Automatic Detection of Mind-Wandering Based on Body and Hand Movements from “Mementos” Dataset

Andrejs Kārkliņš
Supervisor(s): Bernd Dudzik, Xucong Zhang, Hayley Hung
EEMCS, Delft University of Technology, The Netherlands

June 17, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

The aim of this research is to discuss if it is possible or feasible enough to detect Mind-wandering of individuals using their hand and body movements from video recordings. The basis for this research is “Mementos”[9] data set, containing over 2000 recordings of people watching music videos. During experiment videos from data set were used to create software, that would automatically determine Mind-wandering. Results of the study have shown that body and hand movements are useful for detection of the phenomenon, however are not self-sufficient indicators for reliable detection. This study has found a reproducible methodology for automatic detection of Mind-wandering from pre-recorded videos, which is contributing towards exploration of this complex phenomenon in the field of Computer Science.

1 Introduction

Watching videos is playing a significant role in the modern person’s daily routine. Videos in the past decade became a reliable tool for the transfer of information and knowledge across the globe. Due to the recent COVID-19 outbreak, digital education became a sudden reality for many students and educational institutions, making them proceed with education through YouTube videos and video conferences [11]. Video creators could benefit from knowing at what moments of the video viewer’s attention shifts, without conducting any surveys. This would give additional feedback for content creators to improve their videos.

Recently the research has been done, which put as an aim to “provide researchers with a corpus of multimodal data that captures the occurrence of personal memories in response to videos”[9]. As a result of the study was collected ”Mementos data set containing 2098 video recordings of people watching different music videos. After each video survey was made on what type of memories, or emotions participants felt while watching the provided content. As a result, research has proven the usefulness of their data set for further research of machine learning automatic affect prediction[9]. The main advantage of “Mementos” data set for further study is variety since the research was conducted “in the wild”.

Mentioned research was focusing on the emotions and memories of the participants during the videos, which is closely related to phenomenon named Mind-wandering. One common definition for Mind-wandering is - a state of mind, when an individual starts producing task-unrelated thoughts, without external stimuli[12]. By being able to detect this phenomenon from viewer’s recording, it would be possible to understand what parts of the videos are causing an individual to have recollections or intense thought processes. This study in particular is interested in using body and hand movements towards detection of Mind-wandering. This decision is supported by previous studies, that have shown that upper-body and hand movements are playing a role in the detection of Mind-wandering[4]. Research will be aimed to make use of previously mentioned ”Mementos” data set and experimentally answer following questions:

1. Are body and hand movements useful for detection of Mind-wandering?
2. What features of hand and body movements are most significant towards detection of Mind-wandering?
3. Are body and hand movements self-sufficient indicators for reliable detection of Mind-wandering?

2 Background Information

Up to this day, automatic detection of Mind-wandering is still a relatively unexplored field in Computer Science. Despite that, there are few studies that made a significant contribution and consequently motivated further investigation of this phenomenon.

One of the researches made by Robert Bixler and Sidney D'Mello [3] has succeeded in creating fully automated software for the detection of Mind-wandering occurences with relatively high accuracy of over 70%. Despite this research has used a similar definition of Mind-wandering, participants of the experiment were reading, while eye movements were used for detection. Since reading is more attention-capturing action than watching a video, behavior of participants in the "Mementos" data set would be less restrictive which could lead to a higher rates of Mind-wandering.

The other research made by Nigel Bosch and Sidney K. D'Mello [4] also involved reading, however during the experiment participants recorded and analyzed upper-body movements and head pose, as well as 4 other features. Despite outcome of the research was not as successful as from research of Robert Bixler, having F1 scores of .478 and .414, results are high enough to conduct that upper-body and head movements are playing a significant role in the detection of Mind-wandering.

Research by Ahmet Cengizhan Dirican and Mehmet Göktürk has shown that "head response and speed exhibited meaningful patterns resulting from task engagement" [8]. During their experiment, 31 participants were playing games with different levels of difficulty while the camera was recording them. The success of this research contributes to the hypothesis that head movements and seated posture can be used as indications of the cognitive states of an individual, and therefore could be used for the determination of Mind-wandering.

One common observation from these researches is that Mind-wandering occurs rare enough to be called an anomaly. Since it is a complex phenomenon, it is hard to capture its actual occurrences. There is no currently reliable way of determining when a person is actually Mind-wandering. Two possible methodologies have been developed: direct sampling from participants (asking participants with certain intervals), used by studies [3] and [4], and external analysis, which was done in [8] and will be also used for this study. However, neither of these methodologies can provide total accuracy, due to underlying complexity of the phenomenon. As a result of Mind-wandering being rare and hard to observe phenomenon, the collected data assumed to have an extremely high imbalance.

Another common approach that was used in all of these researches is the collection of data in controlled environments. This research, in comparison to previously mentioned ones, is using the "Mementos" data set as a basis, meaning all collected video recordings were taken "in the wild". In real-world scenarios, ML algorithms would face a much higher degree of noise and variation in the recordings. Because of these factors, expectations for accuracy in such environmental conditions are lower. However, in case of success, automatic detection of Mind-wandering in an uncontrolled environment, would have a direct application in the real world and make a big step towards understanding the phenomenon.

3 Methodology

This section will give an insight into the methodologies and the experimental setup of the study. In detail will be described the process of labeling, software selection, and techniques that were used for data pre-processing.

3.1 Labeling of “Mementos” data set

For labeling of “Mementos” data set was used “VGG Image Annotator”[1]. The aim was to visually look at the video recordings, and mark the intervals, where individuals were most likely experiencing an episode of Mind-wandering. The annotation of the videos was made collaboratively with fellow researchers[2]. Due to the complicated nature of Mind-wandering phenomenon guidelines were necessary to preserve consistency. Guidelines consist of a list with indicators (Table 1) that are collectively decided to be sufficient for labeling interval as Mind-wandering. Due to the complexity of the phenomenon, many factors surrounded by context are needed to be taken into account for objective labeling. Therefore, the list of guidelines is far from being exhaustive and subjective opinion has been also taken into account during the labeling process.

Signs	**Description**
Smile	Sometimes a smile can be an indication of good memories, so if the smile is very expressive and sudden/genuine smile, it could be a reaction or a response to the video. A very subtle smile could also be a form of reminiscing / remembering a memory so this is also considered a form of Mind-wandering
Looking up / Rolling eyes	Looking up or rolling eyes are interpreted as looking up for a continuous-time which could be followed by movement of gaze to the side. Usually, this is caused by an individual trying to remember/recollect.
Squinting eyes	Can indicate that person is having a focused thought process happening, which is most likely unrelated to the task of watching the musical video.
Sound of person	When an individual is speaking to himself, it could indicate that person is going through a thought process and most likely it could be interpreted as Mind-wandering.
Frown	Sometimes frowning can be an indication of bad or sad memories, so if the frown is very expressive and sudden/genuine, it could be a reaction or a response to the video. A very subtle frown could also be a form of reminiscing / remembering a memory so this is also considered a possible episode of Mind-wandering

Table 1: Mind-wandering indicators

Annotation of Mind-wandering is very subjective, which led to different opinions in the group. To furthermore increase quality, the research group has been divided into 2 teams of 3 and 2 researchers. After a certain number of annotated videos, teams were shuffled to exclude the possibility of bias between group members during data labeling. In cases, where it was not possible to reach a consensus, the interval was labeled as not Mind-wandering.

“Mementos” data set contains videos with a big variance of conditions and environments in which videos were recorded. It was decided that some outliers would be disregarded

[1] `https://www.robots.ox.ac.uk/~vgg/software/via/`

[2] Radek Kargul, Arbër Demi, Iasonas Symeonidis and Max van Dijk

and not annotated. For example, videos, where individuals were not paying attention to the provided music videos due to external factors. Another reason for disregarding was extremely noisy recording with very low quality. On average approximately 10% of the videos were considered invalid for the experiment. As a result, 500 recordings from the "Mementos" data set were annotated, which will be further used for the experiment.

3.2 Selection of software

Currently, there is a significant number of software solutions to extract body and hand movements from video recordings. There are a few factors that were taken into account while finding a suitable solution for extraction: (a) must be open-source, (b) should have API in Python 3+, (c) should have extensive documentation, (d) should have body and hand tracking functionality, (e) must be flexible to output formats.

OpenPose[3] and MediaPipe[4] are two popular solutions, that met all of the requirements. The biggest difference between them is the significant accuracy advantage that MediaPipe has compared to OpenPose when only upper-body is seen, which is the case across whole "Mementos" data set. From OpenPose research it was found that missing body parts are creating a common failure case for detection[5]. In addition MediaPipe is considered to be much faster: "One of the main reasons for its success can be attributed to the ecosystem of re-usable calculators and graphs." [10]. Therefore, MediaPipe became a choice for the extraction of hand and body movements. MediaPipe offers a few ML algorithms, out of which "Holistic" was the most suitable. It includes tracking of human pose, face landmarks, and hand tracking.

There are a few factors, that determine the time it takes to process one video from the "Mementos" data set. The most time affecting parameter is "Model-complexity", which increases the output landmark accuracy, whilst also increasing the time it takes to process one frame. After manual testing, it was established that the most optimal value for the parameter would be "1", which is the average model complexity. The second time affecting factor is the number of frames per second in one video. Original recordings from the "Mementos" data set were using 30 frames per second, which was reduced to 10 frames per second for the sake of performance. The information loss was considered to be admissibly small, compared to the reduction of processing time and the number of samples to a factor of 3.

The output of MediaPipe consists of landmarks, where each one consists of 3 floating numbers, representing x, y, and z coordinates, normalized to [0, 1]. Posture output contains additional float to each landmark representing visibility, which is indicating the likelihood of the landmark being present and not occluded in the image. With regards to the data, format visibility was processed with a decision boundary of 0.5. That means that landmarks with visibility below 0.5 were treated as invisible and above the threshold as visible. That value for decision boundary is specified as default in MediaPipe documentation and therefore considered to be adequate. One of the problems in the output was, that face landmarks included eye and mouth tracking, which would make the experiment invalid, having additional features not addressed by this study. Therefore, eye and mouth landmarks were removed from the list, leaving only landmarks related to face oval. The total output for one

[3] https://github.com/CMU-Perceptual-Computing-Lab/openpose

[4] https://google.github.io/mediapipe

frame of the video consisted of 33 pose landmarks, 36 face landmarks, 21 left hand and 21 right-hand landmarks. The visual output of the MediaPipe can be seen in Figure 1.

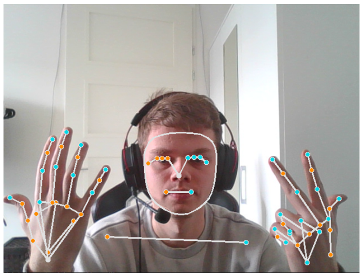


Figure 1: Example of MediaPipe output

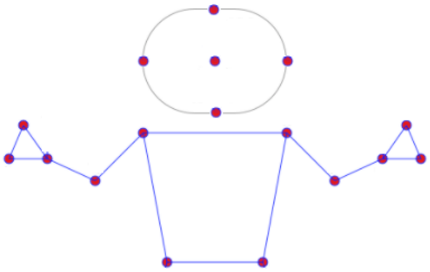


Figure 2: Landmarks after feature reduction

3.3 Data pre-processing

Raw data had very high dimensionality, 111 landmarks with 3 features each have resulted in 333 features. To reduce the dimensionality of the data, landmarks were analyzed for the frequency with which they appeared in the data set. From the analysis, it was conducted that head and shoulders landmarks were present on 99% of the frames. Elbows, wrists, palms and body(excluding legs) were seen on 5% of the frames. Fingers and legs were recognized on $< 1\%$ of the frames. From that analysis, it was decided to remove 42 landmarks that were representing fingers and 8 landmarks representing legs. Out of the remaining landmarks, there were 10 landmarks representing eyes and mouth from the MediaPipe pose that were removed due to providing excess information on face of an individual. The only landmark on the face that was left from the MediaPipe pose solution is nose, since it could potentially be helpful in correctly determining head position. The next step was a correlation analysis, out of all landmarks, the highest correlation was between landmarks of the face oval. As a result, it was decided to leave only 4 landmarks representing the forehead, chin, left ear and right ear, which are all extremums of the face oval. After these feature reduction techniques number of landmarks was reduced from 111 to 17, resulting in 51 floating number features. The resulting selection of the landmarks can be seen in Figure 2.

The next step was to subtract the average position of an individual on each of the videos. This process removes the data bias towards the video and translates absolute positions of landmarks towards the difference from its average position across the video. Only positions of visible landmarks were counted towards average. Since most Machine Learing algorithms are not capable of working with missing data, it was decided to replace missing landmarks with their average when they were visible. Since the data format was changed to the difference from average positions, missing landmarks were replaced with 0.0.

Due to the nature of the problem frames can not be treated as independent samples, therefore they were grouped into time series. Labeled intervals of Mind-wandering have a different lengths with a minimal length of 1 second and an average length of 4 seconds. This leads to a problem of different size intervals, therefore approach of splitting all data into equal intervals was not suitable. Mind-wandering occurrences need to be treated as a whole and can not be divided into multiple time series. All Mind-wandering intervals were granted their own time series, while the rest of the data was split into intervals of random duration ranging from 1 second to 5 seconds to match the approximate length of Mind-wandering episodes.

As a result, 501 videos with 54 intervals of Mind-wandering were split into approximately 10000 (depending on a random state) multivariate time series with labels.

3.4 Time series

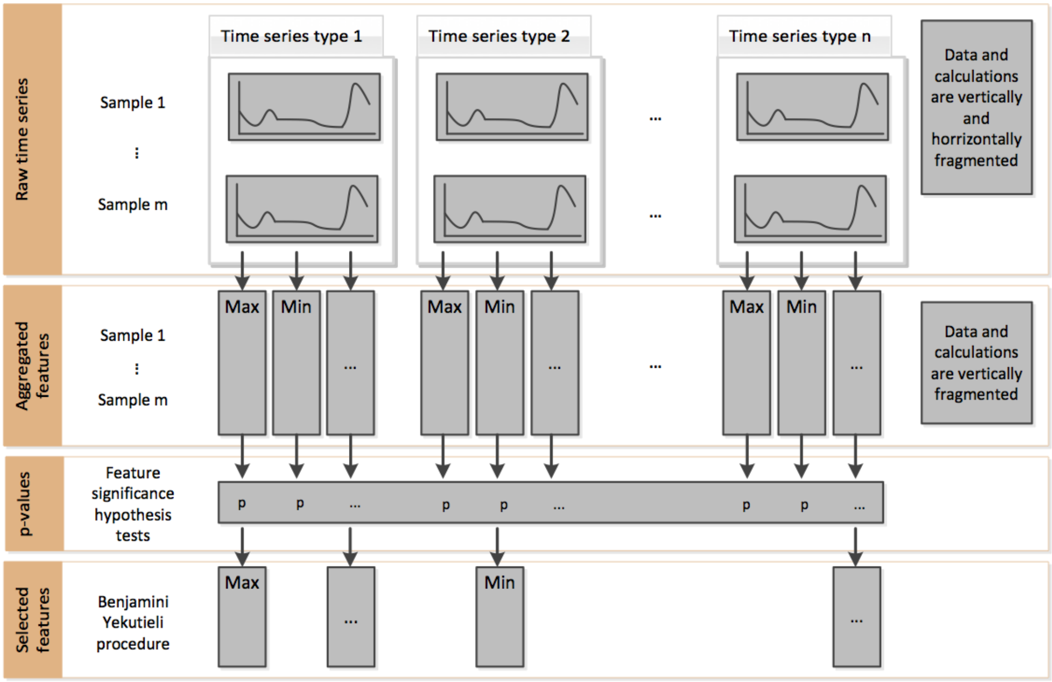

Raw time series
Aggregated features
p-values
Selected features
Time series type 1
Time series type 2
Time series type n
Sample 1
Sample m
Data and calculations are vertically and horrizontally fragmented
Max
Min
...
Data and calculations are vertically fragmented
Feature significance hypothesis tests
p
Benjamini Yekutieli procedure

Figure 3: Relevant feature extraction.

During pre-processing frames were grouped into intervals, therefore problem has been converted to a time series problem. A time series problem implies the extraction of features from ordered by time sequence of measurements by applying functions. In this case, one frame with 17 landmarks is treated as one measurement. For extraction of features from time series was used Python library *tsfresh*[5]. "By identifying statistically significant time series characteristics in an early stage of the data science process, tsfresh closes feedback loops with domain experts and fosters the development of domain specific features early on." [7]. Work with this library can be described in two steps: feature extraction and feature selection, which are depicted in Figure 3. By using this methodology, it was possible to capture the importance of landmarks and functions applied to time series for detection of Mind-wandering. For extraction of relevant features was used whole data set, without splitting on the train and test set. This decision might have led to selected features overfitting the data. The trade-off was made for the sake of efficiency, because of the high computational costs required for feature extraction, filtering and model training for each of the splits.

[5] https://tsfresh.readthedocs.io/

4 Results and Discussion

In this section, there will be discussed the results of feature extraction and filtering from time series. Afterward, will be discussed the results of the classification based on the extracted features.

4.1 Relevant feature extraction

Unfortunately, it was not possible to extract all features that *tsfresh* provides, due to the high complexity and computational costs of some calculations. From all possible set of functions that could be applied to time series the most costly ones were removed. As a result extraction algorithm was able to extract 2496 features.

After that, these features were filtered using the Hypothesis test and Benjamini–Yekutieli procedure [2], which selected only 193 features considered to be the most relevant for classification. From the results it was concluded that the most significant were landmarks representing the head. Out of all extracted features, 166 were used from chin, forehead, ears and nose landmarks. The other 27 were taken from elbows, shoulders, wrists and palms landmarks. In total there were applied 49 functions over each feature in the time series. In Figure 4 can be seen the frequency of function being selected as relevant, where the maximum would be 51, meaning importance for each of the landmarks. Abbreviations: abs - absolute, CID[1] - complexity-invariant distance, CWT - continuous wavelet transform for the Ricker wavelet.

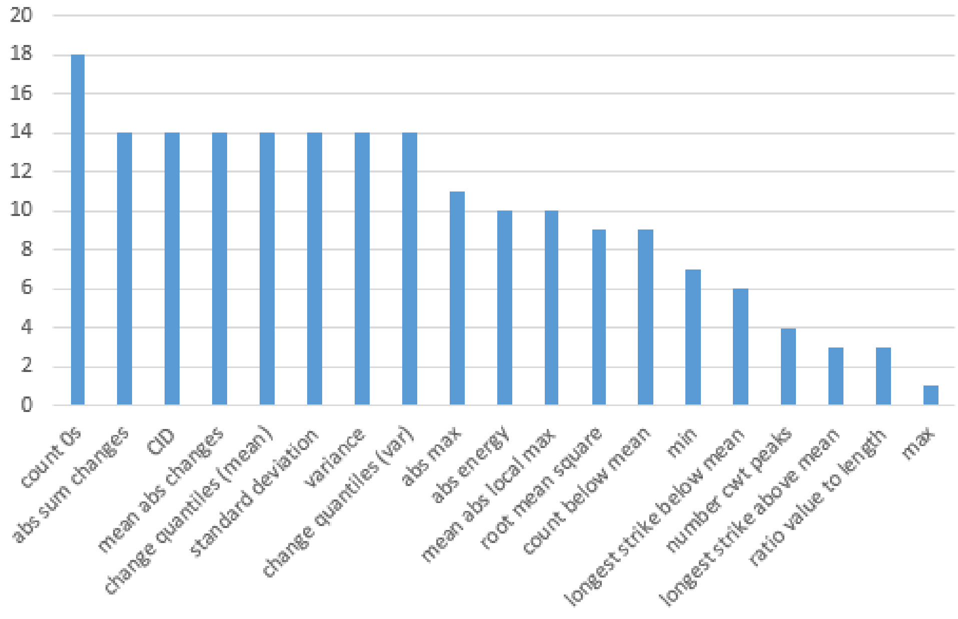


Figure 4: Most relevant functions.

By looking at the frequency of the functions considered to be relevant, the most significant for classification was “count 0s” function. It could be explained by missing landmarks, which were replaced by 0.0 during the pre-processing. In Figure 4 can be seen high importance of variance, standard deviation, mean of absolute changes, CID and changes in quantiles. These functions are representing the amount of activity each time series has had. Since there were much more functions in the *tsfresh* package that could potentially be applied, these results are not entirely representative.

4.2 Classification

Pre-processed data is still heavily imbalanced and therefore the problem of detection of Mind-wandering can be classified as anomaly or outlier detection. For these types of problems SVC, K-nearest neighbors(KNN), and Decision Tree classifiers are common approaches, therefore they will be used for classification. In addition, three outlier/novelty detection algorithms were used: One-Class SVM, Local Outlier Factor(LOF) and Isolation Forest. For comparison was added Dummy classifier which was ignoring inputs and assigning labels randomly with uniform distribution.

For the sake of fairness, each classifier was used with Cross-Validation and Grid Search for parameter tuning. As also, the data set was split into train and test throughout 10 iterations to further support the fairness of the results. The number of samples used for testing was set to 20%. For the scoring metrics, F1 score was considered to be the most representative, because it takes both precision and recall into account. F1 score metrics is known to be good with imbalanced data and high number of actual negatives, which is indeed the case with the Mind-wandering data set. Precision and recall were added to the metrics for easier interpretation of the results. In addition to metrics, the standard deviation for F1 score was provided to give an insight into the stability of algorithms on different train/test splits. 5-Fold Cross-Validation and Grid Search maximizing F1 score were used to find the best parameters for each of the classifiers.

$$precision = \frac{TP}{TP + FP} \quad recall = \frac{TP}{TP + FN} \quad F1 = \frac{2 * precision * recall}{precision + recall}$$

$$TP - TruePositive \quad FP - FalsePositive \quad FN - FalseNegative$$

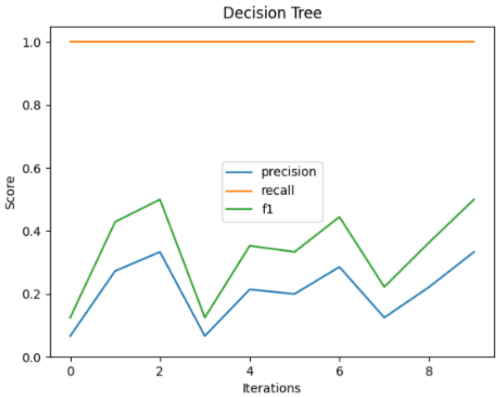
The results for each of the classifiers throughout all iterations can be seen on Figure 5. The average score for each of the classifiers are depicted on Table 2.

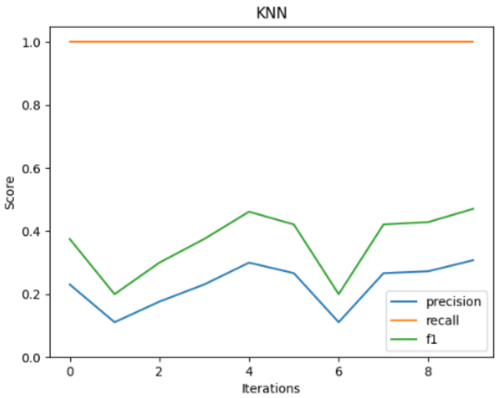
Score Algorithm	Precision	Recall	F1 score	σ_{F1}
Dummy	0.4737	0.0055	0.0108	0.0052
KNN	0.2274	1.0	0.3653	0.0947
DecisionTreeClassifier	0.212	1.0	0.3395	0.1331
SVC	0.3078	0.0241	0.0446	0.014
LOF	0.1315	0.1826	0.1453	0.0745
OneClassSVM	0.7746	0.0085	0.0167	0.0035
IsolationForest	0.0246	0.0261	0.025	0.0415

Table 2: Average classification results

Decision trees and KNN are showing significantly better results than other algorithms. This could be explained by the nature of these algorithms, both of them are well know algorithms for highly imbalanced data. As we can see from the metrics, throughout 10 iterations both algorithms had a recall of 1.0. This means that both algorithms have been able to label all occurrences of Mind-wandering correctly. However, the precision metrics is showing that a lot of not Mind-wandering intervals were classified incorrectly. In other words, Decision tree and KNN algorithms were capable of distinguishing between “Definitely not Mind-wandering” and “Might be Mind-wandering”, which can be interpreted as a decent success. High standard deviation is showing a strong dependence of algorithms on train/test splits. SVC has shown very poor results, compared to the two previously mentioned algorithms, which could be related to a high imbalance in the data set. SVC was tested in different configurations, including applying to samples cost of miss-classification, which was inversely proportional to label frequency. Neither of the configurations were able to increase the recall score of SVC above 0.05. It is possible that Grid Search was not exhaustive enough to find the most optimal parameters.

None of the anomaly detection algorithms have succeeded well on the data set. That could be explained by outliers not standing out enough from the rest of the data. Out of selected anomaly detection algorithms, LOF has shown the best results, with an average F1 score of nearly 0.15. Classification results are highly dependent on train/test split given to the algorithm. The most successful splits are providing F1 scores up to 0.2, whereas the worst ones are giving a score of 0.0. This is also observable by its standard deviation of 0.0745. Due to high dependence on the train/test splits LOF, Isolation Forest and OneClassSVM are considered to be unreliable. All classifiers have shown higher results than Dummy classifier, which proves that selected classifiers were better than guessing.





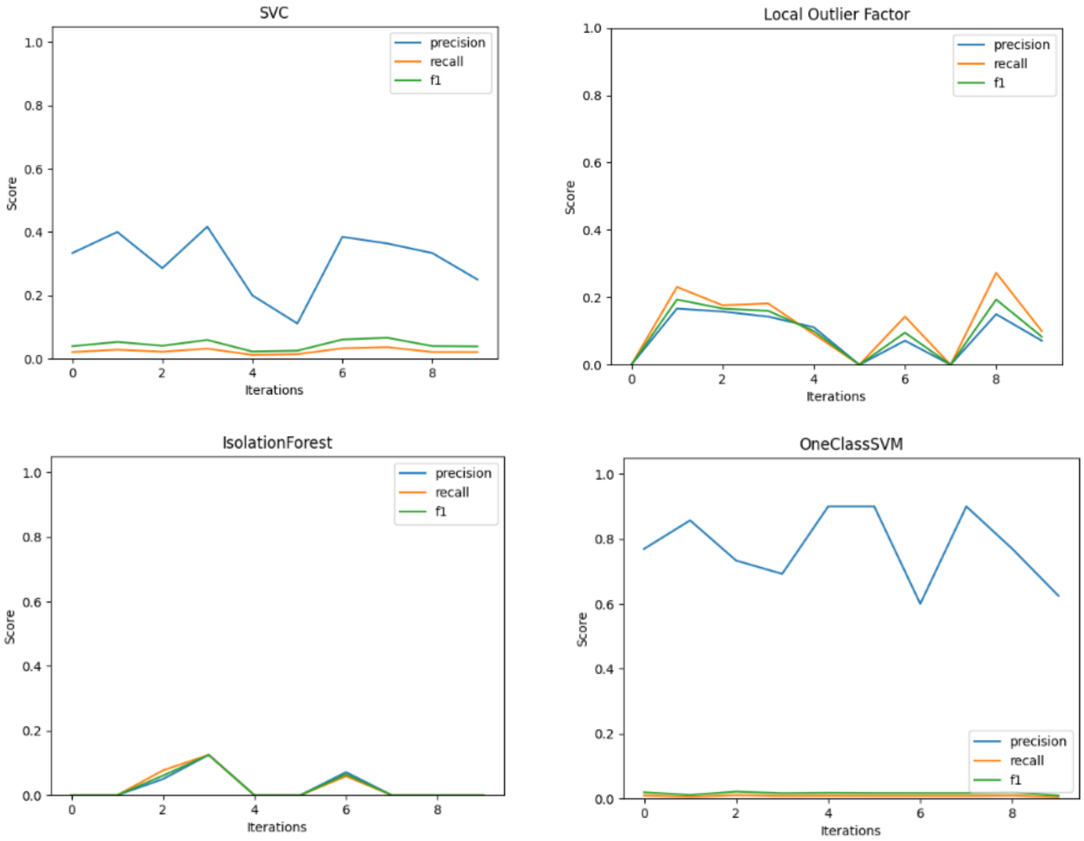


Figure 5: Results over 10 iterations.

5 Responsible Research

This study is using the "Mementos" data set for the detection of Mind-wandering. "Mementos" data set contains sensitive information about the participants and is protected by GDPR[6], therefore must be used responsibly. Videos from the data set are not allowed to be accessible to any third parties. This restriction for the data set is limiting the reproducibility of the study, only researchers that are accepted by the owners of the "Mementos" data set will be able to reproduce the results.

The research on the detection of Mind-wandering in the field of Computer Science is tightly related to Data Ethics. Collected data must be treated with caution, preserving the privacy of the individuals, handling unintentional bias in the outcomes, and avoiding the collection of unnecessary information during the study[6]. In this research privacy of individuals was preserved by processing the videos locally to remove the possibility of a data leak. From the results of the research, the bias towards gender, race, or age is unobserved, therefore considered to be ethically acceptable. All features that have been extracted from the videos

[6] `https://gdpr-info.eu/`

were collected only with the good intention of researching Mind-wandering. All collected information was used for analysis and was important for classification, which proves the necessity of extracted data.

6 Conclusions and Future Work

As a result of this research, there was found a clear methodology to use for further research on the detection of Mind-wandering from pre-recorded videos “in the wild”. The main difficulty of the study was a lack of clear definition of Mind-wandering. The research has successfully created a list of indicators, that could be interpreted as Mind-wandering, but in the end, it is far from being exhaustive. Because of that, the labeling of the data set has been done in a very subjective manner due to the lack of visible indications of Mind-wandering on the recordings. Since currently there is no reliable way of external detection of its occurrences, it was not possible to be always certain that an individual was experiencing an episode of Mind-wandering.

The results of feature analysis have shown, that there is a correlation between head movements and an individual experiencing Mind-wandering. The landmarks, which represented head boundaries, nose and shoulders were recognized as most valuable during the classification process. The results of the classification are clearly indicating that body and hand movements are not self-sufficient factors for reliable detection of Mind-wandering. Despite that, KNN and Decision tree have shown the average F1 score of 0.3-0.4, therefore are providing enough evidence that head and body movements are useful and should be used in addition to other features for detection of the phenomenon.

For further improvements, a bigger part of the “Mementos” data set could be labeled to increase the number of samples. In addition, more people could have participated in labeling to furthermore increase the probability of assigning the correct labels to episodes of actual Mind-wandering. Another big improvement could be a more extensive list of functions applied to the time series extracted from the data. That would require much greater computing power, but would most certainly capture additional information useful for classification. With higher computing power, it would be also possible to try different test/train splits during relevant feature extraction with tsfresh, and for each of them apply GridSearch for parameter tuning, which would remove the possibility of overfitting. Lastly, some of the unsupervised machine learning techniques could be taken into account in further research for the automatic detection of Mind-wandering.

References

[1] Gustavo E. A. P. A. Batista, Eamonn J. Keogh, Oben M. Tataw, and Vinicius M. A. Souza. Cid: an efficient complexity-invariant distance for time series. *Data Mining and Knowledge Discovery*, 28:634–669, 2013.

[2] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[3] R. Bixler and S. K. D'Mello. Toward fully automated person-independent detection of mind wandering. *Lecture Notes in Computer Science*, 8538:37–48, 2014.

[4] N. Bosch and S. K. D'Mello. Automatic detection of mind wandering from video in the lab and in the classroom. *IEEE Transactions on Affective Computing*, 12(4):974–988, 2021.

[5] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 11, 2019.

[6] C. Catherine. 5 principles of data ethics for business. `https://online.hbs.edu/blog/post/data-ethics`, 2021. Accessed: 2022-06-08.

[7] M. Christ, N. Braun, J. Neuffer, and A. Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, 307, 2018.

[8] A. C. Dirican and M. Göktürk. Involuntary postural responses of users as input to attentive computing systems: An investigation on head movements. *Computers in Human Behavior*, 28:1634–1647, 2012.

[9] B. Dudzik, H. Hung, M. A. Neerincx, and J. Broekens. Collecting mementos: A multimodal dataset for context-sensitive modeling of affect and memory processing in responses to videos. *IEEE Transactions on Affective Computing*, (01):1–1, 2021.

[10] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. Yong, J. Lee, W.Chang, W. Hua, M. Georg, and M. Grundmann. Mediapipe: A framework for perceiving and processing reality. page 8, 2019.

[11] S. Pokhrel and R. Chhetri. A literature review on impact of covid-19 pandemic on teaching and learning. *https://doi.org/10.1177/2347631120983481*, 8:133–141, 2021.

[12] J. Smallwood and J. Schooler. The science of mind wandering: Empirically navigating the stream of consciousness. *Annual review of psychology*, 66, 2014.