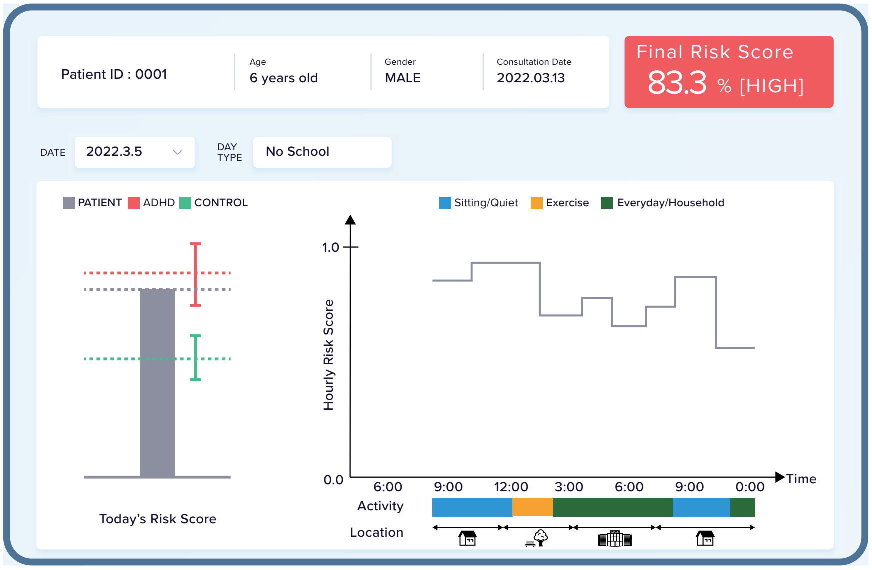
LemurDx: Using Unconstrained Passive Sensing for an Objective Measurement of Hyperactivity in Children with no Parent Input

RIKU ARAKAWA, Carnegie Mellon University, United States
KARAN AHUJA, Carnegie Mellon University, United States
KRISTIE MAK, University of Pittsburgh Medical Center, United States
GWENDOLYN THOMPSON, University of Pittsburgh, United States
SAM SHAABAN, NuRelm, United States
OLIVER LINDHIEM, University of Pittsburgh, United States
MAYANK GOEL, Carnegie Mellon University, United States



Patient ID : 0001

Age: 6 years old

Gender: MALE

Consultation Date: 2022.03.13

Final Risk Score: 83.3 % [HIGH]

DATE: 2022.3.5

DAY TYPE: No School

Fig. 1. LemurDx provides an objective estimate of hyperactivity risk. This visualization mock-up shows an initial prototype informed by need-finding interviews with clinicians and enabled by a hyperactivity risk score from a machine learning model using motion data. The physicians can use such an interface to assess a child's behavior by evaluating changes and trends of hyperactivity risk score across different contexts, such as activity and location.

Authors' addresses: Riku Arakawa, rarakawa@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, United States; Karan Ahuja, kahuja@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, United States; Kristie Mak, makk2@upmc.edu, University of Pittsburgh Medical Center, Pittsburgh, United States; Gwendolyn Thompson, gwendolynth122@gmail.com, University of Pittsburgh, Pittsburgh, United States; Sam Shaaban, sam@nurelm.com, NuRelm, Pittsburgh, United States; Oliver Lindhiem, lindhiemoj@upmc.edu, University of Pittsburgh, Pittsburgh, United States; Mayank Goel, mayankgoel@cmu.edu, Carnegie Mellon University, Pittsburgh, United States.

2474-9567/2023/6-ART46
https://doi.org/10.1145/3596244

Hyperactivity is the most dominant presentation of Attention-Deficit/Hyperactivity Disorder in young children. Currently, measuring hyperactivity involves parents' or teachers' reports. These reports are vulnerable to subjectivity and can lead to misdiagnosis. LemurDx provides an objective measure of hyperactivity using passive mobile sensing. We collected data from 61 children (25 with hyperactivity) who wore a smartwatch for up to 7 days without changing their daily routine. The participants' parents maintained a log of the child's activities at a half-hour granularity (*e.g.*, sitting, exercising) as contextual information. Our ML models achieved 85.2% accuracy in detecting hyperactivity in children (using parent-provided activity labels). We also built models that estimated children's context from the sensor data and did not rely on activity labels to reduce parent burden. These models achieved 82.0% accuracy in detecting hyperactivity. In addition, we interviewed five clinicians who suggested a need for a tractable risk score that enables analysis of a child's behavior across contexts. Our results show the feasibility of supporting the diagnosis of hyperactivity by providing clinicians with an interpretable and objective score of hyperactivity using off-the-shelf watches and adding no constraints to children or their guardians.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Applied computing** → *Life and medical sciences.*

Additional Key Words and Phrases: health sensing, diagnosis support tool, ADHD, hyperactivity, activity recognition

ACM Reference Format:
Riku Arakawa, Karan Ahuja, Kristie Mak, Gwendolyn Thompson, Sam Shaaban, Oliver Lindhiem, and Mayank Goel. 2023. LemurDx: Using Unconstrained Passive Sensing for an Objective Measurement of Hyperactivity in Children with no Parent Input. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2, Article 46 (June 2023), 23 pages. https://doi.org/10.1145/3596244

1 INTRODUCTION

Globally, Attention-Deficit Hyperactivity Disorder (ADHD) affects approximately 5% children and adolescents [17, 43, 47]. ADHD is a neurodevelopmental syndrome that often leads to increased inattentiveness, impulsivity, and hyperactivity. Children with ADHD start showing these signs as early as four years of age. In school-age children, 55% of all ADHD cases show hyperactivity symptoms [53]. The current standard of measurement of hyperactivity in children depends on subjective reports via questionnaires from parents or teachers. These questionnaires are convenient as they save time, money, and effort. However, research has also shown very low inter-rater reliability for these surveys between parents and teachers ($\kappa = 0.11$) [54]. The inherent subjectivity of these tests and physicians' lack of contextual awareness often lead to misdiagnoses. This is problematic as overdiagnosis leads to unnecessary treatment, and underdiagnosis can lead to delayed treatment [5, 24]. Thus, there is a need to add some objectivity to the diagnosis and measurement of hyperactivity.

Researchers have used commodity sensors to measure the amount of body motion and correlate it with hyperactivity. For example, Lin *et al.* [33] tracked children's arm movement using smartwatches, comparing children with and without hyperactivity during seated class activities. They found a significant difference in the measured acceleration signals between the two conditions. Several other works have developed similar machine-learning (ML) models to detect hyperactivity in children from passively sensed data [13, 23, 36, 38, 40]. However, these works assumed the measurements happened while children were in specific activities, such as a 1-hour session at a clinic [40] or taking a test [23]. Finding such settings for children is often not practical, especially given that the condition is chronic and requires frequent measurement and management. Moreover, a clinical examination does not cover a full spectrum of a child's behavior and condition during a typical day or how their behavior changes on different days (*e.g.*, weekends *vs.* school days).

To this end, this paper presents *LemurDx*, a passive sensing approach to identify hyperactivity symptoms in children without putting any behavioral constraints on children. We collected sensor data on an Apple Watch from 61 children (25 with ADHD - hyperactive presentation and rest in the control condition) for up to 7 days. Given that children, irrespective of any underlying behavioral condition, can exhibit a wide range of physical activity in a day, it is important to contextualize the collected motion data. However, providing fine-grained information about a child's day would be impossible for adolescents and very burdensome for their parents. Thus, we asked parents to provide coarse, half-hourly activity labels, such as sitting or exercising, the best they could recall at the end of the day. These labels are almost certain to be noisy, but our analysis showed that they still contained helpful information. Our ML pipeline to estimate the risk of hyperactivity achieves a detection accuracy of 85.2% (F1-score = 81.6%) when using sensor data from moments labeled as "quiet" or "sitting". Without such contextualization, the model accuracy drops to 67.2% (F1-score = 63.0%).

We also interviewed five clinicians to better understand their needs while they make diagnoses. The clinicians stressed the need to contextualize hyperactivity. However, requiring parents to annotate a child's daily life is not practical. Thus, LemurDx provides an automatic context detector that relies on motion and time of the day information to contextualize sensor measurements. Replacing the parent-provided labels with sensed contexts provides comparable performance (accuracy = 82.0%, F1-score = 78.4%) in detecting hyperactivity.

Despite success in lab-based studies, ML systems often fail to be appreciated in clinical practice [57]. Clinicians rarely use black-box systems, and therefore it remains hard to improve healthcare outcomes with such systems [14, 18]. In response, prior research [2, 3] in human-AI interaction has emphasized the importance of the interpretability of ML models for successful collaborations. In our study, too, the clinicians requested a tool that helps them make an informed decision and not force them to rely on a single number or output. Thus, we provide a level of interpretability by estimating a risk score for momentary hyperactivity at a 30-minute resolution for a whole day. The ML model uses this output to make a final inference. However, a physician can visualize the same data and filter it by appropriate contexts (such as location, time of the day, in-person school day, *etc.)* and correlate it with information shared by parents and teachers. We designed an initial interface (Figure 1) that enables clinicians to explore and compare the risk score by different contexts and children, which will help them interpret the model and make final decisions. In the future, we plan to evaluate the effectiveness of such diagnosis-support tools through more clinician interviews and clinical deployments.

In sum, we made the following contributions in this work:

(1) We conducted semi-structured interviews with five clinicians and identified their needs to see a hyperactivity risk score calculated from motion data that can be compared by different contexts (*e.g.*, time, location, activity).
(2) We extended the data collection of the pilot study of 30 children [34] to 61 children (25 hyperactive, 36 control) and collected sensor data using a smartwatch in a completely unconstrained setting.
(3) We developed a machine-learning (ML) pipeline to detect hyperactive children and demonstrated that context filtering based on parent-provided activity labels is effective, achieving 85.2% (F1 score = 81.6%) accuracy.
(4) We then fully automated the pipeline by inferring the child's context automatically from motion data and achieved 82.0% (F1 score = 78.4%) accuracy in detecting hyperactivity symptoms.
(5) We analyzed the result of the model based on the risk score as an objective measure of hyperactivity, which enhances the interpretability of the ML system.

2 BACKGROUND

ADHD is a common mental health disorder. In many children, it causes trouble in paying attention and controlling impulsive behaviors and is a major cause of academic underachievement. The problem becomes more significant when the symptoms last into adulthood, particularly when it is not diagnosed at an early stage. Adults with ADHD encounter many difficulties at work, at home, or with relationships. Therefore, the importance of diagnosing ADHD at an early stage has been emphasized for successful treatment.

There are three presentations of ADHD: (1) inattentive presentation, (2) hyperactive/impulsive presentation, and (3) combined presentation. In school-age children, approximately 55% of all ADHD cases consist of the presentation (2) or (3) [53]. While there are objective assessment tools such as the Conners Continuous Performance Test [28] for measuring inattentiveness, there are no comparable objective tools to assess hyperactivity in clinical practice [34]. The current standard to assess hyperactivity is based on reports via questionnaires, such as the Vanderbilt Assessment Scales [22]. For most children, parents and teachers are supposed to provide data using such questionnaires as it is desirable to assess children in at least two settings [17]. While fundamental for diagnosis, this test has several challenges, especially for teachers [19]. In addition, many primary care settings where children receive treatment for ADHD lack the care coordination infrastructure necessary to obtain such reports [49]. Furthermore, the questionnaire-based approach is subjective, which causes over-diagnosis or under-diagnosis [20, 25]. Thus, we aim to provide physicians with a direct window into a child's daily life using passive sensing. The subjective information from the parents and teachers can serve as an additional layer of information. Taken together, the surveys and passively-sensed hyperactivity measures can be shown to the physician in a single interface to aid in decision making. Given ADHD is a chronic condition, the measurement technique can also be used for long-term drug titration and condition management. Thus, the solution needs to be immediately deployable and usable for the children and their parents, teachers, and physicians.

3 RELATED WORK

3.1 Hyperactivity Detection

Researchers have looked at automating children's hyperactivity detection [35, 52] using MRI, EEG, motion sensors, and activities on social networks. In particular, given that hyperactivity is a presentation of frequent body movement, approaches using motion sensors have been explored most frequently [7]. However, none have looked at ways to contextualize the findings to help physicians make a decision. Until now, the focus has remained on making a binary decision for the presence of hyperactivity. Moreover, several efforts have focused on identifying hyperactivity symptoms in specific contexts. For example, Lin *et al.* [33] tracked children's arm movement using smartwatches during seated class activities. They reported a significant difference in the zero-crossing rate of the gyroscope and accelerometer between children with and without hyperactivity. Similarly, Earnest *et al.* [16] compared an actigraph of children while watching a movie, reporting that higher movement levels were associated with higher parent-rated hyperactivity symptoms. Sleeping is also a popular time to compare the motion data for measuring hyperactivity, as children with ADHD often find it hard to fall asleep. Cortese *et al.* [12] measured children's sleep using actigraphy and found that the children with ADHD had a significantly longer time to fall asleep than the controls. However, research has shown that correlating sleep disturbances with hyperactivity often leads to misdiagnosis [11, 42]. Thus, it is crucial to measure hyperactivity across contexts, assess how different contexts affect a child's behavior, and let the physician make the final diagnosis. The key takeaway from the prior work is that a child with ADHD often moves more than a child with no ADHD when expected to be quiet or relaxed.

However, contextualization of hyperactivity measurement in a child's daily life is still missing. For example, O'Mahony *et al.* [40] conducted a study where children wore two IMU bands on their waist and leg while they spent an hour at a psychiatric consultancy. They developed an SVM model to do the binary classification of 43 children, achieving 95.1% accuracy. Similarly, Gilbert *et al.* [23] measured motion data by two IMUs attached to children's wrists and ankles while they took a medical test. They developed a model based on discriminant function analysis and achieved 86.0% accuracy in hyperactivity detection. Weda is a scale-driven wearable diagnostic assessment system that has ten different tasks during which it measures children's motion data [30]. While showing promise, these works assume the measurements of children's motion data via IMU sensors happen during specific activities in controlled settings (*e.g.*, at a clinic). Here as well, the experiments do not involve hyperactivity measurement during a child's typical day, which is critical since conducting multiple inspections is recommended in clinical practice [17].

Some researchers have conducted experiments in unconstrained real-world settings [1, 13, 36, 38]. These models use datasets that include motion measurements for 24 hours. The models include different parts of the sleep stage as features, and it is unclear what parts of the day ultimately contribute to the model most effectively. Thus, these models may lead to confusion between hyperactivity and sleep disturbance. One way to remove this confusion would be to let the physicians make the diagnosis with passive sensing and machine learning, providing physicians with contextual information and estimated hyperactivity across contexts. Research has also shown that children do not prefer to wear smartwatches while they sleep [9]. Moreover, the data split and validation strategies in these works are questionable. Some efforts have optimized the final decision threshold using all participants' data [13, 36, 38], and others have randomly split windowed data into different cross-validation folds [1]. As highlighted by Hammerla *et al.* [27], such approaches often lead to non-generalizable results.

Therefore, we aim to develop a system for hyperactivity detection from smartwatches' data worn by children freely in their daily lives. In this way, we can leverage existing devices for healthcare purposes without additional user burden, which will enable physicians to inspect children's behavior in the longer term. Naturally, such unconstrained settings will pose challenges for analyzing motion data since the data lacks context.

3.2 Contextualizing Sensor Data in Passive Sensing

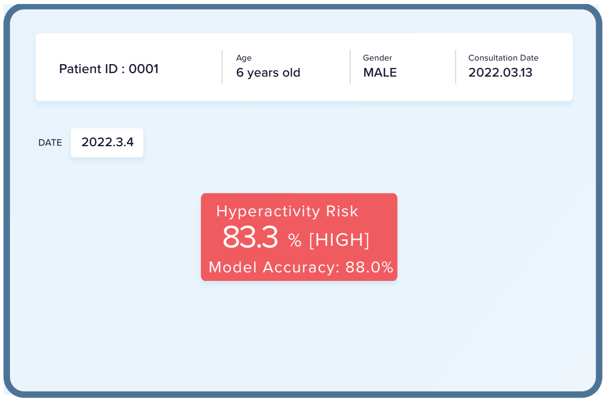
Contextualizing data is a known strategy in different domains of passive sensing. For example, Obuchi *et al.* [39] used a simple time-based contextualization in predicting brain functional connectivity using mobile devices. They divided daily sensor data into three bins corresponding to day, evening, and night, using them as features. Wang *et al.* [51] proposed an approach to predicting depression using passive sensing data from students' smartphones and wearable devices. They leveraged GPS data to contextualize the usage of phones (*e.g.*, on campus or off campus). Xu *et al.* [56] further developed this idea to use multi-sensor data for contextualization, proposing contextually-filtered features for depression detection.

The importance of contextual information for hyperactivity detection is corroborated by clinical descriptions of hyperactivity symptoms, which include "often leaves seat in classroom or in other situations in which remaining seated is expected," and, "often runs about or climbs excessively in situations in which it is inappropriate." Thus, we aim to enable unburdened contextualization of a child's physical behavior to improve the accuracy of the model and to help physicians interpret the model based on clinical and contextual insights. In this work, we start by collecting contextual information from family members by logging half-hourly activity labels such as quiet or exercising. We then demonstrate the effectiveness of utilizing the labels in hyperactivity detection, and also show that it is feasible to train an ML model that estimates such contextual information from sensed data.

3.3 Previous Version of LemurDx

Lindhiem *et al.* [34] published a pilot study with a subset of the data analyzed in this paper (30 *vs.* 61 children) with a similar performance using parent-provided labels. The current paper makes the following novel contributions:

(1) A machine learning approach that generalizes across the two data collection phases. The dataset covers a diverse set of situations. For example, the number of participant days varies from 2 to 7 days, a portion of the second phase of data collection happened during the COVID-19 pandemic, and the data has no school days, virtual and in-person school days.
(2) Need-finding interviews with five clinicians.
(3) Objective measurement of hyperactivity (instead of a binary decision) that a physician can visualize or assess contextually.



Patient ID : 0001	Age	Gender	Consultation Date
	6 years old	MALE	2022.03.13

DATE 2022.3.4

Hyperactivity Risk
83.3 % [HIGH]
Model Accuracy: 88.0%

Fig. 2. Mock-up used in the interview to inform clinicians of potential interfaces based on existing research approaches. Existing works output likelihood of hyperactivity based on a motion sensor. For example, the patient shown in this mock-up is classified as having hyperactivity with a probability of 83.3% by a model with a detection accuracy of 88.0%.

(4) An approach to remove the need of annotated activity labels for reliably measuring hyperactivity without added burden on parents and teachers.

4 PRELIMINARY INTERVIEWS WITH CLINICIANS

Given our aim to build a system that finds adoption in clinical practice, we conducted semi-structured interviews with five clinicians.

4.1 Process

We recruited pediatricians and psychologists who regularly diagnose children's ADHD symptoms. In the end, five clinicians (P1-P3: three pediatricians and P4-5: two psychologists) joined our phone-based interviews. We asked the participants a series of questions: "*What problems, barriers, or obstacles do you have with the diagnostic process?*"; "*What are your thoughts on using smartwatches to quantify children's movement?*". Then, we briefly explained existing research in measuring hyperactivity, as well as a mock-up that would inform them of potential interface ideas (shown in Figure 2). After that, we asked several questions, such as "*How do you interpret the output of the systems from existing research?*": "*What do you think would improve the usefulness of such systems?*"; "*What features would you like to see on a clinical dashboard listing patients and the output we discussed?*" Each interview took approximately 30 minutes.

4.2 Findings

All participants mentioned that the current diagnostic process is biased in response to the first question, "*What problems, barriers, or obstacles do you have with the diagnostic process?*"

> "Yes, it leads to a lot of false positives, it is pretty biased." [P5]

> "It's hard to get the paper back from the teachers and it can be very biased. It's very rudimentary and needs to be updated." [P1]

These comments aligned with our motivation to develop an objective measure of hyperactivity. The latter comment also highlighted the burden of completing questionnaires and validated our approach to measuring hyperactivity using passive sensing. When asked about their thoughts about "*using smartwatches to quantify children's movement?*", four of the five clinicians responded positively.

> "I didn't know such research has been done. It will help with objective diagnosis. I wonder how the systems work and how reliable the output is. It's unlikely just to use the output, but it will probably help us." [P2]

> "It's interesting. I am not sure how you would quantify hyperactivity. By nature, kids can be hyperactive at times, so you would need to distinguish between a normal kid and one that really has ADHD." [P4]

While positive, these participants asked about how motion levels sensed by the watch relate to hyperactivity. The responses suggest a need for interpretability and not using the model as a "black box". This point was further elaborated in the answers to the following questions, "*What do you think would improve the usefulness of such systems?*" and "*What features would you like to see on a clinical dashboard listing patients and the output we discussed?*";

> "Hyperactivity level would be dependent on the context. All children can be active in playing sports regardless of the condition. Exploring data across contexts is necessary for us to rely on the system." [P3]

> "It would be nice to have a scale so that we can compare it across different children and different context. There has to be some way to be able to look at the number and understand more of what it means in terms of the standard patient." [P1]

> "I would like to see the frequency of hyperactive movements, when throughout the day they are occurring (multiple times a day, only after school, etc.) A timeline or a graph would be perfect for this." [P4]

The comments of P3 and P1 highlight the importance of context while measuring hyperactivity. P3 and P4 also suggested that allowing clinicians to explore the data via a dedicated interface (*e.g.*, a timeline with contextual information) would be helpful.

In summary, the clinician responses elucidated the following requirements for a system that aids in ADHD diagnosis:

(1) The algorithm should output a hyperactivity pattern over time.
(2) The clinicians should be able to compare the score by applying several filters such as the time of a day or activity context.

Recently, Sivaraman *et al.* [46] assessed clinicians' interactions with an AI system that predicts the effects of treatment strategies. They found that using tools as a source of additional evidence is critical, such as helping inform comparisons that clinicians may already be making. Our findings corroborate their conclusion by indicating that visualizing the risk score across different contexts can help clinicians compare patients' behavior.

Additionally, P4 and P5 were concerned about the practicality of requiring children to wear a watch for long durations. Thus, we did not expect children to wear the watch at specific times and kept the requirements relaxed overall. We describe our data collection protocol in the next section.

5 DATA COLLECTION

We collected the study data using a custom smartwatch app that recorded data over several days. We also asked participants' family members to log coarse activity labels for the child's day. The study was approved by the Institutional Review Boards at Carnegie Mellon University and the University of Pittsburgh.

5.1 Smartwatch Application

A wide array of specialized wearables have been explored for sensing hyperactivity, such as actigraphy and wearable IMUs. Since smartwatches are now ubiquitous and sensor-laden, we built a custom data collection app to enable immediate uptake of technology. We used the Apple Watch (Series 5) as it provides, apart from motion sensors, an array of other sensors that can be used for context or health sensing, such as heart rate, location, and nearby Bluetooth devices. We built a custom application that recorded accelerometer data at 50 Hz, heart rate, GPS-based location (updated if the change of more than 100 ft is observed), and nearby Bluetooth devices to a local file. The watch's operating system directly controlled the sampling rate for all sensors apart from the accelerometer. We sampled accelerometer data using Apple Watch's CMSensorRecorder[1] class. We specifically mention this class here because we experimented with several hardware platforms and APIs, but every tested alternative optimized for the battery life on the watch and did not reliably record sensor data for 12 hours each day. While specific-purpose devices such as Actigraphs regularly record raw motion, recording such data for extended periods is rare on commodity watches. While the used approach reduces the sampling rate to 50 Hz (which makes it harder to differentiate between jerks, falls, *etc.*), it optimizes for deployability and battery life. Finally, we have yet to utilize the GPS, heart rate, and Bluetooth data in our machine-learning modeling. The collected data can still be shown to clinicians and practitioners to contextualize motion and behavior.

5.2 Activity Context

At the end of each day, we asked a member of the family to log their activity in a spreadsheet-based log file at half an hour intervals. The activity labels consist of seven activities: sleeping, sitting/quiet, everyday/household, exercise, at school, not wearing, or other. For cases where more than one activity occurred in an interval, they were asked to log the most dominant one in terms of time. We gave the family members the option to reduce their annotation burden and provide the activity labels at the end of the day. This approach, no doubt, increases label noise but we optimized user burden and compliance. The logged data provided contextual information to help build the ML models and clinicians make a decision, as found in Section 4.2.

5.3 Participants

We targeted children between the ages of 5 to 12 years as children younger than 5 years would not be comfortable with a watch throughout the day and children older than 12 years often start to transition to adult services [44]. The research staff screened the interested participants and reviewed their eligibility. Children with other alternative mental disorders, such as major depression, autism spectrum, and psychosis, were excluded from both the control and ADHD groups. In total, we recruited 75 participants through a web-based portal advertised at a city-wide hospital. Once selected, the participants and their families were asked to wear the smartwatch and log activities.

Out of the 75 participants, 68 completed the data collection properly; using the watch for the assigned period and reporting activity. Out of this, 32 children had ADHD, and 36 were in the control group. The inclusion criterion for the ADHD group included K-SADS-PL [31], which is a diagnostic interview. We also used the Vanderbilt Assessment Scale-Parent report (VADPRS) [22] to provide a score ranging from 0 to 14, where a higher score indicates stronger hyperactivity.

5.4 Procedure

Our data collection consisted of two phases: one pilot round (Phase 1) followed by a longer data collection round (Phase 2). Phase 1 consisted of data collection across two days with 15 control and 13 ADHD participants, conducted from March 2019 to July 2019. The ADHD group was medicated for one of the two days of participation. Phase 2 data collection was longer and based on the success of the pilot results from phase 1 data [34], conducted

[1] https://developer.apple.com/documentation/coremotion/cmsensorrecorder?language=objc

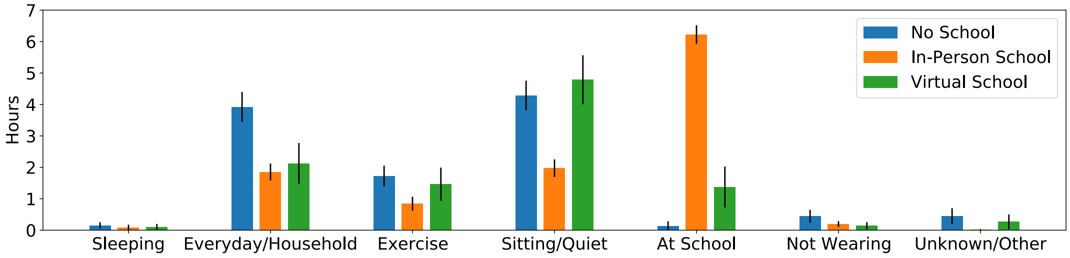


Fig. 3. Distribution of parent-provided activity labels across different types of days. Error bars indicate standard deviation.

from February 2021 to January 2022. Phase 2 collected data across seven days with 40 participants (21 control, 19 ADHD). During Phase 2, children often had virtual school due to the Covid-19 pandemic. Thus, the overall variability in children's behavior between and within the two data collection phases is quite high.

The data collection procedure was similar across the two phases. During each day, the families were asked to have their children wear the watch and run the data collection application. There was no control over the timing or duration of collection. Apart from annotating activity labels, a family member also noted the type of day in terms of school day type. Specifically, participants in Phase 1 selected either in-person school or no school. Participants in Phase 2, on the other hand, selected one of the three categories: in-person school, virtual school, or no school.

5.5 Data Cleaning

Out of the 68 participants that completed our study, four needed to be discarded due to an Apple Watch error while recording their data. Another participant was marked as outlier data due to inadequate contextual data annotated for them. Further, data from two more participants with ADHD were not included in the analyses as they were on medications for the whole data collection period. As a result, we had usable data from 61 total participants (25 ADHD and 36 control). In addition, there are a few days where the data was only logged for a short moment (*i.e.*, less than 1 hour) due to system or user error. As a result, the number of participants remained the same, but 17 out of 244 participant-days were removed.

5.6 Data Exploration

We analyzed the descriptive statistics of the collected dataset. Of the participants in Phase 1, 50% of the collection period was marked as no school, and the rest was marked as in-person school. Of the participants in Phase 2, 45% was no school, 40% in-person school, and the rest was virtual school. On average, each child wore the watch for 11.0 (SD = 1.7) hours per day. The label distribution across different types of days can be seen in Figure 3. At first glance, it seems that the children's behavior varies across the kind of days. While it is entirely possible that the behaviors are indeed different, it is also likely that the parent has no idea how much the child moved at school. There are blind spots for teachers and parents and that is why the inter-rater reliability between teachers and parents is often low for ADHD questionnaires [54]. Thus, LemurDx aims to allow clinicians and practitioners to have an objective window into the lives of children.

We also show the probability of each child wearing a watch for a different time in a day in Figure 4. On average, children wore the watch from the morning until the evening and did not wear it while sleeping at night. The variability is relatively high during day time, indicating that models that use specific time data as inputs (*e.g.*, [36]) might not work properly in an unconstrained scenario over multiple days.

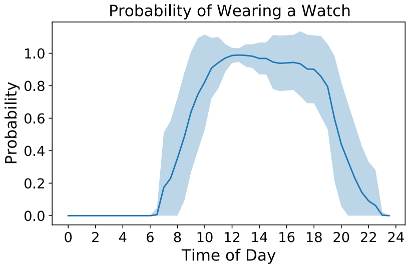


Fig. 4. Probability of wearing a watch throughout the day. The range indicates the standard deviation.

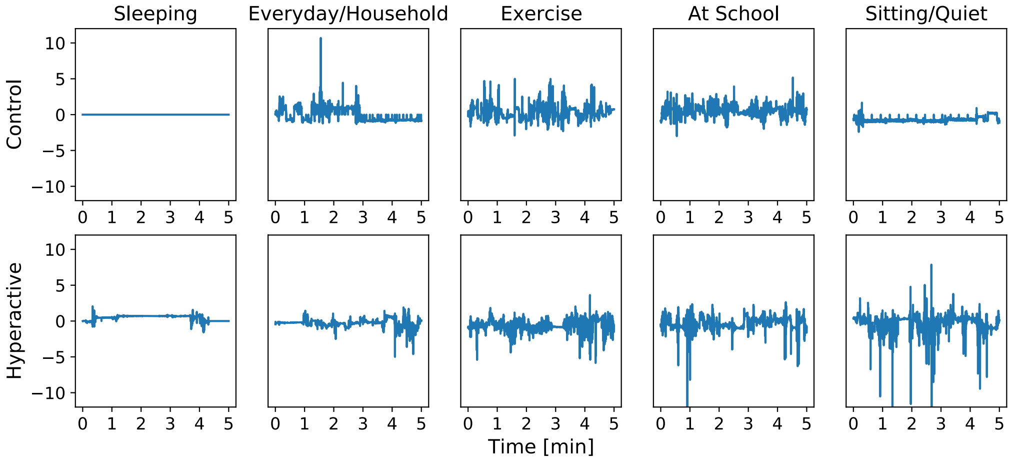


Fig. 5. Example 5-minute raw signals (acceleration in x-axis) of two children in the control and hyperactive condition in different activity contexts. (top) A child with hyperactivity. (bottom) A child without hyperactivity.

When we visualized the collected motion data and compared movement patterns between children with and without hyperactivity, we noticed that there were moments when the children in the two groups behaved similarly. Figure 5 shows randomly-extracted example acceleration plots of two children (one hyperactive and one control) by activity over five minutes. There is not much difference between the two groups during sleeping, everyday/household, exercise, and at school. However, the child with hyperactivity shows larger movement than the child from the control group during sitting/quiet moments. This observation highlights the importance of making ML models aware of activity context, which corroborates the system requirements described in Section 4.2.

6 PROPOSED MACHINE LEARNING PIPELINE

We describe our ML pipeline for estimating hyperactivity risk score. Figure 6 shows the pipeline overview, consisting of featurization, context filtering, feature selection, and predicting daily risk.

The inputs to the ML pipeline are acceleration data and activity labels. The acceleration data is measured using Apple Watch. On the other hand, the activity labels are originally provided by parents to contextualize the

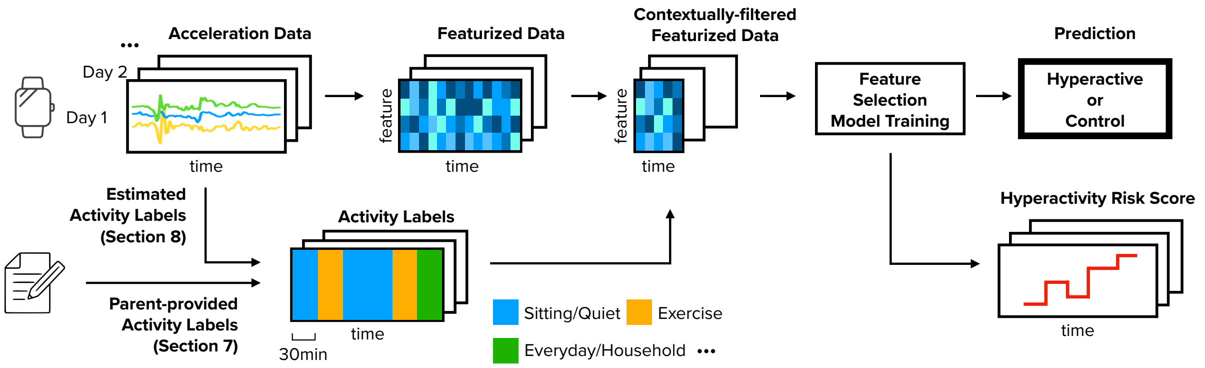


Fig. 6. Overview of LemurDx's pipeline. We use parent-provided activity labels for contextually filtering data in Section 7 and remove this dependency on human input in Section 8 by estimating activity labels from sensed data.

motion data. We first use the parent-provided labels to assess the effectiveness of using contextual data (Section 7). Then, to ease the parents' burden and achieve a fully automated pipeline, we try to estimate the labels without relying on the data the parents provided (Section 8). In either case, we can use the contextual information for each half-hour slot.

7 ESTIMATING HYPERACTIVITY RISK SCORE USING MOTION DATA AND PARENT-PROVIDED ACTIVITY LABELS

We first describe the approach to using parent-provided activity labels. We describe each component and results in order as well as discuss the trained model *i.e.*, selected features and output risk score.

7.1 Featurization

The acceleration data consists of three axes (x, y, z) that were sampled at 50 Hz. We first extracted several features with different window sizes to capture various signal characteristics. Regarding the window size, previous research used various lengths, such as 5s [38] and 2mins [40]. We prepared a set of diverse time windows to calculate features: 5s, 60s, and 600s. These window sizes were not empirically determined or optimized in a hyper-parameter search. We made a design decision to choose three very different window sizes (5s, 1min, and 10min) to capture different behavior patterns. For each window size, we first computed the following 12 features: max, min, the difference between max and min, standard deviation, mean, median, skewness, kurtosis, zero-crossing count, energy, and peak count. Additionally, we applied Fast Fourier Transform (FFT) to each windowed data and computed the same 12 features on the FFT values. We calculated a mean value for these 25 features for each half hour, *i.e.*, 48-time slots per day. This was based on the implication that different children have different average activity levels. In total, we obtained $(12 + 12) \times 3 \times 3 = 256$ features for each of the 48-time slots.

7.2 Context Filtering

Our dataset contains activity labels provided by the parents. Based on this contextual information, we can filter feature data for the model inputs by focusing on time corresponding to specific activities. More specifically, based on the observation made in Section 5.6, we focused on using data during the sitting/quiet activity to train our detection model.

Table 1. Hyper-parameters identified after 5-fold cross-validated grid search for different classification models. ML Models: Linear Model with Stochastic Gradient Descent (SGD), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB)

Model	**Hyper-parameter**	**Without Context Filtering**	**With Context Filtering**
SGD	Max number of Iterations	4	4
DT	Max Depth	5	6
RF	(Max Depth, Number of Estimators)	(4, 600)	(5, 400)
GB	(Max Depth, Number of Estimators)	(4, 500)	(4, 300)

Then, we calculated the mean value of each feature per day. For example, if there were six half-hour slots (*i.e.*, three hours) for the sitting/quiet activity, we calculated the mean value of each feature using the six data points. This way, we obtain features on a daily basis, which we use as one data instance for the ML model described below.

7.3 Modeling

For the classification, we compared multiple ML models that are implemented in scikit-learn [41]: Linear Classifier with SGD training, Decision Tree, Random Forest, and Gradient Boost. We chose these models since they can output not only the final binary output but also a continuous confidence value (0 to 1), which we treat as *hyperactivity risk score* henceforth. Random Forest and Gradient Boost are ensemble ML models that construct a multitude of trees at training time. They are known to be robust against small changes in the input data and have been used as one of the de facto standards in ML tasks with limited samples [32].

These models have hyper-parameters that can be tuned. Using 5-fold cross validation[2], we opted for the hyper-parameters in the models through grid search using F1-score as the target of the optimization. Specifically, we tuned max iteration for the Linear Classifier with SGD, max depth for the Decision Tree, and max depth and a number of estimators for the Random Forest and Gradient Boost. Once tuned, these hyper-parameters were fixed in the subsequent model evaluation. The selected hyper-parameters were summarized in Table 1.

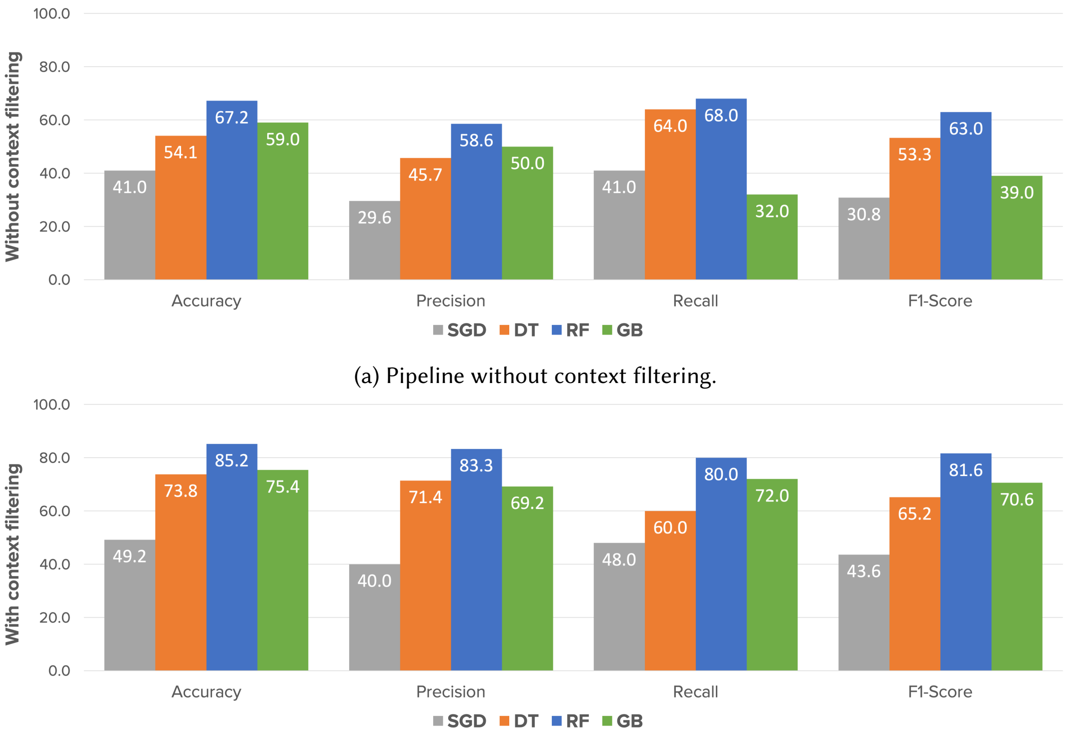
7.4 Feature Selection

We have 256 features for each day of the collection period, while we have 61 participants in total. This means that the sample size is small in comparison to the number of features. Hence, we conducted feature selection to reduce the dimension of the input features so that the ML models could be trained properly and capture the trend in the feature space efficiently. Specifically, we used a similar approach based on the feature importance of the ML model as Wang *et al.* [50]. That is, we selected features with feature importance greater than the average importance value of all features. We repeated the process until the number of the selected features got fewer than a predefined number. We searched the number in a range from 2 to 10 and chose the set of features with the best training accuracy in the leave-one-participant-out cross validation.

7.5 Results

To validate the hyperactivity risk score from the ML models, we assess the risk score's ability to identify children with an ADHD diagnosis. While we do not advocate using LemurDx's output as a black box, assessing accuracy, precision, recall, and F1 score allows us to quantify performance and compare approaches.

[2] We used GridSearchCV in sklearn: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html



(a) Pipeline without context filtering.

(b) Pipeline with context filtering.

Fig. 7. Result of different ML models with and without context filtering. ML Models: Linear Model with Stochastic Gradient Descent (SGD), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB)

7.5.1 Comparison of Different Classification Algorithms. We compared the classification performance with different classifiers in two conditions: *(1)* with context filtering (enabled by parent-provided labels) and *(2)* without context filtering. The first model uses sitting/quiet activity for training and prediction, while the second model uses all activities without selecting moments. These models were evaluated in the leave-one-participant-out cross validation. Across all classification models, the Random Forest model achieved the best performance with 81.6% F1-Score (with context filtering), and 63.0% F1-score (without context filtering). (See Figure 7 for other performance measures).

There is a significant improvement in estimating hyperactivity when using contextual information, which is also apparent in the performance of other classification models. In fact, the precision of the Random Forest model without context filtering is particularly low: there are 12 participants in the control group who are estimated to be hyperactive (*i.e.*, false-positives). Ten of these participants were correctly estimated to be the control in the model with context filtering. This result indicates that these children, even though in the control group, were probably active when expected to be active. However, they were not very active during the sitting/quiet moments. The use of context filtering is, thus, effective for detecting hyperactive children by preventing the model from being confused by the individual difference in children's activeness.

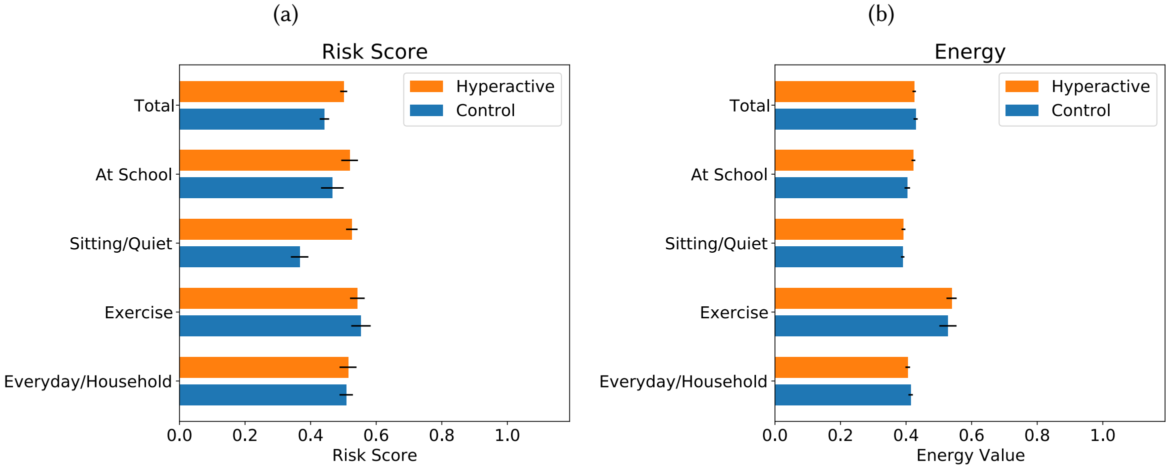


Fig. 8. Different objective measures for different activity contexts between the two conditions. (a) Risk score output by the model and (b) Energy of the acceleration. There was a significant difference between the conditions in the risk score for the sitting/quiet context and total ($p < 0.01$). The error bars are standard error.

7.5.2 Comparison with Prior Approaches to Identify Children with ADHD Diagnosis. As a comparison of the pipeline, we trained and evaluated two existing approaches using our dataset: a Convolutional Neural Network (CNN) architecture by Muñoz-Organero *et al.* [38] and a method using zero-crossing-rate (ZCR) by Lin *et al.* [33]. For the CNN model, we applied FFT to the acceleration data to get spectrograms and use them as input to the CNN. On the other hand, we computed ZCR as a feature to use for the detection of the second approach. In the same evaluation protocol, the accuracies were 70.5% and 65.6%, respectively, resulting in a similar performance to our pipeline without using context filtering. This dip could be attributed to the different settings of the data collection between different datasets. For example, the number of participants was relatively small – 16 participants with 5 hyperactive children in [38] and 30 participants with 15 hyperactive in [33]). In comparison, our dataset consists of a larger population (61 participants, 25 hyperactive) across different periods (especially, involving both before/after Covid-19 pandemic), covering a diverse presentation of hyperactivity over several days. In addition, the result of [33] was based on the data collected during desk and seated class activities. This comparison further highlights the importance of context filtering, which helped the ML pipeline become robust to the data diversity.

7.6 Interpreting Risk Score

As outlined above, our ML models take the features as input and output risk score between 0 and 1 for each day and each patient. Given the requirements unearthed in need-finding interviews with clinicians, we speculated that extending the pipeline to output the score for a more fine-grained time scale would be helpful for clinical decision making. Based on the highest accuracy, we used the Random Forest Classifier as our model in the following analysis.

Figure 8 (a) shows the comparison of the half-hourly risk score in different activity contexts. According to the t-test, there is a significant difference between the two conditions in the sitting/quiet context and total ($p < 0.001$). In addition, given the risk score in exercise is highest in both conditions and the score in sitting/quiet is lowest in the control condition, the score can be interpreted as the level of activity, *i.e.*, high risk score means a high amount

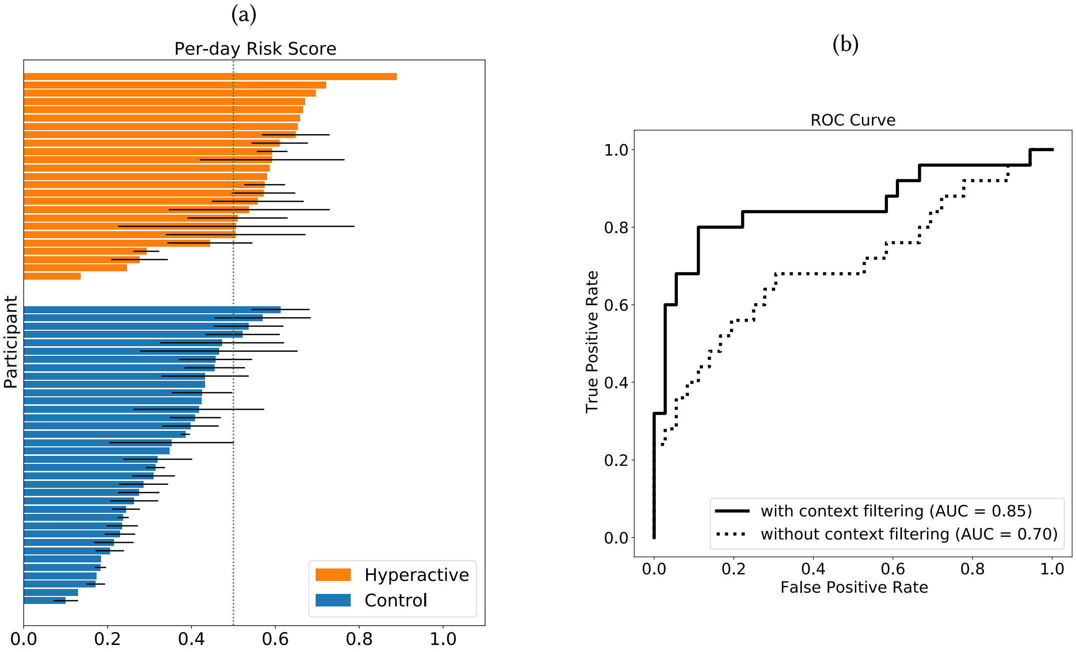


Fig. 9. (a) Average per-day risk score (after context filtering) distribution for each participant in the hyperactive and control condition. The error bars are standard error over the trial days; some participants had one-day data, for whom the error bar is not presented. The green vertical dotted line (risk score is 0.5) indicates the threshold for the final prediction *i.e.*, having hyperactivity or not. (b) ROC curves of the LemurDx pipeline with and without context filtering.

of activity. This is, however, different from simply the total amount of energy of the motion. If we used the energy of the motion as a risk score to classify children into hyperactive or control, the model accuracy was 67.2% even with context filtering. It is also evident from Figure 8 (b) that when observing the total energy of motion over a window, the contextual difference between the control and hyperactive group is not immediately clear. One reason behind this phenomenon is that amount of motion is only one aspect of hyperactivity. Hyperactivity is a combination of the amount of motion and jitteriness. While a formal definition of jitteriness does not exist, it can be thought of as self-acceleration or a non-linear combination of oscillations and jerks. Therefore, the risk score output by an ML model captures a better characteristic of hyperactivity compared to simple hand-crafted features such as signal power.

Figure 9 (a) shows the distribution of the risk score by days of all children with and without hyperactivity. Some participants had one-day data, for whom the error bar is not presented[3]. The result shows a high variability of the risk score, while there is a clear difference between the two conditions for most participants. This result suggests the need for measuring children over time to capture a wide spectrum of their daily lives. As a reference, if we use only data from the first day, the accuracy falls down to 78.7% even with context filtering. This result verifies our approach of using the app for multiple days with no constraints for users. In the future deployment of our system, we plan to collect data for longer periods and evaluate patterns of daily risk scores.

[3] Participants in Phase 1 took medicine for one of the two days, which was removed from the analysis.

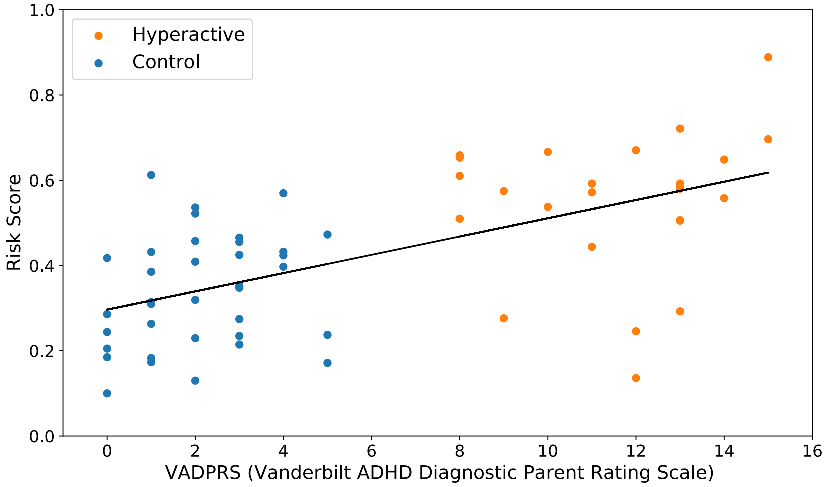


Fig. 10. Relationship between participant's VADPRS and averaged risk score. The black line indicates the regression line.

Figure 9 (b) shows the ROC curves of the performance of LemurDx pipeline with and without context filtering, clearly showing the effectiveness of context filtering. The cut-off point is 0.505 at which the false positive rate and true positive rate are 0.11 and 0.80, respectively.

Figure 10 presents the relationship between participants' VADPRS and final risk score. There is a slight correlation between the output risk score and the VADPRS score. This result implied the possibility of our pipeline to estimate the severity of the hyperactivity, which will be helpful in clinical practice. For example, the model can be used reliably as a screening tool to detect severe hyperactivity by using a higher threshold against the risk score. Additionally, we can monitor the shift of the severity over days and, for instance, we can quantify the effectiveness of medication. However, the correlation is still not very high and it is not clear if it is because of model noise or the noise in children's behavior over multiple days and contexts.

8 ESTIMATING HYPERACTIVITY RISK SCORE USING MOTION DATA WITHOUT PARENT-PROVIDED ACTIVITY LABELS

Our first analysis demonstrated the effectiveness of context filtering, and the proposed ML pipeline successfully detected children with hyperactivity from motion data collected from smartwatches with reasonable accuracy. For fully automated detection, we now aim to remove the labor of providing activity context manually.

8.1 Models

Based on the findings from the previous section, we aimed to develop a filter that finds moments where the difference between children with and without hyperactivity would be significant. Once the filter is developed, it replaces the parent-provided activity labels in the already-developed ML model. We compared several approaches to developing the filter, including using data from other sensors (*e.g.*, location, heart rate), and existing motion-based human activity recognition models. Here, we describe the findings from the approaches that worked best.

For each test participant in the leave-one-participant-out cross validation in hyperactivity detection, we estimate activity labels with an ML model that is trained using the parent-provided activity labels of the training

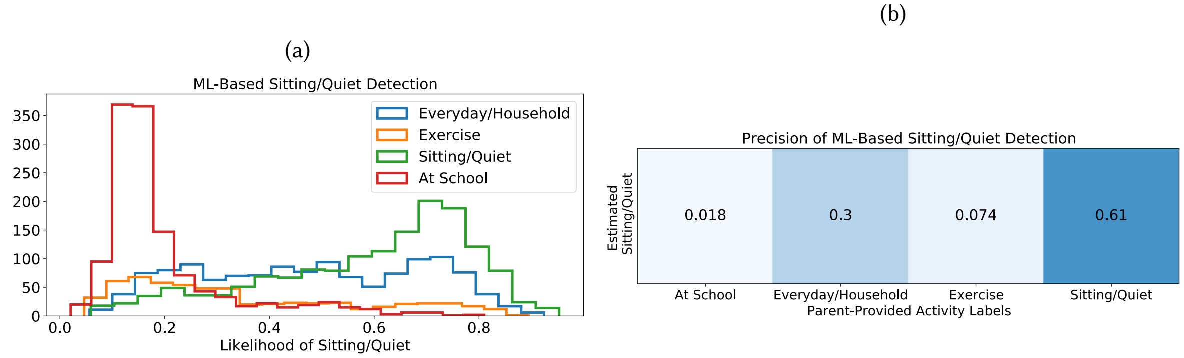


Fig. 11. The results of the ML-based model for the sitting/quiet detection. (a) Distribution of the likelihood output by the model by parent-provided labels. (b) The precision of the model by parent-provided labels.

participants. In this way, we remain agnostic about the activity labels of the test participant while following the same evaluation protocol of Section 7. We trained ML models to estimate whether each child was in the sitting/quiet condition for half an hour when the data was collected. Specifically, we trained Random Forest [21] model that takes motion features as well as time index and day type as inputs. As noted in Section 7.3, Random Forest is known to be a robust classifier and is often used in a similar setting, that is, human activity recognition from wearables [4]. The time index is 48 categorical values: 0 corresponds to 0:00–0:30, and 47 corresponds to 23:30–24:00. The day type is three categorical values: 0 for no school, 1 for in-person school, and 2 for virtual school. The idea behind using the school day type as a feature involves the observation that school day types affect the distribution of activities (Section 5.6). For motion features, we first calculated the correlation between each feature (same as used in Section 7.1) and whether it is sitting/quiet. We then used the top five features that have a high correlation.

Note that several existing approaches use neural networks for human activity recognition, such as using Convolutional Neural Network (CNN) [29] and Long Short-Term Memory Network (LSTM) [26]. These approaches, however, are trained on short segments of raw motion data (*e.g.*, 5s) with precise activity labels. However, the labels in our dataset are coarse and imprecise (recall the parents provided half-hourly labels at the end of the day). We anticipated that, since deep-learning-based approach generally needs a large amount of data, our coarse and noisy dataset would not be suitable. Here, as a comparison point, we prepared the same architecture of the CNN model in [29] and trained it on our dataset. The input window size was set to 10s following their implementation. This model has a different dimension of the outputs (*i.e.*, different classes), so we modified the final layer of the neural networks such that it outputs a two-dimensional vector (*i.e.*, sitting/quiet or not). We then obtained the final half-hourly labels by aggregating the predictions of each window.

8.2 Results

We first report the matching ratio between the predicted and parent-provided labels, that is, the accuracy of the context estimation. Then we discuss the result of the hyperactivity detection that uses context filtering based on the predicted labels.

8.2.1 Sitting/Quiet Detection. The accuracy for the Random Forest model was 74.5% (precision: 60.8%, recall: 60.9%), while the accuracy for the CNN was 63.3% (precision: 41.1%, recall: 45.6%). Here, we analyze the estimation

Table 2. Result of hyperactivity detection models with estimated activity labels.

Context Model	Hyperactivity Model	Accuracy	Precision	Recall	F1-Score
RF	RF	82.0%	76.9%	80.0%	78.4%
CNN	RF	75.4%	69.2%	72.0%	70.6%

result of the Random Forest model. Figure 11 (a) suggests that the distributions of the sitting/quiet likelihood are reasonably different across parent-provided activity labels. The model puts a lower likelihood for the at school and exercise moments. This can be understood as the exercise moments are captured by the motion data, and the moments for children at school can be detected by the day type and time index. However, the model is somewhat confused by the sitting/quiet and everyday/household moments. More specifically, the likelihood corresponding to the everyday/household moments (blue, according to the parents) is distributed almost equally, while the likelihood corresponding to the sitting/quiet moments (green, according to the parents) is skewed to the right side. This result implies that the parents have, understandably, used the label of the everyday/household to represent a broad range of their child's behavior. Thus, it is suggested that future work could refine the label resolution or incorporate a better way of annotating children's behaviors, which we will discuss in Section 9.2.1. Figure 11 (b) shows the matching between the predicted sitting/quiet moments and the parent-provided labels.

8.2.2 Hyperactivity Detection. Finally, we use the estimated activity for context filtering in our ML pipeline, as shown in Figure 6. Apart from the new approach to filtering the sensed data, the rest of the ML model remained the same as in the previous section. Here, we used the Random Forest Classifier as the hyperactivity classifier based on the result of the previous section and compared the different models for context estimation. The results are summarized in Table 2. The automated ML pipeline based on the RF-based sitting/quiet detection achieved 82.0% accuracy *vs.* 85.2% accuracy while using parent-provided activity labels. This result suggests that, even without using parent-provided labels for test participants, the trained filter based on the ML-based model can help reliably estimate the risk score for the children.

Interestingly, there is a gap between the accuracy of the sitting/quiet detection and the accuracy of the final hyperactivity detection. This can be attributed to the fact that the parent-provided labels are noisy and coarse since they were provided by parents at the end of each day. While we cannot access the "ground truth" of the activity, the result implies that the ML-based model using motion data can capture the sitting/quiet context reasonably well to use for context filtering within the pipeline.

9 DISCUSSION

In this paper, we presented an approach to estimating children's hyperactivity risk with data passively sensed at smartwatches in an unconstrained manner. While our ML pipeline successfully detected hyperactivity in children with reasonable accuracy (85.2% with parent-provided labels, 82.0% with estimated labels), several limitations remain. In this section, we discuss them as well as expected future work.

9.1 Limitations

While our system showcases initial real-world feasibility, there are several limitations that need to be overcome before it can be deployed at scale. First is the absence of children with a middle level of hyperactivity. Our study only recruited participants with a VADPRS score of above 8 and below 5. Recruiting a broader range will help in a more representative real-world dataset. Furthermore, while we make a binary distinction between hyperactive and control, there are many subtypes of hyperactivity that would need a more fine-grained analysis. This is further compounded when multiple neurological disorders manifest together. Thus, we need a more holistic data collection and modeling pipeline to tackle these scenarios. In the future, we plan to cover a more diverse range of behavioral, demographic, and clinical factors such as gender, age, ethnicity, medication use, and psychiatric co-morbidity. This would lead to the development of more generalizable learning models, stronger indicators of hyperactivity, contextualization of the user's activity better, and further help improve visualization tools for clinicians.

9.2 Future Work

9.2.1 Refining Estimated Context Filter. While our pipeline using context filtering based on estimated labels showed a better performance than the pipeline without using context filtering (Section 8.2), it falls short of the pipeline using parent-provided labels. In other words, the quality of the context filter is important for LemurDx to work. Here, there are several ways to improve the context filtering. The first is to combine other sensing data at smartwatches, such as GPS, heart rate, and Bluetooth information. For example, heart rate might be helpful to better estimate when a child is exercising. GPS can also be informative for the model to incorporate location information (*e.g.*, at school or at home). Unfortunately, in our data collection, some of the data were not logged correctly due to a bug in the data collection app. Therefore, we continue to collect new data with an updated version of the app.

Secondly, using the trained models for both activity recognition and hyperactivity detection, we are developing an on-device app that proactively asks for parents' input while collecting data. If our activity recognition model estimates that the child is sitting/quiet and our hyperactivity detection model outputs high risk score, the app can send a notification to the parents to annotate their behavior (*e.g.*, usual or unusual). In this way, LemurDx can obtain contextual information interactively from the parents, which will help the prediction. Moreover, we expect parents to experience less cognitive load than the current annotation method, *i.e.*, logging diary at the end of each day (Section 5.2).

9.2.2 Refining Learning Framework. While our pipeline showcases initial feasibility, there are several explorations to further improve our learning and modeling framework. Our current pipeline averages the scores across all the days and weights them equally. In future versions, we can work on dynamic weighting algorithms based on motion-based conditional priors. Consequently, the model can learn temporal artifacts and can make a continuous prediction, rather than a snapshot analysis of each day as a segment. This would require better temporal modeling frameworks such as HMMs, RNNs, and policy-based learning methods. Further, we can leverage the vast availability of general-purpose accelerometer datasets to build a baseline for a "regular" motion data profile. We can then model hyperactivity as a deviation from this baseline, e.g. treating the problem akin to an anomaly detection framework.

We can also incorporate insights from other IMU sources such as gyroscopes and magnetometers that infer not just acceleration but also the orientation of the joints. Further, works such as [37, 45, 48, 55] have also looked at inferring arm pose and the immediate environment from wrist-worn IMU sensors. We can incorporate the inference of such techniques better to contextualize the pose and activity of the wearer.

9.2.3 Expanding Applications of the Objective Measure. We designed our pipeline to output the risk score half-hourly to improve the interpretability of the ML model, based on the findings in the interviews with clinicians Section 4.2. As we showed in Section 7.6, the score captures the characteristic of hyperactivity in the single dimension. In this sense, the score can be regarded as an objective measure of hyperactivity that we can compare across different days, times, contexts, and participants. This score allows us to develop further applications beyond detecting hyperactive children.

We now can calculate a reliable real-time hyperactivity risk score. This will lend itself naturally to assisting clinicians with the accurate titration of medications. The process of medication titration is complex and lengthy, typically taking several weeks to months. For example, it takes more than four months for one in four pediatric patients. Providers start children at a low dose, typically not therapeutic, and gradually increase the dose until a therapeutic dose is reached. This requires several rounds of parent and teacher questionnaires and multiple office visits. LemurDx has the potential to streamline this process by providing an objective, continuous, and low-burden monitoring system to provide clinicians with rapid feedback. This is beneficial as all ADHD medications, including the stimulant classes, have side effects [6, 10].

It can also be used for developing an intervention for children. Dibia [15] developed a smartwatch app to facilitate the self-regulation of people with ADHD. In this regard, Cibrian *et al.* [8] emphasized not distracting them when they are focused, that is, reducing false-positive interventions. To the authors' knowledge, there has not been a system that intervenes in children based on their real-time behavior related to hyperactivity. Thus, LemurDx can be helpful by improving the timing of the intervention based on the estimated risk score.

9.2.4 Deploying a Tool for Clinicians. Our ultimate goal is to develop a diagnostic support tool for clinicians. To this aim, we first conducted semi-structured interviews (Section 4) and found out the requirements for the pipeline, that is, outputting a risk score and enabling context filtering. We successfully detected hyperactivity with a conventional smartwatch in an unconstrained setting. Our next step involves an evaluation of the effectiveness of such a tool. We have come up with several design ideas for the interfaces based on the interview findings and inputs from a clinician on our team. For example, Figure 1 shows one of our current prototypes. Clinicians can inspect the risk score output by our ML pipeline by different filters, such as time and activity context, which fulfills the requirements we obtained in Section 4.2. We plan to deploy the prototype to a local clinic, evaluate how the tool affects the clinicians' workflow, and investigate its effectiveness over trials.

10 CONCLUSION

We reported our series of works on detecting hyperactive children from smartwatches' data collected in an unconstrained setting. We conducted semi-structured interviews with clinicians to gain their perspectives on designing systems and identified two key aspects: having an objective scale of hyperactivity and enabling comparison of the scale by different contexts. Our study involved 61 children (25 with hyperactivity) and spanned 2–7 days, collecting motion data and parent-provided half-hourly activity labels. The first analysis confirms the effectiveness of context filtering based on the provided activity labels, improving the detection performance from 67.2% to 85.2%. Moreover, the second analysis demonstrated that such a context can be learned from the data, meaning we do not have to rely on parents' efforts to log labels. The fully-automated pipeline achieved 82.0% accuracy. Furthermore, our pipeline outputs a risk score by half an hour for each participant, allowing us to do in-depth analysis and room for more applications. While future work remains, our study is a first step to supporting clinicians to diagnose hyperactive children using already existing devices without significant burden to patients or their family members.

ACKNOWLEDGMENTS

This work was partially funded by a National Institute of Mental Health grant (R41MH119644) and a Pitt Innovation Challenge Award (2020) from the University of Pittsburgh. We also thank Prerna Chikersal, Gierad Laput, and Jordan Harris for their contributions to initial pilot studies.

REFERENCES

[1] Patricia Amado-Caballero, Pablo Casaseca de-la Higuera, Susana Alberola-Lopez, Jesus Maria Andres de Llano, Jose Antonio Lopez Villalobos, Jose Ramon Garmendia-Leiza, and Carlos Alberola-Lopez. 2020. Objective ADHD Diagnosis Using Convolutional Neural Networks Over Daily-Life Activity Records. *IEEE Journal of Biomedical and Health Informatics* 24, 9 (Sept. 2020), 2690–2700. https://doi.org/10.1109/jbhi.2020.2964072

[2] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3:1–3:13. https://doi.org/10.1145/3290605.3300233

[3] Riku Arakawa and Hiromu Yakura. 2023. AI for human assessment: What do professional assessors need?. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, CHI EA 2023, Hamburg, Germany, April 23-28, 2023*. ACM, New York, NY, 378:1–378:7. https://doi.org/10.1145/3544549.3573849

[4] Riku Arakawa, Hiromu Yakura, Vimal Mollyn, Suzanne Nie, Emma Russell, Dustin P. DeMeo, Haarika A. Reddy, Alexander K. Maytin, Bryan T. Carroll, Jill Fain Lehman, and Mayank Goel. 2022. PrISM-Tracker. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (Dec. 2022), 1–27. https://doi.org/10.1145/3569504

[5] Laura Batstra, Edo H Nieweg, Sipjan Pijl, Donald G Van Tol, and Mijna Hadders-Algra. 2014. Childhood ADHD: a stepped diagnosis approach. *Journal of Psychiatric Practice®* 20, 3 (2014), 169–177.

[6] Alice Charach, Anna Skyba, Lisa Cook, and Beverley J Antle. 2006. Using stimulant medication for children with ADHD: what do parents say? A brief report. *Journal of the Canadian Academy of Child and Adolescent Psychiatry* 15, 2 (2006), 75.

[7] Franceli L. Cibrian, Gillian R. Hayes, and Kimberley D. Lakes. 2020. *Research Advances in ADHD and Technology*. Morgan & Claypool Publishers. https://doi.org/10.2200/S01061ED1V01Y202011ARH015

[8] Franceli L. Cibrian, Kimberley D. Lakes, Arya Tavakoulnia, Kayla Guzman, Sabrina Schuck, and Gillian R. Hayes. 2020. Supporting Self-Regulation of Children with ADHD Using Wearables: Tensions and Design Challenges. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*. ACM, 1–13. https://doi.org/10.1145/3313831.3376837

[9] Franceli L. Cibrian, Elissa Monteiro, Elizabeth Ankrah, Jesus A. Beltran, Arya Tavakoulnia, Sabrina E. B. Schuck, Gillian R. Hayes, and Kimberley D. Lakes. 2021. Parents' perspectives on a smartwatch intervention for children with ADHD: Rapid deployment and feasibility evaluation of a pilot intervention to support distance learning during COVID-19. *PLOS ONE* 16, 10 (Oct. 2021), e0258959. https://doi.org/10.1371/journal.pone.0258959

[10] David B Clemow and Daniel J Walker. 2014. The potential for misuse and abuse of medications in ADHD: a review. *Postgraduate medicine* 126, 5 (2014), 64–81.

[11] Penny Corkum, Rosemary Tannock, and Harvey Moldofsky. 1998. Sleep disturbances in children with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry* 37, 6 (1998), 637–646.

[12] Samuele Cortese, Stephen V. Faraone, Eric Konofal, and Michel Lecendreux. 2009. Sleep in Children With Attention-Deficit/Hyperactivity Disorder: Meta-Analysis of Subjective and Objective Studies. *Journal of the American Academy of Child & Adolescent Psychiatry* 48, 9 (Sept. 2009), 894–908. https://doi.org/10.1097/chi.0b013e3181ac09c9

[13] P. Casaseca de-la Higuera, D. Martin-Martinez, S. Alberola-Lopez, Jesus Maria Andres de Llano, J. A. Lopez-Villalobos, J. Ramon-Garmendia Leiza, and C. Alberola-Lopez. 2012. Automatic diagnosis of ADHD based on multichannel nonlinear analysis of actimetry registries. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. https://doi.org/10.1109/embc.2012.6346894

[14] Srikant Devaraj, Sushil K. Sharma, Dyan J. Fausto, Sara Viernes, and Hadi Kharrazi. 2014. Barriers and Facilitators to Clinical Decision Support Systems Adoption: A Systematic Review. *Journal of Business Administration Research* 3, 2 (July 2014). https://doi.org/10.5430/jbar.v3n2p36

[15] Victor Dibia. 2016. FOQUS: A Smartwatch Application for Individuals with ADHD and Mental Health Challenges. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2016, Reno, NV, USA, October 23-26, 2016*. ACM, 311–312. https://doi.org/10.1145/2982142.2982207

[16] Tom Earnest, Elizabeth Shephard, Charlotte Tye, Fiona McEwen, Emma Woodhouse, Holan Liang, Fintan Sheerin, and Patrick F. Bolton. 2020. Actigraph-Measured Movement Correlates of Attention-Deficit/Hyperactivity Disorder (ADHD) Symptoms in Young People with Tuberous Sclerosis Complex (TSC) with and without Intellectual Disability and Autism Spectrum Disorder (ASD). *Brain Sciences* 10, 8 (July 2020), 491. https://doi.org/10.3390/brainsci10080491

[17] Fifth Edition et al. 2013. Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc* 21, 21 (2013), 591–643.

[18] Glyn Elwyn, Isabelle Scholl, Caroline Tietbohl, Mala Mann, Adrian GK Edwards, Catharine Clay, France Légaré, Trudy van der Weijden, Carmen L Lewis, Richard M Wexler, and Dominick L Frosch. 2013. "Many miles to go ...": a systematic review of the implementation of patient decision support interventions into routine clinical practice. *BMC Medical Informatics and Decision Making* 13, S2 (Nov. 2013). https://doi.org/10.1186/1472-6947-13-s2-s14

[19] Jeffery N Epstein, Joshua M Langberg, Philip K Lichtenstein, Beth A Mainwaring, Carolyn P Luzader, and Lori J Stark. 2008. Community-wide intervention to improve the attention-deficit/hyperactivity disorder assessment and treatment practices of community physicians. *Pediatrics* 122, 1 (2008), 19–27.

[20] Polly Christine Ford-Jones. 2015. Misdiagnosis of attention deficit hyperactivity disorder:'Normal behaviour'and relative maturity. *Paediatrics & Child Health* 20, 4 (2015), 200–202.

[21] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. The elements of statistical learning. vol. 1 Springer series in statistics. *New York* (2001).

[22] Maalobeeka Gangopadhyay, Heidi Smith, Maryland Pao, Gabrielle Silver, Deepmala Deepmala, Claire De Souza, Georgina Garcia, Lisa Giles, Danica Denton, Natalie Jacobowski, et al. 2017. Development of the vanderbilt assessment for delirium in infants and children to standardize pediatric delirium assessment by psychiatrists. *Psychosomatics* 58, 4 (2017), 355–363.

[23] Hannah Gilbert, Ling Qin, Dandan Li, Xuehua Zhang, and Stuart J. Johnstone. 2016. Aiding the diagnosis of AD/HD in childhood: Using actigraphy and a continuous performance test to objectively quantify symptoms. *Research in Developmental Disabilities* 59 (Dec. 2016), 35–42. https://doi.org/10.1016/j.ridd.2016.07.013

[24] Kimbery Giuliano and Eric Geyer. 2017. ADHD: overdiagnosed and overtreated, or misdiagnosed and mistreated. *Cleveland Clinic journal of medicine* 84, 11 (2017), 873.

[25] C Thomas Gualtieri and Lynda G Johnson. 2005. ADHD: Is objective diagnosis possible? *Psychiatry (Edgmont)* 2, 11 (2005), 44.

[26] Yu Guan and Thomas Plötz. 2017. Ensembles of Deep LSTM Learners for Activity Recognition using Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2 (2017), 11:1–11:28. https://doi.org/10.1145/3090076

[27] Nils Y. Hammerla and Thomas Plötz. 2015. Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2015, Osaka, Japan, September 7-11, 2015*. ACM, 1041–1051. https://doi.org/10.1145/2750858.2807551

[28] Susan Homack and Cynthia A Riccio. 2006. Conners' continuous performance test (; CCPT-II). *Journal of Attention Disorders* 9, 3 (2006), 556–558.

[29] Andrey Ignatov. 2018. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. *Appl. Soft Comput.* 62 (2018), 915–922. https://doi.org/10.1016/j.asoc.2017.09.027

[30] Xinlong Jiang, Yiqiang Chen, Wuliang Huang, Teng Zhang, Chenlong Gao, Yunbing Xing, and Yi Zheng. 2020. WeDA: Designing and Evaluating A Scale-driven Wearable Diagnostic Assessment System for Children with ADHD. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*. ACM, 1–12. https://doi.org/10.1145/3313831.3376374

[31] Joan Kaufman, Boris Birmaher, David A Brent, Neal D Ryan, and Uma Rao. 2000. K-Sads-Pl. (2000).

[32] Yohei Kiguchi, Melvyn Weeks, and Riku Arakawa. 2021. Predicting winners and losers under time-of-use tariffs using smart meter data. *Energy* 236 (Dec. 2021), 121438. https://doi.org/10.1016/j.energy.2021.121438

[33] Lung-Chang Lin, Chen-Sen Ouyang, Ching-Tai Chiang, Rong-Ching Wu, and Rei-Cheng Yang. 2020. Quantitative Analysis of Movements in Children with Attention-Deficit Hyperactivity Disorder Using a Smart Watch at School. *Applied Sciences* 10, 12 (June 2020), 4116. https://doi.org/10.3390/app10124116

[34] Oliver Lindhiem, Mayank Goel, Sam Shaaban, Kristie J Mak, Prerna Chikersal, Jamie Feldman, Jordan L Harris, et al. 2022. Objective Measurement of Hyperactivity Using Mobile Sensing and Machine Learning: Pilot Study. *JMIR Formative Research* 6, 4 (2022), e35803.

[35] Hui Wen Loh, Ooi Chui Ping, Prabal Datta Barua, Elizabeth Emma Palmer, Filippo Molinari, and U. Rajendra Acharya. 2022. Automated detection of ADHD: Current trends and future perspective. *Comput. Biol. Medicine* 146 (2022), 105525. https://doi.org/10.1016/j.compbiomed.2022.105525

[36] D. Martín-Martínez, P. Casaseca de-la Higuera, S. Alberola-López, J. Andrés de Llano, J.A. López-Villalobos, J. Ardura-Fernández, and C. Alberola-López. 2012. Nonlinear analysis of actigraphic signals for the assessment of the attention-deficit/hyperactivity disorder (ADHD). *Medical Engineering & Physics* 34, 9 (Nov. 2012), 1317–1329. https://doi.org/10.1016/j.medengphy.2011.12.023

[37] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*. ACM, New York, NY, 529:1–529:12. https://doi.org/10.1145/3544548.3581392

[38] Mario Muñoz-Organero, Lauren Powell, Ben Heller, Val Harpin, and Jack Parker. 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 (Nov. 2018), 3924. https://doi.org/10.3390/s18113924

[39] Mikio Obuchi, Jeremy F. Huckins, Weichen Wang, Alex daSilva, Courtney Rogers, Eilis Murphy, Elin Hedlund, Paul Holtzheimer, Shayan Mirjafari, and Andrew T. Campbell. 2020. Predicting Brain Functional Connectivity Using Mobile Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1 (2020), 23:1–23:22. https://doi.org/10.1145/3381001

[40] Niamh O'Mahony, Blanca Florentino-Liano, Juan J. Carballo, Enrique Baca-García, and Antonio Artés Rodríguez. 2014. Objective diagnosis of ADHD using IMUs. *Medical Engineering & Physics* 36, 7 (July 2014), 922–926. https://doi.org/10.1016/j.medengphy.2014.02.023

[41] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830. https://doi.org/10.5555/1953048.2078195

[42] Alexandra Philipsen, Magdolna Hornyak, and Dieter Riemann. 2006. Sleep and sleep disorders in adults with attention deficit/hyperactivity disorder. *Sleep medicine reviews* 10, 6 (2006), 399–405.

[43] Guilherme Polanczyk, Maurício Silva De Lima, Bernardo Lessa Horta, Joseph Biederman, and Luis Augusto Rohde. 2007. The worldwide prevalence of ADHD: a systematic review and metaregression analysis. *American journal of psychiatry* 164, 6 (2007), 942–948.

[44] Kapil Sayal, Vibhore Prasad, David Daley, Tamsin Ford, and David Coghill. 2018. ADHD in children and young people: prevalence, care pathways, and service provision. *The Lancet Psychiatry* 5, 2 (2018), 175–186.

[45] Sheng Shen, He Wang, and Romit Roy Choudhury. 2016. I am a Smartwatch and I can Track my User's Arm. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys 2016, Singapore, June 26-30, 2016.* ACM, New York, NY, 85–96. https://doi.org/10.1145/2906388.2906407

[46] Venkatesh Sivaraman, Leigh A. Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023.* ACM, New York, NY, 754:1–754:18. https://doi.org/10.1145/3544548.3581075

[47] Peige Song, Mingming Zha, Qingwen Yang, Yan Zhang, Xue Li, and Igor Rudan. 2021. The prevalence of adult attention-deficit hyperactivity disorder: A global systematic review and meta-analysis. *Journal of global health* 11 (2021).

[48] Jochen Tautges, Arno Zinke, Björn Krüger, Jan Baumann, Andreas Weber, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bernd Eberhardt. 2011. Motion reconstruction using sparse accelerometer data. *ACM Trans. Graph.* 30, 3 (2011), 18:1–18:12. https://doi.org/10.1145/1966394.1966397

[49] Sara L Toomey, Eugenia Chan, Jessica A Ratner, and Mark A Schuster. 2011. The patient-centered medical home, practice patterns, and functional outcomes for children with attention deficit/hyperactivity disorder. *Academic pediatrics* 11, 6 (2011), 500–507.

[50] Rui Wang, Weichen Wang, Min S. H. Aung, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A. Scherer, and Megan Walsh. 2017. Predicting Symptom Trajectories of Schizophrenia using Mobile Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3 (2017), 110:1–110:24. https://doi.org/10.1145/3130976

[51] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1 (2018), 43:1–43:26. https://doi.org/10.1145/3191775

[52] Xiu-Qin Wang, Princess Jane Albitos, Yong-Fu Hao, Hang Zhang, Li-Xia Yuan, and Yu-Feng Zang. 2022. A review of objective assessments for hyperactivity in attention deficit hyperactivity disorder. *Journal of Neuroscience Methods* 370 (March 2022), 109479. https://doi.org/10.1016/j.jneumeth.2022.109479

[53] Erik G Willcutt. 2012. The prevalence of DSM-IV attention-deficit/hyperactivity disorder: a meta-analytic review. *Neurotherapeutics* 9, 3 (2012), 490–499.

[54] Mark L Wolraich, E Warren Lambert, Leonard Bickman, Tonya Simmons, Melissa A Doffing, and Kim A Worley. 2004. Assessing the impact of parent and teacher agreement on diagnosing attention-deficit hyperactivity disorder. *Journal of Developmental & Behavioral Pediatrics* 25, 1 (2004), 41–47.

[55] Chengshuo Xia, Xinrui Fang, Riku Arakawa, and Yuta Sugiura. 2022. VoLearn: A Cross-Modal Operable Motion-Learning System Combined with Virtual Avatar and Auditory Feedback. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2 (2022), 81:1–81:26. https://doi.org/10.1145/3534576

[56] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K. Villalba, Janine M. Dutcher, Michael J. Tumminia, Tim Althoff, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Jennifer Mankoff, and Anind K. Dey. 2019. Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3 (2019), 116:1–116:33. https://doi.org/10.1145/3351274

[57] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, 238. https://doi.org/10.1145/3290605.3300468