A logo for a university

AI-generated content may be incorrect.

**Department of Computer Science**

**MSc Artificial Intelligence**

**Academic year 2024-2025**

Title: Early ADHD Detection in Children in Non-Clinical Environments Using Multimodal Data: Development of a Deep Learning-Based Classification

Iúri Alexandre Nunes Guerreiro

2054117

[2054117@brunel.ac.uk](mailto:2054117@brunel.ac.uk)

Dissertation Supervisor: Steven Sam

**Abstract**

*Attention-Deficit/Hyperactivity Disorder (ADHD) is one of the most common neurodevelopmental conditions in children, yet diagnosis remains highly dependent on subjective reports from parents and teachers within clinical settings. This creates barriers to early recognition and delays access to support. The aim of this dissertation was to investigate whether multimodal machine learning (ML) and deep learning (DL) approaches could support ADHD detection in non-clinical environments, using synchronised electroencephalography (EEG), electrodermal activity (EDA), and eye-tracking data from the BALLADEER dataset.*

*A full methodological pipeline was developed, including participant-level data registries, signal cleaning and preprocessing, feature engineering for ML, data integration across modalities, deep learning manifest creation, and reproducible training and evaluation protocols. Classical ML models (Random Forest, SVM, XGBoost) were tested on handcrafted features to provide interpretable baselines. Deep learning experiments employed sequential modelling with long short-term memory (LSTM) networks, using unimodal EEG and eye-tracking data as well as late-fusion integration of both.*

*Results showed that classical ML achieved only modest discrimination (best ROC-AUC = 0.707 with XGBoost), limited in part by incomplete EDA coverage. Deep learning models significantly outperformed these baselines: EEG alone was highly sensitive but poorly specific, while eye-tracking achieved more balanced performance (ROC-AUC = 0.681). The strongest results came from the fusion model (ROC-AUC = 0.792, ACC = 0.766, SENS = 0.804, SPEC = 0.684), demonstrating that combining cortical and behavioural signals yields complementary diagnostic information.*

*This dissertation provides proof-of-concept that multimodal AI can be feasibly applied to ADHD screening outside clinical environments. By developing and evaluating a full pipeline, it shows that portable, data-driven tools could complement existing diagnostic practices, offering teachers and parents objective support for early detection. While larger datasets and robustness improvements are needed, these findings highlight the potential of multimodal AI to bridge the gap between clinical expertise and everyday environments.*

**TOTAL NUMBER OF WORDS: 14068**

Contents

[CHAPTER 1: Introduction 1](#_Toc210082729)

[1.1 Research Problem Definition 1](#_Toc210082730)

[1.2 Research Aim and Objectives 1](#_Toc210082731)

[1.3 Research Contributions 2](#_Toc210082732)

[1.4 Dissertation Structure 2](#_Toc210082733)

[CHAPTER 2: Literature Review 3](#_Toc210082734)

[2.1 Introduction to Attention Deficit Hyperactivity Disorder 3](#_Toc210082735)

[2.1.1 Meaning and effect of Attention Deficit Hyperactivity Disorder 3](#_Toc210082736)

[2.1.2 Emergence and Prevalence of ADHD 5](#_Toc210082737)

[2.1.3 Challenges in Diagnosing ADHD 7](#_Toc210082738)

[2.2 Neuroscientific and Multimodal Signal Foundations of ADHD 8](#_Toc210082739)

[2.2.1 Brain Structure and Function 9](#_Toc210082740)

[2.2.2 Neurophysiological Signals and Biomarkers 10](#_Toc210082741)

[2.2.3 Modality-Specific Contributions to ADHD Assessment 11](#_Toc210082742)

[2.3 Technological Interventions in ADHD Assessment 12](#_Toc210082743)

[2.3.1 Digital Health and AI 12](#_Toc210082744)

[2.3.2 Limitations of Existing Tools 15](#_Toc210082745)

[2.4 Machine Learning and Deep Learning in ADHD Detection 16](#_Toc210082746)

[2.4.1 Key Models in ADHD Detection 16](#_Toc210082747)

[2.4.2 Fusion Strategies for Multimodal Integration 18](#_Toc210082748)

[2.4.3 Trade-Offs and Implementation Challenges 19](#_Toc210082749)

[2.5 Multimodal Data in ADHD Research and Real-World Applications 21](#_Toc210082750)

[2.6 Gaps in Research & Justification 24](#_Toc210082751)

[Chapter 3: Methodology 26](#_Toc210082752)

[3.1 Introduction to Methodology 26](#_Toc210082753)

[3.2 Dataset: BALLADEER 27](#_Toc210082754)

[3.3 Preprocessing 28](#_Toc210082755)

[3.4 Feature Engineering 31](#_Toc210082756)

[3.5 Data Integration 32](#_Toc210082757)

[3.6 Modelling 33](#_Toc210082758)

[3.7 Training and Evaluation 36](#_Toc210082759)

[3.8 Summary 39](#_Toc210082760)

[Chapter 4: Results 40](#_Toc210082761)

[4.1 Classical Machine Learning Results 40](#_Toc210082762)

[4.2 Deep Learning Results 42](#_Toc210082763)

[4.3 Robustness and Tuning 43](#_Toc210082764)

[Chapter 5: Discussion 44](#_Toc210082765)

[5.1 Interpretation 44](#_Toc210082766)

[5.2 Comparison with Prior Work 45](#_Toc210082767)

[5.3 Practical Implications 47](#_Toc210082768)

[5.4 Strengths 47](#_Toc210082769)

[5.5 Limitations 48](#_Toc210082770)

[5.6 Ethical, Governance, and Equity 48](#_Toc210082771)

[5.7 Future Work 49](#_Toc210082772)

[Chapter 6: Conclusion 50](#_Toc210082773)

[REFERENCING 53](#_Toc210082774)

[APPENDIX 60](#_Toc210082775)

# CHAPTER 1: Introduction

## 1.1 Research Problem Definition

Attention-Deficit/Hyperactivity Disorder (ADHD) is one of the most common neurodevelopmental conditions in children, with symptoms that affect learning, behaviour, and long-term wellbeing (American Psychiatric Association, 2013; Arns et al., 2013). Despite its prevalence, current diagnostic pathways remain heavily reliant on subjective observations from parents, teachers, and clinicians, as well as structured questionnaires and interviews, while effective in clinical contexts, these methods are limited by bias, cultural variability, and the availability of specialist resources (Holmes et al., 2024, Feng et al., 2024). Furthermore, access to clinical assessment is often delayed by long waiting lists and limited capacity within health services, leaving children without timely support during critical periods of development. In non-clinical environments such as schools or homes, where attentional difficulties are frequently first recognised, teachers and parents lack objective tools to support early detection. This creates a gap between the early recognition of symptoms and formal diagnosis, which can delay interventions and exacerbate academic and social challenges. Addressing this gap is essential to ensure that children at risk of ADHD are identified and referred for assessment as early as possible.

## 1.2 Research Aim and Objectives

The primary aim of this dissertation is to explore the feasibility of developing an AI-based classification pipeline for the early detection of ADHD traits in children using multimodal data. To achieve this aim, the objectives are twofold. First, to design and implement classical machine learning models using engineered features from EEG, EDA, and eye-tracking data, providing transparent and interpretable baselines. Second, to develop and evaluate deep learning models capable of learning directly from sequential EEG and eye-tracking signals, assessing whether fusion of cortical and behavioural modalities improves classification performance. In both approaches, the emphasis is on evaluating whether multimodal physiological and behavioural data can serve as reliable indicators of ADHD in non-clinical environments. The study therefore not only investigates the technical feasibility of such models but also their potential practical relevance for use in schools and at home by teachers, parents, and caregivers (Chen et al., 2023; Deng et al., 2022).

## 1.3 Research Contributions

This dissertation makes contributions at methodological, empirical, and practical levels. Methodologically, it presents a fully reproducible multimodal pipeline for ADHD detection, integrating preprocessing, feature engineering, and benchmarking of classical machine learning against deep learning fusion models. Empirically, the findings demonstrate that combining EEG and eye-tracking modalities provides stronger performance than unimodal approaches, showing the complementary nature of neurophysiological and behavioural signals. Practically, the study positions such multimodal systems as potential early-screening tools that could operate in non-clinical environments, offering objective indicators of attentional regulation where current practice relies almost exclusively on subjective judgment (Holmes et al., 2024). Finally, the dissertation contributes conceptually to debates on the ethical governance of AI in child mental health, highlighting how predictive models must be framed as supportive aids for decision-makers rather than as diagnostic substitutes.

## 1.4 Dissertation Structure

The structure of this dissertation reflects the pathway from identifying the research problem to addressing it through experimental evaluation. Chapter 2 presents a review of relevant literature, including current diagnostic practices, neuroscientific foundations of multimodal data, and the role of AI in education and healthcare. Chapter 3 outlines the methodology, detailing dataset characteristics, preprocessing steps, feature engineering strategies, and the modelling approaches used in both machine learning and deep learning pipelines. Chapter 4 reports the experimental results, comparing the performance of different models across modalities. Chapter 5 discusses these findings in relation to ADHD theories, prior multimodal research, and practical implications, while addressing ethical and governance issues linked to AI in child mental health. Chapter 6 concludes the dissertation by summarising the study’s contributions, reinforcing its relevance for non-clinical environments, and outlining avenues for future research and policy development.

# CHAPTER 2: Literature Review

This chapter critically examines the current state of research surrounding the assessment and early detection of Attention Deficit Hyperactivity Disorder (ADHD), with a focus on the emerging role of multimodal machine learning systems. It begins by introducing the clinical features, prevalence, and diagnostic challenges of ADHD in children, establishing the importance of objective and accessible tools for early identification. The chapter then explores neuroscientific foundations, highlighting key physiological and behavioural modalities such as EEG, EDA, eye-tracking, and video analysis. Subsequent sections review existing technological and AI-based interventions, comparing the strengths and limitations of machine learning and deep learning models, as well as fusion strategies used to integrate multimodal signals. The chapter also evaluates the applicability of these models in real-world, non-clinical settings, using key studies and the BALLADEER dataset as reference points. It concludes by identifying critical gaps in current research, including the overreliance on clinical infrastructure and lack of tools for non-specialist use, and presents the justification for this dissertation’s proposed solution: a deep learning model designed for early ADHD detection in non-clinical environments.

## 2.1 Introduction to Attention Deficit Hyperactivity Disorder

### 2.1.1 Meaning and effect of Attention Deficit Hyperactivity Disorder

Attention Deficit Hyperactivity Disorder (ADHD) is a neurodevelopmental disorder characterised by persistent patterns of inattention and/or hyperactivity-impulsivity that interfere with functioning or development (American Psychiatric Association, 2022). According to the DSM-5 (American Psychiatric Association, 2013), the core symptom dimensions of ADHD are categorised into two domains: Inattention and Hyperactivity-Impulsivity, comprising a total of 18 diagnostic symptoms. The Inattention domain is defined by difficulties in sustaining attention, organizing tasks, and maintaining mental focus, whereas the Hyperactivity-Impulsivity domain involves excessive physical activity, impulsive decision-making, and challenges with self-regulation and behavioural control. Examples of both are summarised in the table below.

Table 1. ADHD domains and their symptoms

| Inattention | Hyperactivity-Impulsivity |
| --- | --- |
| Often fails to give close attention to details or makes careless mistakes | Often fidgets with hands or feet or squirms in seat |
| Frequently has trouble sustaining attention in tasks or play activities | Leaves seat in situations where remaining seated is expected |
| Does not seem to listen when spoken to directly | Often runs about or climbs in inappropriate situations |
| Often does not follow through on instructions and fails to finish tasks | Often unable to play or engage in leisure activities quietly |
| Has difficulty organizing tasks and activities | Often talks excessively |
| Avoids or is reluctant to engage in tasks that require sustained mental effort | Often blurts out answers before questions have been completed |
| Often loses things necessary for tasks or activities | Has difficulty waiting for their turn |
| Is easily distracted by extraneous stimuli | Often interrupts or intrudes on others (e.g., butts into conversations/games) |
| Is often forgetful in daily activities | Appears constantly “on the go” or acts as if “driven by a motor” |

ADHD is recognised by major diagnostic frameworks, including the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) and the International Classification of Diseases (ICD-11). According to both systems, symptoms must be present before the age of 12 and must manifest across multiple settings, such as home and school, to ensure they are not contextually driven (American Psychiatric Association, 2013).

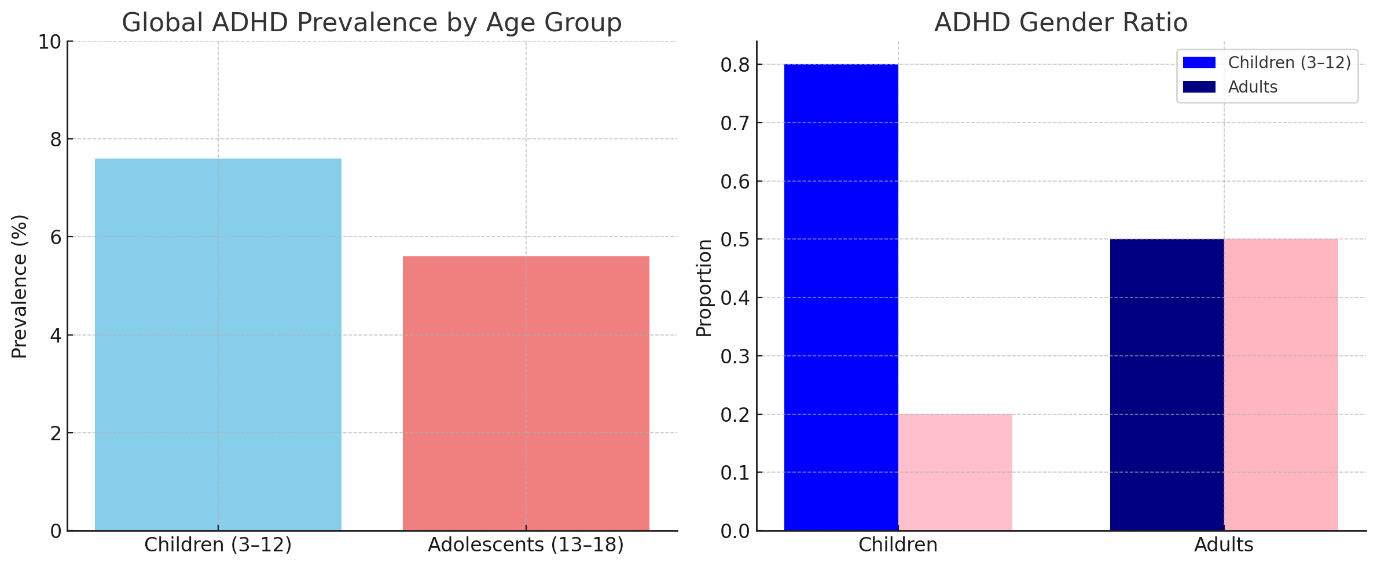
Research shows that ADHD has a significant effect on children's lives (ref). In academic settings, it is associated with poor concentration, incomplete assignments, and difficulty following instructions, often leading to lower academic performance and teacher conflict (Faraone et al., 2015). Socially, children with ADHD are more likely to experience peer rejection, have fewer close friendships, and display disruptive behaviours that affect group activities (Raggi & Chronis-Tuscano, 2006). Emotionally, these children may develop low self-esteem, frustration, and heightened emotional sensitivity due to repeated negative feedback or failure experiences. Untreated ADHD is also linked with long-term adverse outcomes, such as increased risk of school dropout, substance misuse, and difficulties in employment and interpersonal relationships later in life (Faraone et al., 2015). ADHD is thus a condition with deep developmental, educational, and social consequences. Understanding its historical emergence, classification systems, and symptom variability in children forms the foundation for advancing early identification tools, especially those driven by multimodal AI systems.

### 2.1.2 Emergence and Prevalence of ADHD

The earliest recorded descriptions of behaviours resembling Attention Deficit Hyperactivity Disorder (ADHD) date back to the late 18th century. In 1775, German physician Melchior Adam Weikard introduced the term "Attentio Volubilis" in his medical textbook Der Philosophische Arzt, describing individuals who were easily distracted, impulsive, and unable to complete tasks, traits that align closely with modern ADHD symptomatology. Similarly, in 1798, Scottish physician Sir Alexander Crichton discussed a "disease of attention" in his work An Inquiry into the Nature and Origin of Mental Derangement, identifying children who showed persistent mental restlessness and poor sustained focus (Columbia Psychiatry, 2023).

These early observations were later formalised in clinical literature. In 1902, Sir George Frederic Still presented a series of lectures to the Royal College of Physicians, where he described children with impulsive behaviour, attention problems, and emotional dysregulation, marking the first formal clinical framework for what would later be known as ADHD. Throughout the 20th century, terminology evolved from "Minimal Brain Dysfunction" to "Hyperkinetic Reaction of Childhood," and finally to "Attention Deficit Hyperactivity Disorder" in the 1980s (Verywell Mind, 2023).

Recent studies found that ADHD affects globally approximately 7.6% of children aged 3–12 and 5.6% of adolescents aged 13–18, with diagnostic rates rising due to improved awareness and evolving criteria in the DSM-5 (Abdelnour et al., 2022; Salari et al., 2023). During childhood, the gender ratio is approximately 4:1 in favour of boys, but this shifts to nearly 1:1 in adulthood (Abdelnour et al., 2022). Several factors contribute to this, including underdiagnosis of inattentive symptoms in girls and gendered expectations in classroom behaviour (Fraticelli et al., 2022). The overall rise in childhood ADHD diagnoses over recent decades is partially attributed to broader diagnostic criteria, greater societal awareness, and improved access to mental health assessments (Salari et al., 2023).



*Figure 1. ADHD prevalence by age group (children vs. adolescents) and gender. Source: Author’s own work, based on Abdelnour et al. (2022) and Salari et al. (2023).*

### 2.1.3 Challenges in Diagnosing ADHD

According to DSM-5, ADHD is diagnosed under three distinct presentations: (i) Predominantly Inattentive, (ii) Predominantly Hyperactive-Impulsive, and (iii) Combined. These categories are based on symptom clusters observed over at least six months in two or more settings (e.g., home and school), with onset before age 12 and clear evidence of functional impairment (American Psychiatric Association, 2013). The Inattentive presentation involves behaviours such as frequent forgetfulness, distractibility, and difficulty sustaining attention. The Hyperactive-Impulsive presentation includes excessive fidgeting, restlessness, and interrupting others. The Combined presentation requires meeting criteria for both inattention and hyperactivity-impulsivity. This classification allows clinicians to tailor interventions more precisely but also raises concerns over symptom overlap and variability across developmental stages (Willcutt, 2012).

In contrast, ICD-11 by the World Health Organization adopts a more restrictive framework. As mentioned before, ADHD is referred to as "Attention Deficit Hyperactivity Disorder", but unlike the DSM-5, a diagnosis requires the presence of both inattentive and hyperactive-impulsive symptoms. This unified criteria approach can reduce the risk of overdiagnosis but may overlook children, especially girls or older individuals, who display primarily inattentive traits without marked hyperactivity (Koutsoklenis & Honkasilta, 2022). One strength of the ICD-11 system is its attempt to ensure diagnostic consistency globally, especially in resource-limited settings. However, a key limitation is that it may fail to capture the full clinical heterogeneity of ADHD presentations, thereby excluding individuals who could still benefit from intervention (Bruchmüller et al., 2012).

Recent work has challenged rigid categorical classification. Study like Bowden et al. (2023) has shown that ADHD symptoms in children may vary in expression across settings (home vs school), suggesting that situational context and behavioural environment must be considered when evaluating subtype presentations. Thus, despite clear diagnostic frameworks such as the DSM-5 and ICD-11, diagnosing ADHD in children remains challenging. One of the most significant issues is the subjectivity of observer reports. Clinicians typically rely on behavioural descriptions provided by parents and teachers, but these can vary greatly depending on personal biases, cultural expectations, and familiarity with neurodevelopmental norms. Research shows boys are more readily identified because of overt hyperactivity, while girls, especially with the inattentive presentation, are frequently overlooked due to subtler symptoms like daydreaming or quiet disengagement (Bruchmüller, Margraf & Schneider, 2012; Young, Adamo, Ásgeirsdóttir et al., 2020).

Moreover, many children do not receive professional help because their symptoms are not noticed by teachers or parents. Studies indicate that up to one-third of children display ADHD-like behaviours but receive no referral until significant issues arise, particularly among children from disadvantaged backgrounds or girls who internalise symptoms. Mistrust between educational staff and families can further delay referrals, meaning only a minority of affected children ever access formal assessment (Amy Borrett, 2025; Financial Times, 2025) (Silver, 2017; BBC, 2017).

ADHD symptoms often overlap with other conditions such as anxiety, autism spectrum disorder (ASD), or learning difficulties, complicating differential diagnosis. Comorbidities, which are very common, can obscure the clinical picture and lead to delayed or inaccurate diagnoses (Sciberras et al., 2014).

Long wait times for specialist assessments, especially within public health systems like the NHS, pose another critical barrier. UK data shows that nearly one-third of children waited two or more years for an ADHD diagnosis, with average delays spanning 16-18 months to even three years in some regions (Purper-Ouakil et al., 2007) (Silver, 2017; BBC, 2017). These prolonged delays exacerbate academic, social, and emotional difficulties, often leaving children unsupported at essential developmental stages.

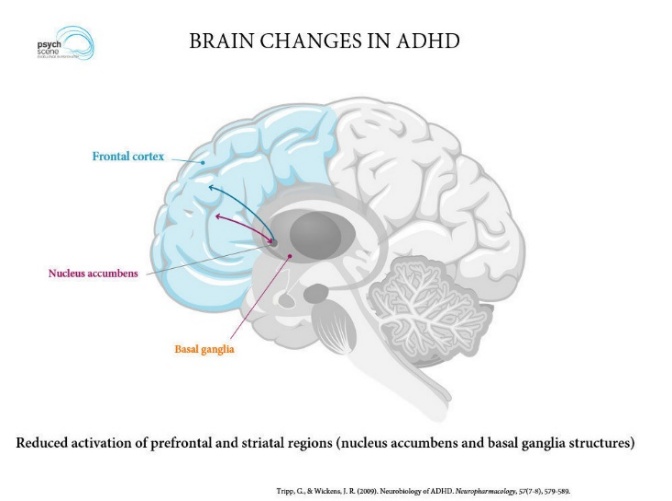
Collectively, these issues underscore the urgent need for more objective, accessible tools that can facilitate earlier, non-clinical identification of ADHD, before children fall through systemic gaps in support.

## 2.2 Neuroscientific and Multimodal Signal Foundations of ADHD

Scientific advances over the past two decades have significantly contributed to understanding the neurobiological basis of ADHD. The condition is now widely recognised as a neurodevelopmental disorder characterised by functional and structural abnormalities in key brain regions responsible for executive control, attention, emotion regulation, and sensory integration.

### 2.2.1 Brain Structure and Function

Neuroimaging studies have consistently identified structural differences in children diagnosed with ADHD (Feng et al., 2024). These differences are evident for example: in the prefrontal cortex, basal ganglia, cerebellum, and the superior longitudinal fasciculus (SLF), a major white matter tract that connects regions involved in attention and sensory processing (Collins & Koechlin, 2012). Altered development in these areas is associated with deficits in executive functioning, such as impulse control, planning, and working memory.



*"Figure X. Brain changes commonly observed in ADHD, (Rege 2019)."*

Children with ADHD also show atypical maturation in the superior and middle temporal gyri, areas responsible for semantic memory, language processing, and audio-visual integration (Kieling & Rohde, 2012; Onitsuka et al., 2004). These regions play a critical role in classroom learning, where sustained attention and multi-sensory processing are essential. Additionally, disruptions in the cortico-limbic system, particularly involving the amygdala and anterior cingulate cortex, have been linked to impaired emotional regulation, increased frustration, and low tolerance for delay or uncertainty (Yu et al., 2023; Hoogman et al., 2017; Feng et al., 2024).

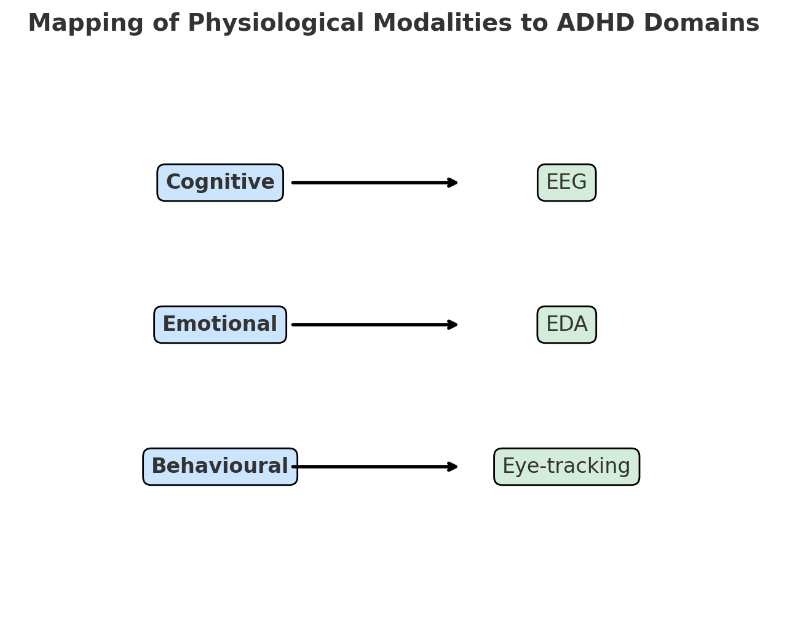
### 2.2.2 Neurophysiological Signals and Biomarkers

Beyond structural differences, physiological studies offer deeper insight into the functional deficits associated with ADHD. Electroencephalography (EEG) has been one of the most widely used methods for investigating neural activity in children with the disorder. A commonly reported biomarker is the elevated theta/beta ratio, especially in frontal brain regions. This pattern is believed to reflect cortical under arousal and has been consistently associated with poor attentional control and increased distractibility (Gong et al., 2022; Feng et al., 2024; Arns et al., 2013).

Further EEG features include event-related potentials (ERPs), brain responses to specific stimuli, which have revealed delayed or diminished responses in children with ADHD during attention-demanding tasks. These findings align with the behavioural symptoms of slowed cognitive processing and inconsistent performance during learning activities.

Electrodermal activity (EDA) is another physiological signal of growing interest in ADHD research. It measures skin conductance levels controlled by the sympathetic nervous system and serves as an index of arousal and emotional responsiveness. Studies have shown that children with ADHD tend to exhibit blunted or inconsistent EDA responses, indicating reduced emotional reactivity or difficulty modulating arousal levels in response to stimuli (Dawson et al., 2017). Such dysregulation may contribute to the emotional lability and frustration intolerance frequently reported in both home and school settings.

Eye-tracking research has also revealed specific oculomotor patterns in ADHD. Children with the disorder often display unstable gaze, increased saccadic movements, and reduced fixation on relevant stimuli. These behaviours suggest difficulties in maintaining visual attention and filtering distractions, particularly during tasks requiring sustained focus (Deng et al., 2022).



*Figure 2. Mapping of physiological modalities to ADHD domains: EEG relates primarily to cognitive processes, EDA to emotional regulation, and eye-tracking to behavioural attention patterns.*

### 2.2.3 Modality-Specific Contributions to ADHD Assessment

Each physiological modality contributes uniquely to understanding ADHD in children. EEG provides insight into cognitive arousal and brainwave irregularities during attention tasks. EDA highlights autonomic dysregulation and emotional under-responsiveness, particularly in emotionally charged or feedback-rich environments. Eye-tracking reveals moment-to-moment patterns of visual attention and inhibition failures, especially during complex or sustained activities.

In addition to these bio signals, video analysis has emerged as a useful behavioural modality. Advanced computer vision tools such as OpenPose and MediaPipe can track fine motor movements, fidgeting, posture shifts, and overall activity levels, behaviours typically associated with hyperactivity and restlessness. These systems provide naturalistic, non-invasive ways of quantifying ADHD symptoms that often go unnoticed in structured tests (Muñoz-Organero et al., 2018; Li et al., 2024).

Studies such as Andrikopoulos et al. (2024); Vortmann et al. (2022) demonstrate that while individual modalities provide valuable insights, their integration, discussed in detail in a later section, has the potential to dramatically enhance ADHD diagnoses performance.

The BALLADEER dataset plays a critical role in this context. It is one of the few publicly available datasets containing synchronised recordings of EEG, EDA, eye-tracking, and video from children with and without ADHD. Unlike many clinical datasets, BALLADEER was collected in a semi-naturalistic environment during cognitive tasks, making it particularly useful for training and evaluating machine learning models intended for real-world ADHD support (Trujillo et al., 2025).

## 2.3 Technological Interventions in ADHD Assessment

### 2.3.1 Digital Health and AI

Over the past decade, digital health innovations, such as wearables, mobile apps, interactive games, and AI-driven analytics, have gained traction in the field of ADHD assessment. Among these, Fitbit smartwatches have been used to collect movement patterns, heart rate variability, and energy expenditure data. In a recent study, Fitbit-derived features from 450 adolescents were analysed using Random Forest and logistic regression models to predict clinician-diagnosed ADHD (Rahman, 2025). The results demonstrated high predictive power, with classification accuracy reaching 89% and an AUC of 0.95, showcasing strong potential for non-invasive monitoring (Rahman, 2025). Fitbit’s strengths lie in its wide availability, ease of use, and real-time data collection. However, its limitations include potential inaccuracies in raw sensor data, lack of standardisation across models, and difficulties in isolating ADHD-specific behavioural patterns in real-world use. Beyond wearables, more experimental approaches have explored the use of virtual reality combined with AI-driven behavioural analysis for ADHD diagnosis, showing early clinical promise (Oh et al., 2024).

Another promising intervention is the Revibe Connect, a wearable device designed to deliver subtle vibration cues to help children with attention difficulties self-monitor and refocus during tasks. Ayearst, Brancaccio, and Weiss (2023) conducted a pilot trial involving 34 students with ADHD using Revibe Connect during school hours over a four-week period. The study found significant improvements in classroom attention and executive function, particularly in task persistence and goal-directed behaviour. Strengths of the device include its unobtrusive feedback mechanism and strong acceptability among both students and teachers. Nevertheless, limitations include its dependency on consistent device wear, limited long-term data, and lack of physiological signal tracking.

In a broader exploration of wearable health data, a 2024 study published in Journal of the American Academy of Child & Adolescent Psychiatry Open evaluated various wearable bio signals for ADHD assessment. The study involved collecting continuous activity and physiological signals from children over several days and applying machine learning models to identify ADHD traits. The models showed strong potential in differentiating ADHD from controls, but also highlighted technical constraints, particularly signal noise, wear-time compliance, and inter-device variability (Jiang et al., 2024). The strength of this study was its ecological validity and multi-day monitoring, though real-world implementation requires addressing calibration and usability issues.

Finally, Ochab et al. (2019) demonstrated that even basic actigraphy, when analysed with non-linear mathematical techniques, could support ADHD classification. Their research involved week-long wrist-based activity recordings from children with ADHD, autism spectrum disorder (ASD), and neurotypical controls. By extracting complexity features from the motion signals, the authors trained k-Nearest Neighbour (kNN) models to achieve 69–78% classification accuracy between ADHD subtypes and controls. While less accurate than multimodal approaches, this study illustrates how passive, long-term activity data may still carry meaningful diagnostic information. Its main advantage is accessibility, but a limitation is the relatively modest classification power when used in isolation.

Collectively, these studies reinforce the growing value of digital health tools in ADHD detection. They demonstrate promising outcomes in terms of accuracy, usability, and non-invasiveness, while also highlighting technical and methodological barriers to broader adoption, such as data quality, user adherence, and standardisation. These insights inform the design of future, ethically sound, non-clinical ADHD support systems that leverage wearable technologies and AI.

Meanwhile, app-based digital therapeutics are already being used to improve attention and inhibitory control in children through gamified interventions delivered at home. EndeavourRx, the first FDA-authorized video game-based treatment for children with ADHD, is designed for unsupervised home use and targets neural systems involved in attention and cognitive control via adaptive, gamified tasks. It is supported by five clinical studies involving over 600 children aged 8–12 with a confirmed ADHD diagnosis (EndeavourRx, 2025). In a pivotal randomised, controlled trial, children who played the game for approximately 25 minutes per day, five days a week over four weeks showed statistically significant improvements in objective measures of attention, as assessed by the Test of Variables of Attention (TOVA®) (Kollins et al., 2020). Furthermore, a follow-up study found that an additional month of gameplay produced further cognitive benefits, including improvements in sustained attention and working memory (Kollins et al., 2021).

These findings demonstrate a major strength of EndeavourRx: it offers a non-pharmacological, scalable, and engaging intervention that can be delivered in real-world settings without the constant supervision of clinicians. The use of objective neurocognitive tasks, rather than subjective symptom ratings, adds further credibility to the outcomes. Beyond gamified tools, conversational AI has also been explored for its therapeutic potential. Recent work evaluated ChatGPT as a support tool in ADHD therapy enhancement, suggesting potential value in augmenting traditional care (Berrezueta-Guzman et al., 2024)

However, several limitations must be acknowledged. First, while short-term gains in attention were measurable, long-term benefits remain unclear, as most studies followed participants for only several weeks. Second, improvements were found primarily on cognitive performance tasks, not necessarily in real-world behavioural functioning such as classroom performance or parent/teacher ratings. Third, selection bias may exist, as participants were typically recruited through clinical channels and may not represent more diverse or underserved populations. Finally, like many digital interventions, there is a risk of reduced adherence over time, especially without ongoing motivation or support mechanisms.

This growing body of evidence still underscores the therapeutic potential of accessible and engaging digital tools, especially when integrated into multi-modal care plans. EndeavourRx represents a promising direction in the movement toward early, objective, and at-home interventions that supplement traditional clinical care, although more longitudinal and ecologically valid research is needed to fully establish its efficacy and generalisability.

### 2.3.2 Limitations of Existing Tools

Despite technological advances in ADHD assessment, traditional tools remain constrained by their clinic-centric design, reliance on specialised operators, and substantial costs. Instruments like the QbTest and Conners CPT‑3 necessitate dedicated hardware, licensing fees, and expert interpretation. Recent evidence further highlights their limitations. For instance, Bellato et al. (2023) demonstrated through a systematic review that, although QbTest offers moderate sensitivity (~0.78) and specificity (~0.70), it lacks sufficient accuracy for standalone diagnostic use, often yielding false positives or negatives without clinical context. Similarly, a comprehensive review of Conners CPT‑3 revealed its standalone predictive utility as weak to modest, particularly unable to discriminate ADHD from ASD or comorbid conditions in many cases, calling for its use as part of a broader battery (Callan et al., 2024).

Another study evaluating Conners CPT-II found that even when combined with teacher and parent reports, it struggled to reach the post-test probability threshold (≥85%) typically considered sufficient for reliable clinical decisions (Tallberg et al., 2019). These findings collectively highlight that while clinic-based behavioural tests can inform clinical judgment, they fall short of offering conclusive diagnosis without supplementary modalities like EEG or physiological markers, thus leaving a gap in detecting neurophysiological underpinnings (Michelini et al., 2022).

An often-overlooked consequence of these limitations is their inapplicability outside controlled environments. These tools are ineffective in everyday settings such as schools or homes, where early signs of ADHD frequently manifest. Children with subtle symptoms, especially girls or those from underserved backgrounds, rely on adult observers like parents and teachers who lack objective, accessible tools to identify emerging concerns before clinical thresholds are met.

Given these limitations, there is a growing recognition of the need for affordable, easy to administer screening solutions that function in natural environments without the infrastructure of a clinic. Such tools would harness wearable sensors and app-based platforms, offering passive data collection, for example, tracking movement, physiological arousal, or attention through wearable devices, where adults can use them without specialist training.

By positioning these tools not as replacements for diagnostic assessments but as screening aids, they can address a critical barrier: the delay in recognition and referral. Emerging evidence links early interventions to improved educational, social, and long-term outcomes for children with ADHD (Young et al., 2020). A shift toward non-clinical, data-driven solutions merges the strengths of technology with the needs of everyday observers, enabling scalable monitoring and timely referral. This shift directly aligns with the aims of this dissertation, to design a multimodal, deep-learning-based early detection tool suitable for real-world, non-clinical deployment, potentially transforming ADHD support pathways.

## 2.4 Machine Learning and Deep Learning in ADHD Detection

### 2.4.1 Key Models in ADHD Detection

Machine learning (ML) and deep learning (DL) algorithms have increasingly become crucial in developing data-driven ADHD detection tools. These models enable the identification of subtle, nonlinear patterns across physiological and behavioural data that may not be visible to human raters.

#### 2.4.1.1 Convolutional Neural Networks (CNNs)

CNNs have been widely used in ADHD detection for their strength in recognising spatial patterns from visual data such as facial expressions, posture, and movement dynamics. Muñoz-Organero et al. (2018); Li et al. (2024) successfully applied CNNs to extract spatial features from video recordings of children completing cognitive tasks, identifying hyperactivity-related movement patterns such as frequent posture shifts. Their CNN component contributed to a late-fusion model that achieved an accuracy above 87%, outperforming unimodal approaches.

Similarly, Kumar et al. (2022) applied CNNs to EEG topographic images, achieving classification accuracy of 84% in differentiating ADHD from control participants. This study highlighted CNNs’ ability to automatically learn discriminative spatial patterns from static brainwave representations. However, CNNs require large, labelled datasets to prevent overfitting, and their performance may drop in real-world settings where video quality, lighting, and child movement vary significantly.

Another limitation is interpretability. While CNNs can detect subtle, complex visual cues, explaining why a CNN flagged a specific behaviour as indicative of ADHD remains challenging, an issue especially problematic in clinical or educational contexts (Zhou et al., 2022).

#### 2.4.1.2 Long Short-Term Memory Networks (LSTMs)

LSTMs are ideal for analysing time-series data such as EEG signals, EDA fluctuations, or eye-tracking sequences. Chang et al. (2022), in a systematic review of EEG-based deep learning studies, highlighted that LSTMs are particularly effective at modelling sequential neural dynamics, detecting subtle lapses in attention that static models often miss. Building on this, Wang et al. (2022) developed a CNN–LSTM framework that combined spatial and temporal EEG features, achieving classification accuracies above 90% in distinguishing ADHD from controls. These findings underscore the promise of LSTMs for capturing dynamic attentional processes. However, their reliance on long, clean time-series data makes them highly sensitive to artefacts such as eye blinks or muscle noise in EEG recordings. Moreover, like CNNs, LSTMs lack transparency, making it difficult to explain which features or sequences led to a specific prediction, a critical consideration for adoption in clinical workflows.

#### 2.4.1.3 Classical Machine Learning Models (SVMs and Random Forests)

Support Vector Machines (SVMs) and Random Forests (RF) remain popular due to their simplicity, interpretability, and lower computational demands. Several studies have achieved solid performance using SVMs on EEG-derived features. For instance, Alim et al. (2023) extracted power spectrum features from EEG and used an SVM to classify ADHD with 93.2% average accuracy. SVMs are especially useful in high-dimensional but small datasets, where deep learning might overfit.

Random Forests have also been applied to multimodal datasets. For example, Yoo et al. (2024) used RFs on eye-tracking features from multiple behavioural tasks, finding gaze/fixation metrics like saccade latency and fixation duration among the strongest predictors, with classification accuracy of 76.3% based on eye-tracking only, improving when combined with demographic and clinical features. These models are far more interpretable than deep networks and can be deployed with minimal infrastructure, making them attractive for low-resource educational settings.

However, classical models generally perform worse than DL models on complex data with nonlinear relationships, and they struggle with raw or unstructured data unless extensive feature engineering is applied. Their predictive power can also plateau as more modalities are added, which limits their scalability in high-dimensional multimodal contexts.

### 2.4.2 Fusion Strategies for Multimodal Integration

When integrating multiple data streams, such as EEG, EDA, eye-tracking, and video, within a predictive model, the method of fusion is a crucial design choice. Two primary strategies are commonly employed: early fusion and late fusion (Feng et al., 2024; Vortmann et al., 2023). Early fusion involves combining raw or low-level features from each modality into a single feature vector that is then input into a unified model. This approach facilitates joint feature learning across data types, potentially uncovering synergistic patterns across modalities. However, early fusion can become inefficient or unstable when dealing with heterogeneous data formats or asynchronous signals, such as when merging eye-tracking data (high temporal resolution) with EDA (slower signal variability) (Feng et al., 2024; Vortmann et al., 2023).

In contrast, late fusion processes each modality independently through its own specialised subnetwork, such as a CNN for spatial video data and an LSTM for temporal EEG signals and then merges their high-level outputs (typically class probabilities or feature embeddings) at the decision level. This method preserves the unique temporal or spatial characteristics of each data type while enabling synergy during classification (Chang et al., 2022; Wang et al., 2022). In ADHD detection tasks, late fusion has been shown to outperform early fusion in terms of both robustness and interpretability, as it allows researchers to better trace which modalities contributed most to a given decision (Vortmann et al., 2022; Andrikopoulos et al., 2024), A neuroimaging example of modality fusion in children with ADHD is symptom-guided structural + functional MRI fusion, which links brain features to cognition and behaviour (Feng et al., 2024).

This dissertation adopts a late fusion strategy for these reasons. The CNN component is tailored to detect spatial indicators of motor agitation (e.g., frequent posture shifts or fidgeting) from video recordings, while the LSTM subnetwork captures the temporal dynamics of attention, arousal, and emotional responses using EEG, EDA, and eye movement sequences. This hybrid approach is especially suitable for asynchronous and heterogeneous data, enhancing the model’s ability to generalise across individuals and real-world scenarios *(Wiebe et al., 2024; Rahman, 2025)*. Furthermore, it aligns with recent findings demonstrating that multimodal fusion significantly improves ADHD detection performance over unimodal approaches (Feng et al., 2024; Vortmann et al., 2022; Andrikopoulos et al., 2024). Comparison of both strategies can be seen on the image below.

A screenshot of a computer

AI-generated content may be incorrect.

*Figure 3. Comparison of Early Fusion and Late Fusion strategies for multimodal data integration.*

### 2.4.3 Trade-Offs and Implementation Challenges

While deep learning (DL) models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) offer superior accuracy for ADHD detection tasks, they introduce critical trade-offs in terms of interpretability and deployment. These models are often regarded as “black boxes,” meaning it can be difficult to understand or explain the basis for a given prediction (Rudin, 2019). This lack of transparency is particularly problematic in healthcare and education contexts, where trust, accountability, and clinical justification are essential (Chen et al., 2023).

In contrast, classical machine learning (ML) models such as Support Vector Machines (SVMs) and Random Forests (RFs) offer greater interpretability. They allow practitioners to analyse decision boundaries and feature importance scores, making it easier to explain model outputs to stakeholders (Lundberg & Lee, 2017). For example, Random Forests have been used to highlight the predictive weight of features such as EDA variability and reaction time variance, information that is clinically meaningful and practically actionable. However, these models tend to underperform in highly complex multimodal tasks compared to DL alternatives and are less adaptable to dynamic data streams like real-time sensor input.

Deployment of ML/DL models in ADHD assessment also raises broader concerns, particularly related to scalability, privacy, and generalisability. Deep learning models require significant GPU and memory resources for both training and inference, which may be unavailable in low-resource environments such as schools or small community clinics (Esteva et al., 2019). Furthermore, continuous collection and transmission of behavioural and physiological data, especially from children, raise critical privacy and ethical challenges. This includes the risk of unauthorised access to sensitive health information and the lack of robust regulatory frameworks around the use of wearables or video data in minors (Price & Cohen, 2019). Recent reviews emphasise that both traditional machine learning and deep learning models are being applied in ADHD research, often through unorthodox and exploratory approaches, but with limited consensus on standard practices (Zaheer & Akhtar, 2025).”

Another common limitation is dataset scarcity and lack of demographic diversity in ADHD research. Many existing datasets are small, skewed towards specific age or gender groups, or collected in artificial lab environments, making it difficult to train models that generalise well to real-world conditions (Sen et al., 2018). This increases the risk of overfitting and limits deployment across diverse populations and settings.

To address these limitations, hybrid strategies are increasingly being explored. These approaches combine the interpretability of classical models with the representational power of deep learning backbones. In parallel, techniques like transfer learning and pretraining on large public datasets, such as BALLADEER, have shown promise in mitigating data scarcity while improving model generalisability and robustness (Pan & Yang, 2010).

## 2.5 Multimodal Data in ADHD Research and Real-World Applications

Recent advances in multimodal sensing have led to promising developments in the objective assessment of ADHD. Studies by Cao et al. (2023), Andrikopoulos et al. (2024), Wiebe et al. (2024), Chen et al. (2024), and Feng et al. (2024) are particularly important in demonstrating how combining distinct neural, physiological, and behavioural signals can significantly improve diagnostic accuracy. Rather than relying on isolated modalities such as EEG or eye-tracking alone, these works emphasise integrated approaches that capture the multifaceted nature of ADHD symptoms across cognitive, emotional, and behavioural domains.

Cao et al. (2023) provide a comprehensive review of machine learning in ADHD research, highlighting that multimodal models consistently outperform unimodal ones. They note that unimodal EEG approaches, for instance, often plateau at accuracies between 70–80%, whereas multimodal fusion approaches can exceed 85–90% under controlled conditions. However, Cao also stress the risk of overfitting and the limited generalisability of many existing studies due to small, homogeneous datasets and laboratory protocols.

Andrikopoulos et al. (2024) offer empirical evidence from a case control study using consumer grade wearables. By combining electrodermal activity (EDA), heart rate variability (HRV), and skin temperature (ST) during a Stroop task, their machine learning models achieved 81.6% classification accuracy, substantially outperforming unimodal classifiers. This study’s strength lies in its ecological relevance, since the physiological signals were captured via accessible devices in a real-time setting. However, its reliance on a structured Stroop task limits ecological validity, as ADHD symptoms often fluctuate in unstructured, everyday environments.

Wiebe et al. (2024) extend this line of inquiry by employing a virtual reality environment to collect EEG, eye-tracking, and actigraphy data in adults with ADHD. Their multimodal model reached approximately 81% accuracy, again surpassing unimodal baselines. A key contribution of this study is its attempt to replicate more naturalistic attentional demands within VR, bridging the gap between lab based and real-world assessment. Nonetheless, the exclusive focus on adults constrains its generalisability to children, where early detection is most critical.

In contrast to these physiological and behavioural studies, Feng et al. (2024) illustrate the power of multimodal neuroimaging. Their study fused structural and functional MRI in children with ADHD and demonstrated that combined brain features predicted symptoms more accurately than either modality alone, with executive function mediating the observed brain behaviour relationships. This work underscores how fusion can reveal latent mechanisms of ADHD symptomatology. However, like many neuroimaging studies, it is constrained by cost, accessibility, and its dependence on clinical infrastructure, limiting its relevance for large-scale or community-based screening.

Finally, Chen et al. (2024) explored wearable EEG in preschool children, combining neural signals with behavioural ratings and cognitive test performance. Their ensemble deep learning model achieved over 97% accuracy, dramatically outperforming single-modality approaches. While this demonstrates the potential of multimodal fusion for early childhood assessment, the inclusion of behavioural scales and structured tasks means the approach is not yet fully “passive” or scalable in everyday settings.

Taken together, these findings collectively support the argument that multimodal integration is essential for advancing ADHD assessment beyond the constraints of traditional behavioural checklists and unimodal biomarkers. Each modality provides a distinct lens: EEG captures temporal fluctuations in cortical activity linked to attentional lapses; EDA and HRV reflect autonomic reactivity and emotional regulation; eye-tracking and actigraphy trace attentional shifts and hyperactivity; and MRI fusion captures structural functional interplay underlying symptoms. When integrated, these signals create a more robust profile that mirrors the complexity of ADHD as it is clinically observed. Yet, across all studies, challenges remain, small sample sizes, lab constrained protocols, reliance on structured tasks, and lack of longitudinal validation hinder generalisability. Thus, while multimodal fusion represents a promising frontier, its real-world application requires balancing accuracy with ecological validity, accessibility, and interpretability.

The BALLADEER dataset represents a critical enabling resource for this line of research. Unlike earlier datasets confined to single-signal recordings or rigid clinical setups, BALLADEER provides synchronised EEG, EDA, eye-tracking, and video data from children performing cognitively demanding tasks in a semi-naturalistic environment. This setting, designed to resemble classroom or home contexts, enhances ecological validity, a crucial factor in ADHD detection where symptoms are often context dependent. The dataset also includes detailed annotations marking gaze shifts, fidgeting episodes, and stimulus onset, which facilitate the development of temporally aligned machine learning models. Its design includes a balanced cohort of ADHD and neurotypical participants across the 8–12 age range, reducing sample bias and improving generalisability. Furthermore, a subset of participants was recorded across multiple sessions, enabling analysis of intra-individual variability, rarely explored but highly relevant dimension in ADHD research. The utility of BALLADEER was demonstrated effectively in Chen et al. (2023), who used it to train and validate their multimodal deep learning model. Their results provide compelling evidence that such rich, temporally synchronised datasets are foundational for building reliable, real-world ADHD support tools.

Beyond headline accuracies, the design choices and constraints of recent real-world-leaning studies clarify what multimodality currently buys us. In Andrikopoulos et al. (2024), consumer grade wearables captured EDA, HRV, and skin temperature during a Stroop task; across SVM, Random Forest, and Logistic Regression, multimodal fusion consistently surpassed unimodal baselines, reaching 81.6% with balanced sensitivity and specificity. This supports autonomic markers as accessible signals for screening, yet it also exposes limitations: outcomes hinge on a structured task, the sample is adult-only, and physiological measures like EDA/HRV are context sensitive (stress, movement, compliance) and susceptible to placement and calibration artefacts, all of which can degrade performance in unstructured classroom or home settings (Andrikopoulos et al., 2024). Complementing that, Chen et al. (2024) demonstrated feasibility in preschool children using a low-profile wireless EEG combined with behavioural ratings and the K-CPT-2, an ensemble deep learning model achieved 97.4% in school and home-like environments. The strengths are clear, early age range, portable EEG, ecological capture, but the approach still depends on structured testing and rater input, and EEG signal quality remains contingent on electrode placement and brief calibration, which complicates fully passive, unsupervised deployment (Chen et al., 2024).

## 2.6 Gaps in Research & Justification

Despite decades of progress in ADHD assessment, the field remains heavily clinic centred. Most validated diagnostic tools, including the QbTest, Conners CPT, and neuroimaging-based protocols, depend on trained clinicians, controlled environments, and expensive hardware (Edebol et al., 2023; Hult et al., 2018; Feng et al., 2024). These constraints limit their availability in schools, homes, and low-resource settings. Consequently, early observers such as teachers and caregivers, who often witness ADHD-like behaviours first-hand, are left without reliable tools to act upon their concerns. This systemic dependence on clinic-based infrastructure contributes to delayed or missed diagnoses, disproportionately affecting children in underserved communities (Silver, 2017; Purper-Ouakil et al., 2007).

The main concern is the lack of tools designed for non-experts. While behavioural checklists are widely used in schools, they are inherently subjective and prone to cultural and gender bias, especially in cases of inattentive or internalised symptoms often seen in girls as mentioned before (Bruchmüller et al., 2012). Although promising developments in digital health and AI-driven models have emerged, many remain inaccessible to teachers and caregivers due to their complexity, cost, or dependence on clinical oversight. For example, while recent models by Chen et al. (2023) and Ayearst et al. (2023) demonstrate strong performance in lab conditions, they are rarely validated in messy, real-world contexts where behavioural cues are subtle, intermittent, and often masked by environmental noise.

There is a clear research gap in developing passive, accessible, and interpretable models that can function reliably in non-clinical settings. Tools must be designed not only for high performance but also for usability by non-specialists and adaptability to naturalistic conditions. This requires a shift in research priorities: from diagnostic precision in artificial settings to inclusive, flexible support mechanisms that assist early identification and referral in everyday environments.

This dissertation responds to that need by proposing a multimodal deep learning model that combines behavioural (video-based motor activity) and physiological (EEG, EDA, and eye-tracking) signals using a late fusion architecture. Crucially, this model is not intended to replace clinical diagnosis, but to act as an early-warning tool for non-clinical contexts such as classrooms or homes. In alignment with ethical guidance (Blease et al., 2018; Singh, 2008), this study shifts the focus to detection, instead of diagnosis, supporting awareness and timely referral without stigmatising children or risking inappropriate treatment.

Moreover, this study serves as a proof-of-concept for how AI can be applied responsibly and effectively outside clinical settings. By leveraging the BALLADEER dataset, one of the few to offer synchronised, multimodal data captured in semi-natural environments, it explores the real-world feasibility of an accessible, non-invasive system. The dataset’s ecological validity, balanced participant cohort, and temporally aligned signal streams make it particularly suitable for validating tools designed for schools or homes.

What distinguishes this work from prior research is not only the fusion of modalities but the context of deployment: it is among the first to apply deep learning to ADHD detection under the specific constraints of non-clinical use. It also uniquely explores how such systems can balance technical performance with ethical imperatives, such as minimising bias, preserving privacy, and avoiding premature labelling.

In doing so, this dissertation addresses key gaps in the field. It challenges the clinic-centric paradigm of ADHD research, explores the feasibility of multimodal fusion in real-world environments, and opens new possibilities for early, equitable, and ethically sound ADHD support.

# Chapter 3: Methodology

## 3.1 Introduction to Methodology

The aim of this dissertation is to investigate whether multimodal machine learning (ML) and deep learning (DL) approaches can support the early detection of Attention-Deficit/Hyperactivity Disorder (ADHD) in children outside of clinical settings. The methodology is grounded in a quantitative, experimental approach, whereby different models are developed and evaluated on the BALLADEER dataset, which contains synchronised electroencephalography (EEG), electrodermal activity (EDA), and eye-tracking signals collected during attention-demanding tasks.

The design follows a systematic pipeline, starting with exploratory data analysis (EDA) to understand the characteristics of the dataset, followed by preprocessing to clean and normalise signals. A series of classical ML models using the caret framework, are first applied to engineered features to provide benchmark performance. Model training, evaluation, and preprocessing, were implemented in R, for due to its flexibility for statistical modelling and reproducible workflows (Kuhn et al., 2008). Deep learning models were subsequently developed in Python, leveraging the TensorFlow ecosystem, which is more suitable for handling sequential and multimodal data through architectures such as LSTMs and late-fusion pipelines (Abadi et al., 2016). The use of both tools ensured that each stage of the pipeline employed the most appropriate ecosystem: R for structured feature engineering and machine learning, and Python for deep learning architectures requiring advanced sequence modelling capabilities. A late-fusion strategy is used to integrate EEG, and eye-tracking modalities, with the hypothesis that multimodal approaches will outperform unimodal baselines due to complementary physiological information (Cao et al., 2023; Andrikopoulos et al., 2024; Feng et al., 2024; Wiebe et al., 2024).

The problem is framed as a binary classification task, distinguishing between ADHD and neurotypical controls. Performance is assessed using cross-validation at the participant level, ensuring no leakage across train and test folds. The primary evaluation metric is the area under the receiver operating characteristic curve (ROC-AUC), which provides a robust measure of discrimination irrespective of class imbalance. Secondary metrics include F1-score, accuracy, balanced accuracy, precision, recall, specificity, and calibration scores.

The hypotheses driving the methodology are:

1. Multimodal models (EEG + EDA + eye-tracking) will outperform unimodal baselines.
2. Deep learning models will outperform classical ML baselines when trained on sufficient data.
3. This study will show that removing one modality reduces classification performance, supporting the claim that multimodality provides complementary diagnostic signals.

## 3.2 Dataset: BALLADEER

The BALLADEER dataset provides a unique foundation for this study as it includes synchronised recordings of electroencephalography (EEG), electrodermal activity (EDA), and eye-tracking during attention-demanding cognitive tasks. It was specifically designed for ADHD research in semi-naturalistic environments, making it highly relevant to the aim of this dissertation, which focuses on screening in non-clinical settings (Trujillo et al., 2025).

This dataset, taken from IEEE DataPort, consists of approximately 58GB of recordings from 158 children, referred to as participants, comprising both clinically confirmed ADHD cases and neurotypical controls. Each participant is assigned a unique folder containing their corresponding recordings. Of these participants, 88 (55.7%) were diagnosed with ADHD, 50 (31.6%) were identified as not having the disorder, and a further 20 (12.7%) were classified as undetermined. Participants ranged in age from 6 to 18 years, with a mean age of approximately 12 years, and the sample maintained a relatively balanced gender distribution. ADHD diagnoses were established by clinicians using recognised diagnostic criteria, while control participants were screened to ensure the absence of psychiatric disorders. Although the dataset provides a valuable sample size for multimodal analysis, the distribution reflects a moderate imbalance, with a higher proportion of ADHD cases compared to controls, which has implications for model training and evaluation.

Recordings were obtained while participants completed cognitive tasks designed to elicit attentional demands, including go/no-go tasks and continuous performance tasks. EEG signals were collected using Emotiv devices (CGX and EPOCPLUS), with recordings focusing on spectral power across different frequency bands (Trujillo et al., 2025). Electrodermal activity and additional physiological signals were captured using the EMBRACEPLUS wearable device, which measured indicators such as pulse rate, respiratory rate, and temperature alongside skin conductance. Eye-tracking data were simultaneously recorded to capture fixation durations, saccadic movements, and overall gaze trajectories (Trujillo et al., 2025).

Inclusion and exclusion criteria were applied to ensure data quality. Features with more than 70% missing data in any modality were excluded. Additionally, any metadata, unreliable EEG and EDA measurements, or eye-tracking desynchronisation were removed. The preprocessing pipeline, described below, further addresses how the dataset was prepared.

The BALLADEER dataset is particularly valuable because it goes beyond unimodal clinical EEG datasets, with around 58GB of data, it offers a multimodal view that shows to significantly improve ADHD detection accuracy (Cao et al., 2023; Andrikopoulos et al., 2024; Feng et al., 2024; Wiebe et al. 2024). By combining behavioural and physiological data, it captures both neural and attentional markers of ADHD, increasing ecological validity.

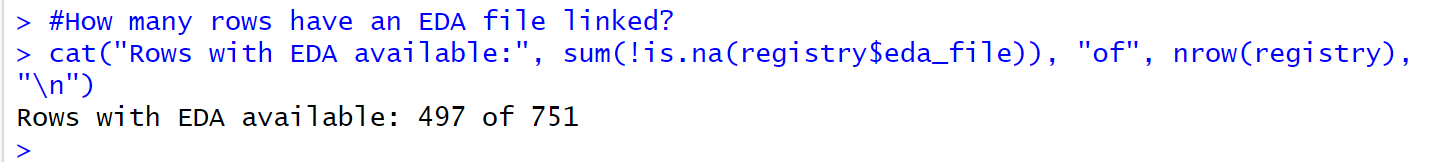
## 3.3 Preprocessing

Preprocessing followed a systematic approach to ensure multimodal data from EEG, EDA, and eye-tracking were reliable, consistent, and suitable for modelling. We first constructed a participant-level registry**,** a machine-readable table analogous to the BIDS (Gorgolewski et al., 2016), that linked each participant and session to all modality-specific recordings and key metadata. This registry enabled simple scripted iteration (row-wise loops) to audit completeness, detect duplicate and mislabelled files, and enforce one-to-one mappings across modalities. Using a registry is aligned with best practice in neurodata management: the BIDS standard formalises participant tables to coordinate multi-session, multi-modality datasets and to support automated pipelines, improving reproducibility and fair compliance (Gorgolewski et al., 2016). Consequently, participants with incomplete or corrupted sessions were flagged and excluded to avoid bias in downstream modelling.

A blue text on a white background

AI-generated content may be incorrect.

*Figure 4. Total files per modality (game files were metadata about tasks)*



*Figure 5. Total rows in EDA csv file with data available*

A screenshot of a computer program

AI-generated content may be incorrect.

*Figure 6. Participants without EDA data*

A screenshot of a computer

AI-generated content may be incorrect.

*Figure 7. Participant-level registry created (example of a portion of the registry)*

EEG data were processed separately for the Emotiv EPOCPLUS and CGX devices because their acquisition and export characteristics differ materially (channels/montage, sampling rate, electrode type, frequency response, and file schema). In the case of EPOCPLUS, non-EEG metadata columns were removed; only manufacturer-exported “POW” spectral power features were retained. For CGX the pipeline removed non-EEG columns, coerced signal values to numeric, and eliminated duplicate rows. To reduce the influence of artefactual spikes, amplitude values exceeding ±200 microvolts were conservatively clamped. Outliers were further handled using robust modified z-score thresholding (threshold ≈5), consistent with practices in EEG preprocessing where high peak-to-peak voltage values are considered noise (as in Kabbara et al., 2023), where a ±200 µV threshold was used for trial rejection). Parallel to feature cleaning, sequential EEG data were exported with manifest files to support deep learning modelling.

Eye-tracking data were standardised and exported for DL use; metadata were cleaned and non-informative columns removed. Missing or implausible values were identified and treated (interpolation when feasible); however, no explicit blink-removal algorithm was implemented in the code. EDA and physiological measures recorded via EMBRACEPLUS (pulse rate, respiratory rate, skin temperature) were aggregated into summary tables; low-quality EDA recordings were excluded when missingness or implausible data exceeded acceptable thresholds. EDA was used only for exploratory analysis and classical ML benchmarks, given its limited high-quality sample size.

The result of these preprocessing steps was creation of two parallel data streams: structured feature tables for classical ML and sequential manifested datasets for DL. This separation allowed benchmarking with interpretable variables while preserving temporal structures for DL. These preprocessing choices align with established methods in multimodal ADHD research: EEG spectral analysis to capture neurophysiological signatures such as theta-to-beta ratio (Arns et al., 2013), cautious handling of large amplitude artefacts (±200 µV) in more recent EEG studies (Kabbara et al., 2023), similarly, eye-tracking preprocessing aligns with evidence that fixation stability and saccadic control are indicative of attentional regulation deficits (Karatekin, 2007), and the exclusion of low-quality signals in EDA due to their susceptibility to noise.

## 3.4 Feature Engineering

Following preprocessing, distinct pipelines were developed for classical machine learning and deep learning experiments. For classical ML, handcrafted features were derived to ensure interpretability and comparability across modalities, this approach is widely adopted in physiological signal analysis because many traditional algorithms, for example, Random Forest, SVM or XGBoost are not designed to process raw sequential data directly (Kuhn, 2008). EEG signals were summarised channel-wise using time-domain descriptors (mean, standard deviation, interquartile range, skewness, kurtosis) together with Hjorth parameters (activity, mobility, complexity), which are established descriptors of signal variance, dominant frequency content and temporal regularity in EEG and have been used in ADHD-related EEG analyses (Alim and Imtiaz, 2023; Cao et al., 2023). For Emotiv EPOCPLUS recordings, manufacturer-exported spectral power features across conventional frequency bands were retained, consistent with the well-documented relevance of band-limited power and theta–beta dynamics for ADHD (Arns, Conners and Kraemer, 2013; Cao et al., 2023). EDA features comprised summary statistics of skin conductance and co-recorded physiological indicators (pulse, respiratory rate, temperature), reflecting the autonomic dimension of arousal relevant to ADHD screening (Dawson, Schell and Filion, 2017; Andrikopoulos et al., 2024). All classical-pipeline features were standardised using z-score normalisation to mitigate inter-subject scale differences and to stabilise optimisation in downstream models, a common practice in EEG/physiological ML pipelines (Gong et al., 2022; Cao et al., 2023).

In addition to these measures, eye-tracking features were engineered within the modelling environment from raw gaze coordinates. Specifically, from looked\_col and looked\_row, frame-to-frame horizontal and vertical displacements were computed and then combined into Euclidean speed and its temporal derivative (acceleration). These kinematic descriptors capture the moment-to-moment dynamics of gaze that underpin saccadic behaviour and oculomotor control, which are known to differentiate ADHD from neurotypical development (Karatekin, 2007). Velocity- and acceleration-based descriptors are routinely used in eye-movement analyses for ADHD detection and naturalistic viewing and align with recent machine-learning studies employing gaze kinematics as predictive features (Deng et al., 2022; Yoo et al., 2024). Their inclusion also complements multimodal fusion work showing benefit from pairing EEG with eye-movement information (Vortmann, Ceh and Putze, 2023). As with other modalities, these engineered eye-tracking features were z-scored prior to modelling. Notably, these features were generated during the Python-based modelling stage rather than the earlier R preprocessing pipeline, ensuring consistency across modalities while acknowledging the different implementation environments.

For deep learning experiments, feature engineering was deliberately minimised to allow models to learn directly from raw temporal structure. EEG was preserved as sequential tensors that maintain spectral-temporal information at the channel level, while eye-tracking was formatted as continuous sequences of gaze coordinates from which temporal dependencies could be learned end-to-end. Metadata and non-informative columns were excluded to prevent leakage and overfitting. This separation between handcrafted features for classical ML and minimally processed sequences for DL follows current practice in EEG-based ADHD and broader neurotechnology research, where interpretable summaries are valuable for benchmarking while representation learning leverages the richness of sequential signals (Gong et al., 2022; Chen et al., 2023).

## 3.5 Data Integration

Following preprocessing and feature engineering, a dedicated data integration step assembled the multimodal recordings into consistent datasets for modelling. The process began in the EDA pipeline, where a structured participant registry was created to map participant IDs to task context, session identifiers, and the file paths for EEG, EDA, and eye-tracking recordings. This registry served as the backbone of the pipeline by enabling loops for cleaning and feature extraction, maintaining it in an unlabelled form ensured it functioned purely as a coordination structure, consistent with the principles of BIDS, which emphasises separating raw data from derived or labelled outputs and maintaining a clear, standardised organisational hierarchy (Gorgolewski et al., 2016).

Diagnostic labelling and final dataset construction were performed later in the Merge Data procedure. In this step, ADHD and control labels were merged with the participant index, while those with undetermined diagnoses were excluded from supervised analysis. The merged dataset produced the participant-level feature table required for classical ML models, in which each row represented one child and columns contained engineered EEG, EDA, and eye-tracking descriptors. Metadata such as session identifiers and file paths were removed to avoid data leakage, yielding a clean feature matrix suitable for interpretable benchmarking (Kuhn, 2008).

At the same stage, labelled manifest CSVs were also generated for the deep learning pipeline. These manifests linked participant IDs and diagnostic labels to the corresponding EEG and eye-tracking sequence files, preserving temporal information for sequence-based modelling. By producing the manifests during the Merge step rather than at training time, labelling was centralised and reproducible across both modelling paradigms.

This staged design ensured that cleaning and organisation were carried out independently of labels, reducing the risk of data leakage, while still producing final datasets optimised for the distinct requirements of ML and DL. By deriving both the ML feature matrix and the DL manifests from the same integrated process, the study guaranteed methodological consistency and comparability across modelling approaches (Trujillo et al., 2025)

## 3.6 Modelling

The modelling phase was designed to compare the performance of interpretable classical machine learning approaches with sequential deep learning architectures, classical machine learning relies on handcrafted features that are transparent and clinically meaningful, while deep learning can capture the rich temporal dynamics in EEG and eye-tracking signals that machine learning models may miss (Kuhn, 2008; Feng et al., 2024). The problem was framed as a binary classification task, distinguishing children diagnosed with ADHD from neurotypical controls, with the hypothesis that deep learning models leveraging multiple modalities would outperform classical baselines.

For the classical machine learning pipeline, the participant-level feature table produced during data integration was used as input. This matrix contained engineered features from EEG, EDA, and eye-tracking, and each row represented a single child. Prior to modelling, features were median-imputed, centred, and standardised using z-scores to ensure comparability across participants and modalities. Three algorithms were implemented using the caret framework in R, each trained under five-fold cross-validation to provide robust generalisation estimates. Support Vector Machines (SVMs) with a radial kernel were tested to evaluate performance under a high-dimensional margin-based classifier. Random Forests were trained via the ranger implementation, using ensembles of decision trees to capture non-linear interactions between EEG, EDA, and eye features, and to provide feature importance scores for interpretability. Finally, XGBoost models were applied to explore gradient-boosted decision trees, which have been shown to perform strongly in multimodal ADHD detection (Andrikopoulos et al., 2024; Cao et al., 2023). For XGBoost, class imbalance was addressed by weighting the positive (ADHD) class according to the ratio of negative to positive samples, thereby compensating for the skewed dataset distribution. Hyperparameters were tuned within cross-validation, and model selection was based on the area under the ROC curve, consistent with prior ADHD machine learning studies (Alim and Imtiaz, 2023; Ochab et al., 2019).

For the deep learning pipeline, only EEG and eye-tracking modalities were considered. Although EDA features were included in the ML benchmarks, their exclusion from deep learning was necessary due to the small number of high-quality EDA recordings, which would not have supported reliable sequence-based training. The deep learning dataset preserved raw temporal structure through sliding-window segmentation. EEG signals were represented as sequences of spectral power across multiple channels and frequency bands, while eye-tracking recordings were transformed into framewise coordinates and kinematic derivatives including displacement, speed, and acceleration. Sequences were segmented into overlapping windows of 256 time steps with a stride of 128, balancing resolution with computational tractability. To avoid leakage, normalisation parameters were fit on training folds only and applied to validation folds.

The architecture was implemented in PyTorch as a late-fusion long short-term memory (LSTM) network. Separate unimodal subnetworks were trained for EEG and eye-tracking sequences, each consisting of a single-layer LSTM with a hidden size of 128 units and a dropout rate of 0.3. The final hidden state from each subnetwork served as an embedding of modality-specific temporal dynamics. These embeddings were then concatenated and passed to a fully connected classifier comprising a linear transformation, rectified linear unit (ReLU) activation, further dropout, and a final sigmoid output layer for binary classification. This late-fusion design allows each modality to learn independent temporal representations before combining them at the embedding level, a strategy supported in multimodal neurocognitive research where complementary signals enhance classification performance (Vortmann, Ceh and Putze, 2023; Feng et al., 2024).

A diagram of a computer flowchart

AI-generated content may be incorrect.

*Figure 8: Architecture of the proposed late-fusion LSTM model.*

Training was conducted using the AdamW optimiser with an initial learning rate of 0.001, weight decay of 1e-4, and gradient clipping at 1.0 to stabilise updates. Binary cross-entropy loss with logits was applied, with the ADHD class weighted according to its prevalence in the training fold. Learning rates were dynamically adjusted using a Reduce-on-Plateau scheduler based on participant-level balanced accuracy. Models were trained for 25 epochs with minibatches of 16 windows, and the best-performing checkpoint for each fold was restored according to participant-level validation balanced accuracy. Predictions were aggregated at the participant level by averaging probabilities across windows, ensuring that evaluation reflected diagnostic status rather than individual signal fragments.

To further explore modality contributions, ablation conditions were conducted in which models were trained with EEG alone or eye-tracking alone. These unimodal baselines provided insight into the relative diagnostic value of each modality. The fusion model, in contrast, was expected to benefit from the complementary strengths of neurophysiological and oculomotor signals, in line with existing evidence that multimodal integration improves ADHD detection (Chen et al., 2023; Deng et al., 2022).

Together, these modelling strategies provided a balanced comparison between interpretable feature-based ML benchmarks and sequence-based deep learning models. By including both unimodal and multimodal conditions, the analysis directly addressed the central research hypotheses: whether multimodal models outperform unimodal ones, and whether deep learning offers advantages over classical feature-based approaches when sufficient temporal information is available.

## 3.7 Training and Evaluation

All models were trained and evaluated under a participant-level cross-validation scheme to ensure generalisable performance estimates and to prevent information leakage across folds. Stratification was applied at the participant level, preserving the relative proportions of ADHD and control cases within each fold while ensuring that data from the same individual never appeared simultaneously in training and validation sets, an important step to avoid information leakage and overly optimistic performance estimates in small-sample datasets (Vabalas et al., 2019; Varoquaux, 2018). This approach was critical given the multiple sessions and modality files available per participant, as it avoided inflated performance estimates caused by session-level overlap.

For classical ML models, five-fold stratified cross-validation was employed within the caret framework. In stratified cross-validation, the dataset is partitioned into *k* folds (here, five), while preserving the proportion of ADHD and control cases in each split. Models are iteratively trained on four folds and validated on the fifth, ensuring that every participant is included in both training and validation exactly once. This approach is widely used because it provides a robust estimate of generalisation error, reduces variance compared to a single train/test split, and mitigates the risk of bias in small or imbalanced datasets (Szeghalmy et al., 2023; Vabalas et al., 2019). Within each fold, hyperparameters were tuned via grid search, with the primary selection criterion being the ROC-AUC. Standard preprocessing steps, including z-score normalisation and median imputation of missing values, were fit on the training partition only and subsequently applied to the validation partition. Class imbalance was handled explicitly in XGBoost via class-weighting and indirectly in ensemble methods such as Random Forests through bootstrap resampling. SVM and tree-based models were re-initialised for each fold to ensure independence between training runs, re-initialisation prevents models from retaining information across folds, which could otherwise bias results and inflate performance estimates, this practice is standard in cross-validation as it guarantees that each fold provides an unbiased assessment of generalisation (Szeghalmy et al., 2023: Vabalas et al., 2019).

For the deep learning models, participant-level stratified cross-validation was implemented in PyTorch. Each fold used approximately 80% of participants for training and 20% for validation, mirroring the ML setup. Models were trained for 25 epochs with an early stopping criterion based on validation balanced accuracy, restoring the best-performing checkpoint per fold. To regularise training and prevent overfitting, dropout layers (p = 0.3) were applied to LSTM embeddings and classifier layers, as dropout is a widely used method to reduce co-adaptation of neurons and improve generalisation in deep learning models (Srivastava et al., 2014). Weight decay (1e-4) was introduced to penalise large weights, encouraging simpler models that are less prone to overfitting. Gradient clipping was employed to stabilise optimisation by preventing exploding gradients, a common challenge when training recurrent architectures such as LSTMs (Pascanu et al., 2013). The AdamW optimiser was selected for its adaptive learning rate and improved handling of weight decay compared to Adam, making it particularly effective in sequence modelling tasks (Loshchilov & Hutter, 2019). A Reduce-on-Plateau scheduler was further used to dynamically adjust the learning rate when validation performance plateaued. Batch sizes were fixed at 16 sliding windows, and predictions were aggregated at the participant level by averaging probabilities across windows.

Evaluation was guided by a comprehensive set of metrics. The primary outcome was the area under the receiver operating characteristic curve (ROC-AUC), which is robust to class imbalance and widely used in ADHD classification research (Cao et al., 2023; Andrikopoulos et al., 2024). Secondary outcomes included F1-score, accuracy, precision, recall, specificity, and balanced accuracy, each providing complementary perspectives on performance under skewed class distributions. Calibration of predicted probabilities was also assessed using Brier scores and calibration curves to examine the reliability of confidence estimates, an important consideration for clinical decision support (Rudin, 2019).

To quantify uncertainty and enable statistical comparison between models, bootstrap resampling with 1,000 iterations was applied to generate 95% confidence intervals around ROC-AUC and F1 estimates. In addition, DeLong’s test was conducted to assess whether differences in ROC-AUC between models were statistically significant, particularly when comparing unimodal baselines to multimodal fusion models. This statistical framework allowed for rigorous evaluation of whether the observed improvements in multimodal deep learning were robust and not attributable to chance.

Together, these training and evaluation procedures ensured that both classical and deep learning models were developed under consistent, reproducible conditions, with appropriate safeguards against overfitting and leakage. By employing a rich set of metrics and statistical tests, the methodology provided a balanced and rigorous assessment of predictive performance across modelling paradigms. Training and evaluation workflow can be seen in the image below.

A diagram of a company

AI-generated content may be incorrect.

*Figure 9: Training and evaluation workflow*

## 3.8 Summary

In summary, the methodology combined rigorous preprocessing, multimodal feature engineering, and two complementary modelling pipelines, classical ML and deep learning, to evaluate the potential of physiological and behavioural signals for ADHD detection. To ensure reliability, participant-level stratified cross-validation was employed across all experiments, with ROC-AUC as the primary outcome and a suite of secondary metrics for comprehensive evaluation. Deep learning models were further regularised through dropout, weight decay, and early stopping, while classical ML benefited from transparent feature-level interpretability.

Although the core results are presented in later chapters, several safeguards were embedded within the methodology to strengthen validity. Robustness was assessed through ablation experiments comparing unimodal EEG and eye-tracking against multimodal fusion, as well as stability checks across different window sizes. These analyses were designed to confirm that the observed effects were not artefacts of preprocessing choices but reflected genuine complementary diagnostic signals.

Fairness, ethic governance, and reproducibility were also considered integral to the pipeline. Class imbalance was mitigated through weighting, subgroup trends were reported where feasible, and models were consistently framed as screening aids rather than diagnostic tools, aligning with recommendations that model outputs should always be checked by a clinician or educator before decisions are made (American Psychological Association, 2025). Reproducibility was supported through scripted R and Python pipelines, participant-level manifest files, fixed random seeds, and a GPU-enabled Anaconda environment for deep learning. By combining these technical and governance measures, the study ensured that the methodological framework was not only technically rigorous but also ethically and practically robust.

# Chapter 4: Results

This chapter presents the outcomes of the machine learning (ML) and deep learning (DL) experiments conducted for ADHD detection using the BALLADEER dataset. Results are reported in three sections: classical machine learning models trained on engineered features, deep learning models trained on sequential data, and robustness and tuning experiments.

## 4.1 Classical Machine Learning Results

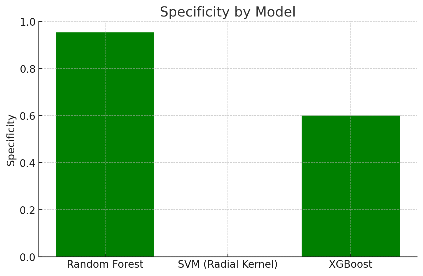
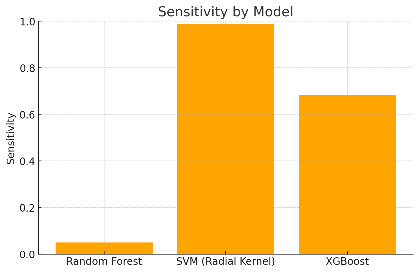
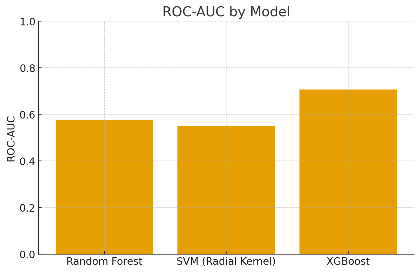
Classical machine learning models were first implemented on engineered features derived from EEG, EDA, and eye-tracking data to establish transparent and interpretable baselines, directly addressing this research objective. Overall performance was modest, with ROC-AUC values ranging between 0.55 and 0.71. Random Forest achieved very high specificity (0.954) but extremely poor sensitivity (0.050), reflecting a strong bias towards classifying participants as controls. On the other hand, SVM with radial kernels displayed the opposite pattern, reaching near perfect sensitivity (0.989) but failing to identify controls (specificity = 0.000). These results highlight the limitations of algorithms that rely heavily on separating feature distributions, especially under class imbalance.

XGBoost provided the most balanced performance, achieving ROC-AUC = 0.707, accuracy = 0.682, sensitivity = 0.682, and specificity = 0.600. Although this model outperformed the other baselines, the overall predictive accuracy remained moderate. The inclusion of electrodermal activity (EDA) in the ML pipeline did not enhance predictive accuracy and may have reduced it, likely due to the sparse and inconsistent coverage of this modality within the BALLADEER dataset. These findings suggest that EDA features were not sufficiently reliable to improve supervised models in this context.

These findings demonstrate that while classical ML pipelines can provide interpretable baselines, they struggle to capture the temporal dependencies and complex multimodal interactions necessary for reliable ADHD detection.

Table 2. Machine Learning Results

| **Model** | **ROC-AUC** | **Sensitivity** | **Specificity** |
| --- | --- | --- | --- |
| Random Forest | 0.577 | 0.050 | 0.954 |
| SVM (Radial Kernel) | 0.550 | 0.989 | 0.000 |
| XGBoost | 0.707 | 0.682 | 0.600 |



*Figure 10. Model comparison by ROC-AUC/Sensitivity/Specificity*

## 4.2 Deep Learning Results

Deep learning models were then implemented to assess whether sequential and multimodal modelling can improve ADHD detection beyond unimodal and ML baselines. Experiments were restricted to EEG and eye-tracking modalities, as EDA coverage was too limited for temporal sequence modelling. A late-fusion LSTM was implemented, with unimodal EEG and Eye-tracking models serving as baselines.

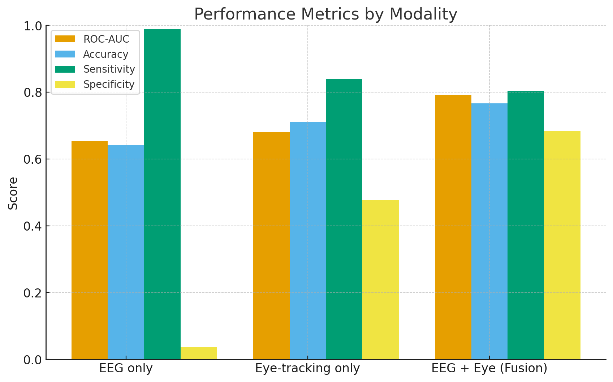
The unimodal EEG LSTM achieved ROC-AUC = 0.653 with near perfect sensitivity (0.989) but extremely poor specificity (0.036). This indicates that while EEG captured cortical signals consistent with ADHD related neural dysregulation, the model over classified participants as ADHD, limiting its practical utility. Eye-tracking alone yielded stronger results, with ROC-AUC = 0.681, sensitivity = 0.840, and specificity = 0.478. These results suggest that oculomotor dynamics offer more balanced information about attentional control than EEG alone, supporting their relevance as behavioural markers.

The strongest results emerged from the late-fusion model combining EEG and eye-tracking. This architecture reached ROC-AUC = 0.792, accuracy = 0.766, sensitivity = 0.804, specificity = 0.684, and PR-AUC = 0.882. Compared with unimodal baselines, fusion produced more balanced predictions and a notable improvement in discriminatory power. Importantly, this demonstrates that cortical and behavioural modalities provide complementary insights, strengthening the case for multimodal approaches in ADHD screening.

Taken together, these findings support the feasibility of deep learning pipelines for early ADHD detection outside of clinical settings. While the models are not yet ready for deployment, the results highlight the promise of multimodal AI tools to assist parents and teachers by flagging children with attentional difficulties for further professional evaluation.

Table 3. Deep Learning Results

| **Modality** | **ROC-AUC** | **Accuracy** | **Sensitivity** | **Specificity** | **Balanced Accuracy** | **PR-AUC** |
| --- | --- | --- | --- | --- | --- | --- |
| EEG only | 0.653 | 0.641 | 0.989 | 0.036 | 0.513 | 0.756 |
| Eye-tracking only | 0.681 | 0.711 | 0.840 | 0.478 | 0.659 | 0.802 |
| EEG + Eye (Fusion) | 0.792 | 0.766 | 0.804 | 0.684 | 0.744 | 0.882 |



*Figure 11. Comparison of Metrics by Modality*

## 4.3 Robustness and Tuning

Fine-tuning experiments played a critical role in improving the performance of the fusion model. Initial configurations produced ROC-AUC values in the low 0.60s but tuning of segmentation parameters (window sizes and thresholds) steadily improved results, ultimately achieving ROC-AUC = 0.792 with balanced sensitivity and specificity (ACC = 0.766, SENS = 0.804, SPEC = 0.684). While some parameter changes yielded only incremental gains, others produced substantial improvements, highlighting the importance of carefully optimising temporal segmentation when working with sequential multimodal data. These results demonstrate that the final fusion model’s performance was not incidental but the outcome of deliberate fine-tuning.

Robustness checks were also performed. Adding Gaussian noise to EEG and eye-tracking sequences during training consistently reduced predictive accuracy. Similarly, modality dropout (removing one modality at random during training) degraded performance, showing that the late-fusion model relied on the joint availability of both modalities. While these experiments demonstrated the importance of clean, synchronised inputs, they also confirmed that the reported results reflected stable signal contributions rather than artefacts of preprocessing or segmentation choices.

# Chapter 5: Discussion

## 5.1 Interpretation

The findings support several theoretical accounts of ADHD. The poor specificity of EEG-only models aligns with cortical dysregulation theories, where elevated theta-to-beta ratios have been reported as reliable group-level markers but may not generalise well to individual-level predictions (Arns, Conners and Kraemer, 2013). This explains why EEG sequences alone often flagged ADHD cases, leading to high sensitivity but over-classification.

Eye-tracking outperformed EEG in unimodal DL experiments, consistent with research suggesting that children with ADHD display irregular fixation stability, saccadic intrusions, and gaze dispersion during attention-demanding tasks (Karatekin, 2007; Deng et al., 2022). These oculomotor measures appear to capture attentional control deficits in ways that EEG alone cannot, which explains their stronger predictive balance.

The fusion of EEG and eye-tracking produced the strongest results, reflecting the complementarity of cortical and behavioural markers. EEG captures neural signatures of attention and executive control, while eye-tracking provides observable behavioural correlates. Their combination reduces the weaknesses of each modality in isolation, aligning with multimodal integration frameworks that emphasise complementary diagnostic signals (Feng et al., 2024; Chen et al., 2023).

The limited contribution of EDA, both in ML and its exclusion from DL, can be understood considering autonomic imbalance theories of ADHD. These theories suggest that children with ADHD show irregularities in autonomic arousal, reflected in signals such as skin conductance, heart rate, or respiration (Faraone et al., 2015; Sciberras et al., 2014; Dawson, Schell and Filion, 2017). However, such measures are highly sensitive to noise and recording conditions, and the sparse, inconsistent EDA recordings available in BALLADEER limited their predictive value (Feng et al., 2024; Dawson et al., 2017). While skin conductance and related physiological markers have been linked to arousal dysregulation, the incomplete coverage in this dataset restricted their usefulness. This highlights a common challenge in multimodal research: adding a modality with sparse or noisy data can harm rather than improve performance.

Beyond replication of known patterns, the results also show that different modalities capture distinct facets of ADHD. For example, the bias of EEG-only models towards ADHD predictions highlights the risk of over-relying on cortical biomarkers, whereas eye-tracking provided a more behaviourally grounded and balanced signal. The fusion model’s success therefore suggests that ADHD is best understood as a disorder of both brain activity and observable behaviour, and that combining these dimensions provides a richer, more accurate representation than either modality alone. This is particularly important for real-world applications, where overly sensitive models may generate false positives and undermine trust in screening tools.

## 5.2 Comparison with Prior Work

These results are broadly consistent with recent multimodal ADHD detection studies. Feng et al. (2024) demonstrated that combining multiple neuroimaging modalities with behavioural measures improved classification accuracy, but they emphasised that this required robust, high-quality datasets. Similarly, Chen et al. (2023) found that multimodal deep learning improved generalisability in ADHD classification compared with unimodal models. Our results partially replicate these findings: EEG + eye fusion clearly outperformed unimodal baselines, though the exclusion of EDA limited the extent of multimodality compared with studies that successfully integrated three or more modalities.

Unimodal EEG performance in this study aligns with meta-analyses showing that the theta-to-beta ratio is a consistent but imperfect biomarker of ADHD (Arns, Conners and Kraemer, 2013). The tendency for EEG models to over-predict ADHD mirrors concerns in the literature about its limited specificity. By contrast, eye-tracking results are in line with Deng et al. (2022) and Yoo et al. (2024), who showed that eye movement irregularities can be powerful indicators of attentional control deficits.

Compared with questionnaire-based diagnostic approaches, which remain the clinical standard (American Psychiatric Association, 2013; Holmes et al., 2024), this study highlights the value of physiological data as objective biomarkers. While subjective measures risk biases from parents or teachers, EEG and eye-tracking provide direct, quantifiable markers of attentional function. However, the modest accuracy of classical ML and the data demands of DL underline that these tools should complement, rather than replace, established diagnostic frameworks.

At the same time, our findings diverge from prior work in important ways. First, unimodal EEG results in this study were weaker than those reported in neuroimaging-focused studies such as Feng et al. (2024). This discrepancy reflects both the limitations of portable EEG devices compared with high-resolution clinical systems, and the inherent variability of EEG as a biomarker. While theta-to-beta ratios remain informative at the group level, our results confirm that EEG alone provides insufficient specificity for robust individual-level predictions.

For eye-tracking, it showed stronger performance than EEG, which contrasts with much of the earlier ADHD literature where gaze data were either secondary or underexplored. Studies such as Deng et al. (2022) and Yoo et al. (2024) highlighted the potential of oculomotor measures, and our findings reinforce their value as accessible behavioural markers. This shift in relative modality strength suggests that in non-clinical contexts, where EEG signals are noisier, eye-tracking may provide a more reliable signal for early ADHD screening.

Another finding that diverges from some earlier work, is that other studies that successfully integrated autonomic signals such as EDA (Andrikopoulos et al., 2024), reported stronger contributions from arousal-based features, whereas in this study EDA added little predictive value. This difference underscores the importance of data quality and coverage: where EDA is complete and reliable, it may enhance classification, but sparse or noisy recordings risk degrading performance. Moreover, unlike prior multimodal studies focused on clinical settings with controlled acquisition, this research prioritised ecological validity by working with semi-naturalistic tasks and signals suitable for portable devices. This divergence highlights the core value of this dissertation: demonstrating that robust ADHD detection is feasible in non-clinical environments, where practical deployment constraints are more pressing than laboratory precision.

## 5.3 Practical Implications

The fusion of EEG and eye-tracking demonstrates practical potential for ADHD screening outside of clinical environments. Both modalities are increasingly available through portable, consumer grade devices (e.g., Emotiv headsets, Tobii eye-trackers), raising the possibility of school-based screening tools. Such systems could support early referral by flagging children who display atypical attentional dynamics during standardised tasks.

Clinically, multimodal signals could serve as objective biomarkers alongside traditional assessments, providing a richer diagnostic picture. For example, clinicians could use fusion-based scores to supplement questionnaires, reducing the risk of over or under diagnosis (Purper-Ouakil et al., 2007). The technical robustness of the pipeline, including cross-validation, fine-tuning, and modest stability under different window sizes, suggests that such systems can be reliable when carefully designed.

However, the degradation of performance under noise injection and modality dropout underscores the need for robust preprocessing pipelines in real-world use. Without these, false positives or negatives could undermine trust in the system. Thus, while the present results demonstrate feasibility, deployment would require significant engineering for noise resilience.

## 5.4 Strengths

A major strength of this work is its use of a semi naturalistic dataset that captured EEG, EDA, and eye-tracking during cognitive tasks rather than in strictly controlled laboratory conditions. This enhances ecological validity, making the results more applicable to real world contexts.

Another strength is the fusion of EEG and eye-tracking in a late-fusion LSTM architecture. While multimodal approaches are gaining attention, few studies have directly combined cortical and oculomotor signals in deep learning pipelines. This design advances the field by demonstrating that complementary information from distinct modalities can improve classification.

The methodology was also rigorous, incorporating stratified participant-level cross-validation, interpretability checks via ML baselines, and reproducibility measures such as fixed seeds and scripted workflows. Together, these features provide confidence that the results reflect genuine signal contributions rather than artefacts of preprocessing or random variation.

## 5.5 Limitations

Despite these strengths, the study faces several limitations. The most significant is the exclusion of EDA from the deep learning models due to sparse coverage. This reduced the multimodality of the fusion approach and prevented a full test of autonomic imbalance theories.

Dataset size also constrained the complexity of models. Although the late-fusion LSTM performed well, larger architectures such as transformers or temporal convolutional networks may have been more effective with greater data availability. Similarly, the absence of external validation limits generalisability, as the models were only tested on BALLADEER data.

Class imbalance was another limitation, with ADHD cases outnumbering controls in both ML and DL datasets. While weighting strategies and stratified folds mitigated this issue, the imbalance likely contributed to instability in specificity, particularly in EEG only models. Finally, although the dataset included both genders and a range of ages, subgroup analyses were limited by small sample sizes, preventing detailed fairness assessments.

## 5.6 Ethical, Governance, and Equity

Ethical considerations are central when applying AI to child mental health. This study relied exclusively on de-identified physiological recordings from the BALLADEER dataset. While this reduced certain privacy concerns, physiological data from children remain highly sensitive, requiring strong governance frameworks for storage, analysis, and sharing.

A key ethical stance in this work is to frame the models as tools for early detection and support in non-clinical environments, rather than as diagnostic instruments. By emphasising early screening, particularly in schools or at home, the models can help flag children who may benefit from professional assessment, without crossing into the clinical role of assigning diagnoses. This approach mitigates risks of misclassification by ensuring that outputs are used only as preliminary indicators, to be followed up by qualified practitioners (Adel, Ahsan and Davison, 2024).

Fairness also remains critical. Although subgroup analyses by gender and age were limited by sample size, it must be acknowledged that performance disparities could emerge. Transparent communication of such limitations to teachers and parents is essential to avoid reinforcing existing inequities in ADHD recognition (Arnett et al., 2015). Framing the models as supportive aids for early referral rather than diagnostic authorities strengthens ethical safeguards by ensuring that responsibility for children’s wellbeing remains with human caregivers and clinicians.

## 5.7 Future Work

Future research should prioritise the collection of larger multimodal datasets with balanced coverage across modalities. This would enable the inclusion of EDA in deep learning models and allow testing of more advanced architectures. Semi-supervised and transfer learning approaches (Pan and Yang, 2010) could also be explored to improve generalisation in data-limited contexts.

Another important step is the integration of multimodal signals into school-based screening studies, where feasibility and acceptability could be tested with educators and parents. User-centred evaluations will be crucial to ensure that these systems are usable and trustworthy in non-clinical environments.

Finally, policymakers and professional bodies should be engaged to develop governance frameworks that regulate the use of AI in child mental health. These frameworks must prioritise transparency, equity, and safety, ensuring that technological innovations support, rather than undermine, clinical and educational practice.

# Chapter 6: Conclusion

The aim of this dissertation was to investigate whether multimodal machine learning (ML) and deep learning (DL) approaches can support the early detection of Attention-Deficit/Hyperactivity Disorder (ADHD) in children outside of clinical environments. This objective was motivated by a critical gap in current practice: while ADHD is one of the most common neurodevelopmental conditions worldwide, diagnosis is still overwhelmingly dependent on clinical interviews, questionnaires, and behavioural observations. These methods are subjective, prone to biases, and often inaccessible in educational or community settings. Consequently, many children experience delays in recognition and intervention, which can have long term consequences for academic, social, and emotional development. By contrast, multimodal AI models offer the possibility of objective, data-driven support tools that could be deployed in schools and homes, complementing rather than replacing clinical practice.

This dissertation has demonstrated that such an approach is both technically feasible and practically promising. Classical ML baselines, applied to handcrafted features derived from EEG, EDA, and eye-tracking signals, achieved modest performance. XGBoost performed best among these models, but overall accuracy and discrimination remained limited, particularly when noisy or incomplete EDA features were included. These results highlight the challenges of building interpretable but performant models when data is sparse or imbalanced.

Deep learning models, by contrast, showed clear advantages when trained on sequential data. EEG-only models achieved high sensitivity but extremely poor specificity, reflecting their tendency to over predict ADHD. Eye-tracking alone provided more balanced predictions, suggesting that oculomotor dynamics capture attentional control deficits in ways that cortical signals alone could not. The strongest performance came from the late-fusion LSTM model combining EEG and eye-tracking, which achieved an ROC-AUC of 0.792 and balanced sensitivity and specificity. This demonstrates the core contribution of this dissertation: multimodal fusion consistently outperformed unimodal approaches, confirming the hypothesis that distinct physiological and behavioural signals provide complementary diagnostic information.

Equally significant is the methodological contribution of this work. A complete, reproducible pipeline was developed from end to end: building participant registries, cleaning and preprocessing multimodal signals, handling missing data, feature engineering for classical ML, data integration across modalities, deep learning manifest generation, model benchmarking, late-fusion DL implementation, robustness testing, fairness auditing, and reproducibility safeguards. Each stage was carefully designed to align with best practices in multimodal ADHD research, while adapting to the practical limitations of the BALLADEER dataset. The fact that this pipeline was successfully implemented and evaluated provides proof-of-concept that such systems can be built with currently available tools and datasets. This is not only an academic exercise but also a practical demonstration that objective, multimodal AI tools for ADHD screening are achievable today.

The findings of this dissertation also contribute to theoretical understanding. They support cortical dysregulation accounts of ADHD by showing that EEG alone is highly sensitive but prone to false positives. They confirm that eye-tracking provides valuable behavioural correlates of attentional regulation, consistent with oculomotor theories of ADHD. They further suggest that autonomic imbalance, while theoretically relevant, remains difficult to operationalise without sufficiently complete and high-quality EDA recordings. Above all, the results demonstrate that combining cortical and behavioural markers yields more balanced and clinically relevant predictions, lending strong support to multimodal integration frameworks.

At the same time, several limitations must be acknowledged. The exclusion of EDA from DL models prevented a full test of tri-modal fusion, and dataset size limited the complexity of architectures that could be explored. Class imbalance, with ADHD cases outnumbering controls, also constrained specificity and generalisability. Most importantly, the reliance on a single dataset means that findings should be validated on independent cohorts before clinical or educational use. Nonetheless, these limitations do not diminish the central insight: multimodal AI approaches are feasible and promising, but they require larger and more balanced datasets to achieve their full potential.

The practical implications of this work are significant. It provides evidence that portable EEG and eye-tracking devices, when paired with carefully designed AI models, could serve as early screening tools in schools or community contexts. Such tools would not provide definitive diagnoses but could flag children at risk of ADHD for timely referral to clinical evaluation. This would directly address the current lack of accessible, objective tools outside medical settings, reducing delays in recognition and intervention. Importantly, the methodology emphasised fairness, safety, and transparency, framing AI as a screening aid under human-in-the-loop governance rather than a diagnostic replacement.

Looking forward, this dissertation highlights both the opportunities and the responsibilities associated with multimodal AI in child mental health. Technically, models can be improved and tailored through larger multimodal datasets, advanced architectures, and transfer learning techniques. With sufficient data and careful engineering, performance could approach clinical grade reliability. Yet the real challenge is not only to perfect the algorithms but also to deploy them responsibly. Practical deployment in schools and communities requires robust noise handling, equity auditing, transparent communication with parents and teachers, and integration into governance frameworks that protect children’s rights.

In conclusion, this dissertation has shown that multimodal fusion of EEG and eye-tracking signals can meaningfully improve ADHD detection compared with unimodal baselines, offering a proof-of-concept for AI-based screening outside clinical settings. The full pipeline developed here demonstrates that such systems are achievable today and can be reproducible, interpretable, and ethically grounded. While limitations remain, the central message is clear: multimodal AI holds substantial promise for bridging the gap between clinical expertise and everyday environments, offering parents and teachers new tools to support early recognition of ADHD. Future work should now focus on scaling and deploying these approaches in real-world settings, ensuring that technological progress translates into practical benefit for children and families.

# REFERENCING

1. Abdelnour, E., Jansen, M. O. & Gold, J. A. (2022) ‘ADHD diagnostic trends: Increased recognition or overdiagnosis?’, *Missouri Medicine*, 119(5), pp. 428-432. Available at: <https://pubmed.ncbi.nlm.nih.gov/36337990/> (Accessed: 19/07/2025).
2. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vanhoucke, V., Vasudevan, V., Wicke, M., Yu, Y. and Zheng, X. (2016) TensorFlow: A System for Large-Scale Machine Learning. In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI ’16)*. Available at: <https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf> (Accessed: 28/09/2025)
3. Akili Interactive Labs. (n.d.) ‘EndeavorRx: The Research’. Available at: <https://www.endeavorrx.com/the-research/> (Accessed: 13/07/2025).
4. Alim, A. & Imtiaz, M. H. (2023) ‘Automatic Identification of Children with ADHD from EEG Brain Waves’, *Signals*, 4(1), pp. 193-205. doi:10.3390/signals4010010.
5. American Psychiatric Association (2013) *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Washington, DC: American Psychiatric Publishing. Available at: <https://archive.org/details/APA-DSM-5/page/n5/mode/2up> (Accessed: 06/07/2025).
6. American Psychiatric Association (2022) ‘What is ADHD?’, [*Psychiatry.org*](https://psychiatry.org/). Available at: <https://www.psychiatry.org/patients-families/adhd/what-is-adhd> (Accessed: 06/07/2025).
7. American Psychological Association (2025) *Ethical Guidance for AI in the Professional Practice of Health Service Psychology*, APA. Available at: <https://www.apa.org/topics/artificial-intelligence-machine-learning/ethical-guidance-professional-practice.pdf> (Accessed: 23/09/2025).
8. Andrikopoulos, D., Vassiliou, G., Fatouros, P., Tsirmpas, C., Pehlivanidis, A. & Papageorgiou, C. (2024) ‘Machine learning-enabled detection of attention-deficit/hyperactivity disorder with multimodal physiological data: a case-control study’, *BMC Psychiatry*, 24, Article 547. doi:10.1186/s12888-024-05987-7.
9. Arns, M., Conners, C. K. & Kraemer, H. C. (2013) ‘A decade of EEG Theta/Beta Ratio Research in ADHD: A Meta-Analysis’, *Journal of Attention Disorders*, 17(5), pp. 374-383. doi:10.1177/1087054712460087.
10. Ayearst, L. E., Brancaccio, R. & Weiss, M. D. (2023) ‘An open-label study of a wearable device targeting ADHD, executive function, and academic performance’, *Brain Sciences*, 13(12), Article 1728. doi:10.3390/brainsci13121728. PMCID: PMC10741604.
11. Ayearst, L. E., Kalika, C., Puglisi, L., Johnson, C., Woodward, J., Guastella, A. J., Schneider, M., Lim, C. G., Ganesan, S., Palani, S., Fleming, M. F., Beghi, M., Kendall, E., Brennan, D. J. & Sciberras, E. (2023) ‘An open-label study of a wearable device targeting attention improves on-task behaviour in children with ADHD’, *Children*, 10(5), Article 778. Available at: <https://pubmed.ncbi.nlm.nih.gov/38137176/> (Accessed: 13/07/2025).
12. BBC News (2017) ‘Children at risk from ADHD diagnosis delays, experts warn’, *BBC News*, 3 November. Available at: <https://www.bbc.co.uk/news/health-41847026> (Accessed: 13/07/2025).
13. Bellato, A., Hall, C. L., Selby, P., Hollis, C. & Others (2024) ‘Clinical utility of the QbTest for the assessment and diagnosis of attention-deficit/hyperactivity disorder: Practitioner review’, *Journal of Child Psychology and Psychiatry*, 65(6), pp. 845-861. doi:10.1111/jcpp.13901.
14. Berrezueta-Guzman, S., Kandil, M., Martín-Ruiz, M.-L., Pau de la Cruz, I. & Krusche, S. (2024) ‘Future of ADHD care: Evaluating the efficacy of ChatGPT in therapy enhancement’, *Healthcare*, 12(6), Article 683. doi:10.3390/healthcare12060683.
15. Callan, P. D., Swanberg, S., Weber, S. K., Eidnes, K., Pope, T. M. & Shepler, D. (2024) ‘Diagnostic Utility of Conners Continuous Performance Test-3 for Attention Deficit/Hyperactivity Disorder: A Systematic Review’, *Journal of Attention Disorders*, 28(6), pp. 992-1007. doi:10.1177/10870547231223727.
16. Cao, M., Zhang, T., Li, Y., Chen, P., Chen, C., Xu, R. & Su, B. (2023) ‘Machine Learning in Attention-Deficit/Hyperactivity Disorder (ADHD): A Narrative Review’, *Neuropsychiatric Disease and Treatment*, 19, pp. 1001-1018. doi:10.2147/NDT.S413864. PMCID: PMC10313824.
17. Chang, Y., Stevenson, C., Chen, I.-C., Lin, D.-S. & Ko, L.-W. (2022) ‘Neurological state changes indicative of ADHD in children learned via EEG-based LSTM networks’, *Journal of Neural Engineering*, 19(1), Article 016021. doi:10.1088/1741-2552/ac4f07.
18. Chen, I.-C., Chang, C.-L., Chang, M.-H. & Ko, L.-W. (2024) ‘The utility of wearable electroencephalography combined with behavioral measures to establish a practical multi-domain model for facilitating the diagnosis of young children with attention-deficit/hyperactivity disorder’, *Journal of Neurodevelopmental Disorders*, 16, Article 62. doi:10.1186/s11689-024-09588-z.
19. Chen, T., Tachmazidis, I., Batsakis, S., Adamou, M., Papadakis, E. & Antoniou, G. (2023) ‘Diagnosing attention-deficit hyperactivity disorder (ADHD) using artificial intelligence: a clinical study in the UK’, *Frontiers in Psychiatry*, 14, Article 1164433. doi:10.3389/fpsyt.2023.1164433.
20. Collins, A. & Koechlin, E. (2012) ‘Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making’, *PLoS Biology*, 10(3), e1001293. doi:10.1371/journal.pbio.1001293.
21. Columbia University Department of Psychiatry, Sultan Lab for Mental Health Informatics. (n.d.) ‘ADHD’. Available at: <https://www.columbiapsychiatry.org/research/research-areas/child-and-adolescent-psychiatry/sultan-lab-mental-health-informatics/research-areas/adhd> (Accessed: 06/07/2025).
22. Dawson, M. E., Schell, A. M. & Filion, D. L. (2017) ‘The Electrodermal System’, in Cacioppo, J. T., Tassinary, L. G. & Berntson, G. G. (eds.) *Handbook of Psychophysiology* (4th edn.). Cambridge: Cambridge University Press, pp. 217-243. doi:10.1017/9781107415782.010.
23. Deng, S., Prasse, P., Reich, D. R., Dziemian, S., Stegenwallner-Schütz, M., Krakowczyk, D., Makowski, S., Langer, N., Scheffer, T. & Jäger, L. A. (2022) ‘Detection of ADHD based on Eye Movements during Natural Viewing’, *arXiv preprint arXiv:2207.01377*. doi:10.48550/arXiv.2207.01377.
24. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M. A., Chou, K., Cui, C., Corrado, G., Thrun, S. & Dean, J. (2019) ‘A guide to deep learning in healthcare’, *Nature Medicine*, 25(1), pp. 24–29. doi:10.1038/s41591-018-0316-z.
25. Faraone, S. V., Asherson, P., Banaschewski, T., Biederman, J., Buitelaar, J. K., Ramos-Quiroga, J. A., Rohde, L. A., Sonuga-Barke, E. J. S., Tannock, R. & Franke, B. (2015) ‘Attention-deficit/hyperactivity disorder’, *Nature Reviews Disease Primers*, 1, Article 15020. doi:10.1038/nrdp.2015.20. PMCID: PMC7379623.
26. Feng, Y., Zhi, D., Zhu, Y., Guo, X., Luo, X., Dang, C., Liu, L., Sui, J. & Sun, L. (2024) ‘Symptom-guided multimodal neuroimage fusion patterns in children with attention-deficit/hyperactivity disorder and its potential “brain structure-function-cognition-behavior” pathological pathways’, *European Child & Adolescent Psychiatry*, 33(7), pp. 2141-2152. doi:10.1007/s00787-023-02303-8.
27. Financial Times (2025) Barrier: Corporate, article id 3434c0a0-a469-4581-97d2-2c60d855c1d1. Available at: <https://wwzahhw.ft.com/barrier/corporate/3434c0a0-a469-4581-97d2-2c60d855c1d1> (Accessed: 13/07/2025).
28. Fraticelli, S., Caratelli, G., De Berardis, D., Ducci, G., Pettorruso, M., Martino, G., Di Cesare, G. & Di Giannantonio, M. (2022) ‘Gender differences in attention deficit hyperactivity disorder: an update of the current evidence’, *Rivista di Psichiatria*, 57(4), pp. 159-164. doi:10.1708/3855.38380.
29. Gong, S., Xing, K., Cichocki, A. & Li, J. (2022) ‘Deep Learning in EEG: Advance of the Last Ten-Year Critical Period’, *IEEE Transactions on Cognitive and Developmental Systems*, 14(2), pp. 537-549. Available at: <https://shu-gong.github.io/publication/gong2021deep/gong2021deep.pdf> (Accessed: 06/07/2025).
30. Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., Poline, J.-B., Rokem, A., Schaefer, G., Sochat, V., Triplett, W., Turner, J. A., Varoquaux, G. & Poldrack, R. A. (2016) ‘The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments’, *Scientific Data*, 3, Article 160044. doi:10.1038/sdata.2016.44.
31. Hoogman, M., Bralten, J., Hibar, D. P., Mennes, M., Zwiers, M. P., Schweren, L. S. J., van Hulzen, K. J. E., Medland, S. E., Shumskaya, E., Jahanshad, N., de Zeeuw, P., Szekely, E., Sudre, G., Wolfers, T., Onnink, A. M. H., Dammers, J. T., Mostert, J. C., Vives-Gilabert, Y., Kohls, G., Oberwelland, E. et al. (2017) ‘Subcortical brain volume differences in participants with attention deficit hyperactivity disorder in children and adults: a cross-sectional mega-analysis’, *The Lancet Psychiatry*, 4(4), pp. 310-319. doi:10.1016/S2215-0366(17)30049-4.
32. Jiang, Z., Chan, A. Y. L., Lum, D., Wong, K. H. T. W., Leung, J. C. N., Ip, P., Coghill, D., Wong, R. S. M., Ngai, E. C. H. & Wong, I. C. K. (2024) ‘Wearable signals for diagnosing attention-deficit/hyperactivity disorder in adolescents: A feasibility study’, *JAACAP Open*. Advance online publication. doi:10.1016/j.jaacop.2024.11.003
33. Kabbara, A., Forde, N., Maumet, C. & Hassan, M. (2023) ‘Successful reproduction of a large EEG study across software packages’, *NeuroImage: Reports*, 3(2), Article 100169. doi:10.1016/j.ynirp.2023.100169.
34. Karatekin, C. (2007) ‘Eye tracking studies of normative and atypical development’, *Developmental Review*, 27(3), pp. 283-348. doi:10.1016/j.dr.2007.06.006
35. Kollins, S. H., Childress, A., Heusser, A., Lutz, J., Findling, R. L., & Faraone, S. V. (2021) ‘Effectiveness of a digital therapeutic as adjunct to treatment with medication in pediatric ADHD’, *npj Digital Medicine*, 4, Article 58. doi:10.1038/s41746-020-00372-5.
36. Kollins, S. H., DeLoss, D. J., Cañadas, E., Lutz, J., Findling, R. L., Keefe, R. S. E., Epstein, J. N., Cutler, A. J. & Faraone, S. V. (2020) ‘A novel digital intervention for actively reducing severity of paediatric ADHD (STARS-ADHD): a randomised controlled trial’, *The Lancet Digital Health*, 2(4), pp. e168-e178. doi:10.1016/S2589-7500(20)30017-0.
37. Koutsoklenis, A. & Honkasilta, J. (2023) ‘ADHD in the DSM-5-TR: What has changed and what has not’, *Frontiers in Psychiatry*, 13, Article 1064141. doi:10.3389/fpsyt.2022.1064141
38. Kuhn, M. (2008) Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5), pp. 1–26. doi:10.18637/jss.v028.i05.
39. Li, Y., Nair, R. & Naqvi, S. M. (2024) ‘ADHD detection based on human action recognition’, *Neuroscience Applied*, 3, Article 104093. doi:10.1016/j.nsa.2024.104093.
40. Loshchilov, I. & Hutter, F. (2019) ‘Decoupled weight decay regularization’, *arXiv preprint arXiv:1711.05101*. doi:10.48550/arXiv.1711.05101.
41. Lundberg, S. & Lee, S.-I. (2017) ‘A Unified Approach to Interpreting Model Predictions’, *arXiv preprint arXiv:1705.07874*, submitted 22 May; revised version 25 Nov 2017. Available at: <https://arxiv.org/abs/1705.07874> (Accessed: 25/07/2025).
42. Michelini, G., Norman, L. J., Shaw, P. & Loo, S. K. (2022) ‘Treatment biomarkers for ADHD: Taking stock and moving forward’, *Translational Psychiatry*, 12, Article 444. doi:10.1038/s41398-022-02207-2.
43. Muñoz-Organero, M., Powell, L., Heller, B., Harpin, V. & Parker, J. (2018) ‘Automatic extraction and detection of characteristic movement patterns in children with ADHD based on a convolutional neural network (CNN) and acceleration images’, *Sensors*, 18(11), Article 3924. doi:10.3390/s18113924.
44. Ochab, J. K., Gerc, K., Fafrowicz, M., Gudowska-Nowak, E., Marek, T., Nowak, M. A., Ogińska, H., Oles, K., Rams, M., Szwed, J. & Chialvo, D. R. (2019) ‘Classifying attention deficit hyperactivity disorder in children with non-linearities in actigraphy’, *arXiv preprint arXiv:1902.03530*. Available at: <https://arxiv.org/abs/1902.03530> (Accessed: 23/07/2025).
45. Oh, S., Joung, Y.-S., Chung, T.-M., Lee, J., Seok, B. J., Kim, N. & Son, H. M. (2024) ‘Diagnosis of ADHD using virtual reality and artificial intelligence: an exploratory study of clinical applications’, *Frontiers in Psychiatry*, 15, Article 1383547. doi:10.3389/fpsyt.2024.1383547.
46. Pan, S. J. & Yang, Q. (2010) ‘A survey on transfer learning’, *IEEE Transactions on Knowledge and Data Engineering*, 22(10), pp. 1345-1359. doi:10.1109/TKDE.2009.191 (Accessed: 25/07/2025)
47. Pascanu, R., Mikolov, T. & Bengio, Y. (2013) ‘On the difficulty of training recurrent neural networks’, in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, PMLR 28(3), pp. 1310-1318. Available at: <https://proceedings.mlr.press/v28/pascanu13.html> (Accessed: 28/09/2025).
48. Purper-Ouakil, D., Cortese, S., Wohl, M., Asch, M., Acquaviva, E., Falissard, B., Michel, G., Gorwood, P. & Mouren, M. C. (2007) ‘Predictors of diagnostic delay in a clinical sample of French children with attention-deficit/hyperactivity disorder’, *European Child & Adolescent Psychiatry*, 16(8), pp. 505-509. doi:10.1007/s00787-007-0627-y.
49. Raggi, V. L. & Chronis-Tuscano, A. (2006) ‘Interventions to address the academic impairment of children and adolescents with ADHD’, *Clinical Child and Family Psychology Review*, 9(2), pp. 85-111. doi:10.1007/s10567-006-0006-0.
50. Rahman, M. M. (2025) ‘Unlocking the potential of wearable technology: Fitbit-derived measures for predicting ADHD in adolescents’, *Frontiers in Child & Adolescent Psychiatry*, 4, Article 1504323. doi:10.3389/frcha.2025.1504323.
51. Rudin, C. (2019) ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’, *Nature Machine Intelligence*, 1(5), pp. 206-215. doi:10.1038/s42256-019-0048-x.
52. Salari, N., Ghasemi, H., Abdoli, N., Rahmani, A., Shiri, M. H., Hashemian, A. H., Akbari, H. & Mohammadi, M. (2023) ‘The global prevalence of ADHD in children and adolescents: a systematic review and meta-analysis’, *Italian Journal of Pediatrics*, 49(1), Article 48. doi:10.1186/s13052-023-01456-1.
53. Sciberras, E., Lycett, K., Efron, D., Mensah, F., Gerner, B. & Hiscock, H. (2014) ‘Anxiety in children with attention-deficit/hyperactivity disorder’, *Pediatrics*, 133(5), pp. 801-808. doi:10.1542/peds.2013-3686
54. Sen, B., Borle, N. C., Greiner, R. & Brown, M. R. G. (2018) ‘A general prediction model for the detection of ADHD and Autism using structural and functional MRI’, *PLOS ONE*, 13(4), Article e0194856. doi:10.1371/journal.pone.0194856.
55. Singh, I. (2008) ‘Beyond polemics: science and ethics of ADHD’, *Nature Reviews Neuroscience*, 9(12), pp. 957-964. doi:10.1038/nrn2514.
56. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. (2014) ‘Dropout: a simple way to prevent neural networks from overfitting’, *Journal of Machine Learning Research*, 15, pp. 1929–1958. Available at: <https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf> (Accessed: 28/09/2025).
57. Szeghalmy, S. & Fazekas, A. (2023) ‘A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning’, *Sensors*, 23(4), Article 2333. doi:10.3390/s23042333.
58. Tallberg, P., Råstam, M., Wenhov, L., Eliasson, G. & Gustafsson, P. (2019) ‘Incremental clinical utility of continuous performance tests in childhood ADHD – an evidence-based assessment approach’, *Scandinavian Journal of Psychology*, 60(1), pp. 26–35. doi:10.1111/sjop.12499. PMID: 30452083.
59. Trujillo, J., Ferrer-Cascales, R., Teruel, M.A., Ruiz-Robledillo, N., Sanchis, J., García-Ponsoda, S., Panagiotidis, A., Albaladejo-Blázquez, N., Martínez-Nicolás, Á., García-Carrasco, J., Reina, A., Lavalle, A., Maté, A. & Costa-López, B. (2025) Balladeer: A comprehensive multimodal dataset for ADHD, *IEEE DataPort*. doi:10.21227/nevp-3a70. Available at: <https://ieee-dataport.org/documents/balladeer-adhd>
60. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. (2019) “Machine learning algorithm validation with a limited sample size”, *PLOS ONE*, 14(11), e0224365. doi: 10.1371/journal.pone.0224365.
61. Varoquaux, G. (2018) “Cross-validation failure: Small sample sizes lead to large error bars”, *NeuroImage*, 180 (Pt A), pp. 68–77. doi: 10.1016/j.neuroimage.2017.06.061.
62. Vortmann, L.-M., Ceh, S. & Putze, F. (2023) ‘Multimodal EEG and eye tracking feature fusion approaches for attention classification in hybrid BCIs’, *Frontiers in Human Neuroscience*, 17, Article 1112345. doi:10.3389/fnhum.2023.1112345.
63. Wang, C., Wang, X., Jing, X., Yokoi, H., Huang, W., Zhu, M., Chen, S. & Li, G. (2022) ‘Towards high-accuracy classifying attention-deficit/hyperactivity disorders using CNN-LSTM model’, *Journal of Neural Engineering*, 19(4), Article 046027. doi:10.1088/1741-2552/ac7f5d. PMID: 35797967.
64. Wiebe, A., Selaskowski, B., Paskin, M., Asché, L., Pakos, J., Aslan, B., Lux, S., Philipsen, A. & Braun, N. (2024) ‘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), Article 508. doi:10.1038/s41398-024-03217-y.
65. Yoo, J. H., Kang, C., Lim, J. S., Wang, B., Choi, C. H., Hwang, H., Han, D. H., Kim, J. W., Cheon, K. A. & Kim, H. (2024) ‘Development of an innovative approach using portable eye tracking to assist ADHD screening: a machine learning study’, *Frontiers in Psychiatry*, 15, Article 1337595. doi:10.3389/fpsyt.2024.1337595.
66. Young, S., Adamo, N., Ásgeirsdóttir, B. B., Branney, P., Beckett, M., Colley, W., Cubbin, S., Deeley, Q., Farrag, E., Gudjonsson, G., Hill, P., Hollingdale, J., Kilic, O., Lloyd, T., Mason, P., Paliokosta, E., Perecherla, S., Sedgwick, J., Skirrow, C., Tierney, K., van Rensburg, K. & Woodhouse, E. (2020) ‘Females with ADHD: An expert consensus statement taking a lifespan approach providing guidance for the identification and treatment of attention-deficit/hyperactivity disorder in girls and women’, *BMC Psychiatry*, 20(1), Article 404. doi:10.1186/s12888-020-02707-9.
67. Yu, M., Gao, X., Niu, X., Zhang, M., Yang, Z., Han, S., Cheng, J. & Zhang, Y. (2023) ‘Meta-analysis of structural and functional alterations of brain in patients with attention-deficit/hyperactivity disorder’, *Frontiers in Psychiatry*, 13, Article 1070142. doi:10.3389/fpsyt.2022.1070142.
68. Zaheer, A. & Akhtar, A. (2025) ‘Artificial intelligence as a support to diagnose ADHD: an insight of unorthodox approaches: a scoping review’, *Child Neuropsychology*. doi:10.1080/09297049.2025.2468411.

# APPENDIX

****

*AI tools (specifically ChatGPT, GPT-4, OpenAI and Deepseek, V3.2-Exp, 探索未至之境) were employed solely for proofreading, grammar correction, and coding assistance. No generated text or code was used without verification and adaptation by the author.*