Assessing learner's distraction in a multimodal platform for sustained attention in the remote learning context using mobile devices sensors

Giuseppe Di Giuseppe Deininger

Supervisors: MSc. Yoon Lee and Prof. Dr. Marcus Specht

June 27, 2021

Abstract

During a learning task, keeping a steady attentive state is detrimental for good performance. A person is subject to distraction from different sources, among which distractions originating from within him or herself or from external sources, such as ambient sound. The detection of such distraction can improve the effectiveness of a task by giving feedback when necessary. Existing researches tried to measure performance on specific activities with the use of mobile devices such as smartphones and smartwatches, and a study showed a correlation between changes of posture and distraction. This paper tackles a main question "How mobile devices sensors can indicate learner's distractions in the remote learning context?". The process to do so included the recording of raw data from the movement sensors from a smartphone and smartwatch during a reading task, which was processed to highlight movements and then used to train a Convolutional Long Short Term Memory (LSTM) model. The final produced result showed a F1 score of 0.919 on validation data and was also combined with an external model to detect distraction from ambient noise to create a multimodal model, which showed better performance than both models individually. The limitations of the data collected during the experiment and improvements for future work are also discussed.

1 Introduction

In the current world scenario with the COVID-19 pandemic but also due to the advancements of technology on the scope of e-learning tools, remote learning is a very relevant topic in the learning context. This brings up the needs to understand the student's engagement even without the close contact otherwise available in a lecture room. One of the lacking points of interest concerns the attention of the students during a computer mediated learning session.

An important part of the learning process, the attention, can highly influence the success of the education process (Lodge & Harrison, 2019), with the lack of it being responsible for the failure of many students in achieving their goals and relating to mental health disorders, such as attention deficit hyperactivity disorder (ADHD) (Jangmo et al., 2019). It is therefore of importance to better understand how to identify the lack of attention to enable actions in accordance.

To better understand the state of the art in the use of mobile devices to identify distraction, a literature study was held to determine how such devices can be used to identify loss of attention and which specific sensors give meaningful data on the topic.

In a systematic literature study, Noroozi et al. found that subjective data, which includes interviews, self-reports, and observations were mostly used in the analysis of the learning process, while objective data, such as facial expressions recognition, eye-tracking, screen recordings and heart rate variability were much less common with some metrics among with temperature and accelerometer data not being found at all (Noroozi et al., 2020). At the same time, Noh et al. found that the the frequent change in posture during an activity is directly related to the existing degree of distraction (Noh, Seo, & Jeong, 2019).

The existence of a correlation between the learner's movement and posture with the distraction level together with the current lack of significant research highlights the importance of better understanding how it is possible to assess the learner's distraction in a learning activity. Research on using wearable devices such as smartwatches in order to identify different aspects of the learning process has already been done. Zhou et al. attempted to categorize a student's learning activity among specific categories: Writing, Talking/Discussing, Listening to the teacher, Mind-wandering/Sleeping, Others (Zhou et al., 2019, 2020). Dimitri et al. tried to predict performance in a self-regulated learning context, but with mixed results (Di Mitri et al., 2017). From the literature study it followed that identifying which task was being executed among options from different contexts showed good results, finding the measure to specific metrics such as productivity had poor results in its initial attempts.

A characteristic found in many of the previously mentioned studies that did achieve a satisfactory result was the fact that the wearable data was implemented in a multimodal context with other types of input. This happens because each component of the multimodal system would be specialized on a specific characteristic of the problem in hand, and the combination of these components would yield a better result than the individual ones or possibly a generalist alternative. This raises the suggestion that the best way to measure a student's sustained attention is by using multiple sensors with different responsibilities. This means that the mobile device sensors could be used to indicate the learner's attention individually and then be connected to other sensors able to measure environmental distractions, emotional changes and attention fluctuation in order to have a more precise estimation of the sustained attention. These other sensors were developed in parallel researches and their results were combined at a later stage of this research.

2 Methodology

The information found during the literature study led to the formulation of the main research question "How mobile devices sensors can indicate learner's distractions in the remote learning context?", which in turn can be detailed with three sub questions:

- (RQ1) What is the most effective way to use and position a mobile device to gather data related to a learner's distraction?
- (RQ2) How can the mobile device data be individually analyzed for a learner's distraction?
- (RQ3) How can mobile device sensor data improve multimodal learning analytics of sustained attention?

From the research questions, three hypotheses were defined:

- (H1) A smartphone can be placed in a user's front pant pocket in order to effectively identify posture changes.
- (H2) A neural network can be trained using recorded mobile devices' sensor data to identify distraction moments.
- (H3) The trained neural network model can produce better results when paired with other models that do not only detect learner's distraction.

To allow for further investigation of the research questions, a user study was assembled involving three of the members of the research group, comprised by three experiments. They aim on collecting data on moments in which the subject is attentive and distracted during a reading text, which can later be processed by each researcher for his own topic.

2.1 Approach

Each experiment held had a different goal allowing the research group to collect all required data. All of them were executed in the same controlled environment, represented in Figure 1. The smartwatch and smartphone are collecting data from their sensors, explained in detail in Section 3. The computer is recording a video with its webcam for processing on the subject's emotion and eye gaze, the microphone is recording the environment sounds and the temperature sensor is measuring the subject's facial temperature. Finally the environment can be subjected to controlled sound stimuli of different nature, such as songs, machine sounds, tones with varying frequencies, etc.

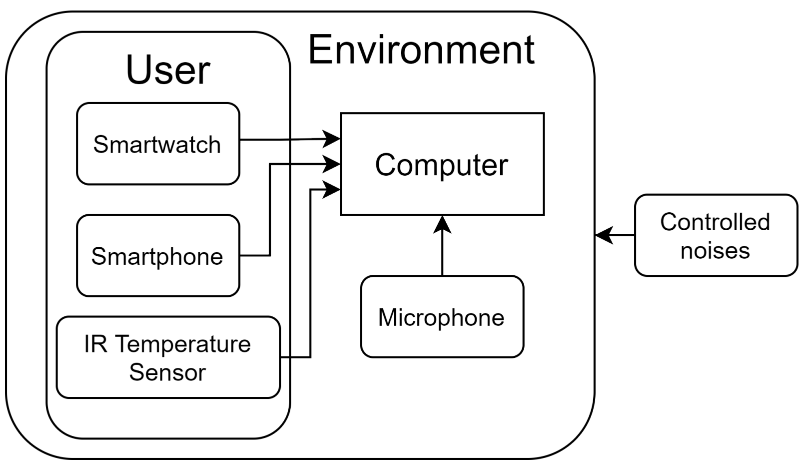


Figure 1: User study organization

2.1.1 Experiment 1: Calibration of the basis deblur time

This experiment follows the results found by Huang et al. which indicates the correlation between internal thought and eye movements in which they point either away from or to each other, known as eye vergence. This correlation can be exploited by blurring the screen at random intervals, and the subject would be asked to react to this blurring as soon as possible. Longer reaction times would then strongly indicate mind wandering, while short reaction times would be inconclusive (Huang, Ngai, Leong, Li, & Bulling, 2019).

To circumvent the inconclusiveness of the fast reaction time, the subject is asked to ensure he is attentive during the recording. His task was to read a number of short texts (under a minute of reading time), and to react as fast as possible to screen blurrings during the process. The collected reaction time is used to give an indication of the blur discrimination time (Td), a threshold time over which a subject can be considered inattentive.

2.1.2 Experiment 2: Gathering of purely attentive data

Because the reaction time is not enough to indicate attentiveness, a specific experiment is conducted to gather purely attentive data. The subject is asked to read small texts while ensuring attentiveness, similarly to Experiment 1. The differences are that the screen is not being blurred at any moment and all sensors are gathering data. The recorded data can then be post processed by each researcher to use for training their models.

2.1.3 Experiment 3: Gathering of inattentive data

In this experiment the user will also wear the same devices and have the same data collected as Experiment 2 but a microphone recording the ambient noise is added to the sensor array. During the experiment the different sound stimuli are played at random moments in an attempt to distract the subject. The goal of this experiment is to collect moments in which the user is inattentive, and to increase the chances of this event happening, a longer text is read.

During the reading process, the screen will be blurred similarly to Experiment 1, to which the participant also needs to react by pressing a button as soon as possible. If the reaction time is higher than the discrimination time, the moment is labeled as inattentive. When the participant notices he is distracted, he is asked to press another button to indicate it. The data collected is then stored to be analyzed similarly to what was described for Experiment 2.

3 Identification of distraction with mobile devices

The mobile devices used during the experiment, iPhone XR and Apple Watch Series 4 contained the following sensors considered relevant for the study.

Table 1: Metrics of models for individual sensors.

Sensor	Description
Acceleration (G)	Device acceleration along phone's X, Y and Z axis
Orientation (rad)	Device orientation on Roll, Pitch and Yaw axis
Rotation Rate (rad/s)	Device rotation speed around X, Y and Z axis
User Acceleration (G)	Device acceleration along phone's X, Y and Z axis with gravity filtered out
Quaternion (R)	Device orientation represented in quaternions
Gravity (G)	Impact of gravity on X, Y and Z axis

Some of those sensors, like Rotation Rate, measure instant changes and therefore have a steady state centralized at 0. Others, such as the Orientation, have values dependant on the device positioning. This means that small changes on how the device is positioned can cause a difference in the reading between experiments that can compromise the final result.

This is why the use of High-pass filters is common on Human Activity Recognition (HAR) tasks. This type of filter only allows for changes that happen in a small time frame. This means that the movement of a posture change is detected, but the initial and final steady positions are filtered out. As a consequence, every signal is then centered around 0, allowing for more consistency amongst the data set. An example of such filtering can be seen at Figure 2

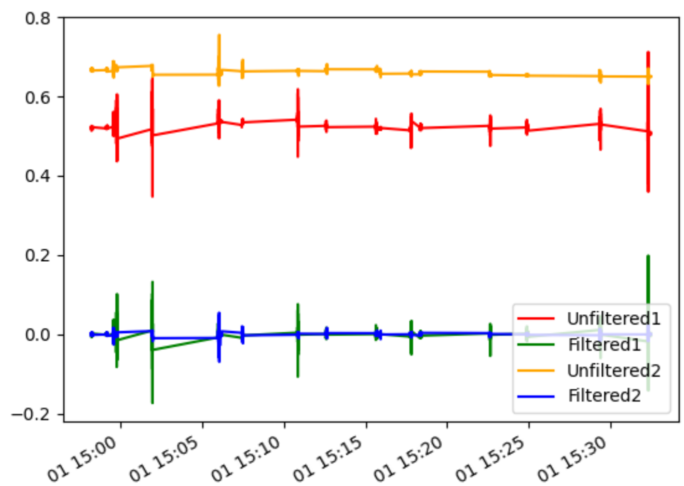


Figure 2: Example of filtered signal

At the example above, two different test subjects were doing the same task, but due to the position the phone was placed in the pocket, the steady state reading of the same sensor was around 0.55 for the first and 0.7 for the second case. After passing through the filter however, both signals are centralized at 0, and the spikes resulting from movements are still present.

The filter used was of Butterworth type because of its property of very low ripple between the pass and stop band, and good balance between phase and attenuation response (Modi & Parashar, 2018). These properties allows for very little dampening on the spikes that can represent the posture changes and therefore keeps it closer to the actual changes read in the raw data. It had an order of 6, found to have a sharp response at the cutoff frequency. Finally the cutoff frequency used was of 0.2Hz chosen as a consequence of the 5 second window being analyzed, which filters out any movement that requires more time than the window size to be completed.

With all the data filtered, it is ready to be applied to the model responsible for detecting distraction. The model chosen was a Long Short Term Memory (LSTM) paired with convolution layers at its input. This combination has shown very high accuracy on HAR tasks on common datasets such as UCI-HAR and WISDM (Xia, Huang, & Wang, 2020). Another advantage of such model is that their ability to learn from the time series data directly prevents the need to manually identify behaviors from the data and consequently increase generalization. An example of such neural network can be seen at Figure 3

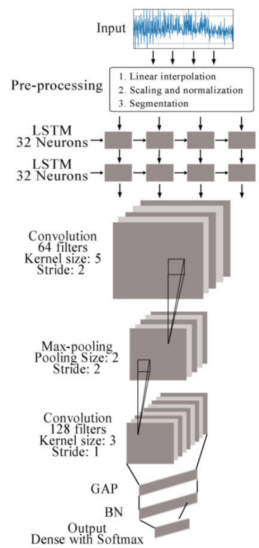


Figure 3: Example of LSTM-CNN (Xia et al., 2020)

To make data compatible with the model, the first step was to divide the filtered data into windows of data. Because each time window was of five seconds, the phone sampled at 80Hz and the watch at 24Hz, it means that for each sensor axis the a phone window data will contain 400 measurements, and the watch window data will contain 120. Each window is represented in a single row of data at the input, and each sensor axis is represented in a different file. As an example, if only the accelerometer of both the phone and watch are used, the input for the model will consist of six files, one for each of the three axis from the devices' sensors, and all files would contain n rows, with n being the number of windows analyzed, with each row of a phone sensor file containing 400 readings and the watch 120. Details on how to generate such dataset can be found on its GitHub page[1].

The chosen file organization allows for experimental modularity thanks to the parallelization of the input since it is possible to compare models using single sensors and a combination of them by just changing which files should be considered as input. After the input files are loaded, the model is ready to be trained. The model created breaks each window of data into four sections of 1.25 seconds that can be passed through the convolutional layers. Afterwards, the LSTM model needs to be flattened to one long vector so that it can go through a dense layer and finally an activation layer using ReLU. The training happens using 25 epochs and a batch value of 64. The program developed to train and test the neural network can be found on GitHub[2]

Once the model is trained, it can be evaluated with validation data organized in the same format as the training one. This evaluation can produce important metrics for the

[1] `https://github.com/MultimodalLearningAnalytics/HighPass-filter-and-dataset-generator`
[2] `https://github.com/MultimodalLearningAnalytics/MobileDevice-distraction-model-trainer`

model analysis, such as accuracy, precision, recall and F1. With all the sensors individually trained, it is then possible to decide which ones have significant correlation with the problem being solved and then more robust models can be generated accordingly. The combination of multiple sensors from the same device can happen by just defining them as input, but when a combination of phone and watch is desired it is not possible to do the same as each device contains a different amount of samples within the same window. Consequently, a model needs to be created for each device and a multimodal model is created for both of them, taking each individual output and generating a combined one using a weighted average of their output in relation to their F1 score.

4 Experimental Setup and Results

To prepare the environment necessary for the gathering of the data, multiple subsystems were prepared. The main platform for that was a Lenovo Yoga 530-14ARR laptop, with a 14 inch Full HD screen and a 720p built-in webcam, running a software developed by the research group. It was connected to an earphone containing a microphone and an ESP32-DevKitC-32D responsible for processing the data from a MLX90614ESF-DCI-000-SP IR temperature sensor, mounted to the user using a tiara. The user was also wearing an Apple Watch Series 4 on the wrist of the hand not used to control the laptop's cursor and kept an iPhone XR in one of his pant's front pockets. Both the smartphone and smartwatch have the SensorLog app installed for collecting raw data from their sensors. The laptop was placed on an empty table set to the height most comfortable for the user. An example of an user during the experiment can be seen at Figure 4.

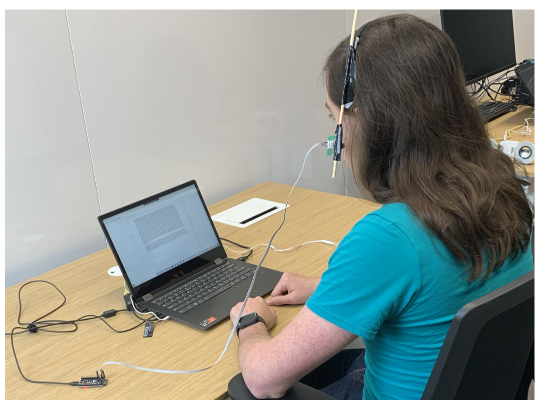
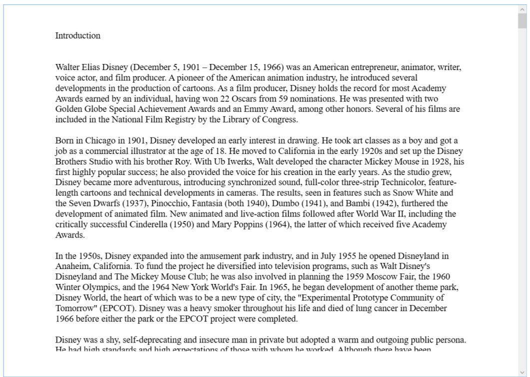


Figure 4: User study environment

4.1 Experiment 1: Calibration of the basis deblur time

This experiment is the only one that does not require the data collection devices to be used. The participant is asked to read 12 short texts, each with reading duration in the 10-30 seconds range. He is also asked to ensure he is fully attentive during the process and in case of distraction, the data collected would be discarded ant the experiment restarted. During the experiment, the screen randomly blurs at intervals of 2-5 seconds and the subject needs to press a button as soon as the blurring is noticed. The average of the reaction time to the blurring will then give and indication of the blur discrimination time (Td). A comparison of a regular and blurred screen can be seen at Figure 5.



Introduction

Walter Elias Disney (December 5, 1901 – December 15, 1966) was an American entrepreneur, animator, writer, voice actor, and film producer. A pioneer of the American animation industry, he introduced several developments in the production of cartoons. As a film producer, Disney holds the record for most Academy Awards earned by an individual, having won 22 Oscars from 59 nominations. He was presented with two Golden Globe Special Achievement Awards and an Emmy Award, among other honors. Several of his films are included in the National Film Registry by the Library of Congress.

Born in Chicago in 1901, Disney developed an early interest in drawing. He took art classes as a boy and got a job as a commercial illustrator at the age of 18. He moved to California in the early 1920s and set up the Disney Brothers Studio with his brother Roy. With Ub Iwerks, Walt developed the character Mickey Mouse in 1928, his first highly popular success; he also provided the voice for his creation in the early years. As the studio grew, Disney became more adventurous, introducing synchronized sound, full-color three-strip Technicolor, feature-length cartoons and technical developments in cameras. The results, seen in features such as Snow White and the Seven Dwarfs (1937), Pinocchio, Fantasia (both 1940), Dumbo (1941), and Bambi (1942), furthered the development of animated film. New animated and live-action films followed after World War II, including the critically successful Cinderella (1950) and Mary Poppins (1964), the latter of which received five Academy Awards.

In the 1950s, Disney expanded into the amusement park industry, and in July 1955 he opened Disneyland in Anaheim, California. To fund the project he diversified into television programs, such as Walt Disney's Disneyland and The Mickey Mouse Club; he was also involved in planning the 1959 Moscow Fair, the 1960 Winter Olympics, and the 1964 New York World's Fair. In 1965, he began development of another theme park, Disney World, the heart of which was to be a new type of city, the "Experimental Prototype Community of Tomorrow" (EPCOT). Disney was a heavy smoker throughout his life and died of lung cancer in December 1966 before either the park or the EPCOT project were completed.

Disney was a shy, self-deprecating and insecure man in private but adopted a warm and outgoing public persona.

(a) Regular text screen

Introduction

Walter Elias Disney (December 5, 1901 – December 15, 1966) was an American entrepreneur, animator, writer, voice actor, and film producer. A pioneer of the American animation industry, he introduced several developments in the production of cartoons. As a film producer, Disney holds the record for most Academy Awards earned by an individual, having won 22 Oscars from 59 nominations. He was presented with two Golden Globe Special Achievement Awards and an Emmy Award, among other honors. Several of his films are included in the National Film Registry by the Library of Congress.

Born in Chicago in 1901, Disney developed an early interest in drawing. He took art classes as a boy and got a job as a commercial illustrator at the age of 18. He moved to California in the early 1920s and set up the Disney Brothers Studio with his brother Roy. With Ub Iwerks, Walt developed the character Mickey Mouse in 1928, his first highly popular success; he also provided the voice for his creation in the early years. As the studio grew, Disney became more adventurous, introducing synchronized sound, full-color three-strip Technicolor, feature-length cartoons and technical developments in cameras. The results, seen in features such as Snow White and the Seven Dwarfs (1937), Pinocchio, Fantasia (both 1940), Dumbo (1941), and Bambi (1942), furthered the development of animated film. New animated and live-action films followed after World War II, including the critically successful Cinderella (1950) and Mary Poppins (1964), the latter of which received five Academy Awards.

In the 1950s, Disney expanded into the amusement park industry, and in July 1955 he opened Disneyland in Anaheim, California. To fund the project he diversified into television programs, such as Walt Disney's Disneyland and The Mickey Mouse Club; he was also involved in planning the 1959 Moscow Fair, the 1960 Winter Olympics, and the 1964 New York World's Fair. In 1965, he began development of another theme park, Disney World, the heart of which was to be a new type of city, the "Experimental Prototype Community of Tomorrow" (EPCOT). Disney was a heavy smoker throughout his life and died of lung cancer in December 1966 before either the park or the EPCOT project were completed.

Disney was a shy, self-deprecating and insecure man in private but adopted a warm and outgoing public persona.

(b) Blurred text screen

Figure 5: Comparison between regular and blurred text

4.2 Experiment 2: Gathering of purely attentive data

In this experiment the participant reads 12 different short texts of similar reading duration as Experiment 1. All data collected is later labeled as attentive for the later stage of model training, so in case the subject claims to have lost attention during the experiment, the process needs to be restarted.

4.3 Experiment 3: Gathering of inattentive data

In this experiment the user was also wearing the same devices and had the same data collected as Experiment 2 but the environment also had a lapel microphone recording the ambient noise during the whole process and the laptop was connected to another system, controlled by a research group member to generate noises during the reading task. This auxiliary environment can be seen at Figure 6.

Figure 6: Auxiliary environment

The text read has a reading duration of 45-60 minutes. During the reading process, the screen will be blurred similarly to experiment 1 within the period of 30-90 seconds, to which the participant also needs to react by pressing a button. The sound stimuli was executed at random moments and was part of one of the following categories:

- Song
- Machine sound
- Construction sound
- Background noise
- Sounds with varying properties
  - Frequency
  - Time interval
  - Frequency interval

For the moments with a reaction time to a blurring was larger than the Td defined in Experiment 1, an inattentive label was given at the deblur moment. Similarly a distracted label was given to the moments in which the user claimed to be distracted. There were no labels applied to the remaining collected data during this experiment, because although the long reaction time correlates to inattentiveness and serves as an indication of mind-wandering, nothing can be said about short reaction times (Huang et al., 2019).

After the experiment is done, all data collected from the mobile devices are parsed and converted into stores compatible with Microsoft Psi[3]. After this step, all data can be seen using the Psi Studio and the experiment can be played with every sensor value available for examination. An example of such environment set up with a subset of the available sensors can be seen at Figure 7.

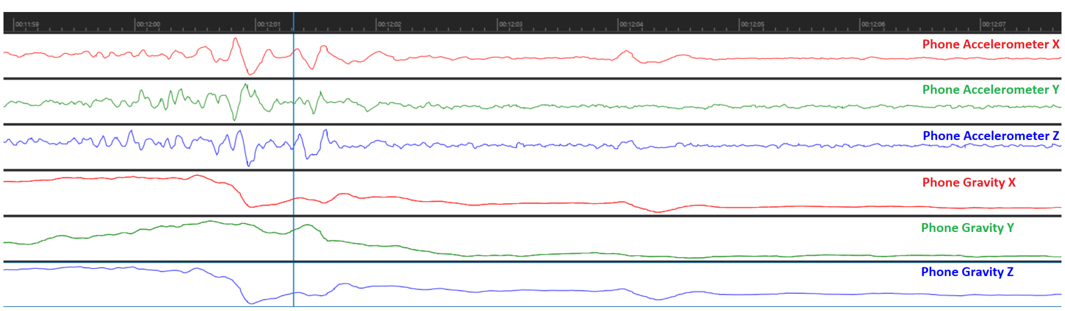


Figure 7: Data visualization on Psi Studio

With all the data available organized in Psi, it is then processed with the filters and used to train and test the model as mentioned in Section 3. A model was generated for each sensor component and the results on Accuracy and F1 scores for these 12 models can be found in Table 2. All the metrics computed for these models can be found on Appendix A

[3]`https://github.com/MultimodalLearningAnalytics/SensorLog-to-PSI-store-parser`

Table 2: Metrics of models for individual sensors.

	Phone		Watch	
	Accuracy	F1	Accuracy	F1
Acceleration	86.25%	84.46%	78.12%	75.50%
Orientation	68.12%	67.69%	71.88%	69.40%
Rotation Rate	91.25%	90.29%	83.12%	82.04%
User Acceleration	84.38%	80.82%	77.50%	74.75%
Quaternion	65.00%	66.67%	74.38%	70.44%
Gravity	67.50%	66.67%	76.25%	72.52%

From this result two main conclusions can be reached: firstly, the phone has sensors with significantly higher metrics than the watch, which can be linked to the movements the body parts they were attached to were subject to during the task. Secondly, the sensors that measure instant movement (Acceleration, Rotation Rate and User Acceleration) had significantly better results than the absolute positioning sensors (Orientation, Quaternion and Gravity). The second finding led to the creation of models using different combination of such sensors, and their results can be seen on Table 3.

Table 3: Metrics of models for combinations of sensors.

	Phone		Watch	
	Accuracy	F1	Accuracy	F1
Acceleration Rotation Rate User Acceleration	92.50%	91.89%	81.88%	80.64%
Acceleration Rotation Rate	91.25%	90.29%	81.88%	80.54%
Acceleration User Acceleration	87.50%	85.25%	78.75%	76.25%
Rotation Rate User Acceleration	89.38%	88.00%	81.25%	79.68%

The results of the combined sensors models achieved a better result than any individual sensor for the phone but the same didn't happen for the watch. Because of this poorer result in a multimodal scenario and because of the added complexity on data processing and experimental setup, it was decided that the use of a smartwatch would not be pursued at the next phase of the research. This last phase entails the integration of the model discussed in this paper with models able to detect distraction and attention fluctuation using environmental noise, eye gaze and emotions data. This final multimodal model can then be applied to the data collected in order to compare how it relates to each individual one.

Each model has a binary output, which can relate to distraction or inattentiveness. The former is used by the model discussed in this research and the model that uses the microphone input to detect distraction based on environment sound. The latter is used by two models that use the webcam data to detect eye gaze and emotion fluctuation. They were combined using weights based on their metric scores. This way, it is possible to easily turn components on and off on to check how they affect the multimodal model. Towards the end of the research, the only parallel research that produced a model with meaningful correlation with its respective topic was the one for measuring distraction from ambient noise. The comparison of its model with the one generated in this research can be seen in Table 4.

Table 4: Metrics comparison of models using mobile device sensor and ambient sound data.

	Accuracy	Precision	Negative Precision	Recall	F1
Mobile Device	92.50%	98.89%	87.78%	86.25%	91.89%
Ambient Sound	60.71%	62.07%	59.26%	62.07%	62.07%

The comparison of the two models made it clear that the mobile device model should have significantly more impact on the final model due to its overall result. Due to the excellent precision, it was decided that when the mobile device detected distraction, its output would also be the output of the multimodal model. When that would not be the case, a weighted average would be applied, using either the precision or negative precision as weights, depending on each model output. A summary of the multimodal calculation can be seen in Table 5.

Table 5: Rules for multimodal model calculation (MD: output of the mobile device model, AS: output of the ambient audio model)

Case	Multi-modal score
MD >= 0.5	MD
MD < 0.5 and AS >= 0.5	((87.78 * MD) + (62.07 * AS)) / (87.78 + 62.07)
MD < 0.5 and AS < 0.5	((87.78 * MD) + (59.26 * AS)) / (87.78 + 59.26)

Using these rules, a new model was generated, which produced the results seen in Figure 8.

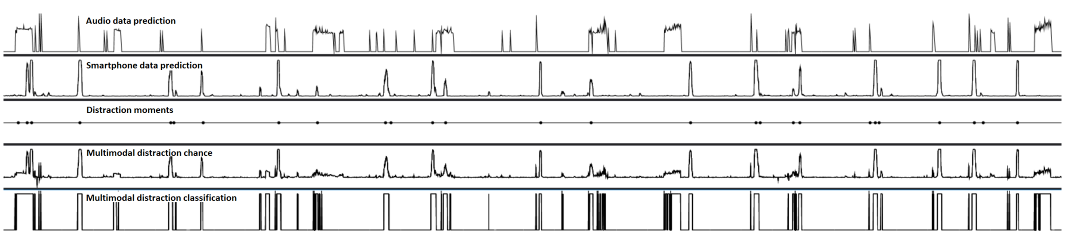


Figure 8: Multimodal model result visualization on Psi Studio

5 Responsible Research

The main concern that is raised on any research involving collection of data of a person is a possible ill intent on the side of the researchers. It has been shown that it is possible to determine a person's feature such as gender, age and level of intoxication with accelerometer data from a smartwatch, and with precise enough sensors, the detection can go as far as keystroke detection that can lead to identification of passwords and other sensible data (Kröger, Raschke, & Bhuiyan, 2019).

During this research it was made sure that none of the participants were doing a task that included any kind of confidential and/or private information so that such concerns could be cleared out. Additionally, all the raw data was processed tailored to the identification of posture changes and then discarded to added an extra layer of security. All the processed data is published in a GitHub organization[4] together with the scripts and programs used to record and process the data and evaluate the models generated under open source license. This allows anyone interested to reproduce the results and execute the experiment for the use with own recordings.

The research was held with the hopes of in the future provide a solution for learners to self assess for sustained attention in order to increase own efficiency, but it can also be used by institutions such as schools to track their students' attention during lectures. Such behaviour would involve collecting long recordings of students personal data and would characterize as a breach in privacy.

6 Discussion

The results found suggest different levels of correlation between devices and sensors, and different reasons can be found to justify it. Firstly the phone and watch difference can be justified by the nature of the task. Because the phone never needs to be removed from the subject's pocket and it doesn't have a lot of freedom of movement, it can reproduce the user's movement without significant noise and therefore could have a high correlation to the change of posture. On the other hand, the user needed to press two different buttons throughout the experiment, and even though it was asked that the hand used for that should not be the same one wearing the watch, some influence may still have been felt by the sensitive sensors. Additionally, it is still possible that changes of posture were not felt if the hand was not moving. Finally the hand movement on a reading task on a computer is expected to be very different as reading on other platforms such as books, so the model found is less generalizable as the phone counterpart.

The model with best metrics was the one using the phone's Accelerometer, Rotation Rate and User Acceleration values. Its confusion matrix can be seen at Figure 9.

[4] `https://github.com/MultimodalLearningAnalytics`

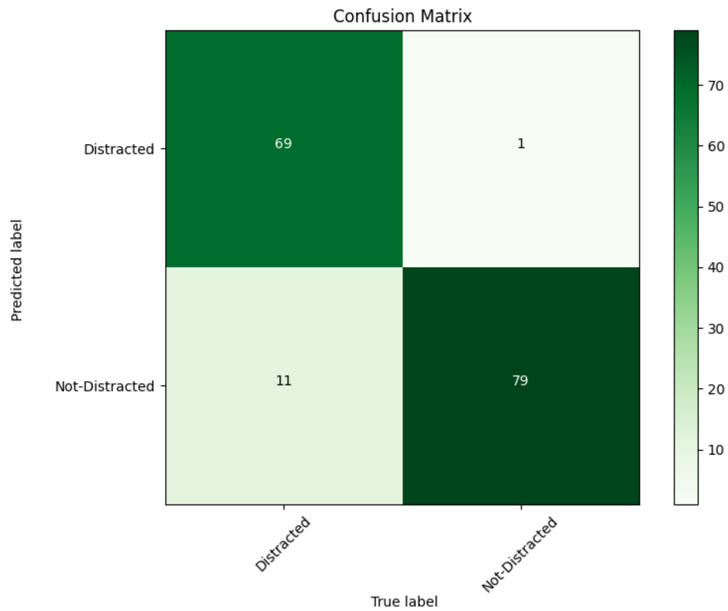


Figure 9: Resulting confusion matrix

The model had the following metrics:

- Accuracy: 92.50%
- Precision: 98.89%
- Recall: 86.25%
- F1 91.89%

One relevant aspect of the metrics is the difference between precision and recall. It can also be seen in the confusion matrix that 91.67% of the wrong predictions happened when the subject had not declared to be distracted and the model find him to be. A likely reason for this behavior is that during the experiments there were sounds as source of distraction and some subjects reported the existence of moments in which they were fully attentive and the only reason for the distraction was the sound, which in turn could cause a distraction moment without any sensors being able to gather data that would indicate it. This could then suggest that when it comes purely measuring distraction originating from the learner, the model could have a better result, but an experiment in isolation would be needed to confirm it.

The relation with sound originating distractions is also supported by analyzing the results from the multimodal model in Figure 8. It is possible to see a couple of distraction moments in which the smartphone prediction does not show a spike, which indicates low chances of distraction, while the multimodal model does identify the distraction thanks to the output of the audio data prediction. Due to the weight assigned to each model, it was possible to generate a final model better than either individually, which shows the potential of the use of a multimodal platform for learning analytics.

7 Conclusions and Future Work

This research aimed to identify how mobile devices could be used to identify a learner's perspective on a reading task. It tailored the use of two devices, a smartphone and a smartwatch, the first positioned in the pant's pocket and the second on the wrist not responsible for using the computer's mouse. The data was then processed and used to generate a model able to identify distraction based on changes of posture using an Convolutional LSTM network. Finally it was applied to a multimodal platform with models responsible for identifying distraction coming from sounds, change of emotions and eye gaze in order to track the subject's sustained attention changes.

It was found that the phone sensors' data could generate a more robust model while the watch can be significantly affected by the reading platform and therefore is less generalizable. A combination of phone sensors could reach up to 92.50% of accuracy and precision of 98.89%, while the best smartwatch model could only reach an accuracy of 83.23%.

These results, although very promising, cannot guarantee strong correlation in any scenario. This is due to the number and distribution of participants on the study. There were only three male participants aged between 20-24 involved, all with the same background in Computer Science, which can suggest similar behavior in a reading task. It is therefore important that an experiment with more participants and with more diversity is held in order to confirm that the results found applies to more generalized human traits.

On the data collection side, the experiment can be improved in a way that it is possible to distinguish the source of the distraction so that each model of the multimodal platform can be trained with data relevant to its scope. The increase in the amount of buttons the subject needs to keep track of during the experiment can lead to a less natural behavior, and therefore isolated experiments can be a viable solution.

With all aspects in mind, the results found show relevant correlation of the data collected by mobile devices to a learner's distraction and after the collection of more data with higher diversity for training, the model generated can be embedded into an application able to detect distraction in real time by just having the phone placed in the pocket.

8 Acknowledgements

This research would not have been possible without the cooperation with my peers from my research group: Jeffrey Pronk, Jurriaan Den Toonder and Sven van der Voort. Thank you for all the discussions on how to tackle the problem and the support throughout the research. The help from the research supervisors MSc. Yoon Lee and Prof. Dr. Marcus Specht was also detrimental for the result achieved. Thank you for all the feedback and continuous attention given to the project.

A Metrics from all trained models

In this section you can find all the computed metrics for the generated models troughout the research

Table 6: Metrics for phone sensors' individual models.

	Accuracy	Precision	Recall	F1
Acceleration	86.25%	95.78%	77.50%	84.46%
Orientation	68.12%	85.00%	66.25%	67.69%
Rotation Rate	91.25%	100.00%	82.50%	90.29%
User Acceleration	84.38%	100.00%	68.75%	80.82%
Quaternion	65.00%	80.00%	70.00%	66.67%
Gravity	67.50%	85.00%	65.00%	66.67%

Table 7: Metrics for watch sensors' individual models.

	Accuracy	Precision	Recall	F1
Acceleration	78.12%	84.40%	68.75%	75.50%
Orientation	71.88%	76.55%	63.75%	69.40%
Rotation Rate	83.12%	86.13%	78.75%	82.04%
User Acceleration	77.50%	84.23%	67.50%	74.75%
Quaternion	74.38%	83.00%	61.25%	70.44%
Gravity	76.25%	82.78%	65.00%	72.52%

Table 8: Metrics for phone sensor's combined models.

	Accuracy	Precision	Recall	F1
Acceleration Rotation Rate User Acceleration	92.50%	98.89%	86.25%	91.89%
Acceleration Rotation Rate	91.25%	100.00%	82.5%	90.29%
Acceleration User Acceleration	87.50%	100.00%	75.00%	85.25%
Rotation Rate User Acceleration	89.38%	100.00%	78.75%	88.00%

Table 9: Metrics for watch sensor's combined models.

	Accuracy	Precision	Recall	F1
Acceleration Rotation Rate User Acceleration	81.88%	85.83%	76.25%	80.64%
Acceleration Rotation Rate	81.88%	85.77%	76.25%	80.54%
Acceleration User Acceleration	78.75%	84.58%	70.00%	76.25%
Rotation Rate User Acceleration	81.25%	85.54%	75.00%	79.68%

References

Di Mitri, D., Börner, D., Scheffel, M., Ternier, S., Drachsler, H., & Specht, M. (2017). Learning pulse: A machine Learning approach for predicting performance in self-regulated learning using multimodal data. In *Acm international conference proceeding series.* doi: 10.1145/3027385.3027447

Huang, M. X., Ngai, G., Leong, H. V., Li, J., & Bulling, A. (2019). Moment-to-moment detection of internal thought during video viewing from eye vergence behavior. In *Mm 2019 - proceedings of the 27th acm international conference on multimedia.* doi: 10.1145/3343031.3350573

Jangmo, A., Stålhandske, A., Chang, Z., Chen, Q., Almqvist, C., Feldman, I., ... Larsson, H. (2019, 4). Attention-Deficit/Hyperactivity Disorder, School Performance, and Effect of Medication. *Journal of the American Academy of Child and Adolescent Psychiatry*, *58*(4), 423–432. doi: 10.1016/j.jaac.2018.11.014

Kröger, J. L., Raschke, P., & Bhuiyan, T. R. (2019). Privacy implications of accelerometer data: A review of possible inferences. In *Acm international conference proceeding series.* doi: 10.1145/3309074.3309076

Lodge, J. M., & Harrison, W. J. (2019, 3). *The role of attention in learning in the digital age* (Vol. 92) (No. 1). Yale Journal of Biology and Medicine Inc.

Modi, R. N., & Parashar, P. (2018). Design and Implementation of Butterworth, Chebyshev-I Filters for Digital Signal Analysis. *International Research Journal of Engineering and Technology*, *0072*(1).

Noh, Y. H., Seo, J. Y., & Jeong, D. U. (2019). Development of distraction limit estimation index using posture change monitoring system. In *Lecture notes in electrical engineering* (Vol. 514). doi: 10.1007/978-981-13-1056-0{_}3

Noroozi, O., Pijeira-Díaz, H. J., Sobocinski, M., Dindar, M., Järvelä, S., & Kirschner, P. A. (2020). Multimodal data indicators for capturing cognitive, motivational, and emotional learning processes: A systematic literature review. *Education and Information Technologies*, *25*(6). doi: 10.1007/s10639-020-10229-w

Xia, K., Huang, J., & Wang, H. (2020). LSTM-CNN Architecture for Human Activity Recognition. *IEEE Access*, *8*. doi: 10.1109/ACCESS.2020.2982225

Zhou, Z. X., Tam, V., Lui, K. S., Lam, E. Y., Hu, X., Yuen, A., & Law, N. (2020). A sophisticated platform for learning analytics with wearable devices. In *Proceedings - ieee 20th international conference on advanced learning technologies, icalt 2020.* doi: 10.1109/ICALT49669.2020.00097

Zhou, Z. X., Tam, V., Lui, K. S., Lam, E. Y., Yuen, A., Hu, X., & Law, N. (2019). Applying deep learning and wearable devices for educational data analytics. In *Proceedings - international conference on tools with artificial intelligence, ictai* (Vol. 2019-November). doi: 10.1109/ICTAI.2019.00124