Predicting ADHD Risk from Touch Interaction Data

Philipp Mock
Leibniz-Institut für Wissensmedien
Tübingen, Germany
p.mock@iwm-tuebingen.de

Maike Tibus
Hector Research Institute of
Education Sciences
University of Tübingen
Tübingen, Germany
maike.tibus@uni-tuebingen.de

Ann-Christine Ehlis
Department of Psychiatry and
Psychotherapy
University Hospital of Tübingen
Tübingen, Germany
ann-christine.ehlis@med.
uni-tuebingen.de

Harald Baayen
Department of Linguistics
University of Tübingen
Tübingen, Germany
harald.baayen@uni-tuebingen.de

Peter Gerjets
Leibniz-Institut für Wissensmedien
Tübingen, Germany
p.gerjets@iwm-tuebingen.de

ABSTRACT

This paper presents a novel approach for automatic prediction of risk of ADHD in schoolchildren based on touch interaction data. We performed a study with 129 fourth-grade students solving math problems on a multiple-choice interface to obtain a large dataset of touch trajectories. Using Support Vector Machines, we analyzed the predictive power of such data for ADHD scales. For regression of overall ADHD scores, we achieve a mean squared error of 0.0962 on a four-point scale (R^2 = 0.5667). Classification accuracy for increased ADHD risk (upper vs. lower third of collected scores) is 91.1%.

CCS CONCEPTS

• **Human-centered computing** → **User models**; **Touch screens**; *User studies*;

KEYWORDS

Multi-Touch, Machine-Learning, User Modeling, ADHD

ACM Reference Format:
Philipp Mock, Maike Tibus, Ann-Christine Ehlis, Harald Baayen, and Peter Gerjets. 2018. Predicting ADHD Risk from Touch Interaction Data. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3242969.3242986

1 INTRODUCTION

Attention-deficit/hyperactivity disorder (ADHD) is one of the most commonly diagnosed psychiatric disorders for children. In recent decades, a number of sources have reported an increase in prescribed ADHD medications, as well as ADHD diagnosis – especially for young school children (e.g., [3, 24]).

Core symptoms of ADHD are inattention, hyperactivity, and impulsivity. The fact that these symptoms are non-specific and only present in certain situations complicates a correct diagnosis.

Before the disorder is clinically diagnosed, screenings are performed to evaluate a person's risk of ADHD. For this, self-reporting scales (such as ADHD Rating Scale-IV [7]), which assess behavioral patterns that are considered ADHD risk factors, are calculated using short questionnaires (aimed either at the parents, teachers, or at the children themselves).[1]

An often criticized problem of self-reporting questionnaires is that they are not directly linked to the specific situational context that is most relevant for a diagnosis. The questionnaires describe situations for which participants need to predict or evaluate their own behavior. This can be problematic, because individuals may not be aware of how they react in certain situations, may have difficulties in evaluating oneself objectively, or may differ in their ability for introspection. The problem of context dependence can be aggravated when different external raters (e.g., teachers and/or parents) are involved. In this line of argumentation, it has been suggested that additional behavioral data could allow for more objective measures [17, 23].

In this paper, we propose an alternative approach to subjective ADHD scales that makes use of touch interaction data combined with pattern recognition techniques. We recorded a rich set of data from 129 fourth-grade students during one-hour sessions with a multiple-choice interface on a touch screen device. Instead of using touch interactions as a supplement for a touch-enabled self-reporting questionnaire, we use data recorded directly from an academic setting to train our prediction models. Traditional pen-and-paper questionnaires serve as labels for that data. While this allows us to obtain more 'natural' touch interactions, using math tasks for the data acquisition process furthermore enables us to induce cognitive workload and monitor potential behavioral differences that changes in task difficulty might evoke.

The main contribution of this paper is a detailed analysis of an objective ADHD prediction approach based on models trained from features acquired using a multiple-choice touch interface. In the

ICMI '18, October 16–20, 2018, Boulder, CO, USA
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3242986

[1] In practice, such self-reporting questionnaires are not only used for screening, but also during the diagnostic process or for the evaluation of treatments.

following, we describe the data collection and feature selection process, and assess whether a student's risk of ADHD as estimated with a self-reporting questionnaire can be predicted using supervised machine-learning techniques. We also discuss limitations of our analyses and how the presented approach could be put into practice.

2 RELATED WORK

As stated above, the most commonly practiced form of ADHD screenings are subjective self-reporting questionnaires. Typically, these self-ratings consist of 18 to 20 questions and take about five minutes to finish. Good test-retest reliability, criterion-related validity, and internal consistency have been demonstrated [6, 7, 14].

Another standardized test for ADHD is a so-called continuous performance test (CPT). A CPT typically uses software to present stimuli in rapid succession and measure attention and impulsivity by means of reaction times and error rates. Young et al. implemented such a test within a smartphone application, which also captures accelerometer and gyroscope data [27].

As for technology-assisted automatic prediction approaches, neuroimaging techniques have been well researched for their potential to help understand the disorder, but also for their predictive power. In recent years, technologies such as EEG, fMRI or fNIRS have successfully been used to identify patterns associated with ADHD using machine-learning techniques such as Support Vector Machines (SVMs) or Gaussian processes classifiers (e.g., [13, 19, 21]). Another well-established approach is to assess ADHD using characteristic patterns of eye movements [2]. For example, Fried et al. found that adult ADHD patients were unable to suppress eye blinks and microsaccades while performing a test of variables of attention [8]. Other approaches that aim at assessing ADHD symptoms objectively rely on motion tracking hardware: For example, inertia measurement units were used to classify children diagnosed with ADHD with an accuracy of above 95% [22]. Infrared motion-tracking systems have been applied to identify higher motor activities of ADHD patients [18].

A limitation of the above technology-assisted prediction approaches is the requirement of dedicated hardware, limiting their application for rapid assessment of the disorder for a larger population of students. With the recent increase of distribution of tablets in schools, we consider touch technology to be a promising platform for quick and efficient tests within the classrooms. After all, modern touch sensors provide a rich set of interaction data that can be obtained on most consumer devices without additional hardware or modifications.

To the best of our knowledge, touch interaction data has not been used to model personality traits or mental disorders as of yet. However, the potential of such data was demonstrated for classification of age groups and mental states. Vatavu et al. used touch distance offsets and tap times to distinguish small children from adults with an accuracy of 86.5% after a single touch point [25]. Gao et al. used touch trajectories from a smartphone gaming app labeled with self-reporting questionnaires to train classification models for affective states [9]. Classification accuracies for four emotions (excited, frustrated, relaxed and bored) were between 69% and 77%. In our own prior work, we recorded touch data with a multiple-choice setup and used SVM classifiers to distinguish between different levels of cognitive workload [20]. Average classification accuracy of easy vs. hard addition problems was 90.67%. The results furthermore revealed that the individual classifiers and feature distributions varied considerably between participants. This paper investigates whether such behavioral differences as empirically observable from recorded touch interactions can be used as a basis for cross-person classification and regression models.

3 STUDY DESIGN AND METHODOLOGY

3.1 Study Description

We tested 129 students of seven local primary schools. All of them attended the fourth grade and were between nine and twelve years old (40.83% male). We did not perform a preselection of subjects with regard to risk of ADHD. All participants filled out self-reporting questionnaires in group sessions and completed a set of multiple-choice math tasks on multi-touch devices (two students at a time). The study was approved by the local ethics committee and by the school board. All measures and procedures had to be sent in and the responsible persons at both institutions reviewed data privacy of our participants. With regard to the actual data collection, the participating children invented a code that we used to match their data and were instructed to not write their names on any of the materials. As stated above, this paper mainly focuses on the assessment of the interaction data collected during the individual multi-touch sessions. The self-reporting questionnaires were used to obtain ADHD scores for all participants, which serve as ground truth for the evaluation of our prediction models. The questionnaires contained the German ADHD Symptom Checklist (FBB-ADHS) which assesses the diagnostic criteria for ADHD (DSM–IV criteria and ICD–10 for hyperkinetic disorders). The self-ratings consist of 20 items with a four-point rating scale ranging from "1 = not at all" to "4 = very much" each. Nine of the items assess inattention (e.g., "I find it hard to concentrate."), seven items assess hyperactivity (e.g., "I often wiggle my hands and feet or fidget in my seat.") and four items assess impulsivity (e.g., "I often interrupt or disturb other people."). The scales were proven to be internally consistent (alpha scores between 0.69 and 0.75 for the subscales and 0.87 for the overall symptom scale). Döpfner et al. demonstrated a construct validity of 0.69 [6].

The used multiple-choice interface (as depicted in Fig. 1) looks as follows: For each task, four possible answer boxes appear in each of the four corners of the screen. The answer boxes can be selected using an interactive crosshair-cursor element that appears in the center of the screen at the beginning of each trial. Initially, the only visible element is the crosshair-cursor. When the cursor is touched, the current task instructions and four empty answer boxes appear. The contents of the answer boxes only become visible, when the cursor is dragged over them. When the cursor is released, it snaps back either to the center (no answer box selected) or to the answer box beneath it. A continue button allows to immediately go to the next trial. No feedback is given on the correctness of an answer. The applied interface is not optimized for efficiency but rather designed to evoke drag gestures with high amounts of variance, which can be exploited using machine-learning models.

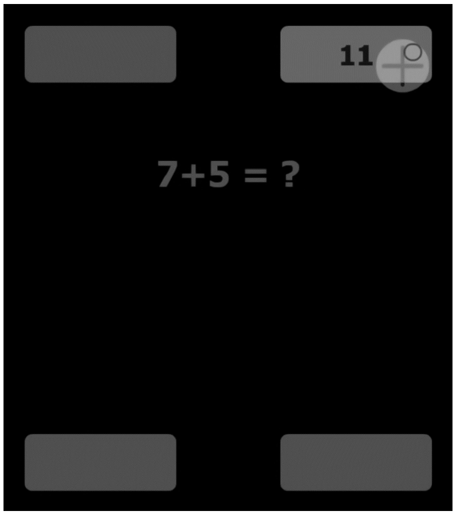

Figure 1: The multiple-choice interface showing an addition problem with low problem size that requires a carry operation (medium difficulty). The answer box at the top right shows an incorrect probe.

Each participant completed a set of 96 math tasks. The tasks were split into two sessions with a short break before the second session. Task difficulty was balanced across sessions. As in prior work [20], we used addition problems with a sum below 99 and with varying difficulty levels, defined by the so called *problem size* (sum > 40) and *carry effect* (sum of unit digits > 10). Both effects have been identified to induce increased response latencies and/or error rates for multi-digit addition problems (e.g., [1, 15]). Since our participants correctly answered 89.34% of the easiest and 61.34% of the hardest tasks, we are confident that the task difficulty levels were a robust modulation of induced cognitive workload for the tested population. Using such a set of addition problems has two major advantages for the presented ADHD prediction approach: first, it allows us to directly compare our methods and data with our own prior results for cognitive workload prediction. Second, using tasks, which robustly induce varying levels of cognitive workload, is just as valuable for the prediction of ADHD scores, since ADHD is linked to deficits in executive functions (e.g., [4, 26]). Using a set of tasks with varying difficulty allows us to capture potentially resulting behavioral differences and utilize them for prediction purposes. This paper consequently includes a comparative evaluation of cognitive workload classification accuracies with a special focus on differences in performances with regard to lower or higher ADHD scores of our study participants.

Problem size and carry-over were manipulated orthogonally in a factorial 2x2 design and problem size was matched between carry and non-carry problems. The three incorrect answer probes (each deviating from the correct answer by 2 or 10 to avoid parity based solution strategies) were randomly but equally distributed across the four answer boxes together with the correct answer.

The multiple-choice study was conducted on a Samsung SUR40 touch table using only a part of the screen space which corresponds to the size of a regular iPad. We used a custom finger tracking application, which allowed us to record all of the occurring touch information. The test persons completed the tasks standing in front of the table. The study was conducted within the schools and students were successively fetched from their regular classrooms to participate in the multiple-choice session, which lasted for about an hour each. After a short introduction and an opportunity to familiarize with the system, each subject completed two sessions with 48 math trials each. Each block of six consecutive trials had the same difficulty level and was concluded with a subjective difficulty rating on a four-point scale (from 'easy' to 'difficult').

3.2 Data Basis

For each trial, we calculate a set of 70 features as suggested in [20]:

- 8 features that are not specific to multi-touch devices (correct answer given, trial duration, time until login, idle time, time until first touch / first movement / first box reached / correct answer reached)
- 19 features that describe finger trajectories (number of strokes (i.e., continuous movement without lifting the finger), number of stroke segments (segmented for every halt of the movement, e.g., for a change of direction), travel distance, average, minimum and maximum segment length and speed
- 3 features that describe movement in relation to the UI (travel distance after first box / correct answer, number of boxes revealed after correct answer)
- 40 features that are extracted from low-level touch sensor data (pixel intensity range, shape descriptors, downscaled 4x4 sensor image and an approximation of applied pressure). These features are specific to optical touch devices and cannot be collected on a capacitive touch screen, as of now. Please note that the used touch sensing technique is susceptible to ambient light. Since our study was performed directly within the school buildings and not under controlled lighting conditions, the ambient light levels vary and thus make the data impractical for models relying on data from multiple sessions and persons.

We performed a comparative classification of cognitive workload, in order to validate our data and methods and to assess coherences between ADHD and cognitive workload (compare 4.1). For this, we applied the above 70 features from the easiest and hardest trials for training and evaluation of individual classification models. The resulting two-class classification models (low vs. high task difficulty) each use trials from a single subject as samples.

For ADHD prediction, we generated models based on data from multiple persons (cross-subject). Consequently, the ADHD score regression and two-class classification models (low vs. high ADHD score) use the subjects as data points. The 40 features extracted from low-level sensor data are not reliable for this case of application, because of changing lighting conditions between sessions. In order to avoid overfitting, we do not apply data from all trials per user as separate features, but use feature averages of the remaining 30 features across all 96 trials instead. Aside from that, we calculate variation measures (variance and standard deviation), which make for another 60 features. The differences between sessions one and two were included as additional features, in order to account for behavioral changes during the course of the experiment. That way, we obtain a set of 180 features for each subject.

The scores of the ADHD self-ratings are used as labels for the touch interaction data. The resulting four constructs that we modelled by means of touch interaction data are: overall ADHD symptom mean score, as well as the subscales inattention, hyperactivity and impulsivity. 20 of the 129 subjects did not complete a substantial number of items of the self-rating questionnaire. Since no ADHD score labels are available for these subjects, they could not be included in the ADHD prediction models.

3.3 Machine-Learning Methods

As stated above, the main contribution of this work is our assessment of possibilities to predict ADHD symptom scores automatically by means of touch interaction data. The questionnaire-based ADHD screening, which we use to label the touch data, gives a continuous score between one and four. Consequently, the most natural way to implement ADHD score prediction is to treat it as a regression problem, that is, learning the mapping function from a dataset of known test persons to their corresponding ADHD scores.

In practice, ADHD screenings are used to identify individuals who show strong symptoms of the disorder in order to decide whether further tests and medical treatment might be appropriate. This selection process requires to apply a threshold to the continuous scale, which corresponds to a classification problem, where a discrete class label (in this case 'increased risk of ADHD') is learned from the data. Since we have successfully applied SVMs for classification of cognitive workload from comparable data before [20], we use SVM classifiers and Support Vector Regression (SVR) to create our ADHD prediction models.

We evaluated the following models: regression of overall ADHD scores and the three subscales for inattention, hyperactivity, and impulsivity, as well as SVM classification for the same constructs. We use RBF kernel functions in both cases. The model parameters were optimized using a grid search approach and we used ten-fold cross-validation for all evaluations.

As for feature selection, we evaluated the approach suggested in [20] for learning of individual cognitive workload models using F-Scores and compared it with two more computationally intensive methods. The three tested conditions were Individual F-Scores (*IF*), Forward Selection (*FS*) and Backward Elimination (*BE*). *IF* uses F-Scores to estimate the discrimination of the two sets for each of the 70 touch features. The features are selected according to the F-Score threshold, which gives the best classification accuracy.

FS starts with only one feature and incrementally adds additional features according to best model performance until classification accuracy can no longer be improved. *BE* starts with the complete set of features and incrementally eliminates those features, which contribute least to or harm the overall model performance. *FS* typically finds smaller feature sets compared to *BE* [16]. However, there is no guarantee that the optimal feature set will be found with either technique [10].

We used a two-step evaluation procedure for both *FS* and *BE*: Initially, a rough hyper-parameter grid search is included in a nested ten-fold cross validation which selects the best / worst feature per iteration for each fold. After this procedure has determined the feature set with best overall performance, the hyper-parameters are optimized with a more refined grid search (intervals from C^{-6} to C^{16} and γ^{-14} to γ^{8}). For *BE*, we used a preselection according to sorted F-Scores in order to avoid that potentially important features become eliminated early in the process. This greatly increased robustness of the selected feature sets and hyper-parameters across multiple runs.

4 RESULTS

4.1 Cognitive Workload and ADHD

Our decision to work with data from an academic setting with tasks of varying difficulty was influenced by the oft-asserted hypothesis that ADHD reflects an executive function deficit [4]. For the purpose of automatic ADHD prediction, it is thus particularly interesting to also measure the impact of varying ADHD levels on automatic prediction of cognitive workload. In this respect, we present our results for cognitive workload prediction (both for the whole population and separately for groups with lower and higher ADHD risk) in the following section. The comparative evaluation also serves the purpose of assessing the suitability of different feature selection approaches. In this line, we trained individual cognitive workload classification models for each of our 129 participants using three different feature selection techniques (*IF*, *FS*, and *BE*). In the following, we discuss the classification results and implications for ADHD prediction that result therefrom.

Cognitive workload classification accuracies with different feature selection techniques averaged across the whole population are illustrated in Fig. 2. For *IF*, average classification accuracy is 89.61% (SD = 7.43) with individual accuracies between 52.67% and 100%. For *FS* and *BE*, average classification accuracies are 94.6% and 94.62% (SD = 5.95 and 5.68). Minimum accuracies are 60.67% and 64.68% and maximum accuracies are 100% in both cases. A repeated-measures ANOVA with classifiers based on *IF*, *FS* and *BE* as the three measures reveals significant differences between the measures ($F_{2,256}$ = 76.05; $p < 0.001$). Bonferroni adjusted post hoc tests reveal that *FS* and *BE* perform significantly better than *IF* (p <

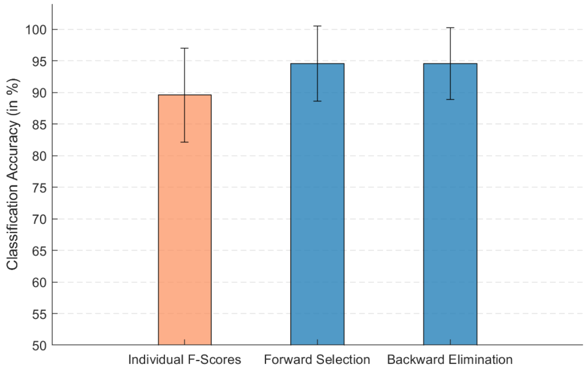

Figure 2: Comparison of cognitive workload classification accuracies for different feature selection techniques. Both *FS* and *BE* (blue bars) outperform feature selection based on individual F-Scores (depicted in orange).

0.001 in both cases). There is no significant difference between the performances of *FS* and *BE* (p = 1.000).

As for the resulting feature sets, the number of used features per classifier ranged from 1 to 48 (∅ = 12.63, SD = 11.77) for *IF*, from 1 to 48 (∅ = 6.94, SD = 8.71) for *FS*, and from 2 to 19 (∅ = 6.83, SD = 4.08) for *BE*. Both *FS* and *BE* have a lower average number of used features compared to *IF*. We deduce from these statistics that a good number of noisy features is removed by both approaches. Our observation that *FS* occasionally leads to bigger feature sets can be explained by the fact that we still added features to the set when they did not have an effect on classification accuracy. That way, it is more likely to preserve potential feature interactions in the final set without harming overall accuracy. However, it leads to the inclusion of a number of irrelevant features.

The mean classification accuracies for subjects with higher ADHD scores do not differ significantly from those for lower scoring subjects: We achieve a mean accuracy of 93.57% for the upper third of subjects regarding ADHD scores and 95.57% for the lower third (compare Fig. 5). However, minimum classification accuracy is lower and standard deviation is higher for the higher third (minimum accuracy: 64.67% vs. 86% and SD: 7.21 vs. 3.70). We found some structural differences in the selected feature sets of these two groups: *number of boxes revealed after the correct answer* is selected more frequently for the higher third of ADHD scores, while *trial duration*, *time until first touch*, and *time until first movement* are selected more frequently for the lower third. Average feature set sizes are relatively constant (6.67 and 7.23 respectively). When looking at the discriminating potential of single features estimated with F-Scores, *trial duration* shows the most apparent differences between the two groups (mean F-Scores 0.44 for the higher and 0.58 for the lower third with SD = 0.41 and SD = 0.58 respectively). Besides, standard deviation of F-Scores for *correctness of the answer* and *time until correct box* was lower for the higher third (differences of -0.28 and -0.37 respectively). Both features have weak negative correlations with mean ADHD score (r = -0.2 and -0.18).

In sum, the above results confirm that our previous results for classification of task difficulties [20] hold for a larger population of students. We moreover found that computationally expensive feature selection techniques, such as *FS* and *BE*, can further improve prediction results compared to the established approach. Both were selected for ADHD score prediction accordingly. Examining the results of subjects with higher ADHD risk separately, we could observe individual differences in behavior when working on tasks of higher difficulty. However, the classification results only differ in details. We did not observe any substantial differences regarding classifier performance and selected feature sets.

4.2 Regression Models for ADHD Scores

As explained above, we trained regression and classification models for ADHD scores and the three subscales using data from all 109 participants, who completed both the touch trials and the ADHD screening questionnaires.

As for regression, we trained models using *FS* and *BE* with all 180 features (mean (∅), standard deviation (σ), and variance (σ^2) of the 30 basic features, as well as differences in between sessions one and two (indicated with a Δ in front of the feature name)). We compared these models to a baseline, which only uses trial duration and correctness of the answer (∅, σ, σ^2, and Δ each). In order to evaluate the features extracted from the swipe gestures in isolation, we also included a reduced set of features in our analysis, which only uses touch and UI related features and no timing information or correctness. ν-Support Vector Regression (ν-SVR) with grid search optimization and ten-fold cross validation was used for all models. Mean squared error of the baseline model is 0.1766 with SD = 0.1854 (compare MSE values in Fig. 3). For *FS* and *BE*, mean squared error of the best model is 0.1343 or rather 0.0962 with SD values of 0.3617 or rather 0.5667. The best reduced set model has a mean squared error of 0.1421 (SD = 0.3425). Generally, *BE* clearly outperformed the other models both regarding mean squared error and goodness-of-fit. All models however performed better than the baseline.

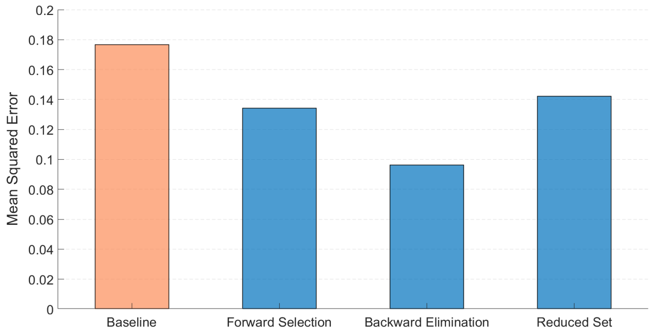

Figure 3: Mean squared error of ADHD regression models with different feature sets (lower is better). *Reduced Set* includes no information about timing and correctness of the answer (only touch and UI related features). The baseline (orange bar) uses only *trial duration* and *correctness of the answer.*

Interestingly, the feature set found by *FS* only includes nine features: *trial duration* (σ), *minimum divergence from the direct path* (∅ and σ^2), *minimum curvature* (σ), Δ *number of stroke segments* (∅), Δ *average curvature* (∅ and σ^2), Δ *time until first touch* (∅), and Δ *maximum divergence from the direct path to the right* (σ^2). Average curvature C of a stroke segment is calculated as the sum of changes in direction Δd_i along the trajectory:

$$C = \frac{1}{n-4}\sum_{i=3}^{n-2} \frac{1}{2}\Delta d_i(p_{i-1}, p_{i+1}) + \frac{1}{2}\Delta d_i(p_{i+2}, p_{i-2})$$

where n is the number of measurements collected for a segment. It should be noted that C is negative for curvatures to the left, thus *minimum curvature* should be read as 'highest curvature to the left'. Average divergence D is calculated as the average distance of all n measured points on a trajectory from P_1 to P_2 to the direct path:

$$D = \frac{1}{n}\sum_{i=1}^{n} dist(p_1, p_2, (x_0, y_0))$$

BE, on the other hand, finds a set of 30 features, including variability measures (σ, σ^2, and Δ) for the *number of strokes* and *segments*, *segment length*, *divergence to the right*, *average*, *minimum*, and *maximum curvature*, *idle time*, *maximum speed*, and *travel distance after the correct answer was revealed.* A direct conclusion from this is that the recorded touch features contain interdependencies, which a feature selection approach that starts with all available features (such as *BE*) can exploit. Due to the above results, we only report the results of *BE* feature selection for the three subscales: The best model for inattention with a total number of 20 features has a cross-validated mean squared error of 0.0942 (SD = 0.6544). For hyperactivity, a model with 15 features achieves a mean squared error of 0.1508 (SD = 0.4803). The best impulsivity model has a mean squared error of 0.1829 (SD = 0.5280).

It is interesting to note that the feature sets for the three subscales do not overlap heavily. While features derived from *average* or *minimum curvature*, *minimum segment length* and some type of timing information (most timing features are highly correlated, so it is to be expected that different features will be selected for different models) are used for all three, other features seem to be beneficial for only one or two of the scales. A more detailed description of which features carry the most weight for each of the scales is given later in this paper.

4.3 Classification Models for ADHD Scores

In practice, automatic ADHD screening from interaction data as proposed in this paper could be applied to identify students with particularly high risk and invite them for further testing. As illustrated in Fig. 4, our data basis does not include a large body of high scoring students. For classification purposes, using only the highest scores (e.g., 2.5+) for the 'high ADHD' class would lead to a highly imbalanced classification problem that would be prone to overfitting the few high scoring participants. Consequently, we were looking for a partition of the data, which includes a maximum number of subjects while still maintaining disjoint ADHD scores and a balanced number of samples per class. We therefore split the dataset into thirds and used the upper vs. the lower third of participants regarding their ADHD scores as classes for our two-class classification problem. In the following, we use 'high ADHD' and 'low ADHD' as class labels to refer to the upper and lower third of collected ADHD scores (rather than absolute values).

The low-scoring third has a maximum score of 1.52 and the high-scoring third has a minimum score of 1.75. Although we would adjust this threshold upwards when data from a clinical study with larger quantities of high ADHD scores becomes available, this partition of the data ensures a balanced distribution of samples across the two classes. While certain interaction patterns might be more pronounced for students with very high ADHD risk, the resulting models for this classification problem can still help us identify patterns that emerge for increasing ADHD scores and validate whether automatic ADHD screening by means of touch interaction data is practically feasible.

Again, we trained a model that only uses *trial duration* and *correctness of the answer* (∅, σ, σ^2, and Δ each) as a baseline for classification of mean ADHD scores. However, a model with all eight baseline features performed significantly worse than a model selected with backward elimination from the subset (60% vs. 70% cross-validated classification accuracy). We consequently only depict this improved baseline model in Fig. 5. It is worth noting that ADHD baseline only uses variability measures derived from *trial duration* – i.e., *correctness of the answer* is not included.

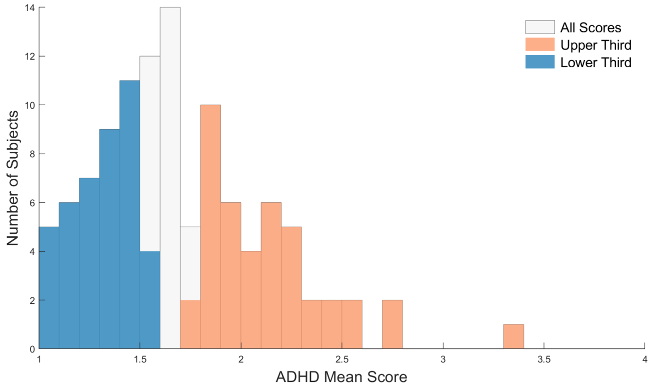

Figure 4: Histogram of ADHD scores. We used the upper third vs. lower third of the subjects for two-class classification.

Using the full set of features and *BE*, we achieve a classification accuracy of 91.1%. Without timing information or *correctness of the answer* included in the feature set (ADHD Reduced in Fig. 5), classification accuracy drops to 85.56%. As for the subscales, classification accuracies for inattention, hyperactivity and impulsivity are 81.1%, 88.9%, and 86.7% respectively. Again, all of these accuracies were achieved using *BE* for feature selection. Misclassifications are mostly balanced across the two classes, with generally more false positives than false negatives. False positive to false negative ratios vary between 2:1 to 1:1.

As with the regression models, feature sets for the different scales are relatively heterogeneous. This has to be expected, because some of the features are highly correlated and can thus be selected interchangeably. Still, there are certain noteworthy differences. In particular, *correctness of the answer* and *minimum segment speed* (∅) are used for overall ADHD score classification, but for none of the subscales. Overall, the best feature sets for hyperactivity and impulsivity are more similar to each other compared to the

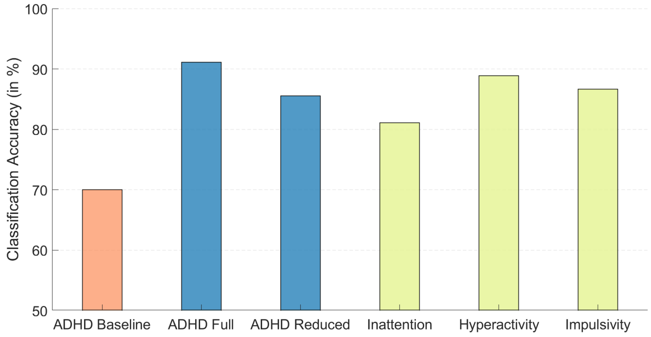

Figure 5: Classification accuracies for ADHD scores and the three subscales (green bars). *ADHD Baseline* (orange bar) only uses *trial duration* and *correctness of the answer* as a starting set for feature selection. *ADHD Reduced* only uses touch and UI related features. Chance level for two-class classification is 50%.

best set for inattention. The former two each contain multiple features related to the shape of the touch trajectories (*segment length*, *curvature*, and *divergence from the direct path*) with a total number of 25 and 29 respectively, whereas only seven features are used in the model for inattention. Despite these differences, all models contain at least one feature from each of the feature groups derived from *trial duration*, *divergence from the direct path* and *curvature*.

In order to estimate the isolated importance of each feature, we also evaluated all of the features in separation. For classification of overall ADHD score, 29 single-feature models achieved a classification accuracy of above 60%. The list includes 22 features related to touch trajectories. The best isolated performance was achieved with *maximum divergence from the direct path to the left* (σ: 74.4%, σ^2: 72.2%), and Δ *maximum curvature* (σ^2: 70%). Other measures for *curvature* and *divergence from the direct path*, *correctness of the answer* and *trial duration* are also among the strongest isolated features. For the baseline features *correctness of the answer* and *trial duration*, session differences show the highest discriminative potential (between 67.8% and 61.1%). Overall, session differences make for 18 of the 29 strongest isolated features.

The impact of single features can also be assessed with a leave-one-out approach using the best performing feature set for each of the scales as a starting point. Table 1 shows the five features that make for the highest decrease in classification accuracy when removed from the respective best model for each scale.

Overall, features related to variability of behavior (i.e., σ, σ^2, and/or Δ features) have a major impact on classification results. Only few absolute measures are among the strongest features according to the above analyses (5 of 29 features with highest isolated performance, and 3 of 20 features with highest accuracy when removed from the best model as depicted in Table 1). Furthermore, our results highlight the importance of features that describe the shape of touch trajectories (*curvature* and *length of segments*, as well as *divergence from the direct path*) for ADHD prediction.

5 DISCUSSION

Primarily, the above results confirm our hypothesis that ADHD scores obtained with self-reporting questionnaires can successfully be predicted by means of touch interaction data. Both the presented regression models (with a mean error of 0.31 on a four-point scale and SD = 0.5667), and the classification models (with 91.1% accuracy for classification of the upper vs. lower thirds of ADHD scores) have shown convincing performances. Our results furthermore demonstrate that touch trajectories contain a substantive amount of information for ADHD score prediction, since regression and classification models that exclusively rely on such data outperform the respective baseline models.

In the following, we have a closer look into how these results can be interpreted and discuss the plausibility of our findings with regard to the features that have been identified to have the strongest impact on automated prediction of ADHD risk. While an exhaustive interpretation of the single features and their relevance for ADHD prediction with a detailed embedding in ADHD theory is out of the scope of this paper, we still believe that some details on the relationships of features and the used scales are important to understand the behavioral patterns which our models are based on.

Table 1: Features with the highest decrease in classification accuracy for each scale when removed from the best model.

Overall ADHD score
Δ *minimum segment length* (σ^2)
Δ *average curvature* ($\emptyset$)
Δ *correctness of the answer*
Δ *minimum curvature* ($\emptyset$)
minimum curvature (σ)
Inattention
Δ *minimum divergence to the right* ($\emptyset$)
Δ *average segment speed* (σ)
trial duration (σ^2)
minimum divergence to the right (σ)
number of boxes revealed after correct answer (σ^2)
Hyperactivity
Δ *number of strokes* (σ)
maximum overall speed ($\emptyset$)
average divergence to the left (σ)
maximum divergence to the right (σ^2)
average curvature (σ)
Impulsivity
trial duration (σ)
travel distance after first answer is revealed ($\emptyset$)
average segment speed ($\emptyset$)
Δ *number of stroke segments* (σ)
Δ *maximum curvature* (σ)

As expected (compare [5]), students with higher ADHD scores showed more variability regarding the trial completion times (correlation coefficients of $0.19 < r < 0.31$ for σ and σ^2 of trial duration with ADHD scores). Correctness of the answer is correlated with overall score and inattention (r = -0.2 and -0.27), but only has minor importance for prediction accuracy.

In general, features that describe the shape of touch trajectories, particularly *divergence from the direct path* and *curvature* of the trajectories, have proven to be of major significance – especially for the prediction of hyperactivity and impulsivity subscales. More precisely, variability measures (σ, σ^2, and Δ) of *divergence* and *curvature* contributed greatly to overall prediction accuracy. While some of these features are not or only weakly correlated with the four scales, others show weak to moderate correlations with at least one of the scales (compare Table 2 for correlation coefficients of selected features). Notably, *minimum* and *maximum curvature* (σ and σ^2) are positively correlated with overall ADHD score and inattention ($0.2 < r < 0.29$), whereas *average curvature* (σ^2) shows the strongest correlation with hyperactivity (r = 0.22). Change of variability of *average curvature* between sessions is, however, negatively correlated with overall score, inattention, and hyperactivity ($-0.28 < r < -0.2$). In summary, we observed more variability in the shape of touch trajectories of students with higher ADHD scores,

Table 2: Pearson correlation coefficients for selected features (some feature names have been abbreviated) with ADHD symptoms. *: at least one feature from the feature group ($\varnothing$, σ, σ^2) is used for regression, †: used for classification

Feature	Overall	Inatt.	Hyper.	Impuls.
trial duration (σ)	$0.30^{\dagger}$	$0.30^{\dagger}$	$0.26^{*\dagger}$	$0.20^{*\dagger}$
correctness ($\varnothing$)	$-0.20^{\dagger}$	-0.27	-0.14	-0.03
# segments (σ^2)	$0.24^{*\dagger}$	0.26	0.18*	0.09
max. segment length (σ)	0.18	0.08	$0.18^{\dagger}$	$0.22^{\dagger}$
Δ *min. segm. length* (σ)	$0.15^{*\dagger}$	0.14	$0.04^{\dagger}$	0.21*
avg. curvature (σ^2)	$0.18^{*\dagger}$	0.15	0.22	0.06
Δ *avg. curvature* (σ^2)	$-0.25^{*\dagger}$	-0.24*	$-0.28^{*\dagger}$	-0.05
min. curvature (σ)	$0.28^{*\dagger}$	$0.28^{*\dagger}$	$0.18^{*\dagger}$	0.19*
max. curvature (σ)	$0.21^{*\dagger}$	0.20	0.20	0.10
max. overall speed ($\varnothing$)	0.20*	0.13	$0.20^{\dagger}$	0.18
# boxes after correct ($\varnothing$)	$0.11^{\dagger}$	$0.28^{\dagger}$	-0.05	0.01
travel distance (σ^2)	0.12	0.21	0.04	0.02
travel dist. after corr. ($\varnothing$)	0.13*	0.31*	-0.01*	-0.07*

and this variability did not change as much between sessions. Regarding *trial duration*, again, students with higher ADHD scores showed higher variability (0.19 < r < 0.31). However, the group difference in session differences is less pronounced.

Students with higher inattention scores interacted more with the system after the correct answer was revealed at least once (r = 0.28 for *number of boxes revealed after the correct answer has been revealed*, and r = 0.31 for the *travel distance after the correct answer has been revealed*). We could not observe this for the other subscales.

The fact that variability measures generally had a higher impact on prediction accuracies suggests that modelling the temporal sequencing of single trials with more longitudinal data could further improve the results. Although we could observe differences in the time sequences for students with higher ADHD scores for certain features, our data basis does not support robust prediction models that make use of the full temporal resolution as of yet.

Our results for workload classification prove that differences in task difficulty induce a significant change in interaction behavior regardless of ADHD scores. To a certain extent, these differences are encoded in the used variability features. We still presume that additional features that put a finer point to these differences could further improve prediction accuracy. However, we found no substantial group differences between subjects with higher and lower ADHD scores regarding change of touch behavior induced by increased task difficulty. Adding features calculated only from the most difficult tasks also did not improve ADHD prediction results. Again, this could change with a larger set of interaction data or for different tasks.

6 LIMITATIONS AND OUTLOOK

Regarding the practicability of our findings, a key limitation is that we have only validated our prediction approach with scores generated from self-reporting questionnaires. Although the used questionnaires have been shown to have adequate construct validity, it remains uncertain whether our approach produces comparable results for subjects actually diagnosed with ADHD. As of now, our results show that touch interaction data contains valuable information for ADHD screening, but a direct comparison with questionnaire-based approaches is not yet possible. By labeling our data with scores from self-reports we cannot become better than the questionnaires used to obtain these scores.

In order to confirm the practicability of our ADHD prediction approach, we have to verify our methods in a clinical follow-up study with children diagnosed with ADHD. Additional data from confirmed ADHD patients is indispensable for future real-world applications, because the dataset that we have used for this work only contains few very high scores. Ideally, a model should be trained with a balanced number of samples across the whole spectrum of ADHD scores.

Although our study was conducted within classrooms, it should still be considered a lab study with controlled conditions and a limited timeframe. A longitudinal study embedded in everyday school life would provide larger quantities of interaction data, which could allow for machine-learning models with a more fine-grained temporal resolution taking greater account of the behavioral variability attributed to ADHD. We believe, however, that it is vital to maintain maximum transparency on the intended purpose of the data collection process and to deal responsibly with the collected personal data – especially when handling sensitive issues such as mental disorders. On that note, a short test session might even be preferable to continuous data collection, although the increased quantity of data could enable precise time series classification (e.g., with Long Short-Term Memory networks [12]).

In practice, a short test could use a combination of touch data and other modalities such as speech or facial expressions. It might even be beneficial to combine our methods with a traditional self-reporting questionnaire integrated into our multiple-choice setup. That way, self-reporting scales could be bolstered with the proposed interaction data models. It should, however, be noted that the reported models are specific to the used combination of tasks and UI. Changing the scenario, the interface or the dimensions of the touch screen could yield different interaction patterns and consequently lead to different models.

Furthermore, we believe that it is necessary to further look into the underlying behavioral patterns that our prediction models are based on. This could give new insights on how students with high ADHD risk interact with a touch screen, which could in turn yield valuable guiding principles for the design of future adaptive systems that foster learning for children suffering from ADHD. Given our results and prior findings about intra-individual variability of ADHD patients [5], we consider statistical methods that are well suited to handle time series data (e.g., generalized additive models [11]) to be particularly well suited for follow-up analyses.

ACKNOWLEDGMENTS

This research was funded by the LEAD Graduate School [GSC1028], a project of the Excellence Initiative of the German federal and state governments.

REFERENCES

[1] Mark H. Ashcraft and John Battaglia. 1978. Cognitive arithmetic: Evidence for retrieval and decision processes in mental addition. *Journal of Experimental Psychology: Human Learning and Memory* 4, 5 (1978), 527.

[2] Richard N. Blazey, David L. Patton, and Peter A. Parks. 2003. ADHD detection by eye saccades. US Patent 6,652,458.

[3] Marie-Christine Brault and Éric Lacourse. 2012. Prevalence of Prescribed Attention-Deficit Hyperactivity Disorder Medications and Diagnosis among Canadian Preschoolers and School-Age Children: 1994–2007. *The Canadian Journal of Psychiatry* 57, 2 (feb 2012), 93–101. https://doi.org/10.1177/070674371205700206

[4] F. Xavier Castellanos, Edmund J. S. Sonuga-Barke, Michael P. Milham, and Rosemary Tannock. 2006. Characterizing cognition in ADHD: Beyond executive dysfunction. *Trends in Cognitive Sciences* 10, 3 (2006), 117–124. https://doi.org/10.1016/j.tics.2006.01.011

[5] F. Xavier Castellanos, Edmund J. S. Sonuga-Barke, Anouk Scheres, Adriana Di Martino, Christopher Hyde, and Judith R. Walters. 2005. Varieties of attention-deficit / hyperactivity disorder-related intra-individual variability. *Biological Psychiatry* 57, 11 (2005), 1416–1423. https://doi.org/10.1016/j.biopsych.2004.12.005

[6] Manfred Döpfner, Anja Görtz-Dorten, and Gerd Lehmkuhl. 2009. *Diagnostik-System für psychische Störungen nach ICD-10 und DSM-IV für Kinder und Jugendliche-II: DISYPS-II; Manual.* Huber, Hogrefe.

[7] George J. DuPaul, Thomas J. Power, Arthur D. Anastopoulos, and Robert Reid. 1998. *ADHD Rating Scale-IV: Checklists, norms, and clinical interpretation.* Guilford Press, New York, NY, US.

[8] Moshe Fried, Eteri Tsitsiashvili, Yoram S. Bonneh, Anna Sterkin, Tamara Wygnanski-Jaffe, Tamir Epstein, and Uri Polat. 2014. ADHD subjects fail to suppress eye blinks and microsaccades while anticipating visual stimuli but recover with medication. *Vision research* 101 (2014), 62–72. https://doi.org/10.1016/j.visres.2014.05.004

[9] Yuan Gao, Nadia Bianchi-Berthouze, and Hongying Meng. 2012. What does touch tell us about emotions in touchscreen-based gameplay? *ACM Transactions on Computer-Human Interaction (TOCHI)* 19, 4 (2012), 31. https://doi.org/10.1145/2395131.2395138

[10] Isabelle Guyon and André Elisseeff. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)* 3, 3 (2003), 1157–1182.

[11] Trevor J. Hastie and Robert Tibshirani. 1990. *Generalized Additive Models*. Vol. 43. CRC Press.

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[13] Blair A. Johnston, Benson Mwangi, Keith Matthews, David Coghill, Kerstin Konrad, and J. Douglas Steele. 2014. Brainstem abnormalities in attention deficit hyperactivity disorder support high accuracy individual diagnostic classification. *Human brain mapping* 35, 10 (2014), 5179–5189. https://doi.org/10.1002/hbm.22542

[14] Ronald C. Kessler, Lenard Adler, Minnie Ames, Olga Demler, Steve Faraone, E. V. A. Hiripi, Mary J. Howes, Robert Jin, Kristina Secnik, and Thomas Spencer. 2005. The World Health Organization Adult ADHD Self-Report Scale (ASRS): a short screening scale for use in the general population. *Psychological medicine* 35, 2 (2005), 245–256.

[15] Elise Klein, Korbinian Moeller, Katharina Dressel, Frank Domahs, Guilherme Wood, Klaus Willmes, and Hans-Christoph Nuerk. 2010. To carry or not to carry – Is this the question? Disentangling the carry effect in multi-digit addition. *Acta psychologica* 135, 1 (2010), 67–76. https://doi.org/10.1016/j.actpsy.2010.06.002

[16] Ron Kohavi and George H. John. 1997. Wrappers for feature subset selection. *Artificial Intelligence* 97 (1997), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

[17] Klaus D. Kubinger. 2009. *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnostizierens.* Hogrefe Verlag.

[18] S. Lis, N. Baer, C. Stein-En-Nosse, B. Gallhofer, G. Sammer, and P. Kirsch. 2010. Objective measurement of motor activity during cognitive performance in adults with attention-deficit/hyperactivity disorder. *Acta Psychiatrica Scandinavica* 122, 4 (feb 2010), 285–294. https://doi.org/10.1111/j.1600-0447.2010.01549.x

[19] Sandra K. Loo and Scott Makeig. 2012. Clinical Utility of EEG in Attention-Deficit/Hyperactivity Disorder: A Research Update. *Neurotherapeutics* 9, 3 (2012), 569–587. https://doi.org/10.1007/s13311-012-0131-z

[20] Philipp Mock, Peter Gerjets, Maike Tibus, Ulrich Trautwein, Korbinian Möller, and Wolfgang Rosenstiel. 2016. Using Touchscreen Interaction Data to Predict Cognitive Workload. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16)*. ACM, 349–356. https://doi.org/10.1145/2993148.2993202

[21] Yukifumi Monden, Ippeita Dan, Masako Nagashima, Haruka Dan, Minako Uga, Takahiro Ikeda, Daisuke Tsuzuki, Yasushi Kyutoku, Yuji Gunji, and Daisuke Hirano. 2015. Individual classification of ADHD children by right prefrontal hemodynamic responses during a go/no-go task as assessed by fNIRS. *NeuroImage: Clinical* 9 (2015), 1–12. https://doi.org/10.1016/j.nicl.2015.06.011

[22] 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 (jul 2014), 922–926. https://doi.org/10.1016/J.MEDENGPHY.2014.02.023

[23] Tuulia M. Ortner, Ralf Horn, Martin Kersting, Stefan Krumm, Klaus D. Kubinger, René T. Proyer, Lothar Schmidt-Atzert, Gernot Schuhfried, Astrid Schütz, Michaela M. Wagner-Menghin, et al. 2007. Standortbestimmung und Zukunft Objektiver Persönlichkeitstests. *Report Psychologie* 32, 2 (2007), 60–69.

[24] Sengwee Toh. 2006. Datapoints: Trends in ADHD and stimulant use among children, 1993-2003. *Psychiatric Services* 57, 8 (2006), 1091.

[25] Radu-Daniel Vatavu, Lisa Anthony, and Quincy Brown. 2015. Child or adult? Inferring Smartphone users' age group from touch measurements alone. In *Human-Computer Interaction – INTERACT 2015*. Springer, 1–9. https://doi.org/10.1007/978-3-319-22723-8_1

[26] Erik G. Willcutt, Alysa E. Doyle, Joel T. Nigg, Stephen V. Faraone, and Bruce F. Pennington. 2005. Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry* 57, 11 (2005), 1336–1346. https://doi.org/10.1016/j.biopsych.2005.02.006

[27] Zoe Young, Michael P. Craven, Maddie Groom, and John Crowe. 2014. Snappy App: a mobile continuous performance test with physical activity measurement for assessing Attention Deficit Hyperactivity Disorder. In *International Conference on Human-Computer Interaction*. Springer, 363–373. https://doi.org/10.1007/978-3-319-07227-2_35