Detection of Mind-Wandering through Sound

Iasonas Jan Symeonidis
Supervisor(s): Bernd Dudzik, Hayley Hung, Xucong Zhang
EEMCS, Delft University of Technology, The Netherlands
22-6-2022

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

Mind-wandering happens when one's current train of thought, related to a specific task, is interrupted, due to internal disconnected thoughts. This phenomenon is highly subjective, and its detection is really important due to the internal understanding of the human mind that can be obtained. Several methods have been used in order to detect mind-wandering, such as thought probes, self-reports, electrophysiological measures or even eye-based tracking methods, however the detection of mind-wandering solely from sound has not been researched about. Therefore, this study is going to investigate if automatic detection of mind-wandering through sound is feasible. In this work, this question is tackled through a machine learning approach, where a linear SVM model is trained through acoustic features. Two methods of oversampling are considered, due to the high data imbalance between the classes, and these two are evaluated and compared. The approach is evaluated through different metrics, such as recall, precision, F1-score and accuracy, but also a comparison with other techniques is done. Results of this work show that sound as its own is not a reliable way of automatically detecting mind-wandering. These results however, might be implementation specific, as the ground truth values of mind-wandering were created through the means of perceived mind-wandering, and the techniques of random oversampling and SMOTE were also used. These could be causes of unreliability of the research. Future work should should take this into consideration and also apply this approach to a different data set, to assess its feasibility.

1 Introduction

Mind Wandering is a phenomenon that 96% of adult Americans say they experience on a daily basis [1], and it is an occurrence that takes up nearly 50% of the day [2–5]. Detecting and evaluating Mind Wandering is consequently really important, as it can be quite useful in deciphering the attention regulation mechanism during specific focal tasks. [6] and it is also a crucial and necessary step toward enhancing the effectiveness of attention training [7] and it may also contribute to exploration of the neural mechanisms underlying the regulation of sustained attention [8]. Research on mind-wandering has thus seen a huge increase in recent years, caused by Smallwood and Schooler's integrated analysis of related issues, which sparked a lot of interest [9]. An immediate consequence of this is that many different definitions of mind wandering have arisen, most of which are closely related to Smallwood's and Schooler's definition of mind wandering [10]. The definition that will be used in this work is : When mind wandering occurs, the executive components of attention appear to shift away from the primary task, not due to external factors or the person interacting with the external environment. The definition selected and formulated to be used in this paper, is in alignment with the existing literature, whilst at the same time reflecting the difference between mind-wandering and distraction.

This study investigates the "Detection of Mind-Wandering through sound", on the Mementos data set, and the research question that will be tried to be answered is: "Is detection of mind-wandering with sound feasible". This data set is the first multi-modal corpus for modeling emotion and memory processing in reaction to music video content [11]. It contains 1995 individual responses collected from 297 unique viewers responding to 42 different segments of music videos, and therefore the task at hand was: comprehending / watching the videos at hand. What was specified in the definition was that the cause of the attention being shifted away from the primary task should not be due to external factors, or due to the person interacting with external factors.

Mind-wandering is a concept that has not been studied or researched about heavily, however more studies have been done on the topic of cognitive task performance. Evidence supports that background sounds have a negative impact on cognitive task performance [12–15], playing an important part in task performance, and also contributing to attention loss. This, has thus, an immediate effect on whether the person is more likely to mind-wander or not. Despite the external factors that play a part in this research, it is important to note that the data set contains both people reacting to the music videos whilst wearing some sort of headphones, and also whilst not wearing headphones. There are instances of people watching the videos whilst wearing headphones, one or both earphones and also instances where they are listening to the audio through external speakers, or the device's built in speakers. Wearing any sort of earphones or headphones has the effect of making the person less susceptible to external sounds, and thus, more likely to not be affected by low volume external sounds. Headphones help to block out outside noise, which in turn will help to keep the focus on the task at hand. Due to these reasons, a working hypothesis is that external sounds do play a role in the existence of mind-wandering or not, and they might be more crucial in detecting cases of non mind-wandering, rather than mind-wandering. The research sub-questions that were formulated were namely:

1. What is the accuracy of mind-wandering detection through sound?
2. Does wearing headphones have an effect on mind-wandering?
3. How well does mind-wandering detection through sound do compared to other techniques?

Due to the fact of there not being many studies regarding the effects of sound on mind-wandering, this paper will provide some more insight in this area, exploring this effect in an "in the wild" environment. This will be done through a machine learning approach, where acoustic features are going to be extracted from the respective audio files. A machine learning algorithm is going to be trained and tested on the data set, to find whether there exists a link between mind-wandering and sound. This will be explained further in section 3.

2 Related work

Mind Wandering is a vague term, which has different interpretations according to the context it is proposed in. A commonly used definition is : When mind wandering occurs, the executive components of attention appear to shift away from the primary task, leading to failures in task performance and superficial representations of the external environment [10]. Other definitions have also arisen, stating that the term "mind wandering" refers to a flow of thought that is unrelated to the current setting [16]. The definition that will be used in the scope of this paper contains parts of the aforementioned definition, with a slight modification. The second part of the definition is really important in this specific case, due to the nature of the data of the Mementos data set.

Studies with the aim of detecting mind-wandering have taken place, mainly detecting it through thought-report methods, namely thought probes [17–20], and self-reports [10, 21]. With the first method, individuals are asked about their subjective attentional states at random. This method however, omits important details like the time of switching states, the commencement time, and the duration of a mind-wandering episode. Furthermore, the mental state of participants following a thought-probe cannot be assessed: whether they continue mind-wandering, commence a new mind-wandering episode, or return their attention to the activity [6].

Participants are asked to record the instant they become aware of their mind wandering in spontaneous self-reports. This approach allows the participant's mind wandering to be tracked in real time. However, because this tracking is subjective, researchers are limited in their capacity to maintain consistency in their evaluations of various individuals. Both approaches share the issue of evaluating a participant's mind wandering completely by themselves, and individuals may not be aware when their attention wanders [6].

A few studies have also been conducted around the relation of sound and mind wandering. A paper studying the effect of sad and happy music on mind-wandering, showed that sad music has the effect of causing stronger mind-wandering, in comparison to happy music, as music "is an effective tool to regulate thoughts via emotion" [22], whilst another study focused on students and how they mind-wander in a classroom, showed that mind-wandering is more common when speech is not clear [23]. The occurrence of mind wandering or not in relation to sound has not been researched about heavily though. For this reason, further research is required and encouraged to show if there exists a correlation between the two.

3 Methodology

To answer the research question formulated, the approach consisted of several steps, all crucial in the process. The first step consisted of preprocessing the video data of the Mementos data set, and then followed the annotation of the video data. These are the initial preparatory processes, which where then followed by the audio feature extraction and finally the process of mind-wandering classification.

3.1 Data preprocessing

The initial step of processing the Mementos data was to split the initial data set into a smaller subset. This was done due to the fact that the process of annotation meant watching every single video, thus, watching all the 1995 videos in the time frame of the project was not feasible. Therefore a smaller subset of 633 videos was taken, however, this subset still needed to be cleaned. The Mementos data set was constructed under an experiment where people were given the task to give their complete attention to watching the music videos and that they should be the only person present in the video recording [10]. However, some of the videos of the data set did not adhere to these conditions, as in some cases people were distracted by other individuals during the whole entirety of the video or in some extreme cases people would simply stand up and leave the frame of the camera. Therefore, these type of videos were removed, in order to have a clean data set, resulting in 45 videos being removed, leaving 588 videos.

3.2 Annotation of videos

The next step that followed after the data preprocessing was to annotate the data set, in order to create the ground truth values for our mind-wandering classification. This annotation consisted of watching every video of the subset, and annotating segments of the video where mind-wandering was expected to be occurring. In order to make this annotation more reliable, a rule book was initially created, stating signs of mind wandering, that should be looked at for when annotating the videos. These signs consisted of :

1. Smiling
2. Gazing
3. Squinting eyes
4. Person making sounds
5. Frowning

A smile can be 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 a form of recognizing mind wandering [24].

Gazing, refers to looking up for a longer time than just a look to a direction that was caused by a distraction. Usually trying to remember/recollect something comes with looking up and to the side for some time, which is also the case for squinting eyes, as it can indicate that that person is trying to remember something [25].

When the person makes a sound, that is not caused by external stimuli, this can be an indication of mind wandering, as it shows the person having some internal thought unrelated to the video they are watching, which are being externalized in the form of speech, or them making some sort of sound.

Frowning can be 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. However, a very subtle frown, could also be a form of reminiscing / remembering a memory. Frowning is thus a sign of negative emotional experience recollection [26].

Thus, some ground rules were created as to what signs indicate mind-wandering, as to guide us in the annotation process. For the annotation, the VGG tool was used, as it is a "a light weight, standalone and offline software package" [27], which seemed like the appropriate choice, given the fact that the work was done on the Mementos data set, which falls under the GDPR regulations of personal data collection and processing [1]. This annotation process consisted of splitting the subset of selected videos across two teams of 2 and 3 members, in order to have a second and third opinion and verification when creating the ground truth data. The teams were regularly mixed up, and breaks were taken during the annotation process, as to ensure the validity of the ground truth creation. In this way all of the videos were mind-wandering occurred were annotated, and the names of these files together with the time period of annotated mind-wandering were exported to csv files.

After the annotation process had finished, it was noticed that out of all the input videos, only 52 videos contained annotations of mind wondering. This meant that our data was imbalanced, and some over/under sampling or data augmentation needed to be done.

3.3 Audio Feature Extraction

After having annotated the data, the step that followed was the audio feature extraction, and therefore the wav files needed to be extracted from the videos. During this process, some of the audio files were corrupted and therefore some more input data was deleted. 8 audio files in total were corrupt, leaving 580 wav files. The signals from all the audio files were not down-sampled or up-sampled, and therefore the sampling rate of the audio input was 44100 Hz. There are two stages in the feature extraction methodology:

1. Short-term feature extraction
2. Mid-term feature extraction

For the short-term feature extraction, the input signal is split into short-term windows (frames) and a number of features for each frame is computed. This process leads to a sequence of short-term feature vectors for the whole signal.

Regarding the mid-term feature extraction, the signal is represented by statistics on the extracted short-term feature sequences described above. A number of statistics, mean and standard deviation are calculated over each short-term feature sequence.

A frame size of 50 msecs and a frame step of 25 msecs (50% overlap) was used and three types of acoustic features were used in this work. One of the features used are spectrum envelope representations used in speech/speaker recognition, namely the typical mel-frequency cepstral coefficients (MFCC) plus the frame energy [28]. They will also be considered together with their first time derivative (the so-called delta features). The second type of features are the so called perceptual features, and are features that are not a part of the above feature set, such as the zero crossing rate, energy and different spectral features. The third type of features are the chroma features, which are descriptors, which indicate the tonal content of a musical audio signal in a compressed format. As a result, chroma characteristics may be thought of as a necessary precondition for high-level semantic analysis such as chord identification or harmonic similarity estimation [29]. These features were used as they are able to grasp the harmonic and melodic qualities of music, and will therefore be useful in this research, as the Mementos data set contains peoples reactions to music videos. Namely, the features considered are the following:

1. Zero Crossing Rate: The rate of sign-changes of the signal during the duration of a particular frame.
2. Energy: The sum of squares of the signal values, normalized by the respective frame length.
3. Entropy of Energy: The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4. Spectral Centroid: The center of gravity of the spectrum.
5. Spectral Spread: The second central moment of the spectrum.
6. Spectral Entropy: Entropy of the normalized spectral energies for a set of sub-frames.
7. Spectral Flux: The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8. Spectral Rolloff: The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9. MFCCs: Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
10. Chroma Vector: A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
11. Chroma Deviation: The standard deviation of the 12 chroma coefficients.

The beats per minute rate (BPM) of the signal were also calculated, and a confidence score was also calculated. The above described features are considered in the experiments described in section 4. The mid-term features were taken for classification, thus, long-term averaged audio features were extracted, and it is important to note that one single feature vector is finally extracted per wav file.

3.4 Machine Learning algorithm

The Support Vector Machine (SVM) paradigm has been proven to be extremely effective in a variety of classification applications. It may use far less data to conduct accurate classification since it discriminates the data by defining borders between classes rather than estimating class conditional densities [28]. SVMs have also already been applied to audio classification and segmentation tasks in the past [30–32]. Due to these reasons, the SVM classifiers are used in this study. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data

[1] https://gdpr-info.eu/

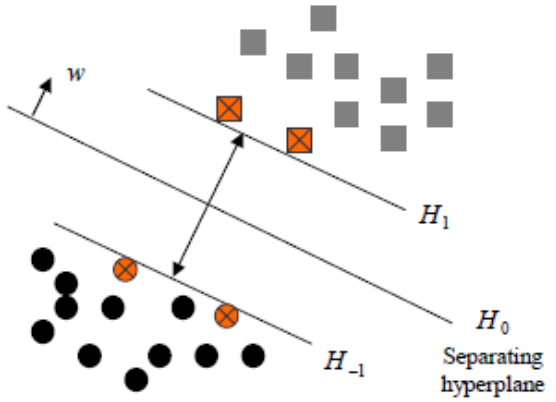

Figure 1: SVM two-class linear classification [28]

points, as also shown in Figure 1. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. The objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Different variants of the SVM technique were tried out, and used in different experiments, as described in section (here reference the experiments section). Not only were different variants of SVM tried out, but also different data balancing techniques were also tried out, due to the fact that the data was also imbalanced.

Both linear and rbf kernel SVM were tried out in this work. Linear SVM is used for linearly separable data, which means if a data set can be classified into two classes by using a single straight line, then such data is termed as linearly separable data. On the contrary, RBF kernel is mostly used for non-linear data, and basically RBF kernels place a radial basis function centered at each point, then performs linear manipulations to map points to higher-dimensional spaces that are easier to separate.

The kernel function, kernel parameters, and the soft margin parameter C all influence the efficacy of an SVM [33]. The error term's penalty parameter is C. It manages the trade-off between a smooth decision boundary and accurately categorizing training points. For large values of C, the optimization will pick a smaller-margin hyperplane if it performs a better job of accurately classifying all of the training points. A very small value of C, on the other hand, will encourage the optimizer to seek for a larger-margin separating hyperplane, even if it misclassifies more points. In order to choose the best value for C, a cross validation procedure was performed in order to select the optimal classifier parameter.

4 Experimental Work

Several different techniques and feature sets were used during the experimentation process, consisting of using different variants of the SVM algorithm, using different oversampling techniques. All of these play an important role in obtaining the final results.

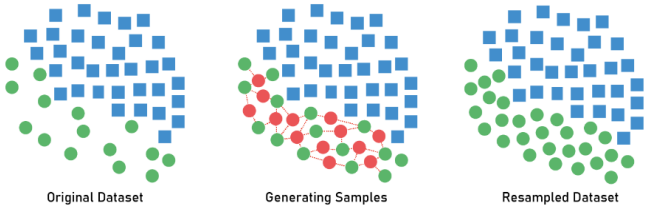

Figure 2: Synthetic Minority Oversampling TEchnique (SMOTE) [34]

4.1 Preparing the experiment

The first step in conducting the experiments is splitting the data into train and test splits, in order for the classifier to be trained, and then to be tested. The ratio of train and test data selected was 80:20, and 10% of the training data was used for validation. Therefore, the training data consisted of 422 wav files with the "No mind-wandering" label, and 41 wav files with the "mind-wandering" label. Each person in the Mementos data set, reacted to 4-7 music videos, therefore meaning that there were multiple instances of every person in the data set. In order for there to not be any bias in the experiment, one single person did not appear in both the train and test split. This means that the data was split randomly into an 80:20 ratio, with the clause that if a person was selected to be in the train split, all of that person's videos would also go to the train split (same goes for a person selected to be in the test split). This was done, as the fundamental objective of testing a model is to estimate how well it will perform predictions on data that the model didn't see.

4.2 Dealing with the data imbalance

Data imbalance was also an issue in this work. The positive class (mind-wandering class) initially consisted of 52 wav files, whilst the negative class (no-min-wandering) consisted of 528 wav files. The remedy to this problem was to over sample the minority class, and two separate strategies were tried: Random Over-Sampling and Synthetic Minority Oversampling Technique (SMOTE). Random oversampling entails picking samples from the minority class at random, replacing them with new ones, and adding them to the training data set until the required ratio is reached. Using the SMOTE approach, new instances are synthesized from the minority class, which is a form of data augmentation. SMOTE works by picking instances in the feature space that are close together, drawing a line in the feature space between the examples, and drawing a new sample at a location along that line, as also shown in Figure 2. To be more specific, a random case from the minority class is picked initially. Then, for that example, k of the closest neighbors are found. A randomly determined neighbor is picked, and a synthetic example is constructed at a randomly chosen position in the feature space between the two instances.

Thus, the next step in preparing the experiment was the process of oversampling the data. Initially Random Over-Sampling was chosen, with which the positive class (mind-

	MW	NO-MW
MW	50.35	0
NO-MW	7.88	41.76
Best Macro f1	92.1	
Best Macro f1 std	2.9	
Selected C parameter	20	

Table 1: Validation Results with oversampling

	MW	NO-MW
MW	2.34	7.66
NO-MW	14.26	75.74
Best Macro f1	52.5	
Best Macro f1 std	7.8	
Selected C parameter	5	

Table 2: Validation Results with SMOTE

wandering label), was oversampled until the ratio between the two classes was 1:1. Initially the linear SVM classifier was trained, and the RBF kernel SVM was also trained on the same data set, taking noticeably more time. Afterwards, the initial training data set was also oversampled using the SMOTE technique of data augmentation. This method of oversampling was more time costly, and both the linear and RBF kernel SVM were trained on this data.

4.3 Selection of Algorithm

The two different SVMs, namely the one trained on the data where random oversampling was used and the other trained on the data where SMOTE was used, were compared against each other through the validation sets. As preliminary tests with the SVM classifier showed a superiority of the linear kernel over the RBF kernel, only the former was used in the evaluation. An RBF kernel SVM is not a parametric model, and the complexity of it grows with the size of the training data. Not only is it more expensive to train an RBF kernel SVM, but the kernel matrix also has to be stored, and projection into this "infinite" higher dimensional space where the input becomes linearly separable is also more expensive during prediction. Furthermore, because there are more hyper parameters to tweak, model selection is more costly, whilst overfitting a complex model is considerably easy.

The aforementioned beat extraction was technique, and preliminary evaluated through the calculated confidence value. The confidence value, was consistently low during the preparation of the experiment, and therefore it was not used in the procedure of obtaining the results. This is due to the fact that the beat extraction technique is mainly useful for music classification tasks, however in the wav files used in this work, the music played through the device, was either of low sound quality, or in many cases also not detectable, due to the person using headphones. Due to these reasons, the confidence value of the BPM was low, and decided not to be used, as it could skew the data. This decision also reduced the time of the feature extraction process, which helped in the following step which was to actually run the experiments.

4.4 Implementation

The approach was implemented as a Python 3.9 project, and the feature extraction and machine learning model training and testing was done with the help of the open source pyAudioAnalysis library [35], which is a library that provides a wide range of audio analysis procedures.

As far as the implementation is concerned, after the data was split into respective train, validation and test sets, ex-

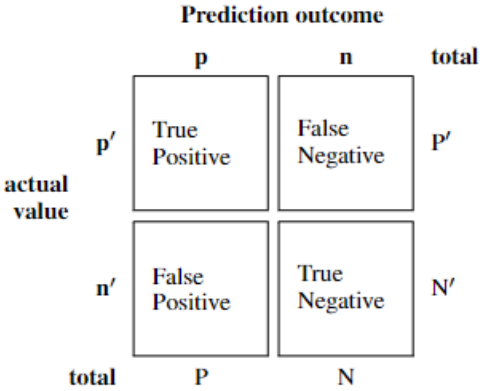

Figure 3: Confusion Matrix

periments were run using both random oversampling and SMOTE. The linear SVM was trained on the train data, and through a 10-fold cross validation, the soft margin parameter C was chosen for both models. For the oversampled data set, C = 20 was chosen whilst for the model where SMOTE was used, C = 5 was chosen. These values were used, as they gave the best results during the validation process, where a confusion matrix was created, of the form showcased in Figure 3. The results of the validation process are shown in Table 1 and also Table 2.

5 Results and Discussion

This section outlines and discusses the results of the study and the experiments that were taken place. Both the results of the model that were trained using oversampling and SMOTE are going to be analysed, and they are going to be compared against each other. Section 5.1 also discusses the answers to the research sub questions, whilst section 5.2 discusses the answer to the main research question.

When evaluating the model, the result will either be a True Positive (TP), True Negative (TN), False Positive (FP) or a False Negative (FN) depending on the result the classification of the model. These result classes are further showcased in Figure 3.

A true positive means that the actual value of the file is mind-wandering, and the model predicted mind-wandering. A false negative means that the actual value of the file is mind-wandering and the model predicted non-mind-wandering, whilst a false positive is when the actual value of the file is non-mind-wandering and the model predicted mind-wandering. Finally a true negative is when the actual

Figure 4: Results with oversampling

Confusion matrix, acc = 79.5%, F1 (macro): 54.1%		
	MW	NO-MW
MW	4	7
NO-MW	17	89

Table 3: Confusion Matrix results with oversampling

value of the file was non-mind-wandering and the model also predicted non-mind-wandering.

For this study 4 measures were used for evaluation:

- Recall: The ratio of accurately predicted positive observations to all observations in the actual class.

$$\frac{TP}{TP + FN} \tag{1}$$

- Precision: The ratio of accurately anticipated positive observations to total expected positive observations is known as precision.

$$\frac{TP}{TP + FP} \tag{2}$$

- F1-score: The weighted average of Precision and Recall is the F1 Score. As a result, this score considers both false positives and false negatives.

$$\frac{2 * Recall * Precision}{Recall + Precision} \tag{3}$$

- Accuracy: This is the simplest intuitive performance metric, which is the ratio of properly predicted observations to all observations.

$$\frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

There metrics where chosen because they are well-known and produce normalized values for the model's accuracy, making comparison easier. The models were tested against the test data set, which contained 117 wav files, consisting of 11 cases of mind-wandering and 106 cases of non-mind-wandering. Below, the results of the respective experiments are shown. In the class-wise Performance measures plots, namely Figure 4 and Figure 5, the green bar indicates recall, the blue bar indicates precision and the red bar the f1-score.

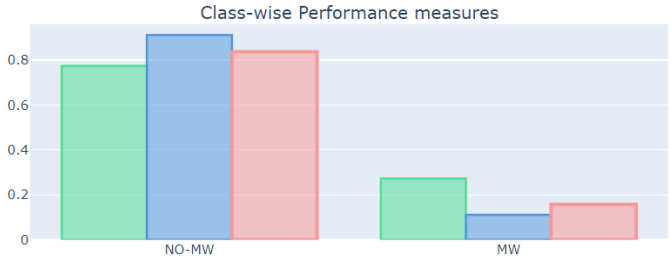

Figure 5: Results with SMOTE

Confusion matrix, acc = 72.6%, F1 (macro): 49.7%		
	MW	NO-MW
MW	3	8
NO-MW	24	82

Table 4: Confusion Matrix results with SMOTE

5.1 Research Sub-questions

5.1.1 Detection of accuracy of mind-wandering through sound. The results of all metric can be found in Figure 4, Figure 5, Table 3 and Table 4. Note that all the metrics have a minimum value of 0 and maximum value of 1, 0 being a complete lack of this metric and 1 being the optimal score.

When oversampling was used Table 3 showcases that the overall accuracy of the model was 79.5%, with an f1 macro score of 54.1%. This high accuracy is due to the fact that the model had many true negatives, precisely 89, where the videos the belonged to the negative class, that of non-mind-wandering, got correctly predicted. However, the model had a very low precision score (0.18) when it came to predicting the correct output of the positive class, namely the mind-wandering class. We see through Table 3 that the model only had 4 true positives, whilst having 7 false negatives and 17 false positives. Similarly, the positive class has also low scores for recall (0.251) and also a low f1-score (0.204). This is most likely due to the fact that due to the little amount of data, and also the quality of the data. Regarding the negative class, high scores were obtained, however the model performs poorly in finding the cases of mind-wandering in the test set. It is difficult to draw conclusions from the obtained results, due to the low amounts of ground truth data regarding the mind-wandering class. As explained above oversampling was used to balance the data set, however the initial imbalance of the data sets still plays a major role in the validity of the results.

When SMOTE was used similar but slightly worse results were obtained. When SMOTE was used, Table 4 showcases that the overall accuracy of the model was 72.6%, and the f1 macro score of 49.7% was obtained. The true positives were also lower, now being 3 instead of 4, whilst the true negatives also decreased to 82. A consequence of this is that more false negatives and false positives were observed, namely 7 and 24. The general accuracy of the model is lower compared to the oversampling technique, and the same goes for the f1-score. SMOTE creates a trade off between a higher recall score and

a lower precision score, which is also what is noticed in the obtained results of the positive class. Through Figure 5 it is shown that the model obtained a slightly higher recall score in the positive class (0.273), whilst having a lower precision score (0.111) and also a lower f1-score (0.159). This is due to the fact that SMOTE has a high chance to oversample the samples with a small amount of information. It can also increase the overlap between different classes around class boundaries. This is because SMOTE can blindly oversample and it can oversample every single sample of the minority class, without any reasoning or justification. This means that only the number of samples in each class and the closeness between the samples in the minority class are considered, and other characteristics of the data are not considered by the algorithm [36].

5.1.2 Effect of wearing headphones on mind- wandering. In order to answer this research sub-question, a quantitative approach was taken. The test split contained some cases of people that were wearing some sort of headphones, whilst also contained people that did not wear headphones, and were listening to the music videos through speakers. When the model was trained and tested against the test data, many false positives were noticed. This means that the model predicted many wav files as mind-wandering whilst in reality they were non-mind wandering, however there was no connection between the use of headphones or not. The results obtained showed no concrete connection between the effect of wearing headphones on mind-wandering. This is probably due to the quantity but also the quality of the data. What is meant by this is that the data used in this study regarding the positive class was very little, and thus techniques such as oversampling and SMOTE were used to remedy this. Regarding the quality of the data, this refers to the cases where people were not using headphones, and the music videos were played through speakers. In these cases, the quality of the music played varied, due to various reasons such as intensity of speaker volume, or microphone used. Thus, there were no signs indicating a strict correlation between the two.

5.1.3 Comparison of mind-wandering detection through sound with other techniques. Other techniques such as thought probes and self-reports, as also mentioned in Section 2, but also using electrophysiological measures [37] and gaze-based eye tracking [38] are methods that have been used to detect mind-wandering. These methods have proven to be more reliable than the method proposed in this study.

Specifically, in the case of electrophysiological measures, it was studies whether these can be used in machine learning models to accurately predict mind wandering states. It was proven that through the recording of scalp EEG from participants, non-linear and linear machine learning models detected mind-wandering, above-above. This suggests that an individuals attention state can reliably be detected based on ERP patterns [37]. Similarly to this, eye-trackers have also been used in order to detect mind-wandering. Through tracking eye-gaze, mind-wandering was able to be predicted with an F1 score of 0.59, considerably better than chance which had an F1 score of 0.24 [38].

The methods of thought probes and self-reports have also been used extensively in the past, and despite having their downsides, they give more reliable results than the use of sound, as experimented in this work.

5.2 Research Question

Feasibility of mind-wandering detection through sound. During this study, it was found that the linear SVM model, using both techniques of random oversampling and SMOTE for dealing with data imbalance, showcases low accuracy in terms of detecting mind-wandering. Moreover, it was found that both techniques provide similar results, with the model using SMOTE for oversampling showing lower general results, however higher recall scores for the positive class, with lower precision, which was expected, due to the nature of the SMOTE algorithm. The hypothesis regarding these results is that this was caused due to the high data imbalance between the two classes, and the use of oversampling methods. Also the initial hypothesis was that external sounds play a more crucial role in detecting cases of non-mind-wandering. This research has not been enough in order to prove this hypothesis, as there where cases where external sounds were present and those audio files were classified correctly as non-mind-wandering cases, however there were also many cases were the testing audio files were in the no-mind-wandering class, however the classifier classified them as mind-wandering. Therefore, the existence of many false positives, as shown in Table 3 and Table 4 shows that many actual non-mind-wandering audio files got classified as mind-wandering. These audio files that were wrongly classified as mind-wandering, contained also audio files that were silent, due to people listening to the music video through headphones, and there were sudden sounds present, indicating external sounds. Therefore, there is not information to back up the hypothesis that was initially proposed, however further research could be helpful in proving more concrete evidence that supports this hypothesis or disproves it.

A comparison was also done between the method of mind-wandering detection through sound and also other techniques that have been previously used. There is definitely a correlation between mind-wandering and sound, as also shown by previous research [22], however through comparison with the other techniques (thought probes, self-reports, electrophysiological measures, eye-based tracking) it was shown that automatic mind-wandering detection through sound is not as effective as the aforementioned methods.

6 Responsible Research

Despite the experiments already described in this work, it is important to also be able to further experiment, however there are some limitations regarding this aspect in this work. The Mementos data set being a sensitive data set, falling under the GDPR regulations of personal data collection, manual annotation and oversampling are all topics that need to be mentioned.

The videos that were used during the annotation and experimenting process are videos from the Mementos data set. This data set contains sensitive data, that can not be accessed, unless signed approval has been given. This means that the data is not accessible to the open public. This complicates

the reproducibility of the research, as a EULA needs to be signed in order to get access to the data. Even after obtaining the data, the data can not be uploaded anywhere, or moved from the local device. This restricts the use of tools and libraries, therefore if one were to reproduce the research, such tools would not be able to be used. However, the library used in this work, is an open source library, therefore if access was given to the Mementos data set, this research could be reproduced. However, similar research could still be done, with the use of this library, and a different data set.

The process of annotating the data set, was done manually by watching the videos one by one, and annotating the time stamps where mind wandering was expected to be occurring. Therefore, the ground truth data was created by the means of perceived mind wandering, and not by physiological data or self reports, which is the case in other existing research [10, 39]. Therefore, this has negative consequences on the reliability of the research. To mitigate this as much as possible, the rule book was created, teams were constantly randomized and breaks were taken during annotating sessions. For any uncertainty, all of the 5 members of the research group were consulted, and a group decision was taken. Another issue arising from this way of creating the ground truth values, is that during the process of annotation, the people watching the videos and annotating them could also be mind-wandering. In order to mitigate this as much as possible, teams were often randomized, breaks were taken, and the mind-wandering annotation sessions were not too long, as also previously mentioned. Despite these efforts to mitigate this happening, it could have still been the case in some sessions.

Another issue that needs to be mentioned is the imbalance of the data. Due to the fact that the data set was imbalanced, oversampling was used in order to balance the data set, in order to be able to train the classifier. This has the advantage of balancing the data, however, because it creates precise replicas of the minority class samples, it may increase the chance of overfitting. A symbolic classifier, for example, can generate rules that appear to be accurate but only cover one reproduced example in this way. Two techniques were tested and tried out for over sampling the minority class, in order to evaluate both methods, however this is an issue that is still worth mentioning.

7 Conclusions and Future Work

This study investigated the ability to recognise mind-wandering through sound. To achieve this a linear SVM model was chosen, and random oversampling and SMOTE were the techniques used to handle the issue of having an unbalanced data set.

An experimental study was conducted to showcase the feasibility of using sound to detect cases of mind-wandering. To achieve this the selected videos from the Mementos data set were annotated, in order to create ground truth values. After this the audio files were extracted from the respective videos in order for the acoustic features to be able to be extracted. With these videos the SVM models were trained and tested to see if the detection of mind-wandering through sound was feasible. The results showcase that through sound good results are obtained for non-mind-wandering cases, however lower metric values are obtained for detecting mind wandering, which is likely due to the imbalance in the non-mind-wandering and mind-wandering test data.

This research suggests a method of binary classification, of mind-wandering and non mind wandering. While it does have the advantage of separating occurrences of mind-wandering and occurrences of non mind-wandering, it does induce a strict binary classification between these two classes, whilst mind-wandering is a more complex state. Therefore, a method of having a confidence score for each video instead of a classification of yes and no could provide a good basis for further extended research in this area. As mind-wandering is a subjective process, which every person can display in different ways, a confidence score could mitigate some of the errors of classification.

Methods for oversampling the minority class were used in this work, due to the low amount of data in the mind-wandering class, and the high amount of data in the no-mind-wandering class. For a future study, a combination between oversampling and undersampling can be used, especially through the use of SMOTE, as the algorithm seems to benefit from a combination with undersampling the majority class. This could potentially also be optimized as a hyperparameter of the pipeline, in order to see what the effects of ratios of undersampling and oversampling are.

References

[1] Jerome L. Singer and Vivian G. McCraven. "Some Characteristics of Adult Daydreaming". In: *The Journal of Psychology* 51.1 (1961), pp. 151–164. DOI: 10.1080/00223980.1961.9916467. eprint: https://doi.org/10.1080/00223980.1961.9916467. URL: https://doi.org/10.1080/00223980.1961.9916467.

[2] Michael J. Kane et al. "For Whom the Mind Wanders, and When: An Experience-Sampling Study of Working Memory and Executive Control in Daily Life". In: *Psychological Science* 18.7 (2007). PMID: 17614870, pp. 614–621. DOI: 10.1111/j.1467-9280.2007.01948.x. eprint: https://doi.org/10.1111/j.1467-9280.2007.01948.x. URL: https://doi.org/10.1111/j.1467-9280.2007.01948.x.

[3] Matthew A. Killingsworth and Daniel T. Gilbert. "A Wandering Mind Is an Unhappy Mind". In: *Science* 330.6006 (2010), pp. 932–932. DOI: 10.1126/science.1192439. eprint: https://www.science.org/doi/pdf/10.1126/science.1192439. URL: https://www.science.org/doi/abs/10.1126/science.1192439.

[4] Jefferson A. Singer and Peter Salovey. "Thought Flow: Properties and Mechanisms Underlying Shifts in Content". In: 1999.

[5] Eric Klinger and W. Miles Cox. "Dimensions of Thought Flow in Everyday Life". In: *Imagination, Cognition and Personality* 7.2 (1987), pp. 105–128. DOI: 10.2190/7K24-G343-MTQW-115V.

eprint: https://doi.org/10.2190/7K24-G343-MTQW-115V. URL: https://doi.org/10.2190/7K24-G343-MTQW-115V.

[6] Yilei Zheng et al. "Detecting Mind Wandering: An Objective Method via Simultaneous Control of Respiration and Fingertip Pressure". In: *Frontiers in Psychology* 10 (2019). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.00216. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2019.00216.

[7] Jonathan W. Schooler et al. "Meta-awareness, perceptual decoupling and the wandering mind". In: *Trends in Cognitive Sciences* 15 (2011), pp. 319–326.

[8] Laura Schmalzl, Chivon Powers, and Eva Henje Blom. "Neurophysiological and neurocognitive mechanisms underlying the effects of yoga-based practices: towards a comprehensive theoretical framework". In: *Frontiers in Human Neuroscience* 9 (2015). ISSN: 1662-5161. DOI: 10.3389/fnhum.2015.00235. URL: https://www.frontiersin.org/article/10.3389/fnhum.2015.00235.

[9] Paul Seli et al. "Mind-Wandering With and Without Intention". In: *Trends in Cognitive Sciences* 20.8 (2016), pp. 605–617. ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2016.05.010. URL: https://www.sciencedirect.com/science/article/pii/S1364661316300523.

[10] Jonathan Smallwood and Jonathan Schooler. "The Restless Mind". In: *Psychological bulletin* 132 (Dec. 2006), pp. 946–58. DOI: 10.1037/0033-2909.132.6.946.

[11] B. Dudzik et al. "Collecting Mementos: A Multimodal Dataset for Context-Sensitive Modeling of Affect and Memory Processing in Responses to Videos". In: *IEEE Transactions on Affective Computing* 01 (June 5555), pp. 1–1. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2021.3089584.

[12] Dylan Jones and William Macken. "Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19 (Mar. 1993), pp. 369–381. DOI: 10.1037/0278-7393.19.2.369.

[13] Jens Gisselgård, Karl Magnus Petersson, and Martin Ingvar. "The irrelevant speech effect and working memory load". In: *NeuroImage* 22 (Aug. 2004), pp. 1107–16. DOI: 10.1016/j.neuroimage.2004.02.031.

[14] Randi Martin, Michael Wogalter, and Janice Forlano. "Reading comprehension in the presence of unattended speech and music". In: *Journal of Memory and Language* 27 (Aug. 1988), pp. 382–398. DOI: 10.1016/0749-596X(88)90063-0.

[15] Catherine Oswald, Sébastien Tremblay, and Dylan Jones. "Disruption of comprehension by meaning of irrelevant sound". In: *Memory (Hove, England)* 8 (Oct. 2000), pp. 345–50. DOI: 10.1080/09658210050117762.

[16] Sandra W. Russ. "Chapter 10 - Mind wandering, fantasy, and pretend play: a natural combination". In: *Creativity and the Wandering Mind*. Ed. by David D. Preiss, Diego Cosmelli, and James C. Kaufman. Explorations in Creativity Research. Academic Press, 2020, pp. 231–248. ISBN: 978-0-12-816400-6. DOI: https://doi.org/10.1016/B978-0-12-816400-6.00010-9. URL: https://www.sciencedirect.com/science/article/pii/B9780128164006000109.

[17] James Cheyne et al. "Anatomy of an error: A bidirectional state model of task engagement/disengagement and attention-related errors". In: *Cognition* 111 (Mar. 2009), pp. 98–113. DOI: 10.1016/j.cognition.2008.12.009.

[18] David Stawarczyk et al. "Mind-wandering: Phenomenology and function as assessed with a novel experience sampling method". In: *Acta psychologica* 136 (Feb. 2011), pp. 370–81. DOI: 10.1016/j.actpsy.2011.01.002.

[19] Paul Seli, James Cheyne, and Daniel Smilek. "Wandering Minds and Wavering Rhythms: Linking Mind Wandering and Behavioral Variability". In: *Journal of experimental psychology. Human perception and performance* 39 (Dec. 2012), pp. 1–5. DOI: 10.1037/a0030954.

[20] Daniel Levinson et al. "A mind you can count on: Validating breath counting as a behavioral measure of mindfulness". In: *Frontiers in psychology* 5 (Oct. 2014), p. 1202. DOI: 10.3389/fpsyg.2014.01202.

[21] Claire Braboszcz and Arnaud Delorme. "Lost in thoughts: Neural markers of low alertness during mind wandering". In: *NeuroImage* 54 (Oct. 2010), pp. 3040–7. DOI: 10.1016/j.neuroimage.2010.10.008.

[22] Liila Taruffi et al. "Effects of Sad and Happy Music on Mind-Wandering and the Default Mode Network". In: *Scientific Reports* 7 (Oct. 2017). DOI: 10.1038/s41598-017-14849-0.

[23] Ian Gliser et al. "The sound of inattention: Predicting mind wandering with automatically derived features of instructor speech". English (US). In: *Artificial Intelligence in Education- 21st International Conference, AIED 2020, Proceedings, Part I*. Germany, 2020, pp. 204–215. ISBN: 9783030522360. DOI: 10.1007/978-3-030-52237-7_17.

[24] Mathieu Arminjon et al. "Embodied memory: unconscious smiling modulates emotional evaluation of episodic memories". In: *Frontiers in Psychology* 6 (2015). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2015.00650. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2015.00650.

[25] Jordana Wynn, Jennifer Ryan, and Bradley Buchsbaum. *Eye movements support behavioral pattern completion*. Sept. 2019. DOI: 10.1101/764084.

[26] Sascha Duken et al. *Reliving emotional memories: Episodic recollection elicits affective psychophysiological responses*. Sept. 2021. DOI: 10.31234/osf.io/ukt5x.

[27] Abhishek Dutta and Andrew Zisserman. "The VIA Annotation Software for Images, Audio and Video". In: *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, Oct. 2019. DOI: 10.1145/3343031.3350535. URL: https://doi.org/10.1145%2F3343031.3350535.

[28] Andriy Temko and Climent Nadeu. "Classification of acoustic events using SVM-based clustering schemes". In: *CHIL* 39 (Apr. 2006). DOI: 10.1016/j.patcog.2005.11.005.

[29] Ayush Shah et al. "Chroma Feature Extraction". In: Jan. 2019.

[30] Stan Li and Guodong Guo. "Content-based audio classification and retrieval using SVM learning". In: (Jan. 2000).

[31] Lie Lu and Stan Li. "Content-based audio classification and segmentation by using support vector machines". In: *Multimedia Systems* 8 (Apr. 2003), pp. 482–492. DOI: 10.1007/s00530-002-0065-0.

[32] Jia-Ching Wang et al. "Content-Based Audio Classification Using Support Vector Machines and Independent Component Analysis". In: vol. 4. Jan. 2006, pp. 157–160. DOI: 10.1109/ICPR.2006.407.

[33] Seeja K.R. and Shweta. "Microarray data classification using support vector machine". In: *International Journal of Biometrics and Bioinformatics (IJBB)* 5 (Jan. 2011), pp. 10–15.

[34] Emilia Orellana. "SMOTE". In: (Dec. 2020).

[35] Theodoros Giannakopoulos. "pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis". In: *PLOS ONE* 10 (Dec. 2015), e0144610. DOI: 10.1371/journal.pone.0144610.

[36] Paria Soltanzadeh and Mahdi Hashemzadeh. "RCSMOTE: Range-Controlled Synthetic Minority Over-sampling Technique for handling the class imbalance problem". In: *Information Sciences* 542 (July 2020). DOI: 10.1016/j.ins.2020.07.014.

[37] Henry W Dong et al. "Detection of mind wandering using EEG: Within and across individuals". In: *PLoS ONE* 16 (2021).

[38] Stephen Hutt et al. "Automated gaze-based mind wandering detection during computerized learning in classrooms". In: *User Modeling and User-Adapted Interaction* 29 (Sept. 2019). DOI: 10.1007/s11257-019-09228-5.

[39] Nathaniel Blanchard, Tera Joyce, and Sidney D'Mello. "Automated Physiological-Based Detection of Mind Wandering during Learning". In: vol. 8474. June 2014. DOI: 10.1007/978-3-319-07221-0_7.