


Article

# PANDAS: Paediatric Attention-Deficit/Hyperactivity Disorder Application Software

Hervé Mukenya Mwamba \*, Pieter Rousseau Fourie and Dawie van den Heever

Biomedical Engineering Research Group (BERG), Department of Mechanical & Mechatronic Engineering, University of Stellenbosch, Cape Town 7602, South Africa; pieter@innovation4life.com (P.R.F.); dawie@sun.ac.za (D.v.d.H.)

\* Correspondence: hervemwamba279@gmail.com; Tel.: +27-60-821-1001

Received: 23 February 2019; Accepted: 28 March 2019; Published: 20 April 2019



**Abstract:** Attention-deficit/hyperactivity disorder (ADHD) is a common neuropsychiatric disorder that impairs social, academic and occupational functioning in children, adolescents and adults. In South Africa, youth prevalence of ADHD is estimated as 10%. It is therefore necessary to further investigate methods that objectively diagnose, treat and manage the disorder. The aim of the study was to develop a novel method that could be used as an aid to provide screening for ADHD. The study comprised of a beta-testing phase that included 30 children (19 non-ADHD and 11 ADHD) between the ages of 5 and 16 years old. The strategy was to use a tablet-based game that gathered real-time user data during game-play. This data was then used to train a linear binary support vector machine (SVM). The objective of the SVM was to differentiate between an ADHD individual versus a non-ADHD individual. A feature set was extracted from the gathered data and sequential forward selection (SFS) was performed to select the most significant features. The test set accuracy of 85.7% and leave-one-out cross-validation (LOOCV) accuracy of 83.5% were achieved. Overall, the classification accuracy of the trained SVM was 86.5%. Finally, the sensitivity of the model was 75% and this was seen as a moderate result. Since the sample size was fairly small, the results of the classifier were only seen as suggestive rather than conclusive. Therefore, the performance of the classifier was indicative that a quantitative tool could indeed be developed to perform screening for ADHD.

**Keywords:** ADHD; screening; machine learning; SVM; children; novel

## 1. Introduction

Attention-deficit/hyperactivity disorder (ADHD) is a brain disorder marked by an ongoing pattern of inattention and/or hyperactivity-impulsiveness that interferes with functioning or development [1]. Its exact origins are uncertain and complex [2] and its diagnosis relies almost exclusively on subjective assessments of perceived behaviour [3]. This presents some unresolved dilemmas. Firstly, there is a potential risk of over-diagnosis. Secondly, males are more likely to be diagnosed compared to females of the same age [4]. Finally, objective diagnostic methods are scarce. Furthermore, it is important to note the significant financial burden associated with the treatment and management of ADHD. It is estimated that for an adult with ADHD, the economic burden is approximately \$3020 per annum [5].

Diagnosis of ADHD is based on clinical criteria defined by the Diagnostic and Statistical Manual of Mental Disorders (DSM 5), or the International Classification of Diseases (ICD 10) [2]. Proper diagnosis involves clinical interviews, patient history, psychometric testing and rating scales [6]. Since comorbidity may occur, diagnosis is patient-specific. It is observed that environmental factors such as peri/pre-natal, psychological and dietary, contribute to the development and severity of ADHD, but these factors may be consequential rather than causal [2]. Furthermore, the consensus is that ADHD is associated with dysfunction of the prefrontal cortex [7–9].

Various methods exist, where objective diagnosis was attempted, with some degree of success. The first category of method is computerized. The most popular and efficient computerized approach is continuous performance testing (CPT). The Conners CPT 3 test [10] is a commercially available test that evaluates attention disorders and neurological functioning. Its aim is to provide an objective evaluation of individuals aged eight years and older. During the 14-minute-long assessment, subjects are to click whenever any letter except 'X' appears on the screen. Using a normative sample of 1400 subjects, representative of an American population, it was found that the classification accuracy was 83.90%, the sensitivity 86% and the specificity 81%.

An example of a tool that integrates the Conners CPT 3 test in a more interactive manner is called MOXO. Instead of presenting the subject with the letter 'X', the subject must click whenever a specific face appears on the screen. Additional studies [11–13] have been done using MOXO to further validate the discriminatory ability of CPT testing. In these studies, comparisons were made between ADHD and non-ADHD groups with regards to the number of omission errors during the test. The difference between the studies was that different age groups were used, and different environmental distractors were used. It was found that when using MOXO, the sensitivity was between 81–91% (depending on the age group) and the specificity was between 85–89%. However, the studies did not focus on measuring the classification accuracy of MOXO and only used the sensitivity and specificity as the performance metrics.

In addition to MOXO, it was also found that objective diagnosis of ADHD could be attempted using inertial sensors that were mounted to subjects and that captured data for one hour through various scenarios [14]. Classification was then done with a linear support vector machine (SVM) model. The results found in the study were that the classification accuracy was 95.12%, the specificity 95.65% and the sensitivity 94.44%.

Additionally, there has been an FDA-approved device called NEBA that uses electroencephalogram signals to aid clinicians to make diagnoses. The most recent study [15] where NEBA was used was done in 2015. The aim of the study was to determine correlation between the diagnosis from NEBA and diagnosis from the consensus of a team of experts. It was found that the combination of NEBA and consensus diagnosis yielded 97% accuracy.

Finally, a study where neuroimaging was used to monitor activity in specific regions of the brain, namely, the pre-frontal cortex, is found in literature [8]. The major conclusions of this study were that a quantitative analysis of neuroanatomical abnormalities in ADHD was provided and that this could be used as the starting point of other studies.

Other challenges of ADHD are its treatment and management. For children between the ages of 6 and 18 years old, symptoms are typically identified in a classroom setting [2]. Specialists such as child psychologists will then interview the whole family and provide the parents and the teacher(s) with questionnaires called rating scales. Furthermore, the patient undergoes a series of psychometric tests. Based on the results of the rating scales as well as those of the psychometric tests, a thorough diagnosis is made, and a treatment plan is drawn up. The patient then goes for follow-up sessions, generally after every 6 months. This process represents the ideal case of identifying and diagnosing ADHD [2]. However, many schoolchildren are not identified as potential ADHD patients. This presents the need for a screening tool that may help identify the disorder from an early stage, such as in a classroom setting.

Since ADHD is defined by the Diagnostic and Statistical Manual for mental health, it can only be diagnosed clinically by a specialist (psychiatrist, psychologist, paediatrician etc.). Any other diagnosis cannot be given. Given this fact, the aim of the study was not to provide a diagnosis, but to provide screening. Ultimately, although there may be various methods and technologies that may claim to provide diagnosis, but clinically speaking, DSM 5 criteria must be met for an individual to be classified as having ADHD.

Thus, the aim of the study was to develop a novel method that provided rapid screening for the hyperactive subtype of ADHD. A concurrent study was done where the screening was done for

the inattentive subtype. The output of the study was a diagnostic aid rather than a diagnostic tool. The final diagnosis was still to be given by a specialist. The study was broken down into the following different phases/objectives: (1) identification of measurable parameters for ADHD based on DSM-5 criteria and psychometric tests; (2) design and development of the tablet-based game; (3) performing beta-tests to gather data; (4) development of SVM classifier.

Since mobile tablets have become popular and very accessible to the general public, playing games on tablets has become a ubiquitous activity. The fact that tablet games are popular and enjoyable made it a good choice for the platform to use for the novel method developed in the study.

## 2. Software Design

The software that was used was in the form of a tablet-based game with an underlying layer of data processing. The main aspect of this two-layer approach was to use the game layer as input to the data-processing layer.

### 2.1. Game Design

A tablet-based game was developed using Unreal Engine v.4.18.0, one of the most popular and reliable platforms for game development for electronic devices. The device that was chosen was the NVIDIA K1 Shield Tablet. The theme that was chosen considered the age group of the subjects that were used for the research. Since the subjects were schoolchildren a jungle/tropical theme was a logical choice.

The objective of the task was to travel on a raft from one end of a river to the other end as quickly as possible. This had to be done while avoiding obstacles and collecting as many gems as possible. The speed of the raft increased as the game progressed, provided that no obstacles were hit. Three straight lanes were present, and the user was able to move to the left or right lane, while the middle lane was the default position. Figure 1 shows a screenshot of the game. The buttons can be seen at the bottom left and right corners of the screen. The buttons were used to navigate between the lanes, jump over incoming obstacles and throw objects at incoming obstacles in order to destroy them. The number of gems collected is displayed in the top left corner and the PANDA character is in the middle bottom section.



Figure 1. Screenshot of Game.

### 2.2. Data Processing

The data processing layer was broken down into three phases: data gathering, data storage and data extraction. This data processing was essential in building the SVM classifier. The raw data that was captured during game-play could be broken down into three categories: 1. personal user data; 2. game-play variables and; 3. accelerometer data. Personal user data was user-specific and included the following: age, gender, race, game enjoyment (“yes” or “no”) and diagnosis (ADHD or non-ADHD). This data was captured manually by the test administrator at the end of a game session and recorded onto an Excel spreadsheet, where each user was given a unique identifier for traceability within the spreadsheet. The use of game-play variables and accelerometer variables was derived from

translating the applicable DSM-V criteria into measurable parameters. Thus, the resulting parameters were the following: mistakes made, task completion time, task termination, distractibility, forgetfulness, sustained attention, sustained attention and device motion. The accelerometer data contained raw time-series data from the tri-axial accelerometer on the device. All this data was then stored on a cloud Firebase database. This is illustrated in Figure 2.

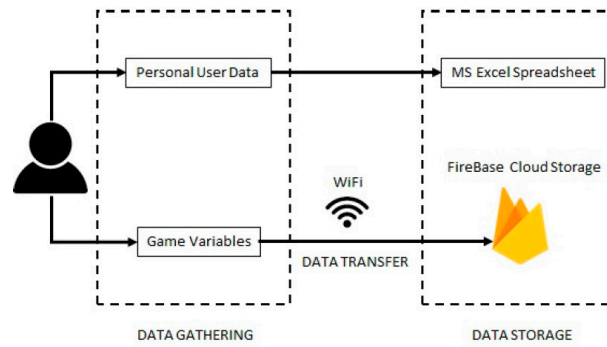


Figure 2. Data Gathering and Data Storage Flow.

Once the data was stored on the cloud, it could be downloaded from the Firebase database into .json files, where each .json file contained individual game session data. File processing was then done to extract the data from the .json files and store them into appropriate structures (matrices) for the SVM classifier. Figure 3 shows the file extraction process, which was implemented in MATLAB.

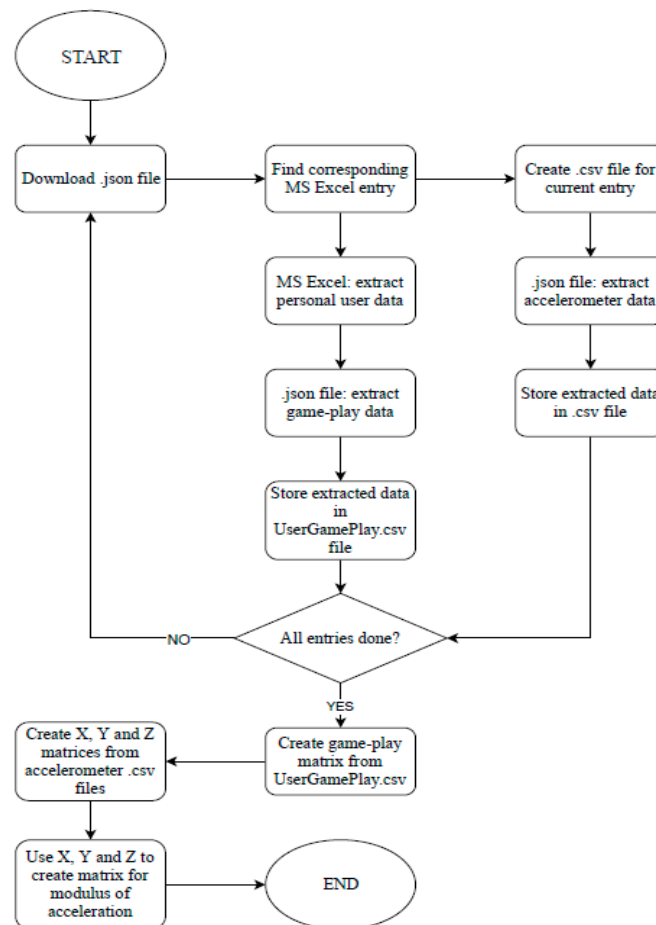


Figure 3. Data Extraction Process.

### 3. Methods

#### 3.1. Subjects

The beta-test consisted of 30 subjects between the ages of 5 and 16 years of old. These subjects had been consulted by a specialist at a private paediatric practice at the Cape Gate Medi-Clinic. Subject participation was completely voluntary and parental consent was sought using information leaflets. The subjects also had to read and sign assent forms. The main inclusion criterion was age, since ADHD is most prevalent in minors. Furthermore, gender ratio was kept as closely as possible to 1:1. Children with a known history of severe mental illness were excluded. Additionally, children that suffered from photosensitive epilepsy were also excluded given the fact that the game images on the screen of the tablet could possibly trigger convulsions. Table 1 shows a breakdown of subject distribution.

**Table 1.** Subject Distribution; ADHD: Attention-deficit/hyperactivity disorder.

Category	Value
Mean age	10 y/o
Maximum age	16 y/o
Minimum age	5 y/o
Number of males	16
Number of females	14
Number of ADHD males	6
Number of ADHD females	7
Number of non-ADHD males	10
Number of non-ADHD females	7
<b>Total subjects</b>	<b>30</b>

#### 3.2. Ethical Approval

The ethical approval process was administered by the Health Research Ethics Committee (HREC) of Stellenbosch University. According to the HREC’s definitions, the research was identified as a clinical trial because its purposes were to test effectiveness and efficacy of a diagnosis-aiding tool. The risk of the research was minimal since the testing only consisted of playing a game on a tablet. Ethical approval was obtained on the 14th July 2017 and is valid until the 13th July 2018. It was subsequently extended to July 2019.

#### 3.3. Power Analysis

A power study was done in the initial stages of sample size estimation with consultation of a statistician. The McNemar test, as defined by [16], was used, as shown in Table 2 below.

**Table 2.** McNemar Test.

	Method 1 Positive	Method 2 Negative	Row Total
Method 2 Positive	A	b	a + b
Method 2 Negative	C	d	c + d
Column Total	a + c	b + d	N

In this case, method 1 referred to the current diagnostic method and method 2 was the proposed method resulting from this study. The POWER analysis was performed using the Statistica software package and the following parameters were used as input:

- Delta: the difference in population proportion when Method 1 was positive and the population proportion when Method 2 was positive as described in McNemar’s test;

- Nuisance parameter: the total proportion of times different events occur for the two methods. This was chosen as 0.4 based on the statistician’s recommendation;
- Type-I error rate: This value is taken as 0.05 and means that one is willing to accept that there is a 5% chance that the null hypothesis is wrong. 0.05 is the standard accepted value.
- Power goal: 0.9

The POWER analysis was applied to various cases where the value of Delta was changed as shown in Table 3. Delta was kept less than 20%. The required sample size that was chosen was 156 subjects. This occurred with Delta = 0:16. In other words, the error in distinguishing between the gold standard method and the new method was 16% (84% accuracy). It was decided, however, that for the beta-testing, a sample size of only 30 subjects would be used to get preliminary results.

Table 3. POWER Analysis.

$\delta$	$\eta$	$\alpha$	Power Goal	Actual Power for Required N	Required Sample Size
0.1	0.4	0.05	0.9	0.9004	412
0.15	0.4	0.05	0.9	0.9003	178
0.16	0.4	0.05	0.9	0.9016	156
0.17	0.4	0.05	0.9	0.9013	137
0.18	0.4	0.05	0.9	0.9009	121
0.2	0.4	0.05	0.9	0.9002	96

### 3.4. Study Design

The study consisted of two main phases: (1) design and development, (2) testing and data collection. The sequence of research activities is illustrated in Figure 4, where activities 5, 7 and 10 were the testing activities and where activities 3, 6, 8, 9, 11, 12 and 13 were the design and development activities.

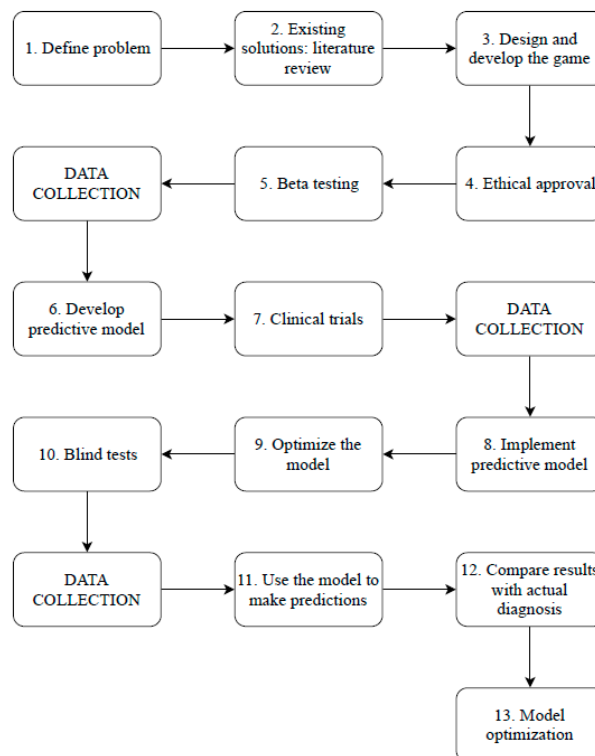


Figure 4. Sequence of Research Activities.



### 3.5. Standard Operating Procedure

Testing took place at Cape Gate Medi-Clinic. The subject was placed in a well-lit room and the tablet was presented to them on a table. The investigator then gave the subjects instructions before launching the game. The subject was given a maximum of 8 minutes to complete the task, unless he/she decided to quit. For each subject, the investigator recorded the name, age, diagnosis (for control group), race and additional comments/observations into an Excel spreadsheet. Data was recorded during game-play and sent to a secure Firebase database via a private WiFi connection.

### 3.6. Feature Extraction

Feature extraction is an important step in building a classifier, as it allows for raw data to be interpreted into meaningful information that can help the classifier distinguish one observation from another. Prior to calculating features to extract, outlier detection was performed. The method used took the interquartile range (IQR) into account where

$$IQR = Q_3 - Q_1 \quad (1)$$

and  $Q_3$  and  $Q_1$  represent the middle values of the first and third half of the dataset respectively. A data-point  $x_i$  was seen as an outlier if it satisfied one of the following two conditions:

$$x_i > Q_3 + 1.5 \times IQR \quad (2)$$

$$x_i < Q_1 - 1.5 \times IQR \quad (3)$$

If one of the conditions was met, then  $x_i$  was replaced by the equation proposed by [17]:

$$x_i := \frac{x_{i-1} - x_{i+1}}{2} \quad (4)$$

Two feature sets were extracted from the datasets. The first one was referred to as the game feature set and the second one was the accelerometer feature set. The game feature set consisted of the following 16 features: number of left button presses, number of right button presses, number of jumps, number of throws, number of obstacles destroyed, number of obstacles hit, number of gems collected, game duration, game enjoyment (boolean: 1 = enjoyed, 2 = did not enjoy), throw efficiency, total button presses, directional button presses, frequency of button presses, age, race and gender. The accelerometer feature set consisted of the following 18 statistical features: mean, standard deviation, minimum, maximum, range, median, sum, variance, skewness, kurtosis, root mean square (RMS), percentiles (10th, 25th, 50th, 75th and 90th), interquartile range (IQR) and crest factor. This resulted in a total of 72 features ( $18 \times 4$ ), where each of these features was calculated for the x, y, z axes and for the modulus which was calculated in the following way:

$$a_{total} = \sqrt{a_x^2 + a_y^2 + a_z^2} \quad (5)$$

The following 10 morphological features were extracted: exact Euclidean distance, autocorrelation coefficient, positive area, negative area, total area, absolute total area, total absolute area, number of zero crossings, latency time and peak-to-peak time-window. This resulted in a total of 39 features ( $10 \times 4 - 1$ ), where each of these features was calculated for the x, y, z axes and for the modulus, except for the exact Euclidean distance which was only calculated for the three possible axis pairings. Table 4 shows the breakdown of the feature set. The user specific features (race, age and gender) were important features in terms of correlating gender and age to the outcome of a diagnosis. Race was not necessarily significant for this sample size, but it was included more out of a speculative point of view. It was later seen that no conclusions could be made from this feature affecting the diagnosis. The game-play features were of significance because they gave insight into how subjects

played the game (i.e., gaming behaviour). Statistical and morphological features were extracted from the multivariate accelerometer time series, as it has been proven to provide good results for this type of data.

**Table 4.** Breakdown of feature set.

Feature Type	Number of Features
Statistical	72
Morphological	39
Game-play	8
Game-play-derived	4
User-specific	3
<b>Total</b>	<b>126</b>

Once the features were extracted, feature normalization was then performed so that the mean of each feature in the set was equal to zero. This helped improve the efficiency of the classifier. Given a feature  $f_i$ , normalization performs the following:

$$f_i : \frac{f_i - \mu_i}{s_i} \quad (6)$$

where  $\mu_i$  is the average of all the observations for that feature and  $s_i$  is the range of the observations of that feature.

### 3.7. Feature Selection

Due to the small sample size, feature ranking could bias the classification accuracy and therefore was deemed unfit for this specific application. As a preliminary feature selection strategy, correlation matrices were used to visually inspect any strong correlations. One feature was removed using this method. Five combinations of feature sets were generated and evaluated. The first feature set was when all the features were used. The second feature set consisted of performing sequential forward selection (SFS) on the full feature set. The third feature set (combine set) was constructed from performing SFS both on the accelerometer features and, on the game, and user features and combining the resulting features. The last two feature sets were constructed from performing SFS on the game and user features and combining that with the morphological features for the one set and the statistical features for the other set. The feature selection sets are shown in Table 5. Ultimately, the feature set that yielded the lowest leave-one-out cross-validation (LOOCV) error was selected. LOOCV was used given the small sample size.

**Table 5.** Feature Selection Sets.

Feature Set	Selection Method	Number of Features
$F_{all}$	All features selected	126
$F_{SFS}$	SFS	10
$F_{combined}$	SFS	21
$F_{man1}$	SFS and manual	83
$F_{man2}$	SFS and manual	55

## 4. Support Vector Machine Model

The approach that was taken in building the SVM classifier was based on the recommendation given by [18] in the following steps:

1. Transform extracted data into the format supported by the SVM package;
2. Perform simple scaling and normalization of the data;



3. Consider a linear kernel as an initial model;
4. Tune the parameter regularization parameter “C” using leave-one-out-cross-validation (LOOCV) method;
5. Use the tuned parameter C to train the whole training set;
6. Test the model on a test set.

The SVM model was implemented in Matlab using the “fitsvm” function. The inputs to this function were the following:

1. A scaled and normalized dataset X;
2. Target values y, where y is part of [0,1];
3. Regularization parameter C. This parameter adds a penalty to features, thus reducing their overall effect. It is tuned to limit over-fitting;
4. The kernel type (linear);

The output of the “fitsvm” function was an SVM model. Since five possible feature sets were evaluated, five models were created and the one with best performance was chosen. The process for creating and choosing the best feature set is shown below:

1. Choose an initial value for C: C = 10;
2. Perform LOOCV on each of the five feature sets;
3. Report LOOCV error for each of the five feature sets;
4. Repeat steps 1 to 3 for different values of C: C = 3, 1, 0.3;
5. Choose the feature set with the lowest LOOCV error;
6. Use the corresponding C value to build a final model;
7. Train the model using the whole training set (unlike for LOOCV that uses a subset of the training set);
8. Use the test set to make predictions with the new model

The results of this process are shown in Table 6 which simply shows that the smallest LOOCV error that was found was with a C value of 0.3. This corresponded to  $F_{combined}$ , where the LOOCV was at a minimum of 0.165. Once the classifier was trained, it was used on the test set to make predictions, using Matlab’s “predict” function.

**Table 6.** Leave-one-out cross-validation (LOOCV) error on different feature sets and various C values.

Feature Set	Features	C			
		10	3	1	0.3
$F_{all}$	126	0.5	0.5	0.5	0.5
$F_{SFS}$	10	0.25	0.1875	0.1875	0.1875
$F_{combined}$	21	0.4375	0.4375	0.415	<b>0.165</b>
$F_{man1}$	83	0.46	0.4375	0.437	0.437
$F_{man2}$	55	0.43	0.4	0.4	0.4

### 5. Results

As mentioned previously, the small sample size that was used for the study induced certain limitations. Chief among them was the validity of the results. However, the results that will be discussed in this section are suggestive rather than conclusive. The main results of the performance of the classifier can be seen in the confusion matrix in Figure 5. The last column shows the percentages correctly classified examples (green) and incorrectly classified examples (red) for each class. The last row shows the same thing for each class. Respectively, the last column may be referred to as the recall, and the last row may be referred to as the false negative rate. These results came from the test set.

Since training accuracy is not a good indication of a classifier’s performance, the training set confusion matrix was not calculated. Various performance metrics were derived from the confusion matrix and are explained next.

Output Class	0	3 42.90%	1 14.30%	75% 25%
	1	0 0%	3 42.90%	100% 0%
		100% 0%	75% 25%	85.70% 14.30%
		0	1	
		Target Class		

Figure 5. Confusion matrix of SVM classifier.

According to the confusion matrix, there were three true positives (TP), three true negatives (TN), one false negative (FN) and no false positives (FP). Table 7 shows the metrics that were used to evaluate the performance of the classifier. The LOOCV accuracy was calculated as 83.5%.

Table 7. Performance metrics of support vector machine (SVM).

Metric	Equation	Value
Test set accuracy (ACC)	$ACC = \frac{TP+TN}{TP+TN+FP+FN}$	0.857
True positive rate (TPR)	$TPR = \frac{TP}{TP+FN}$	0.75
True negative rate (TNR)	$TNR = \frac{TN}{TN+FP}$	1.00
Positive predictive value (PPV)	$PPV = \frac{TP}{TP+FP}$	1.00
Negative predictive value (NPV)	$NPV = \frac{TN}{TN+FN}$	0.75
F1 score	$F_1 = 2 \frac{PPV \times TPR}{PPV + TPR}$	0.857
Type I error	$\alpha = 1 - TNR$	0
Type II error	$\beta = 1 - TPR$	0.25

The reason for choosing 7 children was based on the train-test split when using machine learning, and more specifically SVM. It has been demonstrated that a good split is to use 75% of the full set for training, and 25% of the set for testing. For the sample size of N = 30, this resulted in a test set of 7 children, which is approximately 25%. The proportion of boys to girls is based on the fact that the training and test sets are chosen randomly, so long as the train-test split remains 75:25. The randomness insures that the designer is not biased to pick suitable subjects to yield maximum performance, but that the model will be robust enough to yield reliable results. As a result of the randomness, the proportion of boys to girl was 2:5. Should the study be repeated, a different proportion could be found in the test set, yet the conclusions of the results would still be the same.

## 6. Discussion and Conclusion

### 6.1. Discussion

As a concluding remark to the interpretation of the results, what was seen was that a classifier's performance does not solely rely on its test and cross-validation accuracies. Although cross-validation is a robust way to build classifiers and gives a general indication of how well the model will perform, models should be chosen based on their practical application as well. For example, the classifier built for this study was to be used for screening of ADHD. This means that other metrics become very relevant for assessing the model. Such metrics include sensitivity, specificity and recall.

According to the statistical analysis that was done to estimate a sample size, it was recommended that a total of 200 subjects be used in order to achieve a model accuracy of 84%. The main aim of the study was to conduct a clinical trial, given this sample size. The first step was to perform beta-tests on a smaller population ( $N = 30$ ) in order to demonstrate the validity of the use of machine learning models. Although the beta-test results were seen as preliminary results, they were indicative enough to be used to demonstrate that the research question could be answered. Given the time constraints on the study, it was decided that clinical trials would form part of future work. The aim was to develop a screening tool for ADHD, and the beta-test was able to provide a solution for that.

The machine learning model that was implemented was SVM with a linear kernel. Due to the high dimensionality of the dataset, features were extracted through statistical and morphological analysis. Feature selection was then performed in order to have the most representative feature subset. Due to the small size of the dataset, leave-one-out cross-validation was chosen to determine the generalization error of the classifier, as well as to tune the regularization parameter. The feature set that was chosen consisted of 21 features that were selected using sequential forward selection. This feature selection method outperformed the other 3 methods that were used. The selected features included 11 of the game-play features and 10 of the features extracted from the accelerometer.

It can be seen that the test set accuracy and LOOCV accuracy are both high. This is expected given a small dataset. The sensitivity (TPR) relates to the classifier's ability to classify ADHD test subjects as having ADHD. Sensitivity was therefore an important characteristic of the classifier, especially for screening. Good classifier performance would require for the classifier to correctly identify subjects that are ADHD. Here the sensitivity is 0.75. This means that 75% of the time, the classifier will be able to detect the presence of ADHD. Although ADHD is sometimes difficult to detect, even with classical methods, a sensitivity of 75% is quite low. The specificity here of 1 shows that all non-ADHD test subjects were correctly classified.

Performance metrics of the classifier revealed that although the test and LOOCV accuracies were good (85.7% and 83.5% respectively) care had to be taken when selecting a classifier as being optimal. Important metrics, especially for diagnosing/screening conditions included specificity and sensitivity, which relate to how well a classifier correctly rules out negatives and correctly includes positives. From a screening point of view, the penalty is not as large as for diagnosis, but it is most desirable to have very high sensitivity and acceptable to high specificity. It was seen that the sensitivity was 75% while the specificity was 100%. The sensitivity was seen as low, while the specificity, although being high, was specific to this small dataset and would most likely decrease with a bigger dataset.

The positive predictive value (PPV) relates to the relevance of the outputs that were classified. A precision of 1 means that all the outputs that were classified were relevant. The negative predictive value (NPV) shows that 75% of relevant targets were selected.

The F1 score shows the balance between precision and recall. Values of F1 that are very high or very low, show that precision and recall are not well balanced. This appeared to be the case with this classifier. The high value of 85.7% suggests that the model may have high precision and low recall, or vice versa.

The type I error of 0 suggests that the null hypothesis was true, and accepted. Although this metric is not indicative given the dataset, it would have been approximately equal to 0.05 for a larger

set. The type II error of 0.25 is quite large and suggests that there is 25% probability that the classifier may predict false positives.

In addition to the performance metrics that were discussed, a comparison of the test set distribution and target set distribution was made. The following observations were made: (1) the target values comprised of 4 ADHD subjects and 3 non-ADHD subjects; (2) the predicted values comprised of 3 ADHD subjects and 4 non-ADHD subjects; (3) the test set comprised of 2 boys, 1 of which was ADHD; (4) the test set comprised of 5 girls, 3 of which were ADHD; (6) all the boys with ADHD were classified correctly; (7) all the boys without ADHD were classified correctly; (8) Out of the 3 girls with ADHD in the test set, 2 were classified correctly; (9) all the girls without ADHD were classified correctly.

Although no major conclusions can be drawn from these few observations it is interesting to note that the classifier was able to correctly reject all the boys and girls that didn't have ADHD, as suggested by the 100% specificity. Contrary to the claim that boys are more misdiagnosed than girls, the test set shows that all boys were correctly classified. This observation does not resolve the claim, however, since the dataset was not representative enough of a wider ADHD population.

A comparison of this study's results with other studies and existing tools pertaining to the objective diagnosis of ADHD reveal that the results are close enough, especially considering the small size of the dataset. More specifically, the sensitivity of the proposed method was generally outperformed by the other methods by at least 5%. The specificity found for this method was 100% and this was seen as a biased result, that couldn't be used as representative of the method. The accuracy of the proposed method also performed moderately, although being lower than the other methods by at least 5–7%.

## 6.2. Conclusion

The biggest disadvantage of the method is the small sample size. The significance of this is that the results cannot be treated as conclusive but only indicative. The confidence in the classification is not great, as over-fitting is likely to occur with such a small sample size. However, it has been demonstrated by [2] that SVM can be used for ADHD diagnosis with a sample size of 42, which is the closest study in terms of sample size to date.

What is advantageous about this method is that the method has not yet been explored, in the sense that a game has not been used for screening purposes. Another advantage is the ability to provide screening, without the need of going to a specialist, as this tool could be used by parent's and teachers. The method could curb costs quite significantly, by doing early screening and possible detection, as well as limit over-diagnosis.

Due to the complexity of game development, a simple game with minimal features was implemented. Next, it is recommended that a more interactive and complex game be developed, where more features can be extracted, and more parameters can be monitored. A more complex implementation would give a feature set with higher quality and possibly better classifier performance. Furthermore, many studies have shown that the use of multivariate-time-series (MTS) can help accurately classify diseases such as cancer and even ADHD. Such MTS data is found in the signals of electroencephalograms (EEG), electrocardiograms (ECG) and electromyograms (EMG). These could be implemented into the game by placing sensors and electrodes on subjects. Additional physiological markers could be added, such as eye tracking and heart rate.

The study that was conducted was able to suggest an answer to the research question that was presented, that is: a person can be screened for ADHD using quantitative methods. It was seen that the classifier showed acceptable results, especially considering that those results were only preliminary. It was demonstrated that, given a data acquisition method, in this case being the game tablet, meaningful data could be extracted and used to build a predictive model. The methods that were used to build the model were based on an extensive literature review, where it was shown successfully how those methods were performed with reliability and repeatability. Therefore, the classifier developed for the study was not novel in itself, but it was the whole design process that was novel.

**Author Contributions:** Supervision, P.R.F. and D.v.d.H.; conceptualization, H.M.M. and P.R.F.; methodology, H.M.M. and P.R.F.; writing—review and editing, P.R.F. and D.v.d.H.

**Funding:** This research was privately funded by innovation4life.

**Acknowledgments:** The clinical assistance of Rose-Hannah Brown from the Cape Gate Therapy Centre is acknowledged. The development of the game software by Mark Atkinson is also acknowledged.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. nimh.nih.gov: NIMH: Attention Deficit Hyperactivity Disorder. 2016. Available online: <http://www.nimh.nih.gov/health> (accessed on 29 June 2016).
2. Schellack, N.; Meyer, H. The Management of Attention Deficit-Hyperactivity Disorder in Children: Evidence-Based Pharmacy Practice. *SA Pharm. J.* **2012**, *79*, 12–20.
3. Gualtieri, C.T.; Johnson, L.G. ADHD: Is Objective Diagnosis Possible? *Psychiatry* **2005**, *2*, 44.
4. Bruchmüller, K.; Margraf, J.; Schneider, S. Is ADHD Diagnosed in Accord with Diagnostic Criteria? Overdiagnosis and Influence of Client Gender on Diagnosis. *J. Consult. Clin. Psychol.* **2012**, *80*, 128. [CrossRef]
5. Schoeman, R.; Liebenberg, R. The South African Society of Psychiatrist/Psychiatry Management Group Management Guidelines for Adult Attention-Deficit/Hyperactivity Disorder. *S. Afr. J. Psychiatry* **2017**, *23*, 1–14. [CrossRef] [PubMed]
6. Dopheide, J.A.; Pliszka, S.R. Attention-Deficit-Hyperactivity Disorder: An Update. *Pharmacotherapy* **2009**, *29*, 656–679. [CrossRef] [PubMed]
7. Biederman, J. Attention-Deficit/Hyperactivity Disorder: A Selective Overview. *Biol. Psychiatry* **2005**, *57*, 1215–1220. [CrossRef] [PubMed]
8. Valera, E.M.; Faraone, S.V.; Murray, K.E.; Sideman, L.J. Meta-Analysis of Structural Imaging Findings in Attention-Deficit/Hyperactivity Disorder. *Biol. Psychiatry* **2007**, *61*, 1361–1369. [CrossRef] [PubMed]
9. Toplak, M.E.; Tannock, R. Time Perception: Modality and Duration Effects in Attention-Deficit/Hyperactivity Disorder (ADHD). *J. Abnorm. Child Psychol.* **2005**, *33*, 639–654. [CrossRef] [PubMed]
10. Multi Health Systems. 2019. Available online: <https://www.mhs.com/MHS-Assessment?prodname=cpt3> (accessed on 26 March 2019).
11. Berger, I.; Cassuto, H. The Effect of Environmental Distractors Incorporation Into a CPT on Sustained Attention and ADHD Diagnosis Among Adolescents. *J. Neurosci. Methods* **2014**, *222*, 62–68. [CrossRef] [PubMed]
12. Berger, I.; Slobodan, O.; Cassuto, H. Usefulness and Validity of Continuous Performance Tests in the Diagnosis of Attention-Deficit Hyperactivity Disorder Children. *Arch. Clin. Neuropsychol.* **2017**, *32*, 81–93. [PubMed]
13. Cassuto, H.; Ben-Simon, A.; Berger, I. Using Environmental Distractors in the Diagnosis of ADHD. *Front. Hum. Neurosci.* **2013**, *7*, 805. [CrossRef] [PubMed]
14. O'Mahony, N.; Florentino-Liano, B.; Carballo, J.J.; Baca-Garcia, E.; Rodriguez, A.A. Objective Diagnosis of ADHD Using IMUs. *Med. Eng. Phys.* **2014**, *36*, 922–926. [CrossRef] [PubMed]
15. nebahealth.com: Neba Health. 2015. Available online: <https://nebahealth.com/faq.html#1> (accessed on 20 September 2016).
16. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef] [PubMed]
17. Esmael, B.; Arnaout, A.; Fruhwirth, R.K.; Thonhauser, G. A Statistical Feature-based Approach for Operations Recognition in Drilling Time Series. *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.* **2015**, *5*, 454–461.
18. Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. A Practical Guide to Support Vector Classification. 2003. Available online: <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf> (accessed on 10 August 2018).

