Computer Vision and Machine Learning for Assisted ADHD Screening: A Systematic Review

Eric Pereira[0009−0004−0319−0353], Eduardo Kirchner[0009−0000−2212−7510], Théo Soares[0009−0007−9569−4146], André Vargas[0000−0001−6873−3700], and Rodrigo de Bem[0000−0002−4860−0115]

Center of Computational Sciences, Federal University of Rio Grande (FURG), Rio Grande, Brazil
{eric.pereira,eduardo.kirchner,theocsoares,andre.prisco,rodrigobem}@furg.br

Abstract. Current diagnostic methods for Attention Deficit Hyperactivity Disorder (ADHD) rely heavily on clinical evaluations and behavioral scales, characterizing largely subjective processes. Meanwhile, objective approaches such as EEG and fMRI are invasive, expensive, and not viable for routine use. In this context, computational methods based on computer vision and machine learning emerge as less invasive and potentially scalable alternatives, allowing for the automatic extraction of motor and behavioral patterns from videos. In this context, this work presents a literature review, conducted according to the PRISMA guidelines, on the use of artificial intelligence applied to the assisted screening of ADHD from 2018 to 2026. The results show that approaches based on body movement, eye movement, virtual reality, and multimodal fusion achieve high predictive performance, with a notable emphasis on multimodal models. However, many studies still use small sample sizes and internal validation, limiting clinical generalization and characterizing research gaps for future investigation.

Keywords: Attention Deficit Hyperactivity Disorder (ADHD) · Computer Vision · Machine Learning · Systematic Review

1 Introduction

Traditional methods for diagnosing Attention Deficit Hyperactivity Disorder (ADHD) are predominantly based on clinical interviews, behavioral scales, and neuropsychological tests, all of which heavily depend on interpretation and self-reporting, making these processes highly subjective. Moreover, approaches based on neuroimaging and electrophysiology, such as fMRI (functional magnetic resonance imaging) and EEG (electroencephalography), although objective, are invasive, have high operational costs, and limited applicability on a large clinical scale [7, 22, 30]. In this context, computational methods based on computer vision, machine learning, and more recently, deep learning, are emerging as less invasive, low-cost, and potentially scalable alternatives, exploring motor, postural, and expressive patterns captured by conventional cameras [16, 19, 17, 18].

Evidence from the literature indicates that behavioral manifestations constitute relevant markers of ADHD, supporting the feasibility of screening based on visual data. Studies conducted in real clinical settings demonstrate that body movement patterns automatically extracted from videos show statistically significant differences between individuals with ADHD and controls, particularly with regard to motor variability, movement frequency, and temporal entropy of displacement [4, 26]. Complementarily, action recognition approaches show that behaviors such as fidgeting (repetitive, low-amplitude, high-frequency micromovements of the body), associated with motor restlessness and quantified by temporal variations in joint positions in video [16, 19, 17], rapid postural changes, and limb movements can be quantified with high precision by 3D-CNN and pose-based models, achieving high accuracy and AUC (Area Under the ROC Curve) rates [16, 19, 17, 18]. These results support the hypothesis that core ADHD symptoms have a direct representation in visual movement patterns, which can be captured and computationally analyzed in a non-invasive manner.

Results indicate that methods based on action recognition, pose analysis (computational estimation of human body joints), and multimodal fusion (video, body markers, audio, and questionnaires) achieve accuracies and AUC values frequently exceeding 0.95, especially with dual-stream architectures (with two parallel processing streams) and multimodal models such as AVEN (Attention-based Video and Evaluation Network) [24, 25]. On the other hand, critical bias analysis reveals that many studies use small samples and predominantly internal validation [18, 17, 24], which limits clinical generalizability. Studies with larger samples, such as [20] and [5], show slightly lower performance but greater methodological robustness. Thus, in this review, we aim to answer this research question: *How robust are current experiments on vision-based ADHD screening in terms of generalization and clinical reliability?* In the following sections, we detail the methodology of the systematic review, analyzing the existing approaches with particular attention to sample size, generalization, and clinical feasibility, also shedding light on future research directions in the field.

2 Methodology of the Literature Review

For the identification, screening, and selection of studies included, we adopted the PRISMA 2020 *(Preferred Reporting Items for Systematic Reviews and Meta-Analyses)* methodology [27]. Using PRISMA ensures transparency in the reliability and reproducibility of the findings. Thus, we follow the PRISMA steps comprising: i) Identification of relevant databases; ii) Definition of inclusion and exclusion criteria; iii) Initial screening of titles and abstracts; iv) Full-text reading and final selection of eligible articles.

Thus, we conducted a systematic search across four major databases: Web of Science (WoS), Scopus, IEEE Xplore, and PubMed. Furthermore, the following study inclusion criteria were established for this review: i) Articles published between 2018 and 2026; ii) Studies focused on ADHD diagnosis using computer vision or related techniques; iii) Articles published in peer-reviewed conferences

or journals. The exclusion criteria were defined as follows: i) studies relying on MRI, fMRI, ECG, EEG, or other non-vision-based physiological signals.

These criteria guided the search strategy, with search strings applied to titles, abstracts, and keywords. The search string combined the terms shown in Table 1 using Boolean operators (`AND/OR`) to retrieve literature focused on non-invasive, vision-based ADHD screening. In total, 27 relevant studies were selected from the initial set of 290 records after applying the PRISMA screening process. One study predating the 2018 temporal window, Jones and Klin [12] (2013), was retained as an intentional exception due to its foundational role for the gaze-based approaches reviewed.

Table 1: Systematic search string components.

Category	**Keywords**
Condition	ADHD, "Attention Deficit Hyperactivity Disorder"
Computational Methods	"Machine Learning", "Deep Learning", "Artificial Intelligence", "AI", "Neural Network", "Computer Vision", "Pattern Recognition"
Modalities	"Image", "Video", "Motion Analysis", "Body Movement", "Eye Tracking", "Oculomotor", "Gaze", "Skeleton-based", "Pose Estimation"
Clinical Outcomes	"QbTest", "CPT", "Objective Measurement", "Diagnosis", "Detection", "Assessment", "Screening"
Exclusions (`NOT`)	"Magnetic Resonance", "MRI", "fMRI", "Electrocardiogram", "ECG", "Electroencephalogram", "EEG"

3 Methodological Analysis of the Studies

The reviewed studies show a consistent trade-off between predictive performance and clinical reliability. The highest reported metrics usually come from multimodal or dual-stream systems that combine RGB video, pose, audio, questionnaires, or physiological traces, such as AVEN and autoencoder-based models [24, 25], or video–pose action recognition pipelines [17, 18]. These methods approach or exceed 95–99% accuracy, but most rely on very small, highly controlled samples. In contrast, studies with larger and better-matched cohorts, such as eye-tracking biomarkers [20] and retinal imaging [5], report slightly lower accuracy. Overall, the evidence suggests that progress in this field should not be judged by accuracy alone: sample size, validation strategy, bias risk, sensor cost, and feasibility in real clinical environments are equally decisive.

The studies can be grouped into three main methodological families: body-movement analysis, eye-movement or ocular biomarkers, and multimodal systems. Body-movement approaches are attractive because hyperactivity and motor restlessness are core ADHD manifestations and can be measured from standard videos. Eye-based approaches capture attentional control, inhibition, and

visual exploration, but often depend on specialized hardware. Multimodal approaches are the current frontier because they fuse complementary behavioral, physiological, and cognitive signals; however, they are also the most vulnerable to overfitting when trained on small datasets.

3.1 Approaches Based on Body Movement

Body-movement studies form a well-established line in computer vision-assisted ADHD detection. Most works extract pose, action classes, motion entropy, or angular variability from videos, then classify participants using either classical machine learning or deep spatiotemporal models. Across this group, reported accuracies range from approximately 88% to 96%, with stronger results usually obtained by pose-aware 3D architectures and engineered behavioral indices such as *Hyperactivity Score* (HS), *Attention Deficit Ratio* (ADR/RAD), or *Stationary Ratio* (SR). The main weakness is recurrent: several high-performing studies use fewer than 25 participants.

Li et al. [16] proposed a video-based ADHD diagnostic system using pose estimation from three 4K GoPro cameras and PoseC3D applied to joint heatmaps. The dataset included 17 adults, 7 with ADHD and 10 controls, recorded during structured attention and responsiveness tasks. The method achieved 88.2% accuracy and AUC = 0.83, outperforming C3D, R3D, ST-GCN, and MS-G3D, whose accuracies ranged from 58.8% to 76.4%. The study also introduced HS and ADR, with the final decision based on an empirically defined ADR threshold of 76.5%. Its main limitation is the very small sample and the controlled multi-camera setting, which restricts generalization.

Tang et al. [28] introduced MF-Net for recognizing ADHD-related behaviors during the TOVA assessment. The dataset contained 3,801 video samples of 60 frames each, collected from seven diagnosed children. The model fused joint movement, bone movement, and facial micro-expression streams, combining MSF-AGCN for body cues with MobileViTv2 for facial keypoint pseudo-images. MF-Net achieved 90.6% accuracy, outperforming pose-based and video-based baselines. The result is promising, but the dataset is extremely narrow: only seven children, imbalanced classes, and behaviors tied specifically to TOVA.

Chiu et al. [4] used pixel subtraction from outpatient consultation videos to quantify motor activity without deep models. The study analyzed 85 children, 43 with ADHD and 42 controls, using four-minute video segments at 30 Hz. Three motion descriptors were extracted from the region of interest: mean motion, variance, and Shannon entropy. Among six classifiers evaluated with repeated nested cross-validation, the best configuration reached 90.24% accuracy, 88.85% sensitivity, 91.75% specificity, and AUC = 0.9387. The ADHD group showed markedly higher entropy than controls, but the sample was dominated by combined-type ADHD, and motor behavior may have been affected by medication, fatigue, or clinical context.

Rajendran et al. [21] combined two learning tracks for early ADHD detection: classical machine learning on a clinical dataset of 486 records and 29 variables, and a deep neural network with attention mechanisms for video frames. In the

structured-data track, Random Forest achieved 91.09% accuracy after preprocessing, missing-value treatment, PCA, and feature selection. The video branch used a 105-layer DNN-AM with Local Binary Pattern enrichment for facial and behavioral patterns. Although the reported performance is competitive, the work still lacks broad external validation and clearer generalization tests across populations.

Ouyang et al. [26] proposed an outpatient-video pipeline based on OpenPose, extracting 25 body joints and computing the temporal variance of 11 skeletal parameters. The dataset included 96 children, 48 with ADHD and 48 age- and sex-matched controls. Using the thigh-angle descriptor alone, the model obtained 91.03% accuracy, 90.25% sensitivity, 91.86% specificity, and AUC = 94.00%; combining multiple parameters increased performance to 92.10% accuracy and 95.22% AUC. The study is stronger than many video-based works because of its matched control group, but subtype imbalance and day-to-day variability in movement remain important limitations.

Li et al. [19] presented an action-based video method that classifies three movement classes: still position, limb fidgets, and torso movements. The method uses C3D and summarizes behavior through the *Stationary Ratio* (SR), with a decision threshold of 0.71. On a small dataset including only seven ADHD cases, the method reached 100% sensitivity, 90.9% precision, 94.1% accuracy, and AUC = 0.97, outperforming R3D and R2Plus1D. The performance is high, but the evidence remains fragile because of the small sample, one inconsistent case, long recordings, and dependence on a controlled multi-camera protocol.

Li et al. [17] extended action recognition by combining RGB videos and pose-based representations through a two-stream architecture using C3D and PoseConv3D. The M-ADHD dataset included 22 adults, 10 with ADHD and 12 controls, recorded with three 4K cameras during cognitive tasks, interviews, distractor videos, and auditory attention tests. The full C3D + PoseConv3D system achieved 95.5% accuracy and AUC = 99.2%, with pose information contributing strongly to performance. Nevertheless, the number of participants is very small, performance drops on shorter videos and among women, and the fixed-camera setup limits deployment outside controlled environments.

Isaev et al. [10] analyzed caregiver–child interaction during six minutes of free play using two RGB cameras. The sample contained 78 children across autism, ADHD, autism+ADHD, and typical-development groups. DensePose, OpenPose, and 3DMPPE were used to estimate posture and identify reaching events, which were modeled with Markov and mixture hidden Markov models. The study did not report diagnostic accuracy or AUC, but the automated reaching-event detector achieved Cohen's kappa of 0.85 for children and 0.86 for caregivers. MixMM identified two interaction clusters, including one with a 130% increase in child-to-caregiver reaching transitions. The main weaknesses are the small neurotypical group, limited demographic diversity, and noise from the approximate 3D reconstruction.

Li et al. [18] proposed an RGB-video and pose-sequence system using 90-minute recordings from 10 ADHD adults and 12 controls. C3D processed RGB

clips, while PoseConv3D processed joint heatmaps extracted through HRNet and Faster-RCNN; the method then fused RAD and SR into a *Fusion Ratio* (RF). The model achieved 95.5% accuracy and AUC = 0.99, outperforming conventional action-recognition networks and several prior ADHD-detection methods. As in related work, the result is constrained by the tiny sample and controlled camera arrangement.

The study in [15] also evaluated video-based action characteristics for ADHD diagnosis using three cameras, DSM-5-informed tasks, and deep action-recognition architectures. Among R3D, C3D, MS-G3D, ST-GCN, and PoseC3D, the pose-based model performed best, reaching 88.2% accuracy, 100% precision, 88.9% F1-score, and AUC = 0.83; competing models reached roughly 70–76.4% accuracy and AUCs between 0.70 and 0.72. The dataset had only 7 ADHD participants and 10 controls, so the paper is best interpreted as a proof of concept rather than a clinically validated system.

Amado et al. [1] used wrist actimetry, CNNs, SVM classification, and occlusion maps to study ADHD-related activity patterns in 139 children aged 6–15 years. Accelerometer data were collected for 24 hours, transformed into spectrograms, and classified window-by-window before SVM aggregation. The previous CNN+SVM ensemble reported 98.6% accuracy, 97.62% sensitivity, and 99.52% specificity, while the present work emphasized interpretability through occlusion maps and Gaussian Mixture Models. The analysis suggests a stronger discriminative signal in frequency-related features and in boys' activity patterns, but female and adolescent subgroups were small.

The QbTest is not a pure computer-vision system, but it is relevant because it combines a computerized CPT with infrared motion tracking to quantify inattention, impulsivity, and hyperactivity [9]. Its ten primary parameters are aggregated into QbInattention, QbActivity, and QbImpulsivity scores, normalized by age and sex. Reported discrimination is moderate, with AUC values generally between 0.70 and 0.80.

3.2 Eye Movement-Based Approaches

Eye-movement studies focus on oculomotor control, visual attention, inhibition, and exploratory behavior. Compared with body-movement approaches, they often offer cleaner physiological interpretation, but several depend on expensive eye trackers or controlled tasks. The strongest studies in this group report AUC values above 0.95, while lower-cost tablet or webcam-based alternatives are more deployable but usually less accurate.

Varela et al. [29] validated ocular vergence as an ADHD marker using a two-layer model. The sample included 43 children with ADHD, 19 clinical controls, and 30 neurotypical controls, plus an additional validation population of 67 children. A 30 Hz binocular eye tracker recorded vergence, behavioral, and pupillary responses during cued and non-cued visual tasks. The SVM layer classified ADHD versus neurotypical controls with 96.3% accuracy, AUC = 0.99, 5.12% false positives, and 0% false negatives; the ADHD versus clinical-control layer

reached 85.7% accuracy and AUC = 0.90. The main concern is the false-negative rate in the second layer and the need for independent replication.

Yoo et al. [32] proposed portable eye-tracking screening using a Samsung Galaxy Tab 7+ and the SeeSo SDK. The sample contained 135 children, 56 with ADHD and 79 typically developing, who completed five tasks: pro-saccade, anti-saccade, memory-guided saccade, change detection, and Stroop. From 100 initial features, recursive feature elimination reduced the set to 33, and a RandomForest/ExtraTrees soft-voting ensemble achieved 76.3% accuracy, AUC = 0.724, recall = 0.725, and precision = 0.789. The approach is clinically practical, but performance is modest, and tracking failures reduced usable clinical data to n = 73.

Devkota et al. [6] converted eye-movement time series into images using MTF, GASF, GADF, and recurrence plots, then trained a CNN to compare representation quality. The best accuracies were 92.3% with GASF, 91.2% with MTF, 89.9% with GADF, and 89.7% with RP, depending on the eye-movement variable used. The study is useful because it isolates the effect of time-series imaging rather than architecture complexity. However, it does not report complete AUC values and remains limited by dataset size, interindividual variation, and single-center data collection.

Jayawardena et al. [11] used eye tracking during a Reading Span working-memory task. Fourteen participants were included, 7 with ADHD and 7 controls, and features were extracted from manually defined areas of interest, including fixation duration, saccade amplitude and velocity, and pupil diameter. Classical WEKA classifiers were evaluated, with Random Forest performing best: 86.20% accuracy using sentence-based features and 83.33% with scene-based features. The study supports the relevance of oculomotor measures during cognitive tasks, but its sample is too small, and no AUC is reported.

Liu et al. [20] offered one of the more convincing eye-tracking studies because it used 216 children aged 5–10 years, including 94 with ADHD and 122 typically developing controls, matched by age and IQ. A Tobii Pro Spectrum recorded pro-saccade, anti-saccade, and delayed-saccade tasks at 1200 Hz, generating 28 digital biomarkers and 183 derived variables. XGBoost, trained with repeated 5-fold stratified cross-validation, achieved 0.908 accuracy, 0.877 sensitivity, 0.932 specificity, 0.892 F1-score, and AUC = 0.965. The main limitations are single-city recruitment, gender imbalance, and reliance on high-precision non-portable equipment.

Khan et al. [14] presented EXECUTE, a webcam-based eye-tracking approach for monitoring attention in e-learning. The dataset included 25 participants and 2D gaze coordinates captured through standard webcams. Logistic Regression, SVM, and Polynomial Regression were compared, and the reported accuracy exceeded 89%, showing that low-cost gaze estimation can support real-time attention monitoring. The limitation is clear: the setting targets learning attention rather than clinical ADHD diagnosis, and the dataset is small.

Jones and Klin [12] followed 110 infants from 2 to 24 months and showed that attention to the eye region declines over time in children later diagnosed with

autism. Although the study is not an ADHD-detection method, it is relevant to the broader literature on early neurodevelopmental biomarkers because it demonstrates that longitudinal gaze trajectories can reveal clinically meaningful developmental divergence. Its direct applicability to ADHD is limited.

Choi et al. [5] used retinal fundus imaging as a non-invasive ADHD biomarker in a comparatively large sample of 646 children. Four supervised models were evaluated: Random Forest, XGBoost, Extra Trees, and Logistic Regression, with XGBoost achieving the best screening performance. The work is important because it shifts from behavioral observation to ocular imaging and includes age- and sex-balanced data, increasing methodological credibility. However, data came from two hospitals in South Korea, and the use of two-dimensional retinal images may not capture all relevant ocular or neurodevelopmental variability.

3.3 Multimodal Approaches

Multimodal methods currently define the most ambitious direction in ADHD detection. They combine video, pose, speech, inertial sensors, questionnaires, EEG, actigraphy, retinal data, and/or virtual-reality behavior. Their central advantage is complementary evidence: one modality may capture motor restlessness, another attentional control, and another subjective symptoms. Their central risk is statistical: when many modalities are fused in small samples, excellent metrics may reflect overfitting rather than genuine clinical generalization.

Nash et al. [23] proposed an attention-guided cross-modal model combining RGB video, visual prompts, and ASRS questionnaire responses from the ISAT dataset. The sample included 22 adults, 10 with ADHD and 12 controls, recorded in 4K and preprocessed with Mask-RCNN-based participant segmentation. The method reached 98.2% accuracy and an AUC close to 0.99, indicating strong fusion of behavioral and self-report cues. The result is technically strong but methodologically fragile because of the small sample, synthetic text component, and computational cost of Transformer-based video modeling.

Kaur et al. [13] used the public HYPERAKTIV dataset to classify 103 adults, 51 with ADHD and 52 clinical controls, from real-time motor activity data. The pipeline extracted 788 features, reduced dimensionality with PCA while retaining approximately 95% of variability, and compared six WEKA classifiers. SVM with an RBF kernel achieved 98.43% accuracy, 98.33% sensitivity, 98.56% specificity, F-measure = 98.42%, and AUC = 0.983; Random Forest achieved the highest AUC (0.999) but slightly lower global performance. Limitations are the moderate sample, lack of subtype analysis, and incomplete use of available ECG data.

Bouchouras et al. [3] applied temporal anomaly detection to HYPERAKTIV actigraphy. Isolation Forest first generated anomaly labels, then a SimpleRNN learned 10-step temporal dependencies in normalized activity sequences. The model reached 0.953 average accuracy and 0.124 average loss, suggesting that recurrent architectures can capture atypical activity dynamics. However, the study used only ADHD participants and did not include controls, so the result is more relevant to anomaly characterization than diagnostic discrimination.

Nash et al. [24] introduced AVEN, a dual-stream ensemble combining RGB video with facial, body, and hand landmarks extracted from the ISAT dataset. With only 22 participants, the model processed 5-second clips through a lightweight *ShuffleNet* 3D-CNN and an LSTM for landmarks, then fused predictions with SVM, Random Forest, or Decision Tree ensembles. The best AVEN configuration reached 98.67% accuracy, 98.01% precision, and 98.88% recall; AUC values ranged from 0.68 to 0.93 for ADHD and 0.65 to 0.98 for controls. The main weaknesses are the small sample, predominance of severe adult cases, controlled recordings, and reduced sensitivity to subtle symptoms.

Huang et al. [8] proposed a graph-based multimodal fusion model for ADHD subtype classification using speech and movement. The dataset included 106 children, 56 with ADHD and 50 controls, with 66 movement variables from six IMUs and 28 acoustic variables. A GRU embedding layer and graph attention network performed multilevel fusion. The model achieved 96.2% accuracy for binary ADHD versus control classification and, for subtype classification, Weighted F1 = 0.692 and Micro-F1 = 0.743, outperforming CNN, LSTM, GBDT, SVM, and TextCNN baselines. The main limitation is subtype imbalance, especially the small hyperactive/impulsive group (n = 8), and no AUC is reported.

Nash et al. [25] explored multimodal autoencoders for adult ADHD detection, integrating RGB video, audio, and ASRS responses. The ISAT dataset again included 22 adults, 10 with ADHD and 12 controls. A multivariate autoencoder combined 3D-CNN video encoders, 2D-CNN audio encoders, and MLP questionnaire encoders, while an ICAE compressed all modalities into 144×144 images with only 98,760 parameters. The multivariate model reached 98.9% accuracy, 99.2% sensitivity, and 98.5% specificity; the lighter ICAE still reached 96% accuracy with an MLP classifier. These numbers are impressive, but the sample is too small to establish clinical generalization.

Zhang et al. [33] built an auxiliary diagnostic system combining psychometric tests, eye behavior, facial expression, and body posture from three synchronized cameras. The dataset included 82 children, 71 with ADHD and 11 controls, segmented into 9,116 abnormal and 27,666 normal video fragments. The model combined CNNs, VGGNet, LSTMs, PAF-based 3D pose estimation, and a multimodal BERT classifier. The best multimodal model reached 80.25% accuracy, 0.9357 sensitivity, and 0.6258 specificity. The high sensitivity is useful for screening, but the low specificity, single-hospital dataset, and severe class imbalance limit clinical reliability.

Wiebe et al. [30] evaluated adult ADHD prediction in a virtual seminar room using eye tracking, EEG, actigraphy, CPT performance, and experience sampling. The study used independent training and test samples: n = 50 for training and n = 36 for testing, balanced between ADHD and controls. A linear SVM trained on 76 features reached 0.69 accuracy on the test set, while selecting the 11 most relevant features with MRMR improved performance to 0.81 accuracy, 0.78 sensitivity, and 0.83 specificity. The independent test strengthens the evidence, but the sample remains too small for subtype or demographic analyses, and no AUC is reported.

Yeh et al. [31] developed a VR classroom with head-movement tracking for children performing visual CPT and auditory sustained-attention tasks. The dataset included 68 children, 37 with ADHD and 31 controls, aged 6–12 years. SVM and XGBoost were trained using task performance, neurobehavioral head-movement indicators, and clinical scales. XGBoost reached 83.2% accuracy when combining performance and neurobehavioral data, approaching the 91.5% accuracy obtained from clinical scales alone.

4 Comparative Analysis

Table 2 summarizes the main quantitative results and bias risk. The best numerical performance appears in multimodal models and structured body-representation systems, especially [24, 25, 17, 18], which report accuracies above 95% and, in some cases, AUC values near 0.99. Simpler approaches are not obsolete: pixel subtraction [4], OpenPose-based skeletal variance [26], and eye-movement imaging [6] reach roughly 90–92% accuracy with less complex pipelines. Several of the most impressive studies have high bias risk because they use very small samples, limited external validation, or highly controlled recording setups. In contrast, [20] and [5] have lower accuracies but are stronger methodologically due to larger cohorts and better matching.

5 Discussion

5.1 Synthesis of Evidence and Strengths

The most consistent finding across the 27 selected studies is that ADHD leaves measurable traces in behavior, especially in movement, gaze, posture, and multimodal interaction patterns. However, the strength of this evidence is uneven. The best reported numbers are impressive, with several video-, pose-, and multimodal systems reaching accuracies above 94% and AUC values close to 0.99 [19, 17, 18, 24, 25]. These results confirm that computational models can capture behavioral regularities associated with ADHD. At the same time, they should not be read as evidence that clinical diagnosis has already been solved: many of the highest-performing models were trained and evaluated on small, controlled, and demographically narrow samples. In practical terms, the field has become very good at detecting ADHD-like signals under experimental conditions, but it has not yet demonstrated the same reliability in broad clinical deployment.

Computer vision is the most visible axis of this progress. Pose-based and action-recognition methods such as PoseC3D, MF-Net, AGCN, C3D, and 3D-CNN variants repeatedly identify patterns compatible with motor restlessness, including fidgeting, rapid postural changes, limb fluctuations, and reduced stationarity [16, 28, 19, 17, 18]. The attraction of these methods is clear: they are non-invasive, relatively inexpensive, and can operate on RGB videos or body-keypoint representations. The strongest studies in this group report approximately 88–96% accuracy, with [4] reaching 90.24% accuracy and AUC = 0.9387

Table 2: Comparative performance and bias risk level regarding sample size, control, and validation of state-of-the-art methods.

#	Study	Year	Sample	Accuracy	Methodological Approach	Risk Level[1]
1	[25]	2025	22	**98.9%**	Multimodal autoencoders	High
2	[5]	2025	**646**	–	Retinal imaging + ML	**Low-Medium**
3	[19]	2023	17	94.1%	Video + 3D-CNN	High
4	[17]	2024	22	95.5%	Video + Pose + Deep Learning	High
5	[18]	2024	22	95.5%	Video + Pose	High
6	[28]	2024	7	90.6%	Multimodal MF-Net + TOVA	High
7	[4]	2024	85	90.2%	Clinical video + Classical ML	Medium
8	[21]	2024	486	91.1%	Clinical data + Deep Learning	Medium
9	[26]	2024	96	92.1%	OpenPose + Supervised ML	Medium
10	[32]	2024	135	76.3%	Portable eye tracking	Medium
11	[6]	2024	80	92.3%	Ocular time series + CNN	Medium
12	[20]	2024	216	90.8%	Eye tracking + XGBoost	**Low-Medium**
13	[23]	2024	22	98.2%	ViViT + ASRS	High
14	[24]	2024	22	98.7%	Multimodal AVEN	High
15	[3]	2024	51	95.3%	Actigraphy + RNN	High
16	[30]	2024	86	81.0%	VR + Multimodal	Medium
17	[10]	2024	78	–	Caregiver–child interaction + HMM	Medium
18	[16]	2023	17	88.2%	Movement + PoseC3D	High
19	[1]	2023	139	98.6%	Actimetry + CNN	Medium
20	[14]	2022	25	92.6%	Webcam eye tracking	High
21	[13]	2022	103	98.4%	Actigraphy + SVM	Medium
22	[8]	2022	106	96.2%	Speech + IMUs + GNN	Medium
23	[33]	2021	82	80.3%	Eyes + Face + Body	Medium
24	[11]	2019	14	86.2%	Eye tracking + AOIs	High
25	[31]	2020	68	83.2%	VR + CPT	Medium
26	[29]	2018	92	96.3%	Ocular vergence	**Low-Medium**
27	[12]	2013	36	–	Infant eye tracking	Medium

[1]**Risk Level** (risk of bias) — **High** (very small sample, absent control, or limited/internal validation), **Medium** (moderate sample, adequate controls, cross-validation, no external replication), **Low–Medium** (large, matched cohorts and/or independent external test).

from clinical videos, and [26] reaching 92.10% accuracy and AUC = 95.22% from OpenPose-derived skeletal descriptors. These are not marginal gains; they suggest that observable motor behavior contains clinically relevant information. Still, the interpretation must remain cautious because medication status, task design, camera placement, age, subtype, and comorbidities can all change movement patterns without necessarily changing the underlying diagnosis.

The studies by [10] and [12] are especially useful because they widen the discussion beyond simple binary classification. Instead of treating ADHD-related behavior as a single static label, they show that temporal organization, interaction dynamics, and visual attention trajectories may reveal developmental or behavioral profiles that are not captured by standard accuracy metrics. This point matters for future work: a clinically useful system should not only answer whether a participant belongs to an ADHD or control group, but also explain which behavioral dimensions contributed to that decision and whether those dimensions are stable across contexts.

Multimodal systems currently represent the most ambitious methodological direction. Models such as ADViQAL, AVEN, and ICAE combine RGB video, body and facial landmarks, audio, questionnaire data, and latent cross-modal representations [23–25]. Their reported performance, often in the 98–99% range,

supports the hypothesis that ADHD is better represented as a constellation of motor, attentional, expressive, and cognitive cues than as a single visual marker. This is also where the evidence becomes most vulnerable to overfitting. When a model fuses many signals from only a small number of participants, it may learn the idiosyncrasies of the sample or recording protocol rather than robust ADHD markers. Therefore, multimodal fusion should be considered the most promising direction, but not yet the most clinically validated.

Finally, the reviewed studies show a consistent convergence between ADHD-related behaviors and computational markers, especially movement frequency, postural instability, limb activity, head motion, gaze behavior, and skeletal variance. These patterns suggest that the models capture clinically plausible manifestations of hyperactivity, impulsivity, and attention difficulties rather than arbitrary visual cues. Recent work also moves from isolated signals toward multimodal representations, combining video, pose, gaze, actigraphy, and questionnaire data through fusion, attention, graph-based, or transformer architectures [23–25]. Another positive trend is the gradual use of more realistic settings, such as outpatient consultations, caregiver–child interaction, free play, and virtual classrooms [4, 10, 31, 30]. Although still uneven, stronger validation practices such as matched controls, nested cross-validation, independent test sets, and sensitivity analyses are beginning to appear, making the field more comparable and clinically meaningful [4, 26, 30, 29].

5.2 Main Gaps and Limitations

The main limitation of the reviewed literature is the fragility of its experimental evidence. Several video- and pose-based studies report accuracies above 94% using very small samples, often between 7 and 30 participants, which raises concerns about overfitting, clip-level bias, and poor generalization to new patients, clinics, cameras, or task conditions [16, 19, 17, 18, 24, 25]. The field is also demographically narrow, with limited cultural diversity, scarce Latin American representation, frequent male predominance, and insufficient analysis of ADHD subtype, severity, medication status, sex, age, and comorbidities. In addition, comparisons across studies remain difficult because tasks, capture protocols, preprocessing pipelines, validation strategies, and reported metrics vary widely. Technical constraints further limit clinical translation, since pose estimation and multimodal systems can be sensitive to lighting, occlusion, camera angle, movement, hardware requirements, and acquisition burden. Thus, while computer vision and multimodal learning are promising for assisted ADHD screening, the current evidence is not yet strong enough for clinical replacement; progress depends less on larger neural networks and more on larger, diverse, participant-level, externally validated datasets.

Another remarkable limitation in the literature is the scarcity of data. To our knowledge, among all datasets reviewed, only HYPERAKTIV [3, 13] is publicly available. This fact creates an obstacle for the fair comparison between methodologies and contributes to the reproducibility gap.

Finally, deep models provide ADHD versus control scores but do not explain the reasoning behind them, which is a significant barrier to clinical trust. However, many reviewed systems do offer interpretable markers aligned with DSM-5 inspectable through attention maps and skeletal-joint visualizers serving as explainable clinical decision support [16, 26, 1, 2].

6 Conclusion

This review aimed to answer "*How robust are current experiments on vision-based ADHD screening in terms of generalization and clinical reliability?*". Through the analysis of the state-of-the-art found in the literature, it shows that computer vision, motion analysis, and multimodal deep learning are among the most promising approaches for objective support in ADHD screening. The strongest evidence comes from video- and pose-based methods, particularly when combined with multimodal or *dual-stream* architectures, which have achieved high performance by capturing subtle motor, attentional, and behavioral patterns associated with ADHD.

Among the reviewed methods, the model proposed by [24] stands out as the strongest camera-based approach, with accuracy above 98% and consistent results across multiple metrics. Nevertheless, its clinical relevance remains limited by the use of a small and controlled dataset. This limitation reflects the broader state of the field: the technical potential is clear, but external validation on larger, more diverse, and clinically realistic samples is still insufficient.

Future work should use standardized capture protocols, participant-level validation, diverse samples, and interpretable behavioral markers. Under these conditions, computer vision may become an accessible and clinically useful tool for assisted ADHD diagnosis, complementing rather than replacing expert clinical evaluation.

Acknowledgements This research is supported by the MindFocus project (FURG/INOV-82).

References

1. Amado-Caballero, P., Casaseca-de-la Higuera, P., Alberola-López, S., et al.: Insight into ADHD diagnosis with deep learning on actimetry: Quantitative interpretation of occlusion maps in age and gender subgroups. Artificial Intelligence in Medicine **146**, 102630 (2023)
2. Association, A.P.: Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association Publishing, Washington, DC (2022)
3. Bouchouras, G., Sofianidis, G., Kotis, K.: Temporal anomaly detection in attention-deficit/hyperactivity disorder using recurrent neural networks. Cureus (2024)
4. Chiu, Y.H., Lee, Y.H., Wang, S.Y., et al.: Objective approach to diagnosing attention deficit hyperactivity disorder by using pixel subtraction and machine learning classification of outpatient consultation videos. Journal of Neurodevelopmental Disorders **16**(71) (2024), https://doi.org/10.1186/s11689-024-09588-z

5. Choi, H., Hong, J., Kang, H.G., et al.: Retinal fundus imaging as biomarker for ADHD using machine learning for screening and visual attention stratification. npj Digital Medicine **8**(1), 45 (2025)
6. Devkota, A.: Detecting ADHD from Eye Movements – Comparing Time-Series Imaging Methods. Master's thesis, Oslo Metropolitan University (2024)
7. Drechsler, R., Brem, S., Brandeis, D., et al.: Adhd: Current concepts and treatments in children and adolescents. Neuropediatrics **51**(5), 315–335 (2020)
8. Huang, W., Jiang, X., Gao, C., et al.: A graph-based information fusion approach for adhd subtype classification. In: 2022 IEEE UIC-ATC. pp. 714–723. IEEE (2022)
9. Hult, N., Kadesjö, J., Kadesjö, B., et al.: Adhd and the qbtest: Diagnostic validity of qbtest. Journal of Attention Disorders **XX**(X), 1–7 (2015)
10. Isaev, D.Y., Sabatos-DeVito, M., et al.: Computer vision analysis of caregiver–child interactions in children with neurodevelopmental disorders: A preliminary report. Journal of Autism and Developmental Disorders **54**, 123–145 (2024)
11. Jayawardena, G., Michalek, A., Jayarathna, S.: Eye tracking area of interest in the context of working memory capacity tasks. In: 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI). pp. 208–215 (2019). https://doi.org/10.1109/IRI.2019.00042
12. Jones, W., Klin, A.: Attention to eyes is present but in decline in 2–6 month-olds later diagnosed with autism. Nature **504**(7480), 427–431 (2013)
13. Kaur, A., Kahlon, K.S.: Accurate identification of adhd among adults using real-time activity data. Brain Sciences **12**(7), 831 (2022)
14. Khan, A.R., Khosravi, S., Hussain, S., et al.: Execute: Exploring eye tracking to support e-learning. In: 2022 IEEE Global Engineering Education Conference (EDUCON). pp. 670–676. IEEE (2022)
15. Li, Y., Naqvi, S.M., Nair, R.: Adhd diagnosis based on action characteristics recorded in videos using machine learning. In: Neuroscience Applied. vol. 2, p. 102835. Elsevier (2023)
16. Li, Y., Nair, R., Naqvi, S.M.: Video-based skeleton data analysis for adhd detection. In: 2023 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 1–6 (2023). https://doi.org/10.1109/SSCI52147.2023.10372062
17. Li, Y., Nair, R., Naqvi, S.M.: ADHD detection based on human action recognition. Neuroscience Applied **3**, 104093 (2024)
18. Li, Y., Nair, R., Naqvi, S.M.: Harnessing video intelligence: Intelligent system for ADHD detection. In: IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). pp. 1–5 (2024)
19. Li, Y., Yang, Y., Nair, R., Naqvi, S.M.: Action-based ADHD diagnosis in video. In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN). pp. 453–458 (2023)
20. Liu, Z., Li, J., Zhang, Y., et al.: Auxiliary diagnosis of children with attention-deficit/hyperactivity disorder using eye-tracking and digital biomarkers: Case-control study. JMIR Mhealth and Uhealth **12** (2024)
21. M, S., Rajendran, P.S.: Early detection of ADHD in children using two learning approaches. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). IEEE, Chennai, India (2024)
22. Musullulu, H.: Evaluating attention deficit and hyperactivity disorder (adhd): a review of current methods and issues. Frontiers in Psychology **Volume 16 - 2025** (2025). https://doi.org/10.3389/fpsyg.2025.1466088
23. Nash, C., Nair, R., Naqvi, S.M.: Cross-modal attention for multimodal information fusion: A novel approach to attention deficit hyperactivity disorder detection. In: IEEE International Conference on Information Fusion (FUSION). pp. 1–10 (2024)

24. Nash, C., Nair, R., Naqvi, S.M.: Insights into detecting adult adhd symptoms through advanced dual-stream machine learning. IEEE Transactions on Neural Systems and Rehabilitation Engineering **32**, 3378–3387 (2024)
25. Nash, C., Nair, R., Naqvi, S.M.: Optimising adhd detection: An autoencoder approach for multimodal classification. IEEE Trans. Artif. Intell. (2025)
26. Ouyang, C.S., Yang, R.C., Wu, R.C., et al.: Objective and automatic assessment approach for diagnosing attention-deficit/hyperactivity disorder based on skeleton detection and classification analysis in outpatient videos. Child and Adolescent Psychiatry and Mental Health **18**(60) (2024)
27. Page, M.J., McKenzie, J.E., Bossuyt, P.M., et al.: The prisma 2020 statement: An updated guideline for reporting systematic reviews. BMJ **372**, n71 (2021), https://doi.org/10.1136/bmj.n71
28. Tang, W., Shi, C., Li, Y., et al.: Keypoints-based multi-cue feature fusion network (mf-net) for action recognition of adhd children in tova assessment. Bioengineering **11**(12) (2024). https://doi.org/10.3390/bioengineering11121210
29. Varela-Casal, P., Esposito, F.L., Morata Martínez, I., et al.: Clinical validation of eye vergence as an objective marker for diagnosis of ADHD in children. Journal of Attention Disorders **22**(14), 1317–1329 (2018)
30. Wiebe, A., Selaskowski, B., Paskin, M., et al.: Virtual reality-assisted prediction of adult ADHD based on eye tracking, EEG, actigraphy and behavioral indices: a machine learning analysis of independent training and test samples. Translational Psychiatry **14**(1), 125 (2024)
31. Yeh, S.C., Lin, S.Y., Wu, E.H.K., et al.: A virtual-reality system integrated with neuro-behavior sensing for attention-deficit/hyperactivity disorder intelligent assessment. IEEE Trans. Neural Syst. Rehabil. Eng. **28**(9), 1899–1907 (2020)
32. Yoo, J.H., Kang, C., Lim, J.S., et al.: Development of an innovative approach using portable eye tracking to assist ADHD screening: a machine learning study. Frontiers in Psychiatry **15**, 1337595 (2024)
33. Zhang, Y., Kong, M., Zhao, T., et al.: Auxiliary diagnostic system for ADHD in children based on AI technology. Frontiers of Information Technology & Electronic Engineering **22**(9), 1167–1178 (2021)