improved by integrating a variety of data sources. In addition to video analysis, other modalities could be used, including textual inputs from clinical observations or surveys, wearable sensor data (*e.g.*, movement tracking, heart rate variability), and audio (*e.g.*, speech patterns, tone analysis). ADHD evaluation that is comprehensive and context-aware can be made possible by multimodal learning strategies that integrate these data streams.[29] For example, sensor data can offer continuous movement patterns suggestive of hyperactivity, while audio features may capture impulsivity or hyperactivity in speech. The generalizability of the model would be further enhanced and potential biases would be decreased by enlarging datasets to cover age groups and culturally varied populations.

6.3 Prospects of integrating this system into adaptive educational or behavioral programs

The application of this system in adaptive educational environments could revolutionize support for children with ADHD. Personalized treatments, such reminders to refocus or modifications to teaching strategies, may be made possible by the real-time detection of concentration problems or hyperactivity in classrooms via video feeds. These models could be included into behavioral programs to monitor development, assess the success of interventions, and customize treatment plans. Learning might become more interesting and individualized for each child with the help of gamified adaptive tools that are based on behavioral insights specific to ADHD from video data.[30] For these systems to be widely adopted, cooperation with educators and healthcare professionals to develop safe, privacy-compliant deployment frameworks will be essential. In the long run, these resources might be used as a starting point for early intervention techniques, which could lessen the social and academic difficulties that kids with ADHD encounter.

7. Conclusion

This study introduces a pioneering framework that leverages LLMs for early ADHD detection in children through video data analysis. By incorporating advanced preprocessing techniques such as frame extraction, image processing, and behavioral feature identification, the system successfully captures nuanced indicators of attention patterns and movement behaviors. The integration of LLMs contextualizes these features within established diagnostic frameworks, significantly improving the precision, reliability, and scalability of ADHD risk assessment. The ADHD detection model developed in this research classifies children into high- and low-risk categories, offering clinicians a streamlined, objective, and actionable diagnostic tool. Furthermore, the integration of a clinician-friendly dashboard ensures that insights are accessible and immediately applicable in practical healthcare settings. This novel approach addresses key limitations of traditional observation methods, including subjectivity, time consumption, and limited scalability, thereby marking a paradigm shift toward data-driven early diagnosis. The broader implications of this research extend beyond ADHD detection. This framework sets a foundation for applying LLM-based methodologies to a wider range of behavioral and developmental disorders. It underscores the transformative potential of artificial intelligence (AI) in healthcare, especially in pediatric diagnostics, where early and accurate intervention can lead to significantly improved long-term outcomes. Future research could focus on expanding the diversity and size of datasets, refining the detection models for enhanced generalizability, and exploring multimodal data inputs, such as speech and physiological signals, to create a holistic diagnostic system. Additionally, ethical considerations, such as data privacy and informed consent, should remain central to future developments to ensure responsible deployment of AI in sensitive healthcare applications. By advancing the intersection of AI and behavioral healthcare, this research paves the way for scalable, accurate, and equitable solutions in the diagnosis and treatment of ADHD and related conditions.

Acknowledgements

I sincerely acknowledge the support and guidance of Dr Jayashree Prasad in the completion of this research. I also extend my gratitude to MIT School of Computing, MIT Art Design and Technology University, Loni Kalbhor, Pune, India for providing resources and to my colleagues for their valuable insights and encouragement.

Conflict of Interest

There is no conflict of interest.

Supporting Information

Not applicable.

References

[1] M. Maaz, H. Rasheed, S. Khan, F. S. Khan, Video-ChatGPT: towards detailed video understanding *via* large vision and language models, arxiv preprint arxiv: 2306.05424, 2023, doi: 10.48550/arXiv.2306.05424.

[2] S. Alam, P. Raja, Y. Gulzar, Investigation of machine learning methods for early prediction of neurodevelopmental disorders in children, *Wireless Communications and Mobile Computing*, 2022, **2022**, 5766386, doi: 10.1155/2022/5766386.

[3] H. Lin, A. Zala, J. Cho, M. Bansal, VideoDirectorGPT:

consistent multi-scene video generation *via* LLM-guided planning, arxiv preprint arxiv: 2309.15091, 2023, doi: 10.48550/arXiv.2309.15091.
[4] H. W. Loh, C. P. Ooi, P. D. Barua, E. E. Palmer, F. Molinari, U. R. Acharya, Automated detection of ADHD: Current trends and future perspective, *Computers in Biology and Medicine*, 2022, **146**, 105525, doi: 10.1016/j.compbiomed.2022.105525.
[5] L. Ter-Minassian, N. Viani, A. Wickersham, L. Cross, R. Stewart, S. Velupillai, J. Downs, Assessing machine learning for fair prediction of ADHD in school pupils using a retrospective cohort study of linked education and healthcare data, *BMJ Open*, 2022, **12**, e058058, doi: 10.1136/bmjopen-2021-058058.
[6] D. Yu, J. H. Fang, Using artificial intelligence methods to study the effectiveness of exercise in patients with ADHD, *Frontiers in Neuroscience*, 2024, **18**, 1380886, doi: 10.3389/fnins.2024.1380886.
[7] S. Oh, Y.-S. Joung, T.-M. Chung, J. Lee, B. J. Seok, N. Kim, H. M. Son, Diagnosis of ADHD using virtual reality and artificial intelligence: an exploratory study of clinical applications, *Frontiers in Psychiatry*, 2024, **15**, 1383547, doi: 10.3389/fpsyt.2024.1383547.
[8] N. Alsharif, M. H. Al-Adhaileh, S. N. Alsubari, M. Al-Yaari, ADHD diagnosis using text features and predictive machine learning and deep learning algorithms, *Journal of Disability Research*, 2024, **3**, 20240082, doi: 10.57197/jdr-2024-0082.
[9] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, Y. Qiao, VideoChat: chat-centric video understanding, arxiv preprint arxiv: 2305.06355, 2023, doi: 10.48550/arXiv.2305.06355.
[10] S. Li, R. Nair, S. M. Naqvi, Acoustic and text features analysis for adult ADHD screening: a data-driven approach utilizing DIVA interview, *IEEE Journal of Translational Engineering in Health and Medicine*, 2024, **12**, 359-370, doi: 10.1109/JTEHM.2024.3369764.
[11] T. J. Layton, M. L. Barnett, T. R. Hicks, A. B. Jena, Attention deficit-hyperactivity disorder and month of school enrollment, *New England Journal of Medicine*, 2018, **379**, 2122-2130, doi: 10.1056/NEJMoa1806828.
[12] L. Kazda, K. McGeechan, K. Bell, R. Thomas, A. Barratt, Association of attention-deficit/hyperactivity disorder diagnosis with adolescent quality of life, *JAMA Network Open*, 2022, **5**, e2236364, doi: 10.1001/jamanetworkopen.2022.36364.
[13] W. P. Kim, H. J. Kim, S. P. Pack, J. H. Lim, C. H. Cho, H. J. Lee, Machine learning-based prediction of attention-deficit/hyperactivity disorder and sleep problems with wearable data in children, *JAMA Network Open*, 2023, **6**, e233502, doi: 10.1001/jamanetworkopen.2023.3502.
[14] J. AlShaqsi, M. Borghan, O. Drogham, G. H. Roshani, S. Al Whahaibi, Predicting pandemic fatality based on supervised machine learning methods, *Engineered Science*, 2024, **30**, 1169, doi: 10.30919/es1169.
[15] M. Sudhi, V. K. Shukla, D. K. Shetty, V. Gupta, A. S. Desai, N. Naik, B. Z. Hameed, Advancements in bladder cancer management: a comprehensive review of artificial intelligence and machine learning applications, *Engineered Science*, 2023, **26**, 1003, doi: 10.30919/es1003.
[16] R. D. Jathanna, D. Acharya, L. E. Lewis, K. Makkithaya, Early detection of late onset neonatal sepsis using machine learning algorithms, *Engineered Science*, 2023, **26**, 976, doi: 10.30919/es976.
[17] V. Kiruthik, S. Sathiya, R. M. M, K. Sakthidasan Sankaran, An intelligent machine learning approach for ovarian detection and classification system using ultrasonogram images, *Engineered Science*, 2023, **23**, 879, doi: 10.30919/es8d879.
[18] P. Zhou, L. Wang, Z. Liu, Y. Hao, P. Hui, S. Tarkoma, J. Kangasharju, A survey on generative AI and LLM for video generation, understanding, and streaming, arxiv preprint arxiv: 2309.17444, 2024, doi: 10.48550/arXiv.2404.16038.
[19] L. Lian, B. Shi, A. Yala, T. Darrell, B. Li, Llm-grounded video diffusion models, arxiv preprint arxiv: 2309.17444, 2023, doi: 10.48550/arXiv.2309.17444.
[20] C. Zhang, T. Lu, M. M. Islam, Z. Wang, S. Yu, M. Bansal, G. Bertasius, A simple LLM framework for long-range video question-answering, arxiv preprint arxiv: 2312.17235, 2023, doi: 10.48550/arXiv.2312.17235.
[21] B. He, H. Li, Y. K. Jang, M. Jia, X. Cao, A. Shah, A. Shrivastava, S.-N. Lim, MA-LMM: memory-augmented large multimodal model for long-term video understanding, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, 13504-13514, doi: 10.1109/CVPR52733.2024.01282.
[22] J. Cao, Y. Wu, W. Chi, W. Zhu, Z. Su, J. Wu, Reframe anything: LLM agent for open world video reframing, arxiv preprint arxiv: 2403.06070, 2024, doi: 10.48550/arXiv.2403.06070.
[23] I. de Zarzà, J. de Curtò, C. T. Calafate, Socratic video understanding on unmanned aerial vehicles, *Procedia Computer Science*, 2023, **225**, 144-154, doi: 10.1016/j.procs.2023.09.101.
[24] Y. S. Liu, F. Talarico, D. Metes, Y. Song, M. Wang, L. Kiyang, D. Wearmouth, S. Vik, Y. Wei, Y. Zhang, J. Hayward, G. Ahmed, A. Gaskin, R. Greiner, A. Greenshaw, A. Alexander, M. Janus, B. Cao, Early identification of children with attention-deficit/hyperactivity disorder (ADHD), *PLoS Digital Health*, 2024, **3**, e0000620, doi: 10.1371/journal.pdig.0000620.
[25] L. B. Thorell, J. Burén, J. Ström Wiman, D. Sandberg, S. B. Nutley, Longitudinal associations between digital media use and ADHD symptoms in children and adolescents: a systematic literature review, *European Child & Adolescent Psychiatry*, 2024, **33**, 2503-2526, doi: 10.1007/s00787-022-02130-3.
[26] D. Andrikopoulos, G. Vassiliou, P. Fatouros, C. Tsirmpas, A. Pehlivanidis, C. Papageorgiou, Machine learning-enabled detection of attention-deficit/hyperactivity disorder with multimodal physiological data: a case-control study, *BMC Psychiatry*, 2024, **24**, 547, doi: 10.1186/s12888-024-05987-7.
[27] L. Merzon, K. Pettersson, E. T. Aronen, H. Huhdanpää, E. Seesjärvi, L. Henriksson, W. Joseph MacInnes, M. Mannerkoski, E. Macaluso, J. Salmi, Eye movement behavior in a real-world virtual reality task reveals ADHD in children, *Scientific Reports*, 2022, **12**, 20308, doi: 10.1038/s41598-022-24552-4.
[28] J. Wallace, E. Boers, J. Ouellet, M. H. Afzali, P. Conrod,

Screen time, impulsivity, neuropsychological functions and their relationship to growth in adolescent attention-deficit/hyperactivity disorder symptoms, *Scientific Reports*, 2023, **13**, 18108, doi: 10.1038/s41598-023-44105-7.
[29] W. Lee, S. Lee, D. Lee, K. Jun, D. H. Ahn, M. S. Kim, Deep learning-based ADHD and ADHD-RISK classification technology through the recognition of children's abnormal behaviors during the robot-led ADHD screening game, *Sensors*, 2022, **23**, 278, doi: 10.3390/s23010278.
[30] C. Park, M. D. Rouzi, M. M. U. Atique, M. G. Finco, R. K. Mishra, G. Barba-Villalobos, E. Crossman, C. Amushie, J. Nguyen, C. Calarge, B. Najafi, Machine learning-based aggression detection in children with ADHD using sensor-based physical activity monitoring, *Sensors*, 2023, **23**, 4949, doi: 10.3390/s23104949.

Publisher's Note: Engineered Science Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.