



**HOCHSCHULE  
MITTWEIDA**  
University of Applied Sciences

---


# MASTER THESIS

---

Ms.  
**Alexandra Lengert, B.Sc.**

## **Analyzing Privacy Threats in VR/AR Devices**

Mittweida, February 2024



Faculty of **Applied Computer Sciences and Biosciences**

---

# **MASTER THESIS**

---

## **Analyzing Privacy Threats in VR/AR Devices**

Author:

**Alexandra Lengert**

Course of Study:

Cybercrime/Cybersecurity

Seminar Group:

CY21wC-M

First Examiner:

Dr. Michael Spranger

Second Examiner:

Dr. Ivan De Oliveira Nunes

Submission:

Mittweida, 24.02.2024

Defense/Evaluation:

Mittweida, 2024

Faculty of **Applied Computer Sciences and Biosciences**

---

## **MASTER THESIS**

---

### **Analyse von Bedrohungen für die Privatsphäre durch VR/AR Geräte**

Author:

**Alexandra Lengert**

Course of Study:

Cybercrime/Cybersecurity

Seminar Group:

CY21wC-M

First Examiner:

Dr. Michael Spranger

Second Examiner:

Dr. Ivan De Oliveira Nunes

Submission:

Mittweida, 24.02.2024

Defense/Evaluation:

Mittweida, 2024

## **Bibliographic Description**

Lengert, Alexandra:

Analyzing Privacy Threats in VR/AR Devices. – 2024. – 69 S.

Mittweida, Hochschule Mittweida – University of Applied Sciences, Faculty of Applied Computer Sciences and Biosciences, Master Thesis, 2024.

## **Abstract**

As new sensors are added to VR headsets, more data can be collected. This introduces a new potential threat to user privacy. We focused on the feasibility of extracting personal information from eye-tracking. To achieve this, we designed a preliminary user study focusing on the pupil response to audio stimuli. We used a variation of machine learning models to test the collected data to determine the feasibility of obtaining information such as the age or gender of the participant. Several of the experiments show promise for obtaining this information. We were able to extract with reasonable certainty whether caffeine was consumed and the gender of the participant. This demonstrates the unknown threat that embedded sensors pose to users. A further studies are planned to verify the results.

# Contents

|   |           |
|---|-----------|
| <b>Contents</b>   | <b>I</b>  |
| <b>List of Figures</b>  | <b>IV</b> |
| <b>List of Tables</b>   | <b>V</b>  |
| <b>Acronyms</b>   | <b>VI</b> |
| <b>1 Introduction</b>   | <b>1</b>  |
| 1.1 Purpose and Research Questions . . . . .                  | 1         |
| 1.2 Methodology Overview . . . . .                            | 2         |
| 1.3 Ethics and Data Protection . . . . .                      | 2         |
| 1.4 Limitations . . . . .                                     | 2         |
| 1.5 Outline . . . . .   | 3         |
| <b>2 Background</b>   | <b>4</b>  |
| 2.1 Augmented vs. Virtual Reality . . . . .                   | 4         |
| 2.2 Biological Background . . . . .                           | 4         |
| 2.2.1 Anatomy of the Eye . . . . .                            | 4         |
| 2.2.2 Eye Movement . . . . .                                  | 5         |
| 2.2.3 Pupil Dilation . . . . .                                | 6         |
| 2.3 Machine Learning . . . . .                                | 8         |
| 2.3.1 Machine Learning Algorithms . . . . .                   | 8         |
| 2.3.2 Feature Selection . . . . .                             | 10        |
| 2.3.3 Cross-Validation . . . . .                              | 12        |
| 2.3.4 Performance Measurement . . . . .                       | 12        |
| <b>3 Related Work</b>   | <b>14</b> |
| 3.1 Authentication Schemes for VR Systems . . . . .           | 14        |
| 3.2 Attacks on VR Systems . . . . .                           | 15        |
| 3.3 Privacy Concerns in VR . . . . .                          | 16        |
| 3.4 Eye Measurements . . . . .                                | 17        |
| <b>4 Experiment Design</b>                                    | <b>21</b> |
| 4.1 Original Experiment Setup . . . . .                       | 21        |
| 4.2 First Line of Experiments - Protocol 1 . . . . .          | 22        |
| 4.2.1 Experiment 1 - Stabilizing the Pupil Size . . . . .     | 22        |
| 4.2.2 Experiment 2 - Illumination . . . . .                   | 23        |
| 4.2.3 Experiment 3 - Focus Point . . . . .                    | 23        |
| 4.2.4 Conclusion of Experiments 1-3 . . . . .                 | 24        |
| 4.3 New Experiment Setup and Follow-up Experiment . . . . .   | 24        |
| 4.3.1 Experiment Setup . . . . .                              | 24        |
| 4.3.2 Experiment 4 - Defining the Layout . . . . .            | 25        |
| 4.4 Second Line of Experiments - Protocol 2 . . . . .         | 25        |
| 4.4.1 Experiment 5 - Differentiating Yes/No Answers . . . . . | 26        |

---

|          |   |           |
|----------|---|-----------|
| 4.4.2    | Experiment 6 - Differentiating Agreement and Disagreement . . . . . | 26        |
| 4.4.3    | Conclusion of Experiments 5 and 6 . . . . .                         | 26        |
| <b>5</b> | <b>Methodology of Data Collection</b>                               | <b>28</b> |
| 5.1      | Data Collection . . . . .   | 28        |
| 5.2      | Experiment Design . . . . .   | 28        |
| 5.2.1    | Pre-Study . . . . .   | 30        |
| 5.2.2    | Experiments . . . . .   | 30        |
| 5.2.3    | Post-Study Survey . . . . .   | 31        |
| 5.3      | Participants and Recruitment . . . . .                              | 32        |
| 5.4      | Demographics . . . . .  | 32        |
| <b>6</b> | <b>Methodology of Training</b>                                      | <b>34</b> |
| 6.1      | Collected Data . . . . .  | 34        |
| 6.2      | Data Preprocessing . . . . .  | 34        |
| 6.3      | Missing Data . . . . .  | 35        |
| 6.4      | Cleaning . . . . .  | 35        |
| 6.5      | Feature Extraction . . . . .  | 36        |
| 6.5.1    | Statistic Features . . . . .  | 36        |
| 6.5.2    | Curve Fitting and Gradients . . . . .                               | 37        |
| 6.5.3    | Blink Count . . . . .   | 37        |
| 6.6      | Data Preparation . . . . .  | 37        |
| 6.7      | Training . . . . .  | 38        |
| 6.8      | Evaluation . . . . .  | 38        |
| <b>7</b> | <b>Results and Discussion</b>                                       | <b>40</b> |
| 7.1      | Model Tuning . . . . .  | 40        |
| 7.2      | Gender Classification . . . . .                                     | 41        |
| 7.3      | Detection of Caffeine Consumption . . . . .                         | 43        |
| 7.4      | Age Classification . . . . .  | 43        |
| 7.5      | Decision-Making Classification . . . . .                            | 44        |
| 7.5.1    | Datasets . . . . .  | 44        |
| 7.5.2    | Cross-validation . . . . .  | 45        |
| 7.5.3    | Evaluation . . . . .  | 46        |
| 7.5.4    | Follow-up Experiment . . . . .                                      | 47        |
| 7.6      | Discussion . . . . .  | 48        |
| 7.6.1    | Discussion of Gender Classification . . . . .                       | 48        |
| 7.6.2    | Discussion of Caffeine Detection . . . . .                          | 50        |
| 7.6.3    | Discussion of Age Classification . . . . .                          | 52        |
| 7.6.4    | Discussion of Decision Detection . . . . .                          | 54        |
| 7.6.5    | Evaluation of the Classification Models . . . . .                   | 56        |
| <b>8</b> | <b>Conclusion and Future Work</b>                                   | <b>59</b> |
|          | <b>Bibliography</b>   | <b>61</b> |
|          | <b>Appendix</b>   | <b>70</b> |

---

|  |           |
|--|-----------|
| <b>A Survey Questionair</b>                          | <b>70</b> |
| A.1 Demographic Survey . . . . .                     | 70        |
| A.2 Ground Truth Survey [Digital Appendix] . . . . . | 70        |
| <b>B Study Material [Digital Appendix]</b>           | <b>71</b> |
| <b>Statutory Declaration in Lieu of an Oath</b>      | <b>74</b> |

# List of Figures

|     |   |    |
|-----|---|----|
| 2.1 | Anatomy of the eye, from the outside (upper) and horizontal cross-section (lower) [25]  | 5  |
| 2.2 | The extraocular muscles of the eye [28]   | 5  |
| 2.3 | Pupilar pathways, showing the efferent (blue) and afferent pathways (red) [30]  | 7  |
| 2.4 | Example of a DT for classifying ripe watermelons [41]   | 9  |
| 2.5 | Sigmoid function as graph and formula [43]  | 10 |
| 2.6 | Process of choosing feature selection techniques in the context of the input and output data [47]   | 11 |
| 3.1 | Overview of collectable eye data and the correlated PII [71]  | 18 |
| 4.1 | The original setup of the experiments before improvements, for better visibility taken in a lit room  | 21 |
| 4.2 | Pupil size change during Experiment 2   | 23 |
| 5.1 | Device appearance and its manner of wear.   | 28 |
| 5.2 | The current setup of the experiments, for better visibility taken in a lit room   | 29 |
| 5.3 | Distribution of nationalities in the participant group  | 32 |
| 5.4 | Age distribution of the participants  | 33 |
| 6.1 | Visualization of the cleaning steps starting with the raw data (a) followed by masking the first derivative (b) and the confidence (c), smoothing missing data with interpolation (d), and lastly applying the 3rd order Butterworth filter (e) | 36 |
| 7.1 | Confusion Matrices of the Decision Tree (DT) and the K-Nearest Neighbors (kNN) for gender classification evaluations  | 42 |
| 7.2 | Distribution of average blink rates for both genders  | 42 |
| 7.3 | Samples for gender classification in the feature space  | 49 |
| 7.4 | Samples for age classification in the feature space   | 52 |
| A.1 | Visuals of an example question from the Ground Truth Survey   | 70 |



## List of Tables

|      |  |    |
|------|--|----|
| 2.1  | Confusion Matrix . . . . .   | 12 |
| 4.1  | RGB values of the selected colors . . . . .  | 24 |
| 5.1  | Demographics of the participants . . . . .   | 33 |
| 7.1  | Overview of the parameters used for the classifiers that differ from the default settings . . . . .    | 40 |
| 7.2  | Gender distribution of the datasets . . . . .  | 41 |
| 7.3  | Average Precision, Recall, and Accuracy(acc) for gender evaluation . . . . .                           | 41 |
| 7.4  | Datasets distribution of caffeine consumption . . . . .  | 43 |
| 7.5  | Average Precision, Recall, and Accuracy(acc) for caffeine consumption evaluation                       | 43 |
| 7.6  | Age distribution of the datasets . . . . .   | 44 |
| 7.7  | Average Precision, Recall, and Accuracy(acc) for age classification evaluation .                       | 44 |
| 7.8  | Datasets label distribution . . . . .  | 45 |
| 7.9  | Cross-validation accuracy of Yes/No classification . . . . .   | 45 |
| 7.10 | Average Precision, Recall, and Accuracy(acc) for Yes/No evaluation . . . . .                           | 46 |
| 7.11 | Average Precision, Recall, and Accuracy(acc) for Agree/Disagree evaluation . .                         | 47 |
| 7.12 | Average Precision, Recall, and Accuracy(acc) for Yes/No evaluation of only male participants . . . . . | 47 |
| 7.13 | Average Precision, Recall, and Accuracy for Yes/No evaluation of only female participants . . . . .    | 48 |
| B.1  | Overview for the videos used in the design process . . . . .   | 72 |
| B.2  | Overview for the videos created for the preliminary study . . . . .                                    | 73 |

# Acronyms

|              |       |                                     |
|--------------|-------|-------------------------------------|
| <b>ANOVA</b> | ..... | Analysis of Variance                |
| <b>AR</b>    | ..... | Augmented Reality                   |
| <b>DT</b>    | ..... | Decision Tree                       |
| <b>FN</b>    | ..... | False Negative                      |
| <b>FP</b>    | ..... | False Positive                      |
| <b>HGS</b>   | ..... | Halving Grid Search                 |
| <b>kNN</b>   | ..... | K-Nearest Neighbors                 |
| <b>MLA</b>   | ..... | Machine Learning Algorithm          |
| <b>MLP</b>   | ..... | Multilayer Perceptron               |
| <b>MSB</b>   | ..... | Mean Square Between                 |
| <b>MSW</b>   | ..... | Mean Square Within                  |
| <b>P</b>     | ..... | Precision                           |
| <b>PII</b>   | ..... | Personally Identifiable Information |
| <b>R</b>     | ..... | Recall                              |
| <b>RIT</b>   | ..... | Rochester Institute of Technology   |
| <b>SVM</b>   | ..... | Support Vector Machine              |
| <b>TN</b>    | ..... | True Negative                       |
| <b>TP</b>    | ..... | True Positive                       |
| <b>VR</b>    | ..... | Virtual Reality                     |

# 1 Introduction

Advancements in Virtual and Augmented Reality have opened up new possibilities, changing the way we interact in the digital world and enhancing the immersive experience. Continuous developments, such as the VIVE Wrist Tracker, render hand-held controllers obsolete and allow for a more intuitive and liberating user experience [1]. In addition to entertainment, new fields are being explored through tailored services and developments for business and personal use [2]. Facial expression trackers lend more authenticity to conversations, resulting in online interactions that are true to life [3]. At the same time, eye-tracking technology is introducing more nuanced dimensions to the virtual environment. This technology facilitates user-friendly adjustments for a clearer vision and mitigates discomfort by aligning visual fields with human-like precision [4].

Other notable contributions address security issues that accompany emerging advancements. Zhu *et al.* [5] introduce a new authentication scheme called Soundlock, which addresses security issues such as "shoulder surfing". The program leverages the uniqueness of pupillary response to audio stimuli. The built-in eye-tracking measures pupil dilation while playing a user-selected audio sequence. However, all of this data is continuously collected by the device's integrated sensors. The usage of this data has the potential to transcend enriching the immersive experience.

The question remains: Can more information be extracted if pupillary response is used as a biometric marker?

Recent studies have explored what data can be collected from the various sensors [6]. In particular, eye-tracking data has been a focus of research for over a decade opening up multiple avenues of information extraction. Obtaining personal information has been explored, from gender through iris biometrics [7], to gaze patterns leading to cultural affiliation [8], and native language [9]. Emotion detection was one of the research topics [10–12]. In addition to emotional processes, decisions can also be analyzed through pupillometry [13, 14]. Even health-related indicators can be extracted. Saccadic eye movements provide insights into substance abuse [15, 16], while pupil response to light can be an indicator of Alzheimer's [17] and Parkinson's [18] diseases.

## 1.1 Purpose and Research Questions

This work begins to explore the feasibility of using device-integrated sensors to identify Personally Identifiable Information (PII) that users unknowingly provide, through for example behavioral patterns. PII is defined by the U.S. DEPARTMENT OF LABOR as "any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred by either direct or indirect means"[19]. Direct means include, for example, name or address, while indirect identifiers are a conjunction of elements such as race, age, or gender [19].

Following the main question we designed a preliminary study focused on the extraction of indirect identifiers, which are presented in the following questions:

- Is it possible to infer binary decisions made by the user from eye data?
- Does the eye data indicate whether caffeine has been consumed?
- Is there a difference in eye behavior between males and females?
- Is it possible to extract age?

## 1.2 Methodology Overview

In this thesis, we will examine the extraction of age and gender as well as caffeine consumption. We are also looking into inferring the decision-making process of positive and negative responses. The scope of this work is to design a preliminary study for a possible future user study. To achieve this, several small experiments were conducted to develop the stimuli presentation and to gain a basic understanding of triggers and reactions of the eye.

Following we explore different Machine Learning Algorithm (MLA)s for analyzing the data collected in the preliminary study. This includes the development of a pipeline to clean the raw data and extract morphological and statistical features. Later several MLAs will be tested to determine their proficiency in classifying the data.

## 1.3 Ethics and Data Protection

The Institutional Review Board (IRB) of the Rochester Institute of Technology (RIT) approved our study plan. The principal investigator briefed the participants on the experiment's layout as well as any discomforts they might experience and their right to withdraw. The participants voluntarily agreed to the conditions and gave informed consent. We pseudonymized participants' data for privacy and confidentiality. For further protection, we collected no direct identifiers and encrypted all collected data during cloud storage. The informed consent form was the only link between the participant and the study. After the study's conclusion we securely discarded any identifying information and only kept de-identified data for future experiments.

## 1.4 Limitations

The accuracy of the eye-tracking device and the pupil detection was limited. The pupil detection's confidence fluctuated, along with high blink rates, leading to unusable data (see Section 4.4.3). As the eye pupil diameter naturally fluctuates and the eye shifts, unrelated reactions were also detected (see Section 4.2.1). External interference also affected the data such as lights (see Section 4.2.4) and sounds (see Section 5.2). By conducting our experiment non-verbally, we eliminated the possibility of recording the participant's reaction to their voice.

We also needed to define a general time frame in which the participants thought about their answers or reacted to the stimuli. This affected the amount of data we analyzed for each stimulus. Furthermore, there was no guarantee that the participants were focused on the task. Both resulted in uncorrelated data. This will be discussed in more detail in Section 7.6.4.

We were also dependent on the participant giving a truthful answer, during the experiment and the post-experiment study as well as their consistency. Otherwise, the ground truth for our model would be incorrect. As the study was conducted at RIT and all participants were over 18, the study might not generalize over all cultures, ages, or economic standings. For more information on the participants' demographics, refer to Section 5.4.

## **1.5 Outline**

The following Chapter 2 will provide an understanding of the structure and behavior of an eye focusing on the pupil in Section 2.2. Furthermore, Section 2.3 will cover the basics of machine learning and related techniques used in this thesis. In Chapter 3 related work is discussed, followed by the steps we took to design our experiments in Chapter 4. Chapter 5 defines the layout of the preliminary study. The methodology used to analyze the collected data is described in Chapter 7. Finally, the results will be presented and discussed in Chapter 7 before coming to the conclusion and probable future work in Chapter 8.

## 2 Background

### 2.1 Augmented vs. Virtual Reality

Virtual Reality (VR) and Augmented Reality (AR) are two concepts that describe the seamless transition between reality and virtuality [20]. Scholars have been attempting to define the distinctions between VR and AR for several decades [20]. Benford *et al.* [21] defined four groups. On one hand AR shows virtual objects in a real-world environment allowing for interaction with the object. VR on the other hand presents the user as a virtual avatar that can be controlled within the virtual environment. This control includes a change of viewpoint as well as interaction with other occupants and the virtual objects. The virtual world does not have to be based on the real world [21] but rather tries to replace it for the user, going so far as to imitate sensation [22]. The system's immersiveness is limited by the technology's ability to give sensory feedback [23]. For further information, refer to Benford *et al.* [21] and Milgram and Colquhoun [22].

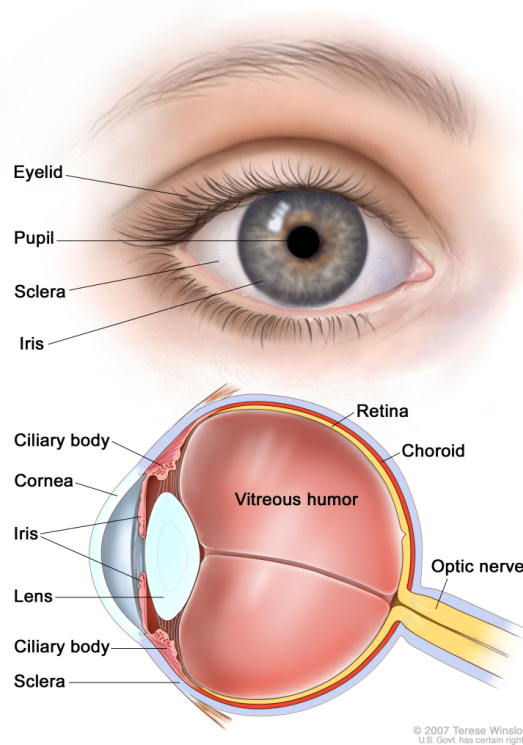
### 2.2 Biological Background

To comprehend the outcomes and facilitate the manipulation of potential triggers, it is essential to first comprehend the biological processes underlying the response. The anatomy in Section 2.2.1 and an understanding of the muscles in Section 2.2.2 will give insight into eye movement. As the primary focus in this thesis is the pupil, Section 2.2.3 delves into the pupillary pathways and briefly outlines possible triggers for pupillary size changes

#### 2.2.1 Anatomy of the Eye

Figure 2.1 depicts the anatomy of the eye depicted with the eyeball situated within the orbital socket [24]. The pupil, located in the center of the front of the eye, allows light to reach the retina. The iris encircles it, and the color of the eye is determined by the iris's pigmentation. Both are covered by the glassy cornea, which transitions into the white sclera. The sclera encapsulates the eyeball in connective tissue, forming the outer wall. It connects to three pairs of muscles responsible for eye movement. These muscles are typically concealed behind the conjunctiva, which folds back from the eyelids.

The inside of the eye is divided into two chambers by the lens [24]. The chamber between the lens and the retina is known as vitreous humor. It is filled with a clear, viscous substance that maintains the shape of the eyeball. The lens is located behind the iris and connects to the ciliary body through fibers [24]. The ciliary body contains circular muscles known as ciliary muscles, which surround the lens and change its shape through contraction [24]. The retina is composed of two groups of light-sensitive photoreceptors responsible for monochromatic vision and color perception [26]. Perceived stimuli are transmitted along the optic nerve. Between the retina and the sclera is the choroid, which is made up mostly of blood vessels. It nourishes the photoreceptors and maintains a stable temperature.

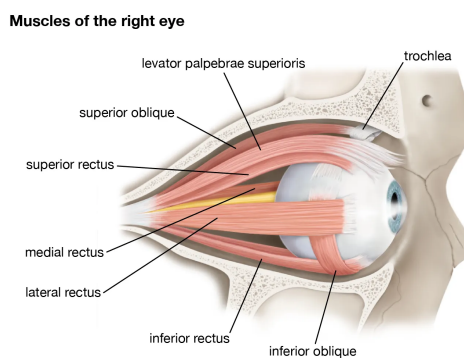


**Figure 2.1:** Anatomy of the eye, from the outside (upper) and horizontal cross-section (lower) [25]

### 2.2.2 Eye Movement

The movement of the eyeball is controlled by three sets of agonist and antagonist muscles arranged in a funnel shape around the optical canal [27]. These muscles can be divided into two groups: rectus muscles and oblique muscles.

The rectus muscles are straight muscles while the oblique muscles are curved [27]. The superior rectus muscle and the inferior rectus muscle run above and below the eyeball respectively, and are responsible for upward (elevation) and downward (depression) movement. The medial rectus muscle is located on the nasal side of the eyeball and the lateral rectus muscle is located on the opposite side. They move the eyeball in their directions, also called adduction and abduction respectively.



© Encyclopædia Britannica, Inc.

**Figure 2.2:** The extraocular muscles of the eye [28]

Movement along the sagittal axis is achieved through the superior oblique muscle and the inferior oblique muscle [27]. These muscles rotate the eyeball inward (intorsion) or outward (extorsion).

There are four distinct types of eye movements that each fulfill a different function [29]. Saccades are swift, ballistic eye movements that quickly shift the point of fixation. Slowly following a moving stimulus is known as smooth pursuit movement. When each eye focuses on targets at different distances it is called vergence movement. Vestibulo-ocular movements mitigate head movements, by keeping the eyes focused on a target. For more detailed information, refer to Purves *et al.* [29].

### 2.2.3 Pupil Dilation

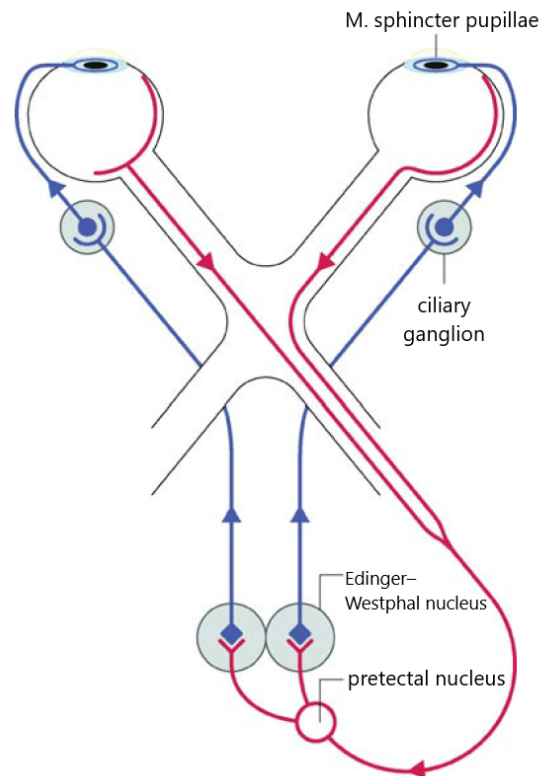
The pupil adjusts the amount of light that enters the retina by changing its size in response to the level of illumination [30]. This is accomplished by the circular (musculus sphincter iridis) and radial (musculus dilator iridis) muscles [31]. The dilation of the pupil is an automatic reaction and therefore cannot be voluntarily influenced. As part of the Automatic nervous system, it has a sympathetic as well as a parasympathetic system which works tonically [31].

The pupillary pathway splits into the efferent and afferent pathways [31]. The efferent pathway, colored blue in Figure 2.3, is the motoric component. It originates from the Edinger-Westphal nucleus and terminates at the muscles. It is interrupted by the ciliary ganglion synapse. Pupil constriction is caused by parasympathetic system stimulation, which constricts the sphincter muscle. Conversely, dilation results from stimulation of the sympathetic system, which constricts the dilator muscle. Although the relaxation of the sphincter muscle occurs because of supranuclear inhibition.

The efferent pathway stimulates the muscles, while the afferent pathway is responsible for sensory input and triggers reflexes such as the pupil light reflex [30]. Figure 2.3 illustrates the similarity between the afferent pathway (shown in red) and the visual pathway, both starting from the retina and following the optic nerve. Afterward, it branches off to the area praetectalis and connects to the Edinger-Westphal nucleus before passing into the efferent pathway.

It is important to note, that the pupil size is not constant. The pupil light reflex is a precise regulation, in addition to being a reflex [31]. The pupil size undergoes a slow (3Hz) oscillation due to recurring corrections. Furthermore, after each blink a small pupil light reflex can be registered. The baseline pupil size varies among individuals, with women generally having larger pupils than men. Another influence on pupil size is age, with newborns in the first year having smaller pupils, due to their weaker dilator muscles. As people age pupil dilation decreases, therefore old people have less dilation in darkness than young people. Additionally, approximately 17% of people have anisocoria, a condition where the pupil size differs between the eyes.





**Figure 2.3:** Pupilar pathways, showing the efferent (blue) and afferent pathways (red) [30]

Besides the pupil light reflex, there is also the pupil near reflex or pupil near response [32]. The pupil constricts when looking at a nearby object while dilating when looking into the distance. Mathôt [32] describes as well, that the depth of field increases as “a small lens [...] suffers less from optical distortions”[32, p. 9]. According to them, this phenomenon is often disregarded in labor settings, where visual stimuli are typically presented in 2D.

Furthermore, pupil size can be influenced by a variety of chemicals and health issues [31]. They can have either a direct influence by reacting with the muscle receptors or an indirect one by for example inhibiting their breakdown or boosting their natural effect [31]. The restriction of the pupil is called miosis and can be caused by opiates, inhaling the smoke of marijuana, and more [33]. Mydriasis is the opposite effect of an enlarged pupil and can be caused by belladonna alkaloids, LSD, and carbon monoxide poisoning. Pupil dilation can occur naturally as an effect of arousal, either through feelings of lust or fright and other moods [31], as well as in response to pain [34].

Additionally, health conditions are another reason for unusual pupil behavior[33]. Hippus, which is the sudden rhythmic change of pupil dilation, can be observed in individuals with conditions such as Multiple sclerosis or an epileptic seizure. However, it is not always connected to an illness. If only one pupil dilates and the sides switch, it is a jumping pupil, which can be the result of toxic influences. There are various illnesses that can affect the pupil's behavior by impacting the nervous system, as well as defects that deform the pupil.

## 2.3 Machine Learning

A variety of machine learning techniques can be used to solve pattern recognition problems. In the context of this thesis, we aim to identify and classify patterns in pupil behavior, to distinguish different PII of the experiment participants. Cross-validation was used to test 5 machine-learning methods and measure their accuracy.

### 2.3.1 Machine Learning Algorithms

MLAs can be divided into two subgroups: supervised and unsupervised, depending on whether the training data is labeled [35]. Supervised learning involves the use of labeled training data and includes machine learning techniques such as classification or regression algorithms. Unsupervised learning involves unlabeled data, using techniques like clustering. The objective of any classification algorithm is to be well-generalized so that new samples can be correctly classified [35].

#### **K-means**

K-means is a clustering algorithm and a form of unsupervised learning. Given a preset number of classes, a random sample is selected for each class, known as a centroid [36]. All other samples are assigned to the nearest centroid. The algorithm then iteratively selects the center of each cluster as the new centroid until the centroid remains unchanged. The algorithm aims to find clusters with a small radius (intra-cluster similarity) while maintaining a high distance from other clusters (inter-cluster similarity).

#### **K-Nearest Neighbors**

The kNN algorithm classifies each sample according to its k nearest neighbors. They are calculated through a distance metric [37], such as the Minkowski distance [36]:

$$Minkowski = \left( \sum_{i=1}^n |x_{2i} - x_{1i}|^p \right)^{\frac{1}{p}}$$

The metric can have different names depending on parameter p. For p = 2, it is referred to as the Euclidean distance. More distance metrics are described in Zhou [36]. A weight can be assigned to the neighbors based on the distance. As a lazy learner, the training samples are simply stored during the training phase. For more detailed information, refer to Zhou [37] Chapter 10 or online Amor and Liu [38].

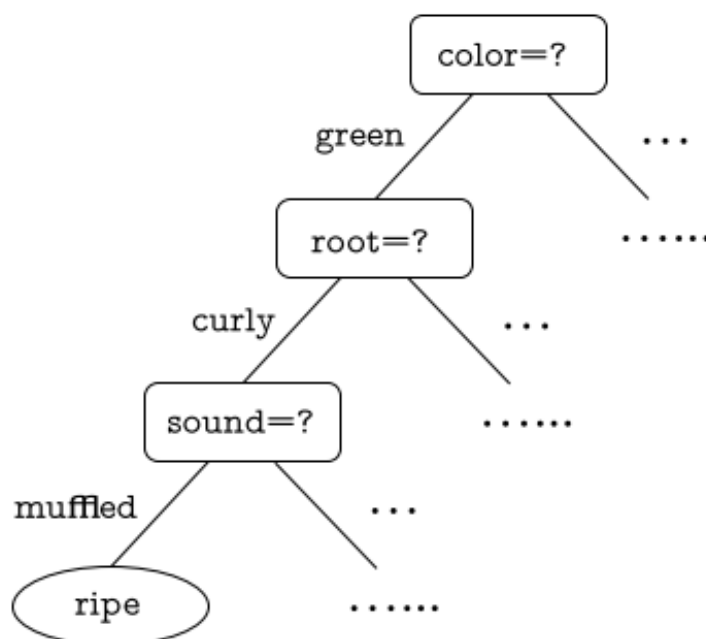
#### **Support Vector Machine**

During the training phase of a Support Vector Machine (SVM), the objective is to construct a hyperplane that can effectively separate two sample classes in the feature space [39]. It is possible to have multiple eligible hyperplanes, but the goal is to select the one with the best generalizing ability. To achieve this, support vectors are chosen. They denote hyperplanes based on sample points on the outer edges closest to the probable hyperplane. The distance between the resulting support vectors is called the margin. Maximizing the margin is a crucial

aspect of optimizing the hyperplane, especially for linearly separable classes. If the classes are not linearly separable, the data can be mapped to a higher dimensional feature space. Generalized if the data “has a finite number of features, then there must exist a higher dimensional feature space in which the samples are linearly separable” [39, p. 135]. The function that maps the features is derived from a kernel function [39]. The quality of the mapped feature space is crucial for the classification and depends on the selected kernel function. Amor and Liu [40] or Zhou [39] can be referred to, for more detailed information.

### Decision Tree

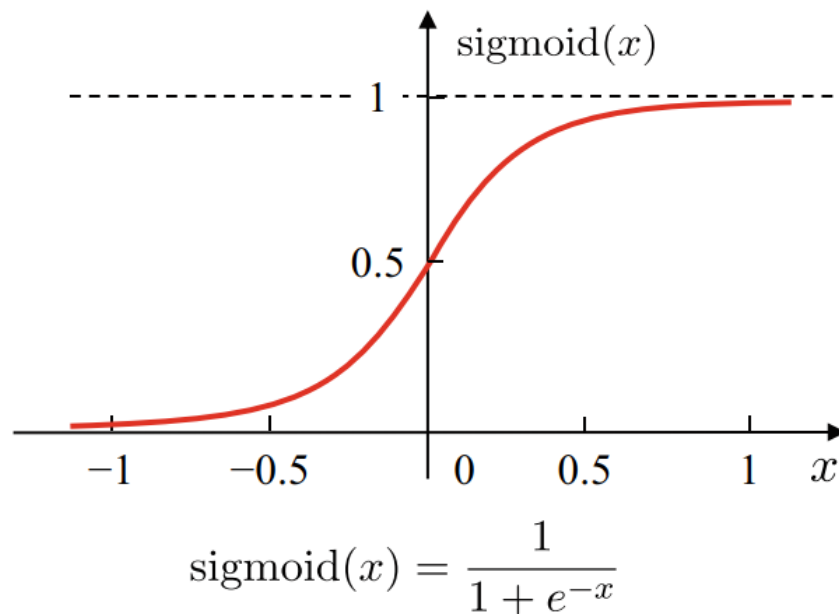
A DT divides a classification problem into sub-decisions that together solve the classification problem [41]. Based on these sub-decisions a binary structure is created [41]. Each split or internal node divides the samples based on a single feature. The end nodes represent a potential classification. Figure 2.4 provides an example based on the classification of watermelons. The square nodes divide the data based on the feature, while the oval shape denotes the classification decision. Here whether the melon is ripe. The complexity of the classification is reflected in the tree depth [42]. Zhou [41] gives a more in-depth explanation, while Amor and Liu [42] explore the implementation and advantages and disadvantages of DTs.



**Figure 2.4:** Example of a DT for classifying ripe watermelons [41]

### Multi-layer Perceptron

The Multi-layer Perceptron (MLP) is a network model composed of artificial neurons and part of unsupervised learning [43]. Similar to a biological neuron, it receives a number of weighted inputs. When the sum of inputs exceeds a threshold, information is passed on. The output is determined by the activation function. One example of this function is the sigmoid function, which is visualized in Figure 2.5. For more functions, see Baheti [44].



**Figure 2.5:** Sigmoid function as graph and formula [43]

A perceptron is created when two input neurons are connected to one output neuron, resulting in a binary classifier [43]. The resulting levels are called layers, referred to as input layer or output layer, depending on their position. The number of neurons in the input layer is determined by the number of features, while the output layer transforms the values of the previous layer into output values [45]. When additional connected neuron layers are added in between, they are called hidden layers, resulting in a MLP [45]. The function assimilated by the network is the basis for classification [45]. This function can be non-linear. For more detailed information, refer to Zhou [43] Chapter 5 or Amor and Liu [45]

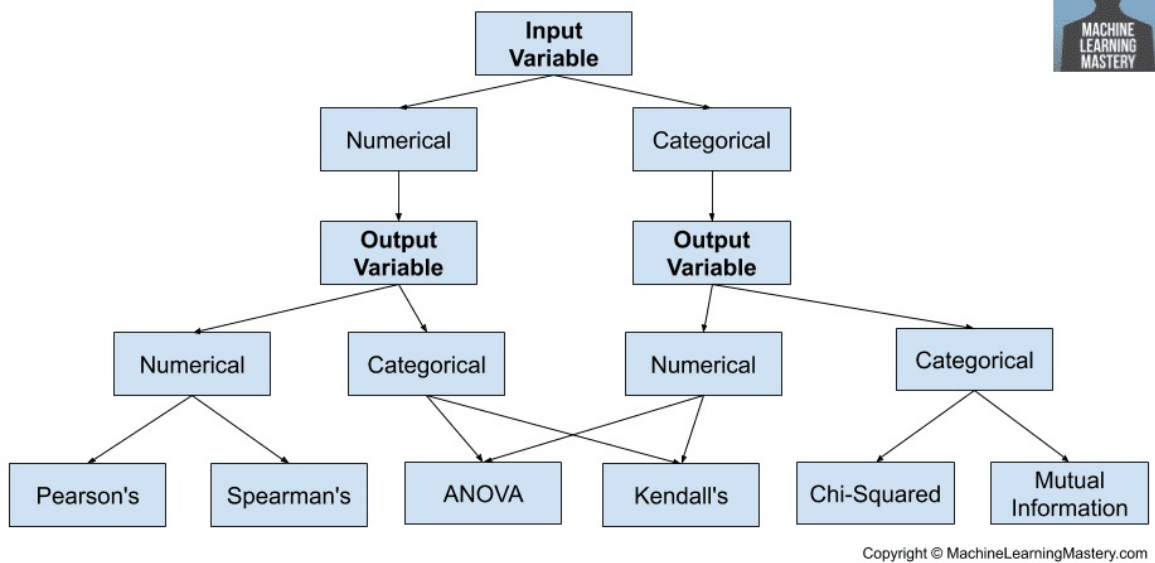
### 2.3.2 Feature Selection

Data samples consist of various attributes, also known as features [46]. However, not all of these features are relevant to the machine-learning process. Therefore, a selection must occur to sort out the irrelevant features. This process is called the feature selection and forms a component of the data preprocessing. It addresses two problems. First, the elimination of features reduces the dimensionality [46]. This approach mitigates the curse of dimensionality, which refers to the issues that arise with high-dimensional data such as data sparsity and problems with distance calculation that all machine learning methods face [37]. Second, it helps the learner to focus on relevant features, increasing the likelihood of uncovering the truth and simplifying the learning process [46].

It is crucial to retain important features, to avoid hindering the learning process [46]. The importance of features depends on the given learning task. Another group of features, which can be derived from others, are known as redundant features [46]. They can be eliminated to reduce workload without losing any additional information. However, if a feature is “an intermediate concept of the learning task” [46, p. 266] keeping the redundant features can be beneficial.

There are different techniques to calculate the optimal feature subset [46]. These techniques can be divided into supervised and unsupervised based on consideration of the prediction goal [47]. Unsupervised methods disregard prediction goals and may remove redundant features based on correlation [47]. Supervised approaches aim to discard irrelevant features and can be categorized as wrapper, intrinsic, or filter [47]. Wrapper methods are used to evaluate the learner's performance in assessing the feature subset [46]. If the feature selection is a part of the learning model, it is referred to as intrinsic [47]. Filter methods may consider statistical relevance but do not take into account the follow-up learner [46]. These techniques are often based on the correlation between input and output [47], Consequently, the chosen method is tied to their data types, as Figure 2.6 depicts.

#### How to Choose a Feature Selection Method



**Figure 2.6:** Process of choosing feature selection techniques in the context of the input and output data [47]

In our prediction problems, samples consist of numerical features that are classified into categorical groups. According to Brownlee [47], the best methods are either the Analysis of Variance (ANOVA) correlation coefficient for linear data or Kendall's rank coefficient for non-linear data. As one of our primary focuses is on classifying positive and negative reactions, we chose to use the ANOVA filter in this work. It should be noted that Kendall's coefficient assumes the ordinality of the output [47], which is not applicable in our case. It may be relevant for classifying PII such as age groups and should be kept for consideration in future work.

ANOVA has several advantages for feature selection [48]. It compares the means of groups by calculating the f-value, which is denoted by the division of the Mean Square Between (MSB) and the Mean Square Within (MSW) [48]. For more detailed information on the calculation, refer to Ding *et al.* [48]. The discriminative capability of a feature is proportional to the f-value [48]. This can be used to rank the features, from which a number of the highest-ranked features are selected for the classification.

### 2.3.3 Cross-Validation

For cross-validation, the dataset is split into  $k$  equal sample groups, while maintaining the original distribution if possible [49]. Then all but one group are used for model training, while the remaining one is reserved for testing. This process is repeated  $k$ -times and the average of all trials is used as the evaluation result. For more information, refer to Zhou [49] Chapter 2 or online Amor and Liu [50] for instructions concerning implementation.

### 2.3.4 Performance Measurement

In this work, we primarily deal with binary classification problems. We use a confusion matrix (see Table 2.1) to visualize the distribution of correctly and incorrectly classified samples [49]. Binary classification generally defines two classes, one as positive and the other as negative. The predicted classes are represented on the horizontal axis, while the correct classes are on the vertical axis. Correctly classified samples are counted as the True Negative (TN) or True Positive (TP) respectively. Misclassified samples are either wrongly in the positive class and therefore False Positive (FP) or inversely the False Negative (FN). The total number of samples can be derived from the sum of all four cases.

**Table 2.1:** Confusion Matrix

|              |          | Predicted Class     |                     |
|--------------|----------|---------------------|---------------------|
|              |          | Positive            | Negative            |
| Actual Class | Positive | True Positive (TP)  | False Negative (FN) |
|              | Negative | False Positive (FP) | True Negative (TN)  |

Accuracy is used to compare the performance of MLAs. It is calculated by dividing the number of accurate predictions the model made by the total number of predictions, using the formula [51]:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Other measurements we use to evaluate the quality of the classification are Precision (P) and Recall (R) [51]. P represents the ratio of TP overall positive classified samples, therefore quantifying the accuracy of positive prediction. R, also known as sensitivity or true positive rate,

---

measures the ability to correctly classify positive samples, by calculating the ratio of TP over all positive samples. Shah [51] and Zhou [49] provide additional examples of performance measurements

## 3 Related Work

### 3.1 Authentication Schemes for VR Systems

New security features are being developed for VR/ AR devices. An important topic is the authentication process, as conventional methods are not practical, due to deficits in their usability such as the difficulties in entering passwords using controllers [52]. Another reason is their vulnerability to attacks like shoulder surfing, due to the user's blocked visual field [52]. Successful attacks on deriving virtual keyboard inputs through hand gestures [53] or head movements [54] have been reported. The use of a wide variety of sensors presents new possibilities for behavioral or head-based biometrics, while the handling of the device leads to new ways for password input [55]. It is also conceivable to combine both.

Stephenson *et al.* [55] systemized and evaluated currently employed authentication schemes and lists new possibilities. Both tested devices required a password-based device login, either on an external PC or on the device using a hologram keyboard. After installation, several apps required additional authentication. The authentication methods in their research fell into three groups: knowledge-based, biometric, and token-based. The incumbent knowledge-based methods were found to be vulnerable to observational attacks or guessing. In contrast, token-based methods, such as those implemented through a QR code, proved to be resilient to both. The same level of resilience was observed for the iris scan, a biometric method. While the collection of proposed works on this topic is extensive, they exceed the scope of this work. Therefore, this thesis focuses on eye-related authentication schemes, however, there are other notable mechanisms. The virtual environment allows for 3-dimensional passwords and pins [56]. Physical biometrics can be extended to include ear geometrics [57] and brain responses as a password [58]. Behavioral biometrics include head movements in response to auditory stimuli [59] and various continuous systems based for example on walking [60]. A proposed combination uses the presented behavioral patterns during password entry, as additional security [61].

In 2020 Zhu *et al.* [52] introduced a novel system for authentication on VR devices called BlinkKey. They developed a two-factor scheme for integrated eye trackers, based on unique blink rhythms and pupillary size changes. The system has two phases: enrollment and login. During the enrollment, the user creates a template called 'blinke', which must be presented multiple times for training. This template consists of a chosen blinking pattern. The first factor is the blink pattern, made up of blink on- and offsets, their interval, and the time between blinks. The second factor involves additional features extracted during the time between features, such as changes in pupil dilation related to the light pupil reflex. As classification methods SVM and kNN performed best due to the small sample size. The kNN reached an Equal-Error-Rate of under 4% with only six samples, making it a promising technique.

Last year, Zhu *et al.* [5] introduced another new authentication process called Soundlock. Unlike their previous method that based the authentication on reproducible blink rhythms, Soundlock focuses on the biometric attribute of natural pupil dilation changes in response



to auditory stimuli. This method proved viable because participants exhibited consistent pupillary responses to the same sound, whereas different users displayed varied responses. The primary advantage lies in the cancelability of the credentials, meaning their ability to be easily changed by selecting different stimuli. This is in contrast to conventional biometrics, such as fingerprints, which are less variable. The program layout is similar to the previous example, with kNN persisting as the classification system with the highest performance, but the features have been altered. They recorded the pupil's reaction to the stimulus and split it into two phases. The excitation phase occurs when a stimulus is present, and the recovery phase follows its offset. Aside from statistical features, several morphological features were calculated, based on the graph resulting from the pupil size changes. Those include for example peak and valley positions and distance as well as the coefficient resulting from curve fitting and pupil size baseline. Other measurements are temporal like the response lag and the needed recovery time. During feature selection, it became apparent that morphological features have a greater impact on classification accuracy.

## 3.2 Attacks on VR Systems

Zhu *et al.* [52] tested the robustness of BlinKey against four different attacks. One was the zero-effort attack, in which participants were given five attempts to guess the pattern without prior knowledge. Attacks on blinkeys consisting of six blinks or less have a low probability of success, while the success rate for more blinks is zero. Statistical attacks assume access to prior user blinkeys to calculate the feature distribution. The most probable of which is used to devise the current blinkey. As the attacker still has to use their own eyes during the authentication, the biometric pupil light response makes the success rate close to zero in the longer blink rhythms [52]. If the attacker knows the credentials, it is called a credential-aware attack. Although the biometric factor still applies, the success rate increases substantially to 14.2% for a blinkey with a length of seven, which shows that short extracts of biometric reflex patterns are not secure enough [52]. Visual observation while seemingly impossible due to the VR headset, is not entirely foolproof, as it has a non-zero success rate. This is due to the ability to guess the blink rhythm and outward signs like unconscious nods while blinking [52].

All of these attacks assume the use of a natural human eye during the login process, and therefore, physical access to the device [52]. This incorporates genuine responses in the authentication mechanism. Zhu *et al.* [5] demonstrated that Soundlock is resistant to this kind of attack with a mere attack success rate of 0.76%. It is not possible to externally record the pupil reaction for use in a relay attack. The previous installation of malware could allow for the injection of internal recordings. However, the timing required for the call response type of authentication provides some protection. Consequently, extracting other eye-related recordings could happen as well. Replication, as observed with other biometrics like fingerprints, is currently not considered feasible [5].

Giaretta [62] highlights that in addition to attacks on the authentication scheme, generic attacks such as Man-in-the-Middle or Denial-of-Service remain a possible threat. Furthermore, side-channel attacks can be used to derive gestures and voice commands [63]. However, there are also specific attacks on Virtual Reality. Casey *et al.* [64] introduce four immersive attacks on OpenVR, which target the virtual environment: the Chaperon, the Disorientation

attack, the Human Joystick, and the Overlay attack. The Chaperon attack manipulates the set virtual boundaries that prevent users from bumping into real-world objects. This could lead to injuries, and therefore compromise the user's safety. The Disorientation attack aims to cause dizziness and confusion in the user by moving the virtual environment without physical movement. Similarly, the Human Joystick manipulates movement in the virtual world; here the goal is to guide the user's avatar to a specific location. Similarly to the Disorientation attack, the Human Joystick can also cause confusion and dizziness. An Overlay attack obstructs the user's virtual visual field by superimposing unwanted content that cannot be removed. Casey *et al.* [64] investigated the option to export the headset's outward camera stream. They pointed out that this combined with information extracted from the system, provides a wealth of data on the environment as well as the user's behavior.

### 3.3 Privacy Concerns in VR

In 2018, Adams *et al.* [65] conducted a study on the perception of security and privacy on VR devices. During their interviews, developers mentioned concerns but did not consider them relevant for their devices. The raised concerns included the collection of data without the users' knowledge, either while the device is seemingly turned off or through the embedded sensors, for example, headset cameras could divulge the location, and the microphone could record happenings in the room. The developers who were questioned mostly disregarded the possibility of privacy issues, since none have arisen so far. A way to raise awareness and mitigate these concerns, one potential solution could be the integration of permission requests for sensors before deployment, as has already been done with phones. There is a request for guidelines concerning privacy and security standards [65, 66].

Stephenson *et al.* [55] state that all biometric-based authentication schemes raise concerns for the protection of user privacy. When deploying these methods sensitive data is stored on the device, which could lead to misuse. The storage location and whether the biometric is cancelable are important factors to consider. However, there are additional privacy concerns. Three primary concerns were defined by Giaretta [62]. The first is informational privacy, which encompasses legal data vulnerability resulting from legal holes or unclear privacy policies. This is dependent on the platform and state laws and is therefore not part of this work. The second concern arises, due to social networks. And lastly, privacy issues related to the physical world.

As in any kind of social networking, virtual reality-based ones also raise concerns for user privacy [62]. As of 2018, user privacy was not an issue in the VR community, which was perceived as exclusive [65]. The users felt secure in their virtual communities and saw them as a safe space. However, with the growth of the user base, this issue may become more relevant. In 2016, Nwaneri [67] highlighted Facebook's history of conducting experiments on its users. As a result of the immersive experience, more personal data could be collected such as eye movement, or hand and head movement. This sentiment is shared by the developers and users in the interviews conducted by Adams *et al.* [65] in 2018. The developers pointed out that possible profits are prioritized over the consequences. The developers even

accuse Facebook of investing in VR technology mainly for the gained data. This sentiment is concerning for users [66], especially considering planned developments such as the tracking of facial expressions [3] that will provide even more usable data.

Physical privacy can take various forms. Buck and Bodenheimer [68] point out that our need for security through personal space is a possible privacy concern. When interacting with virtual reality, a user's preferences can be revealed through their behavior. Personal space is defined by the authors as "the minimum distance that one feels comfortable interacting with an object or person", and this distance can consequently reveal a user's bias. Like Stephenson *et al.* [55], Giaretta [62] also defines leakage of biometric data as a problem. Physical issues related to VR devices include problems that arise from surrounding people [69]. The VR device obstructs the user's real-world view, potentially allowing people to hide from the user [69]. This could lead to possible stalking or recording of the user, which could be used in attacks on the VR device or on the person [62, 66]. Furthermore, collisions with other people pose a health risk [69].

Giaretta [62] highlighted integrated eye-tracking technologies as a significant privacy concern due to the possibility of extracting "emotional information, without explicit knowledge or consent from the users" [62, p. 4]. The following section will delve deeper into how eye-tracking data can be utilized, as well as other sensors that can be exploited for user identification. In 2023, Jarin *et al.* [70] tested the feasibility of identifying a user by their body motion, facial expression, eye gaze, and hand joints. The study found that body motion and facial expression perfectly identify an app user. The used data in the study was collected through developer APIs, which makes these APIs a significant privacy threat. Adams *et al.* [65] also found that users are concerned about the threat posed by sensors such as microphones and cameras. The users are aware of the always-on function and the data that may be collected to some extent. However, not all users share their concerns at the current time, this could change in the future. Concerns were also raised regarding impersonation through voiceprints or manipulation based on brainwaves [66].

### 3.4 Eye Measurements

Research has been conducted on data from eye trackers and extractable private information over the last decade. As indicated in the first two Sections any data that can be used for authentication has the potential to lead to more extractable information. Kröger *et al.* [71] created an overview (see Figure 3.1) of collectable attributes and possible private information that could be derived from them.

One major research question revolves around the detection of mood and emotion. Pupil dilation triggered by emotional arousal was investigated in different studies. Gingras *et al.* [72] examined pupil reactions activated by short music excerpts, which were rated for their subjective arousal and other factors. These ratings, along with the role of music in their lives and gender, were used to predict the pupil response. However, some predictions did not hold such as the expectation that females would have a stronger response than males. The researchers discovered that emotional arousal is independent of pleasantness and that the response is not related to music amplitude. This study does not address questions of

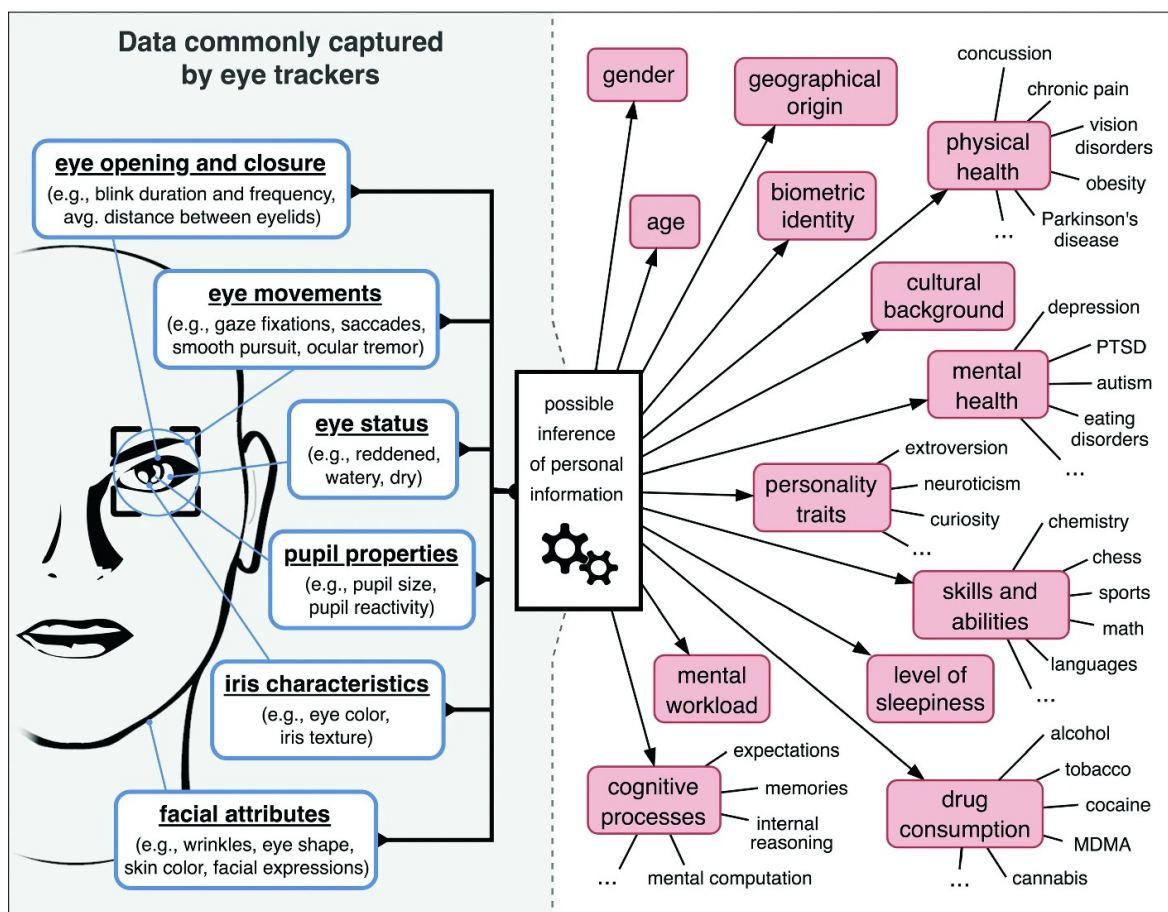


Figure 3.1: Overview of collectable eye data and the correlated PII [71]

pupillary response related to unpleasant sounds and hormone balance. Another study has found a correlation between the music type and the strength of the response [73]. Zekveld *et al.* [73] describe that “largest pupil sizes are observed for aversive stimuli” [73, p. 13], but the opposite was also observed. Babiker *et al.* [12] attempted to differentiate between positive and negative emotions and discovered a greater dilation in correlation to negative stimuli. The studies seem to agree that while neutral responses exist, emotional responses are more significant.

In addition to studying sound activation, researchers also examined the emotional reaction of pupils to visual stimuli. They found that the emotional valence is not related to pupil size [11], contrary to the description by Zekveld *et al.* [73]. Another study was conducted to assess the influence of viewing changes when presenting emotional pictures [74]. The first experiment studied the influence of the viewing duration, and no correlation was found. However, the peak constriction was observed at 800ms. The second experiment did not observe a lesser reaction after repeated viewing. The last experiment determined that active naming of the observed emotion does not change the pupil response. Bradley *et al.* [11] and Snowden *et al.* [74] confirmed that pupillary changes were smaller for neutral pictures compared to emotional pictures.

In order to infer emotional reactions, researchers have studied various movements that can be detected by VR devices, such as head movements or gaze positions, in addition to pupillary reactions. Behnke *et al.* [75] discovered that humans tend to avoid negative stimuli. However, when presented with positive stimuli participants moved towards the stimuli and were more active. Meanwhile, differentiating this from the behavior to neutral stimuli was not possible. The avoidance is consistent with the study by Huijding *et al.* [10] on eye movements in arachnophobia. Participants with arachnophobia avoided looking at a presented spider for a longer duration than those without the fear. Similar behavior was observed in a study conducted by Buck and Bodenheimer [68].

Furthermore, emotional pupil response gaze patterns can also be utilized to detect personality traits, as demonstrated by Berkovsky *et al.* [76] through natural responses to various stimuli like emotional pictures and videos. They defined 16 traits based on three different models, while another study focused on the Big-5 Model [77]. Berkovsky *et al.* [76] focused on features such as blink rate, data derived from saccadic movements as well as fixation. These features were also identified as crucial in the personality study by Hoppe *et al.* [77]. Berkovsky *et al.* [76] discovered that the Naïve Bayes had the best accuracy across all traits, although other methods can reach similar or better results for specific traits. More importantly, almost all traits could be detected with an accuracy of over 80%, which is consistent with the earlier study [77]. One conclusion reached was that video stimuli result in better classification accuracy, but a combination of both image and video stimuli is even more effective [76].

Additionally, Gee *et al.* [13] studied pupillary changes during cognitive processes. They used pupil dilation to determine differences between two options presented to participants in their study, where they had to decide whether a signal was present or not. Approximately 30 participants were presented with noise and uniformly distributed signals at intervals. They had to press a button to indicate their yes or no answer. The study discovered that pupil dilations differ depending on the given response.

Eye movement can even reveal cultural affiliation. The Own-Race bias effect is already known and manifests in different gaze patterns and fixations as well as pupil diameter [8]. A larger diameter was observed when looking at other race faces, indicating a higher cognitive process [78]. Fixation on presented cultural habits can also indicate cultural affiliation [79]. Ito *et al.* [80] found that a person's native language can be determined from their fixation patterns. When given a clear context, participants are more like to fixate on a corresponding word in their native language. Others used fixation patterns on text to assess the skill level in a language [81].

Since eye-tracking tools can also cover the area surrounding the eye, analysis of the wrinkles enables age detection [71]. Another way to derive age through eye measurements is pupil oscillation, as according to Alexandridis [31] the frequency decreases with age. Additionally, the pupil dilation of older people in darkness is reduced to what younger people exhibit [31]. Other researchers use geometric features of the iris to predict broader age groups [7]. Blink rate can also be used as an age indicator because a higher blink frequency was observed in older women [82].

Blink rates can also be used to classify gender. According to Sforza *et al.* [82], women blink on average 19 times per minute, which is higher than the blink rate of men. Fixation patterns have also been explored as a method for gender classification through eye-tracking in various settings [83, 84]. A study by Mercer Moss *et al.* [83], discovered that women tend to have a more explorative pattern while keeping their focus not directly on high information points. Contrary, they found more and briefer eye movements in the pattern of male participants. Differences in viewing patterns were also observed during online shopping [84]. Eye-tracking has also been used to investigate homophobia [85] and pedophilia in forensic settings [86]. Meanwhile, pupil dilation can be an indicator of sexual orientation when presented with sexual stimuli [87].

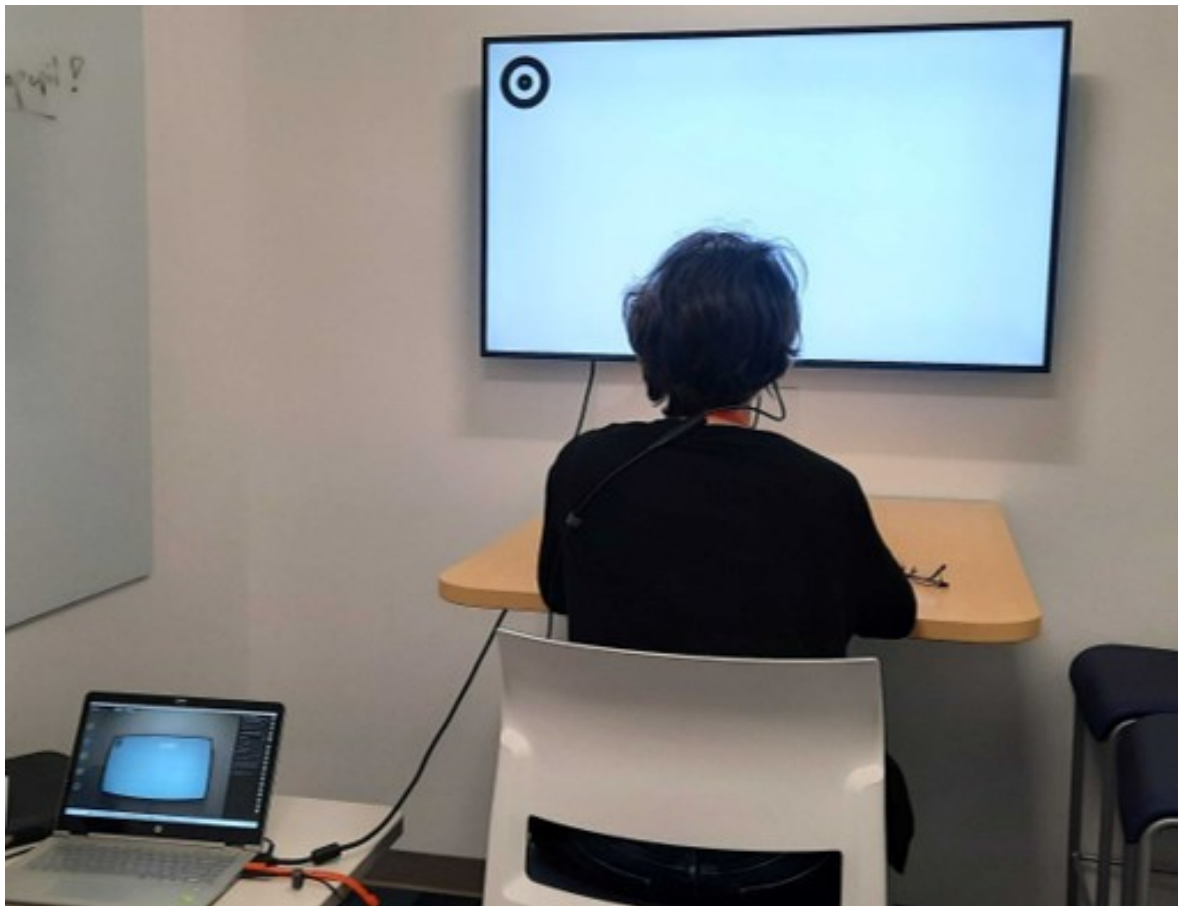
A different field of research is the relation of pupil dilation to health issues such as drug consumption or neuronal diseases and pain response. A study conducted by Jain *et al.* [88] investigated the correlation between pupillary unrest and Parkinson's disease. The researchers found that sleepiness, a common side effect of Parkinson's disease, is associated with pupillary unrest. Additionally, they discovered a correlation between sleep rating and pupil changes that increase with more Parkinsonian motor signs. These findings suggest that pupillary unrest may serve as a marker for disease progression. The pupil light reflex has been identified as an indicator of Alzheimer's [89]. Additionally, it was observed that pain has a dilating effect on the pupil [90]. In the field of drug dependencies, Dias *et al.* [16] studied the attentional bias of cocaine-dependent individuals towards drug-related cues. They found a possible indicator for relapse in saccadic eye movements when presented with drug-related stimuli. Another drug that was studied for its effects on the pupil is caffeine [91]. The study indicates that caffeine consumption leads to an increase in pupil size. Additionally, it was found that caffeine consumption reduces pupillary unrest [92]. In Section 2.2.3 several chemicals as well as health issues were mentioned that can result in atypical behavior. Detecting such behavior can help identify the underlying cause.

## 4 Experiment Design

We developed our methodology through a series of experiments while observing our participants' reactions. Conclusions derived from the results of the data, collected during these experiments as well as interacting and communicating with the participants about their experience, helped us to adjust the experiments.

### 4.1 Original Experiment Setup

The experiments were held in a separate room, to minimize outside disturbances. The participant sat in a high chair in front of a table and a wall-mounted monitor, as shown in Figure 4.1. The brightness was controlled through the monitor after turning off the light. Real-time data was recorded using the Pupil Core, developed by Pupil Labs [93], as our eye-tracking device. For calibration and annotations, we used the built-in functions of the recording program Pupil Capture, refer to Section 5.1 for more information. The laptop used for the data collection was positioned lower and outside of the participant's visual field.



**Figure 4.1:** The original setup of the experiments before improvements, for better visibility taken in a lit room

## 4.2 First Line of Experiments - Protocol 1

The experiments began with the collection of demographic information from the participants. To ensure anonymity, each participant was assigned an ID, consisting of the prefix 'ID', a five-digit sequence of random letters, and a revision number indicating how often they participated in previous experiments. The other collected data were age, gender/ sex, height, and weight as well as whether coffee/ caffeine was consumed that day.

After calibrating the device, we held a 30-second break in complete darkness to allow the pupil to expand and stabilize. This enabled us to measure the baseline of the pupil size. We then conducted the experiments with a 3-minute break between each one. Each experiment was conducted twice. Finally, we discussed the participant's experiences and conducted a short survey regarding their comfort. In our initial series of experiments, we had three objectives:

- To determine the time required to stabilize the pupil
- To identify the best level of brightness, through the background color
- To discover the most comfortable focus point

To achieve each objective, we designed a different experiment. All experiments consist of a video presenting the visual and audio stimuli. All designed experiment videos can be found in the Digital Appendix, refer to Table B.1 in Appendix B for more information.

### 4.2.1 Experiment 1 - Stabilizing the Pupil Size

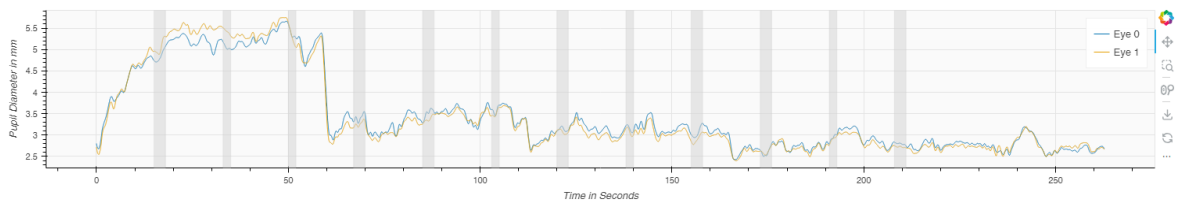
In the first experiment, we tried to determine the time it takes for the pupil to stabilize again after the participant is presented with a stimulus. Our goal was to find a balance between allowing the pupil enough time to return to its baseline, therefore ensuring that the reactions to different stimuli do not influence each other, and keeping the participant's focus on the experiment. The experiment starts with a 15-second pause before the first stimulus is presented. Three different sounds were chosen as stimuli: one high-frequency sine curve, and two human non-verbal sounds, crying and laughter. After each sound, there is a pause. The pause is shortened by five seconds for each repeat of the three sounds. The first pause is 30 seconds long and the last one is five seconds. The overall 5-minute-long video has a completely dark background. Throughout the experiment, all outside interruptions are avoided.

We observed that after the stimulus, it takes less than 10 seconds for the pupil of any participant to return to the baseline. Also, the pupil is never completely stable, exhibiting slight fluctuations in size, known as pupil unrest (see Section 2.2.3 for more information). Therefore, a 10-second pause between the stimuli was chosen for future experiments. According to the participants, waiting for 30 seconds was boring, causing them to lose focus on the task. The experiment yielded measurements of pupil size changes that were unrelated to the stimuli. Consequently, we concluded that five minutes was too long to maintain the participants' focus. This was evidenced by an increase in the fluctuations of pupil size after three minutes. As a result, we decided to limit the length of the individual experiments to two and a half minutes.



## 4.2.2 Experiment 2 - Illumination

This experiment aimed to identify, which background yields the best pupil data. Our focus lay on finding a background illumination that resulted in the most stable pupil dilation. Grayscale was used to ensure a neutral background. Four different levels of illumination were chosen, ranging from the darkest shade of 0 to the brightest shade of 90 based on their respective RGB values. For each of the four grayscale shades, we presented three distinct stimuli, as in Experiment 1. The overall layout followed that of the first experiment, starting with a 30-second break and then presenting stimuli with 15-second breaks in between. The full experiment video was about four minutes long.



**Figure 4.2:** Pupil size change during Experiment 2

Figure 4.2 illustrates an example of the collected pupil size data during the experiment. The presented data has already been cleaned. Each increase in brightness, resulted in the pupil dilation stabilizing further, and the most stable pupil size was achieved in the lightest environment. Although the general stabilization was more evident in the lightest background, the reaction to the presented stimuli also decreased. In the brightest environment, it was difficult to distinguish visually between a state of arousal and pupil unrest. In the lowest illumination, the data is less noisy than in brighter environments. Participants also reported that a lower brightness is more comfortable for the duration of the experiment. Additionally, the results showed a strong reaction to light switches because of the pupil light reflex. The reaction to the changes in illumination is stronger than to any of the presented stimuli, supporting our prerogative of preventing outside distractions from light.

## 4.2.3 Experiment 3 - Focus Point

After the first experiment, participants reported difficulty maintaining their attention on the stimuli and keeping their eyes focused on the monitor, particularly when no stimulus was present. In anticipation of this issue, we designed a third experiment, in which we gave the participant a focus point on which they could reset their gaze. We tried six shades of colors, each in combination with the different backgrounds introduced in the previous experiment. As colors, we selected green, blue, and yellow each in a vibrant and pale version. The colors' exact RGB values are listed in Table 4.1.

These colors were chosen to provide comfort to the participants' eyes and help them concentrate on the stimuli. In this experiment, we reduced the starting pause to ten seconds. Each focus point was presented for 10 seconds in 5-second intervals, resulting in a 6-minute experiment video. After running through the experiment we asked the participants three questions:

**Table 4.1:** RGB values of the selected colors

| Color         | Vivid     | Pale        |
|---------------|-----------|-------------|
| <b>Green</b>  | 0,190,25  | 143,225,124 |
| <b>Blue</b>   | 0,137,191 | 124,197,225 |
| <b>Yellow</b> | 255,210,0 | 255,229,131 |

- Which of the colors did they find the most pleasant?
- Which color helped them concentrate the most?
- Which background focus combination was most comfortable?

From the participants' reports, we gathered disparate answers, with some finding that the color switch, while initially interesting became boring over time. The consensus was that the vibrant green color provides comfort and has a good contrast against a dark background, helping to keep the gaze focused.

#### 4.2.4 Conclusion of Experiments 1-3

During the experiment series, four issues were encountered. Firstly, it became apparent that the current seating position is uncomfortable for smaller individuals. Furthermore, the whiteboard mounted on the sidewall of the room reflected the monitor's light, disrupting the light-controlled environment and therefore negatively influencing the collected data. As a result, the experiment setup was reworked.

Another issue was the illumination and the focus point. Although the experiments narrowed down the brightness to one of the two darker shades of gray, a precise decision could not be made. The same conclusion was drawn for the focus point, as it was narrowed down to a shade of green, but it was unclear which shade would be the most beneficial for future experiments. Therefore, a follow-up experiment was designed to clarify which combination of both should be used later on.

### 4.3 New Experiment Setup and Follow-up Experiment

During the design of the new experiment setup and the follow-up experiment for the video layout, we identified several oversights in the experiment outline. We addressed these oversights accordingly.

#### 4.3.1 Experiment Setup

For the new experiment setup, we opted for a lower table and placed it in front of a white wall so that the participant faces away from the wall-mounted whiteboard. Contrary to the first setup we selected a desktop screen to present the experiments on instead of the wall-mounted monitor. Additionally, we replaced the high chair with a more comfortable office chair. To ensure uniformity in the participants' position relative to the screen, we standardized

the chair's placement. To memorize its position, we marked the floor where the back wheels should be placed. We decided to place the conducting researcher beside the participant, separating their position through a screen to prevent the light emitted by the researcher's laptop from interfering with data collection by triggering the pupil light reaction.

We acknowledged our oversight in standardizing the experiments. To address this we designated a specific headset for future experiments and maintained a consistent volume. During the post-experiment discussions, it became evident that individuals who wear glasses, experience a stronger strain on their eyes. This could be related to the pupil near response, described in Section 2.2.3. This caused us to add a question to the demographic survey asking participants if they wear glasses, and if so, what type and whether they are nearsighted or farsighted.

### 4.3.2 Experiment 4 - Defining the Layout

The objective of this experiment was to determine the optimal video layout. To achieve this, we conducted four variations of the experiment, each combining one of the green shades with one of the brightness levels (0,30) identified in Experiments 2 and 3. The experiment included four different sounds, which were repeated thrice. In addition to the sound we used in the earlier trials (high pitch, laughter, crying), a sensual moaning sound was added. Based on the description of Leydhecker [34], pupil dilation can be a result of arousal which led us to the belief that we could elicit a stronger reaction. We kept the 30-second break at the beginning, allowing the pupil to adjust to the brightness. The sounds were presented in 10-second intervals. As the sounds are shorter, the leftover time presents the break in between to let the pupil reset. We shortened the length to two and a half minutes, as discussed earlier. In addition, we incorporated low-volume white noise into the audio presentation, as a study showed it aids individuals to concentrate [94].

The results indicated that the pupil size stabilizes best when using a black background with a vivid green focus point. According to the participants' feedback, and supported by the resulting curves, the incorporated sound helps to focus on the given stimuli. Another insight of this experiment was, that the strongest pupil reaction occurs in conjunction with sensual moaning.

We decided to add this experiment to the final lineup for the user study. Multiple versions of this experiment were created for use in the study. To ensure a broader range of reactions, we used six sounds, each five seconds long, repeated twice throughout the video. For a more detailed description, please refer to Section 5.2.2.

## 4.4 Second Line of Experiments - Protocol 2

In addition to general demographic information like gender and age, we wanted to find a way to discover more in-depth information such as opinions and political standpoints. To achieve this, we used a variation of the experiments conducted by Gee *et al.* [13]. For our experiments, participants were presented with a series of questions designed to elicit binary responses.

Like in the previous protocol, the experiment series began with a demographic form. The questionnaire remained unchanged and an ID was assigned to the participants again. In the case of participants, who took part in the previous protocol, they were assigned the same ID with an incremented revision number. Based on the findings of the first protocol we kept the length of all experiments limited to two and a half minutes. A 3-minute break was given between each experiment. After the experiments, the participants were given a questionnaire that included all presented questions. Their answers were used as the ground truth in our experiments.

#### **4.4.1 Experiment 5 - Differentiating Yes/No Answers**

The objective was to determine whether there was a difference between the pupil's reaction depending on the answer given by a participant to questions with binary Yes or No answers.

No changes were made to the layout of the experiment or the setup. Indifference to the sounds used in the previous experiments, the audio stimuli consisted of verbally stated questions. In addition, we gave a visual presentation as part of the video. To maintain consistent illumination, the focus point was removed while the question text was displayed. Furthermore, the brightness of the text was adjusted to match the focus point. The experiment video included twelve questions, which were neither personal nor thought-provoking. We also refrained from using controversial topics in this version.

While conducting this experiment, it was repeated three times with a different set of instructions given to each participant. During the first round, participants were instructed to think about their answers without verbalizing them, simulating a typical use case of a VR headset where the user is alone. The second round required participants to verbally state their answers to determine if there was a difference between verbal and non-verbal responses. In the third round, we tested whether there is a possibility of detecting if the participant does not answer truthfully. Here the participant was asked to answer the opposite of what they perceived as the truth.

#### **4.4.2 Experiment 6 - Differentiating Agreement and Disagreement**

In this experiment, the focus was on discerning binary answers, without predefined them as Yes and No. We wanted to figure out whether there is a difference to Experiment 5 when a general agreement or disagreement is given.

Instead of questions, participants were presented with statements and asked to indicate agreement or disagreement. The layout and design were the same as in Experiment 5. There were also variations in the conducting of the experiment identical to those in Experiment 5.

#### **4.4.3 Conclusion of Experiments 5 and 6**

The experiment series yielded inconclusive findings. The use of MLAs to classify the data led us to deduce that we did not have a sufficient amount of data. Therefore, we decided to increase the number of experiments. We extended both experiments to three sets, each

encapsulating twelve questions. Furthermore, we discarded verbal answers. As we set our use case to single users who have no reason to speak we concluded that this variation would not produce findings that align with our experiment goal.

We hypothesized that a limited reaction could be a contributing factor to our lack of findings. According to Alexandridis [33], arousal is one reason for a natural pupil reaction. To elicit a higher emotional involvement from the participants, we added questions and statements on controversial topics. We assumed that these would provoke stronger reactions than the neutral questions.

Another issue we encountered was participants' understanding of the questions. According to several participants, who normally wore glasses, comprehending the verbal questions without the visual representation was difficult. To address this, we tried slowing down the voice. However, as we used a text-to-speech model, an artificial slowing of the voice resulted in a robotic tone which made understanding the question harder. As a solution, we decided to be more selective in choosing the questions for the experiments. Questions that were not articulated well were discarded. Furthermore, we created new questions that were less complicated in content and phrasing.

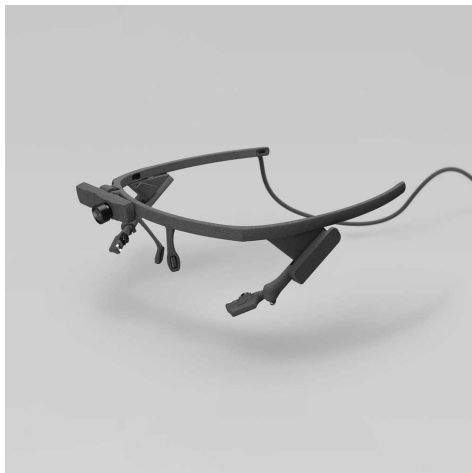
In these defining experiments, we also discovered software problems, rendering some data unusable. For some participants, the software had difficulties in identifying the pupil in the image. The resulting data cannot be co-related to the stimuli with certainty.

## 5 Methodology of Data Collection

The accumulation of our findings from the previous experiments resulted in the design of our preliminary study. We conducted single in-person experiments with twelve participants in a controlled environment. Each experiment was embedded in a 2.5-minute video shown on a monitor. The duration of the entire session was approximately one hour.

### 5.1 Data Collection

We used the Pupil Core, developed by Pupil Labs [93], as our eye-tracking device to record real-time data (see Figure 5.1a). It comes with three accompanying programs. The recording program, Pupil Capture, has built-in functions to calibrate the device as well as make annotations. Pupil Player [95] extracts various measurements for the 2D and 3D representation of the eye collected with Pupil Capture, along with blink count and gaze position, in addition to timestamps and methods. These measurements are presented in a separate table for each recording.



(a) The eye-tracking device: Pupil Core [93]



(b) Eye-trackin device's manner of wear

**Figure 5.1:** Device appearance and its manner of wear.

The eye-tracking device is worn like a pair of glasses. Figure 5.1b illustrates the manner of wear.

### 5.2 Experiment Design

The experiment was set up in an isolated room to minimize external interference. However, some disturbances were still present since soundproofing was not possible.

The participant sat in front of a monitor as depicted in Figure 5.2, with the chair position relative to the monitor being the same for all participants. During the experiment, brightness was controlled through the monitor, using it as the main light source. The only other



**Figure 5.2:** The current setup of the experiments, for better visibility taken in a lit room

light source was the monitor of the administering researcher placed behind the participant. Participants wore the eye-tracking device and a headset during the experiments. Prior to conducting the experiments, participants filled out a survey regarding their demographic information. Subsequently, we carried out a post-study survey to collect the ground truth on their reactions to presented stimuli.

### 5.2.1 Pre-Study

The survey conducted before the study gathered demographic information such as gender, nationality, age, spoken languages, caffeine consumption, whether participants wore glasses, and the type of prescription. The complete demographic survey is listed in Section A.1 in Appendix A. During the experiments, participants were required to remove their glasses, due to the design of the Pupil Core, in order to prevent any influence on the data collection.

The Pupil Core was physically adjusted and the pupil detection parameters were adapted for the participant by the supervising researcher. Following this, the Pupil Capture program's built-in function was used for calibration. The program displays circles, the calibration points, in each corner, and the center of the screen, which the participants are required to look at. During the calibration, the participants were instructed to not move their heads and only move their eyes to look at the calibration points.

### 5.2.2 Experiments

The experiment videos followed a consistent layout, with a black background (RGB (0,0,0)) and a centered green (RGB (0,190,25)) point of approximately 51px in diameter. To aid participants in focusing on the task, white noise was added as an underlay.

Each experiment consisted of twelve stimuli, presented at 10-second intervals. The first 30 seconds were stimulus-free to allow the pupil time to adjust to the lighting conditions. The full length of one experiment video was 2.5 minutes. Together with short breaks between the experiments, conducting the study took approximately an hour per participant. We conducted variations of three distinct types of experiments: sound, questions, and statements. Any conscious reactions to the experiment were non-verbal. It is important to note that we designed more experiment variations than could be conducted within the given time limit for one participant.

All experiment videos created for possible use in the study can be found in the Digital Appendix, refer to Table B.2 in Appendix B for more information.

#### **Sound - Experiment type A**

These experiments consisted of six different sounds, which were repeated in the same order. Each sound lasted five seconds leaving a 5-second break before the onset of the next sound. The first experiment involved different sounds, including bagpipes, buzzing, human nonverbal emotional exclamations [96], and high-frequency tone given by a sine wave. The second experiment consisted of different frequencies and valences of the same sine-wave-formed sound [97]. The objective was to assess the feasibility of extracting demographic data from basic reactions to sounds.

#### **Questions - Experiment type B**

Twelve questions were verbally stated by a text-to-speech model, accompanied by textual representation. We adjusted the text, to ensure that the overall brightness was consistent



with the default screen. It is assumed, that the observed reactions were not due to a change in brightness. The questions were designed to have a binary answer in the form of: 'Yes' and 'No' and fall into one of four categories:

1. Basic general knowledge, for example:
  - a) Is the sky blue?
  - b) Is honey sweet?
  - c) Do cats fly?
2. Debated non-controversial Topics, for example:
  - a) Should pineapple be on pizza?
  - b) Is a hot dog a sandwich?
  - c) Is soccer better than football?
3. Personal Questions, for example:
  - a) Do you have a pet?
  - b) Do you enjoy cooking?
  - c) Are you a morning person?
4. Controversial Questions, for example:
  - a) Is climate change real?
  - b) Should a woman have the right to an abortion?
  - c) Should minors be allowed to have Gender Affirming Surgery?

Each category represented a separate experiment. We chose these categories based on their diverse levels of emotional connections to the participant. The objective was to detect whether the participant chose Yes or No as their answer to an individual question.

### **Statements - Experiment type C**

The aim, categories, and layout of this experiment type were consistent with type B experiments. Unlike the binary design of the questions, the statements were intended to have a broader response spectrum. The objective was to detect agreement without predetermined answering options.

### **5.2.3 Post-Study Survey**

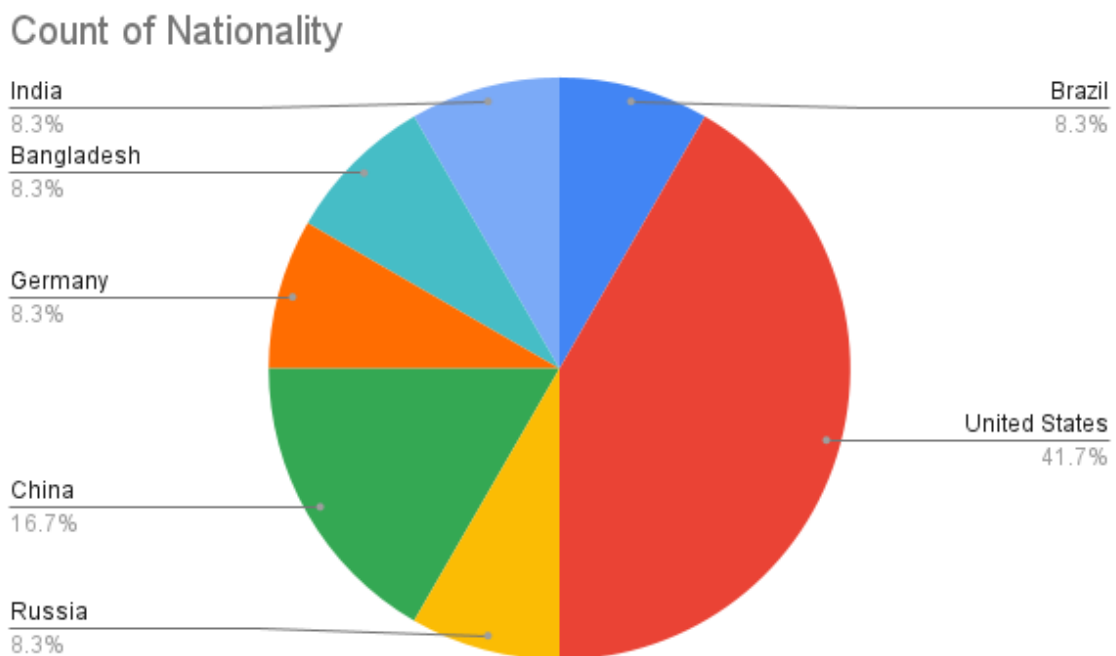
To protect participant's privacy, we did not record their answers to the questions and statements during the experiments. Furthermore, any recordable reply such as verbalizing the answer, could have influenced the eye data. Consequently, a pseudonymized post-study survey, in which the participants answered again, was used to collect the ground truth. A complete list of all questions and statements can be found in the Digital Appendix, refer to Table A.2 in Appendix A for more information.

### 5.3 Participants and Recruitment

We recruited individuals over 18 years old who did not have an eye impairment that would have prevented us from conducting pupil measurements. Recruitment comprised flyer distribution among the university students and word of mouth. Furthermore, we used snowball sampling by asking the participants to distribute the flyer to their social networks. Participants received 30\$ per hour as compensation for taking part in the study.

### 5.4 Demographics

This preliminary study consists of twelve participants with various cultural backgrounds shown in Figure 5.3. Although most of the participants came from the United States, four of the participants originated in Asia, others came from Germany, Brazil as well as Russia.



**Figure 5.3:** Distribution of nationalities in the participant group

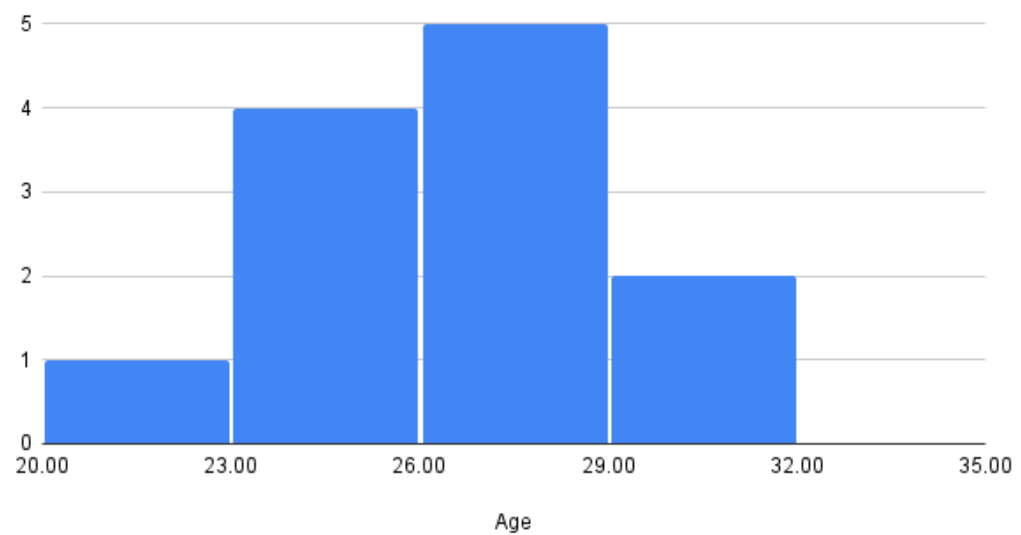
Two of the analyses focus on gender and caffeine consumption, which is presented in Table 5.1. However, the group of participants was not balanced in terms of either consideration point. The collected information about whether participants wear glasses is for future consideration as well as whether the quality of eye data is related to visual impairment.

It is important to clarify that in this work, the terms gender and sex are used interchangeably. The survey asked participants about both their sex and gender, and no differences were reported.

The age range is between 20 and 30 years old, with the majority of participants being in their late twenties as shown in Figure 5.4. As the study was conducted at RIT most of the participants were students. For future studies, we aim for a more diverse age distribution.

**Table 5.1:** Demographics of the participants

| Class | Gender |        | Caffeine Consumption |    | Glasses |    |
|-------|--------|--------|----------------------|----|---------|----|
|       | Male   | Female | Yes                  | No | Yes     | No |
| Count | 8      | 4      | 8                    | 4  | 9       | 3  |

**Age Distribution****Figure 5.4:** Age distribution of the participants

## 6 Methodology of Training

The data analysis used three tables provided by the accompanying programs, `pupil_position.csv`, `blink_timestamps.npy`, `annotation_timestamps.npy`. Additionally, we included the participants' demographics, manually annotated labels for the on-/ & offset of the audio, and, if applicable, the users' labels from the post-study survey given answers.

### 6.1 Collected Data

Pupil Labs' 'Pupil Player' [95] offers a range of collected measurements, from which twenty were selected for further processing. Several possible features were discarded due to the amount of missing data points or irrelevance. Furthermore, the timestamp was normalized to ensure the temporal independence of the data. Although unsuitable as a feature, the confidence rating was used in the data cleaning process. Additionally, to the two given diameter versions, the gaze position was determined using the base features of the x and y coordinates. All available features of the 2D pupil detection were used including the detected pupil ellipse, described by its two centers and axes in pixel as well as its angle. From the available features of the 3D detection, three coordinates that depict the pupil center and its indicated direction were chosen. The center was also represented by the spherical coordinates theta and phi. In total, 20 raw features are used for further analysis. For a complete list of the available data, consult the Pupil Player [95] documentation.

### 6.2 Data Preprocessing

To analyze the collected data, the time sequence is split after cleaning the data. Initially, we attempted to split the time into one-second intervals and label each segment according to the ground truth obtained from the questionnaire given to the participants. Feature extraction was then performed on each segment and the MLAs were tested for prediction. This method of data splitting enabled the utilization of time correlation by including the extracted features of the previous and the following time steps. While the results were satisfactory, two problems arose.

One was the data leakage, as each stimuli sample results in one reaction, it is necessary to account for when that reaction occurred. Since it could be split over several sequences those segments could end up in both the training and the test data. Furthermore, it is important to consider whether the use of temporal data could result in data leakage. The second problem is related to the separation of the pupil reactions. Without the connected reaction, it is difficult to determine if specific changes in pupil size are related to a certain reaction. Consequently, the fixed separation length gave way to mixing data connected to a stimulus with unrelated data.

Hence, we chose a different way of splitting the data. Similar to the previous method, we selected the time frame during which the reaction would occur, based on the start and end of the given stimuli. We allowed for an adjustable offset to the stimuli for the start and end of

the reaction. The parameters were set so that the sample begins at the onset of the stimulus and stops one second after it terminates. According to Zhu *et al.* [5], the reaction to a stimulus decreases four seconds after the onset. The stimulus duration together with the buffer of one second covers this time frame.

All data between the start and the end of a reaction to a specific stimulus is considered one sample. This means that we worked with continuous data instead of several splits for one reaction. This reduced our sample size, to twelve samples for each participant in one experiment. This ensures no leakage due to temporal data and separates unrelated data. All cleaning steps and feature extraction were done separately for each sample.

### 6.3 Missing Data

As stated in the limitations, we discovered that pupil detection may not be effective for all individuals. The program assesses the quality of collected data through measured confidence, which is a value between 0 and 1. 0 indicates that the detected pupil size is uncertain or that it could not be detected at all. Pupil Labs advises discarding any data with a 0-confidence rating and suggests a confidence rating of at least 0.6 [95]. In this thesis, we have chosen to discard samples where less than 80% had an acceptable confidence rating. Smaller gaps were filled using interpolation.

### 6.4 Cleaning

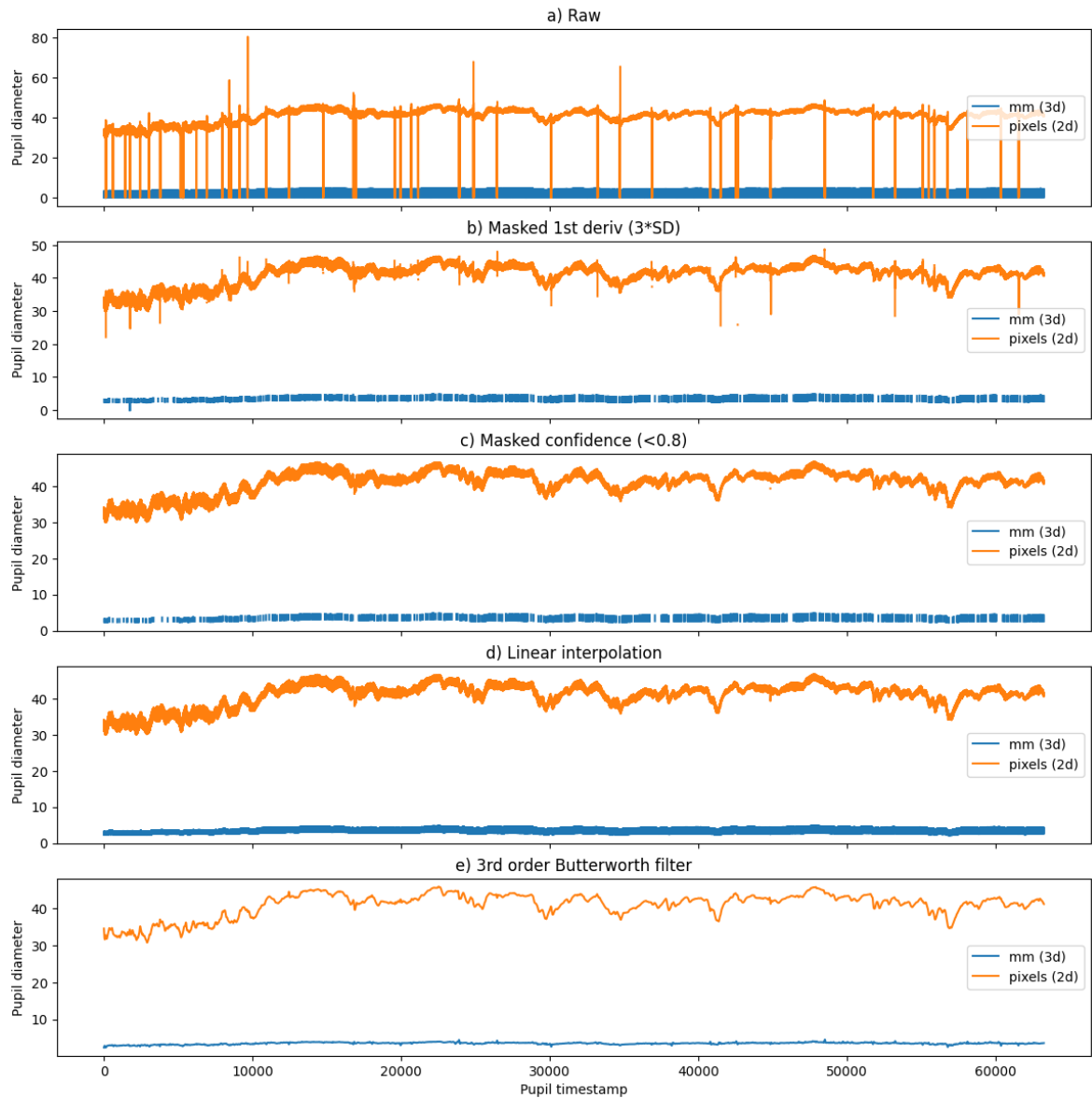
For cleaning we followed with a certain level of discretion the suggested process of `pyplr` [98], a Python package specifically designed for data analysis on Pupil Labs data.

Figure 6.1 visualizes the different cleaning steps not necessarily in order as some are repeated. First, we filtered the confidence (c) of the raw data (a). Before masking missing data points due to low confidence through linear interpolation (d), we masked lost data due to blinking by interpolating the first derivative (b). Then, we removed outliers and smoothed the removed data points through another interpolation. Finally, we removed high-frequency noise via the suggested 3rd order Butterworth filter (e). Through experimentation, we found that the best results were not achieved by using a 4Hz cutoff. Instead, we set the parameter to 0.02.

Before the feature extraction, the data was normalized using z-score normalization. This scaling variant represents the number of standard deviations from the mean [99].

$$z - score = \frac{x - \mu}{\sigma}$$

The resulting feature distribution of the data has a mean of 0 and a standard deviation of 1 [99]. As the data has already been cleaned of outliers, any remaining outliers should not be ignored.



**Figure 6.1:** Visualization of the cleaning steps starting with the raw data (a) followed by masking the first derivative (b) and the confidence (c), smoothing missing data with interpolation (d), and lastly applying the 3rd order Butterworth filter (e)

## 6.5 Feature Extraction

We implemented statistical features, frequency domain features, and signal analysis features.

### 6.5.1 Statistic Features

For each of the 20 selected raw features, 68 statistical features were calculated including the mean, standard deviation, maximum, minimum, and the resulting range of the raw data were also included. The skew, kurtosis, as well as median for the signal excerpts, were considered. Other examples of the used features are the entropy and energy of the signal.

Besides the basic statistical features, two groups of features were calculated. One is the signal analysis features. Here means of different crossing rates were determined. It denotes the mean rate with which a signal crosses zero, so the switch between positive to negative [100]. The second group consists of frequency domain features. Here 18 basic statistical calculations like the mean, are performed on the frequency.

Not all features were explicitly mentioned in this list.

### 6.5.2 Curve Fitting and Gradients

Curve fitting is defined as the process of determining a mathematical function that approximates given data points [101]. Zhu *et al.* [5] used it to approximate the waveform of the pupil size changes. In their experiment, the researchers focused on two distinctive starting waves. They empirically determined a 4-degree polynomial curve and used the corresponding parameters as features.

As the analysis includes the pupil size but also the other given measures, no specific waveform could be detected over the heterogeneous data. Empirical testing showed that a higher degree allows for more freedom to fit the curve. Therefore, we did a six-degree polynomial curve fitting in the form of:

$$a + bx + cx^2 + dx^3 + ex^4 + fx^5 + gx^6 = y$$

We retained the factors b to g, as features for further analysis. Additionally, we calculated the first and second gradients and saved the amplitude along with the minimum, maximum, mean, and standard deviation of both.

### 6.5.3 Blink Count

The blink count is the absolute number of blinks within one sample, rather than the mean blink rate that was used to detect gender differences in Section 3.4. The analysis aimed to determine the presence of blinks during the reaction to the presented stimuli.

Separately, the total blink counts were analyzed as possible features for gender classification as suggested by Sforza *et al.* [82]. To visualize the differences in blink rate distribution a histogram was chosen. There are more male than female participants as shown in Table 5.1 in Section 5.4. Consequently, the total blink count was adjusted relative to the gender groups.

## 6.6 Data Preparation

To begin, we separated a third of the collected data, for future evaluation purposes. The evaluation data consists of the data of four participants, which was not used in training the models to prevent data leakage.

The remaining two-thirds of the data was then balanced according to the classification problem. Since most of the classification problems are binary, we made sure that each class had the same amount of samples. To achieve this, the samples of the less frequent class were counted, and the same amount was randomly selected from the prominent class. This resulted in a smaller sample size. The labels were encoded into numbers and singular occurring 'nan' values were set to zero. Samples without labels were dropped.

Afterward, the data was split into two halves for training and testing of the machine learning models. We build a pipeline applying ANOVA filtering to the training set to, choose the 50 best features. With the exception of the DT, as relevant features are selected during the training process.

The classification of gender, caffeine consumption, and decision-making are binary, while age could be a multi-class problem. In the current dataset, the age ranges within ten years. As the difference in age is small multiple classes would be very specific and not sustainable, so it was decided to split the age range into two classes. The split is denoted by the age median. Consequently, the age classification is also binary.

To classify gender, caffeine consumption, and age groups the collected data from all three experiment types were used. However, only data from specific experiments can be considered for decision-making. The experiments with questions and statements were used for Yes/No and Agree/Disagree, respectively. Both were also analyzed together.

## 6.7 Training

The implementation utilized the sklearn libraries [102] from scikit learn [103]. To tune the models, we employed a combination of empirical testing and a Halving Grid Search (HGS). This parameter search strategy initiates the evaluation process for all candidates on minimal resources [104]. During the iterative selection process, the used resources are increased and the parameters are cross-validated. The parameters for the HGS were set to default. A ten-fold cross-validation was performed on the training set.

Similar to Zhu *et al.* [52] in 2020 and Zhu *et al.* [5] in 2023 we tried kNN and SVM as MLAs. Furthermore, we evaluated k-means, MLP and DT as possible candidates. Each MLA has a wide range of parameters that can influence the quality of the classification. In the following Chapter, the parameters are assumed to be the default setting given by the implementation when not stated otherwise. The full list of possible parameter settings for the classifiers can be found in their documentation [105–109].

## 6.8 Evaluation

To evaluate the quality of the training by calculating the accuracy during cross-validation on both the training and test sets, as well as determining the mean overall steps.



The models were trained on both the training and test datasets, which included the data of eight individuals. These pre-trained models were then tested on the evaluation set. In difference to the training set, the data was not balanced before the prediction. This is because, in real-world use cases, knowledge of the classified data is not available and naturally balanced data is rare. Similar to the training set, 'nan' values in the features were set to zero.

To evaluate the classification P and R were calculated as well as the accuracy. To aid in the understanding of the classification process, the data was presented in a confusion matrix, which helps to visualize possible deficits and classification errors.

## 7 Results and Discussion

This Chapter presents the results of the experiments, including additional tests conducted to better understand the structure of the data beyond the classification problems outlined in the research questions. These additional results, give information that is not necessarily used for the classification, such as the positions of samples within the feature space as well as the distribution of blink counts.

### 7.1 Model Tuning

The HGS calculated that the default parameter settings result in the best classification for all algorithms. However, through empirical testing, another set of possible parameters was found. Table 7.1 gives an overview of the selected parameters that differ from the default settings we used to classify the data. One exception is the k-means where no parameter was changed except for the number of clusters, which was set to 2 as two clusters are expected.

**Table 7.1:** Overview of the parameters used for the classifiers that differ from the default settings

| MLA     | Parameter                               | Value                             |
|---------|---|-----------------------------------|
| SVM     | Weight of Class                         | Inverse to class frequency        |
|         | Regularization                          | 0.005                             |
|         | Random State                            | 42                                |
| kNN     | Number of Neighbors                     | 5                                 |
|         | Weight function                         | Inverse of Distance               |
|         | Power for Minkowski                     | 2 (Euclidean)                     |
| DT      | Split Quality Measurement               | Shannon Information Gain          |
|         | Number of Features Considered for Split | Square root of Number of Features |
|         | Random State                            | 42                                |
| k-means | Number of Clusters                      | 2                                 |
| MLP     | Activation Function                     | Sigmoid                           |
|         | Random State                            | 42                                |

The parameters are consistent for all classification problems. The HGS was employed for all classification problems, but no parameter changes were observed.

For the SVM the class weight was set to be the inverse of the frequency to account for the possible class imbalances. When noise is expected, it is advised to decrease the regularization parameter [40]. For the kNN the number of neighbors is set to 5. Generally, a larger amount of neighbors would be better for noisy data [110]. However, we do not have a lot of data, therefore, a smaller number was chosen. For the distance metric the Euclidean distance was chosen (see Section 2.3.1). In the DT the Shannon Information Gain was determined to result

in the best classification. Shannon minimizes the log loss between the predicted and the true label [42]. For the MLP, only the activation function was changed to the sigmoid function, as described in Section 2.3.1.

## 7.2 Gender Classification

As presented in Table 5.1, there are four female and eight male participants. Both the training and evaluation sets include data from two female participants each. The gender classification utilizes data from the statement, question, and sound experiments, providing more samples for the classifiers during evaluation. As the evaluation set consists of data from two male and two female participants the samples are balanced. This can be seen in Table 7.2

**Table 7.2:** Gender distribution of the datasets

| Label  | Training | Evaluation |
|--------|----------|------------|
| Male   | 528      | 168        |
| Female | 156      | 168        |

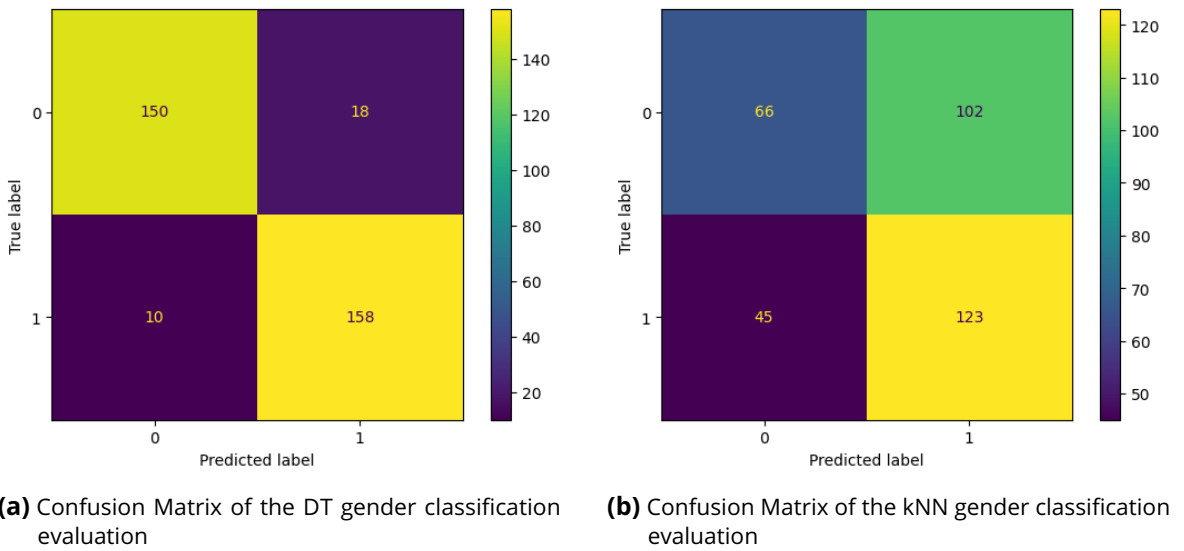
Table 7.3 presents the evaluation results for the gender classification. The kNN achieves an accuracy of almost 60% and is the second-best classifier. The DT classifier, with empirically determined parameters, has an accuracy of 0.92. The P and R are equally high. The accuracy of the other classifiers is approximately 0.5.

**Table 7.3:** Average Precision, Recall, and Accuracy(acc) for gender evaluation

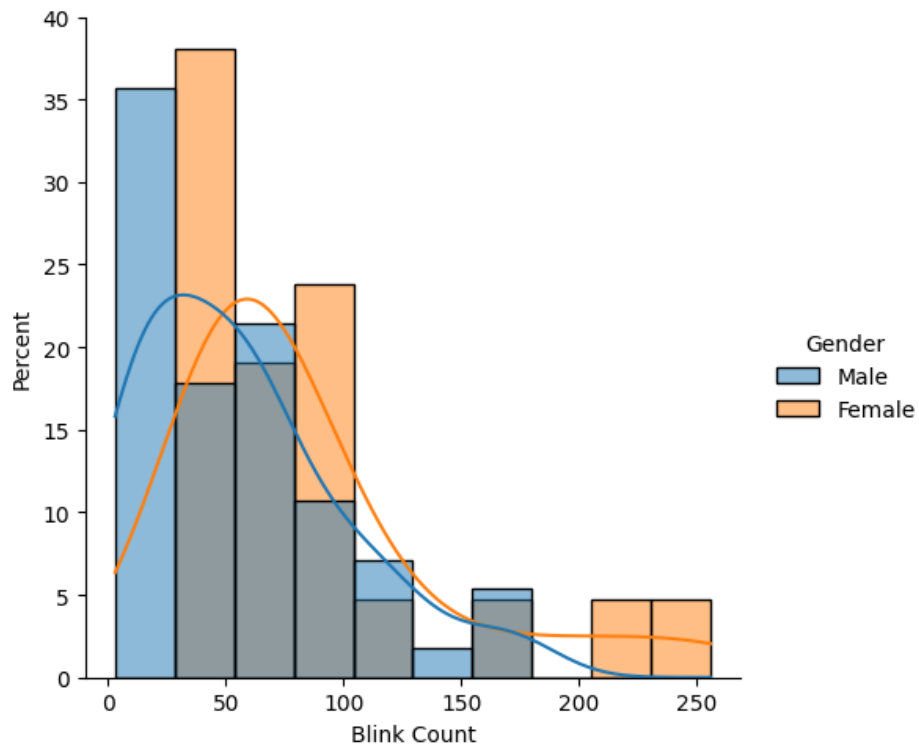
|         | MLA     | avg P | avg R | acc  |
|---------|---------|-------|-------|------|
| SVM     | empiric | 0.46  | 0.47  | 0.47 |
|         | HGS     | 0.47  | 0.47  | 0.47 |
| DT      | empiric | 0.92  | 0.92  | 0.92 |
|         | HGS     | 0.37  | 0.49  | 0.49 |
| kNN     | empiric | 0.58  | 0.58  | 0.58 |
|         | HGS     | 0.58  | 0.57  | 0.57 |
| k-means |         | 0.46  | 0.47  | 0.47 |
| MLP     |         | 0.57  | 0.54  | 0.54 |

Figure 7.1a displays the confusion matrix of the DT classification results. The label '1' denotes male participants and '0' represents female participants. It is apparent that the classification for male participants has a higher R value, while for female participants the P value is higher.

When the test is repeated with new balanced training samples, it becomes evident that the DT is not robust. It may reach a lower accuracy, with a mean accuracy of 0.75 over ten rounds. While not as high, the kNN is more stable, with an accuracy of around 0.56.



**Figure 7.1:** Confusion Matrices of the DT and the kNN for gender classification evaluations



**Figure 7.2:** Distribution of average blink rates for both genders

To test the average blink count as a possible future feature, we calculated the relative distribution for each gender. Figure 7.2 shows that female participants blink more frequently per experiment than male participants on average. For both genders, the majority of individuals blink less than 100 times. The slight difference is due to the female outliers towards the blink counts.

### 7.3 Detection of Caffeine Consumption

Out of the twelve participants, four did not consume caffeine before the experiment. Table 7.4 presents the sample split. In the pre-determined evaluation set, only one participant did not consume caffeine. In this case, it is an identification of the person who did not consume caffeine.

**Table 7.4:** Datasets distribution of caffeine consumption

| Consumed Caffeine | Training | Evaluation |
|-------------------|----------|------------|
| <b>Yes</b>        | 432      | 252        |
| <b>No</b>         | 252      | 84         |

Table 7.5 shows the quality of the classifiers for caffeine consumption classification. The MLP achieves an accuracy of 0.88, making it the best classifier. The SVM with empirically determined parameters is the second-best classifier, with an accuracy of 0.67. The other SVM as well as the kNN both reach nearly 60% accuracy. The DT and the k-means have low accuracy, resulting in almost random classification.

**Table 7.5:** Average Precision, Recall, and Accuracy(acc) for caffeine consumption evaluation

|                | MLA     | avg P | avg R | acc  |
|----------------|---------|-------|-------|------|
| <b>SVM</b>     | empiric | 0.58  | 0.61  | 0.58 |
|                | HGS     | 0.71  | 0.77  | 0.67 |
| <b>DT</b>      | empiric | 0.60  | 0.62  | 0.51 |
|                | HGS     | 0.58  | 0.60  | 0.50 |
| <b>kNN</b>     | empiric | 0.64  | 0.68  | 0.56 |
|                | HGS     | 0.63  | 0.65  | 0.55 |
| <b>k-means</b> |         | 0.51  | 0.51  | 0.49 |
| <b>MLP</b>     |         | 0.83  | 0.87  | 0.88 |

The results indicate that the MLP has an average accuracy of 0.87 when the classification is repeated several times with newly balanced training data in each round, while the SVM is at 0.67. Both kNN and DT have similar results achieving an average accuracy of close to 0.6. It should be noted, that the P and R are higher than the accuracy, with MLP being the exception. P is lower than the R or similar for all classifiers.

### 7.4 Age Classification

The study participants' ages range from 20 to 30 years old. To simplify classification, the samples were divided into two groups based on the median. The 'Older' group includes participants who are older than 26 years. The 'Younger' group consists of samples from participants who are 26 years old or younger. Table 7.6 shows the distribution of age classes in the training and evaluation sets.

**Table 7.6:** Age distribution of the datasets

| Age            | Training | Evaluation |
|----------------|----------|------------|
| <b>Older</b>   | 180      | 168        |
| <b>Younger</b> | 504      | 168        |

Table 7.7 displays the evaluation of age classification, by displaying the average P and R, as well as the accuracy. The DT with parameters determined by HGS achieved the highest accuracy of 0.70, while the kNN and SVM have an accuracy of about 0.6. K-means as well as the MLP have an accuracy close to 0.5, indicating that their classification is near random. For the SVM, kNN and DT the average P is higher than the the R.

**Table 7.7:** Average Precision, Recall, and Accuracy(acc) for age classification evaluation

| MLA            | avg P   | avg R | acc  |      |
|----------------|---------|-------|------|------|
| <b>SVM</b>     | empiric | 0.52  | 0.52 | 0.52 |
|                | HGS     | 0.64  | 0.60 | 0.60 |
| <b>DT</b>      | empiric | 0.59  | 0.58 | 0.58 |
|                | HGS     | 0.72  | 0.70 | 0.70 |
| <b>kNN</b>     | empiric | 0.66  | 0.62 | 0.62 |
|                | HGS     | 0.66  | 0.62 | 0.62 |
| <b>k-means</b> |         | 0.51  | 0.51 | 0.51 |
| <b>MLP</b>     |         | 0.34  | 0.46 | 0.46 |

Repeated testing of the classifiers shows that the DT is not robust, with an accuracy that fluctuates between 0.5 and 0.74, resulting in a mean of 0.59. The accuracy of the SVM also varies, although the mean is 0.59, too. The kNN is more robust, with an accuracy that fluctuates between 0.64 and 0.59 and a mean accuracy of 0.61. No significant changes can be observed for the MLP or k-means.

## 7.5 Decision-Making Classification

For decision-making, there are two concepts: one involves asking direct yes and no questions, while the other involves a broader spectrum of agreement to statements. In this work, we examined not only one classification problem but also tried to learn whether there is a difference in classification quality between the two concepts. Additionally, we explored whether the classification can be improved by combining both.

### 7.5.1 Datasets

Overall, we collected 852 samples for both classification problems. Table 7.8 shows the distribution of the labels in the training and evaluation sets. Samples labeled 'No Opinion' could not be evaluated, due to the lack of a clear reason behind the answer. The samples in the training set were balanced. Consequently, the number of samples per class equals the class with fewer samples.

**Table 7.8:** Datasets label distribution

| Label             | Training | Evaluation |
|-------------------|----------|------------|
| <b>Yes</b>        | 144      | 82         |
| <b>No</b>         | 132      | 62         |
| <b>Agree</b>      | 120      | 67         |
| <b>Disagree</b>   | 132      | 60         |
| <b>No Opinion</b> | 36       | 17         |

### 7.5.2 Cross-validation

All found parameter settings for the MLAs were tested in a ten-fold cross-validation on the training set. Table 7.9 presents the accuracy of each round, as well as the mean accuracy over all rounds. The classification of 'Yes' and 'No' is used as an example. The mean accuracy shows similar results of about 0.6 for all empirically determined models. The only exception is k-means, which has an accuracy of merely 0.49. The SVM shows a slightly better result. For kNN and SVM, the parameter set determined by the HGS gives worse results than the empirically determined ones. Only the DT, does not show a difference.

**Table 7.9:** Cross-validation accuracy of Yes/No classification

| MLA                                    | DT      |      | kNN     |      | SVM     |      | k-means | MLP  |
|--|---------|------|---------|------|---------|------|---------|------|
|  | empiric | HGS  | empiric | HGS  | empiric | HGS  |         |      |
| Cross-validation<br>Rounds<br>Accuracy | 0.56    | 0.59 | 0.57    | 0.56 | 0.67    | 0.49 | 0.49    | 0.66 |
|  | 0.57    | 0.58 | 0.50    | 0.48 | 0.55    | 0.50 | 0.48    | 0.58 |
|  | 0.64    | 0.56 | 0.64    | 0.52 | 0.62    | 0.52 | 0.47    | 0.48 |
|  | 0.61    | 0.55 | 0.63    | 0.54 | 0.67    | 0.48 | 0.52    | 0.58 |
|  | 0.54    | 0.64 | 0.53    | 0.54 | 0.64    | 0.46 | 0.50    | 0.63 |
|  | 0.63    | 0.64 | 0.58    | 0.58 | 0.58    | 0.52 | 0.48    | 0.57 |
|  | 0.54    | 0.58 | 0.58    | 0.55 | 0.64    | 0.50 | 0.51    | 0.55 |
|  | 0.50    | 0.67 | 0.54    | 0.52 | 0.60    | 0.51 | 0.49    | 0.56 |
|  | 0.60    | 0.58 | 0.61    | 0.55 | 0.62    | 0.45 | 0.52    | 0.59 |
|  | 0.71    | 0.55 | 0.58    | 0.51 | 0.48    | 0.57 | 0.48    | 0.56 |
| Mean Accuracy                          | 0.59    | 0.59 | 0.58    | 0.53 | 0.61    | 0.50 | 0.49    | 0.58 |

When considering the individual cross-validation rounds, it becomes apparent that the quality of the classifications is closely related to the split of the training and test data. The k-means algorithm has the smallest range between its highest and lowest results, but all rounds have only an accuracy of approximately 0.5. The DT with empirically determined parameters has the smallest amplitude, with a difference of 0.21. The highest overall accuracy of 0.71 was also achieved by this DT. The empirically determined SVM and MLP also reach nearly 70% accuracy, but they also have high disparities between the scores. The kNN achieves as highest score an accuracy of 0.64, but it is also more stable than the others. Generally, models with empirically determined parameters have more volatile accuracy values, compared to the

models with HGS determined values. This might indicate overfitting, which means that the models are fitted too well on the training data [49]. Conducting the evaluation on data, which is not used for training might provide a clearer picture.

### 7.5.3 Evaluation

Table 7.10 shows the average P and R as well as the accuracy for the Yes/No classification of each model. The highest accuracy was measured for the DT with parameters determined by HGS and the SVM with empirically determined parameters. Both reach nearly 0.6, with the DT model being slightly better. For all other models, the classification is arbitrary, with an accuracy of about 0.5. The same can be observed for P and R.

**Table 7.10:** Average Precision, Recall, and Accuracy(acc) for Yes/No evaluation

|                | <b>MLA</b> | <b>avg P</b> | <b>avg R</b> | <b>acc</b> |
|----------------|------------|--------------|--------------|------------|
| <b>SVM</b>     | empiric    | 0.56         | 0.56         | 0.56       |
|                | HGS        | 0.51         | 0.51         | 0.53       |
| <b>DT</b>      | empiric    | 0.45         | 0.45         | 0.45       |
|                | HGS        | 0.59         | 0.59         | 0.58       |
| <b>kNN</b>     | empiric    | 0.50         | 0.50         | 0.51       |
|                | HGS        | 0.49         | 0.49         | 0.49       |
| <b>k-means</b> |            | 0.49         | 0.49         | 0.46       |
| <b>MLP</b>     |            | 0.45         | 0.45         | 0.46       |

As the evaluation data is unbalanced, the data contains more 'Yes' answers. This slight bias is also reflected in the class-specific P and R. Consequently, it is more probable that samples that are classified as 'Yes' are really 'Yes', even if some are missed. R is higher for 'No' classifications, indicating that more 'No' samples are correctly identified even though some are wrongly in this class.

When comparing the results of the cross-validation and the evaluation, the SVM with empirically determined parameters and the DT with HGS determined parameters achieved similar accuracies. However, the maximum accuracy was not reached.

Table 7.11 displays the average P and R as well as the accuracy for the Agree/Disagree classification for each model. The classification is also nearly random for most models. K-means has the highest accuracy across all models. Meaning it is closer to a clustering problem than a classification. Overall is the classification worse than for the Yes/No classes.

The classification for all decision-making experiments is overall also random. None of the models demonstrate significantly higher accuracy than 0.5. For all three classification problems, the evaluation for most models is close to arbitrary. However, in comparison with the average accuracy measured during cross-validation, where most models reached about 0.6, similar results could be achieved during the evaluation for some models. Generally, the results were slightly better for the Yes/No classification, while Agree/Disagree and both together classified randomly.



**Table 7.11:** Average Precision, Recall, and Accuracy(acc) for Agree/Disagree evaluation

|                | <b>MLA</b> | <b>avg P</b> | <b>avg R</b> | <b>acc</b> |
|----------------|------------|--------------|--------------|------------|
| <b>SVM</b>     | empiric    | 0.47         | 0.47         | 0.49       |
|                | HGS        | 0.47         | 0.47         | 0.49       |
| <b>DT</b>      | empiric    | 0.54         | 0.53         | 0.54       |
|                | HGS        | 0.49         | 0.49         | 0.49       |
| <b>kNN</b>     | empiric    | 0.46         | 0.46         | 0.47       |
|                | HGS        | 0.46         | 0.46         | 0.47       |
| <b>k-means</b> |            | 0.57         | 0.57         | 0.56       |
| <b>MLP</b>     |            | 0.46         | 0.46         | 0.46       |

#### 7.5.4 Follow-up Experiment

The idea for this work is based on the research of Zhu *et al.* [5]. They use pupil reaction as a biometric. It is possible that the data may not be generalizable across multiple individuals. Therefore, we decided to test the models additionally on subgroups of participants. Consequently, the data was divided by gender, to explore whether pupil data generalizes better within one gender.

Table 7.12 shows the average P and R as well as the accuracy for the Yes/No classification of each model on data from male participants. The SVM with empirically chosen parameters reached an accuracy of 0.6, while the SVM with parameters determined by HGS and the MLP come close. The other classifiers are closer to an accuracy of 0.5. The P and R for the SVM are slightly 0.57 for the one with a 0.58 accuracy. This could be a result of the unbalanced data.

**Table 7.12:** Average Precision, Recall, and Accuracy(acc) for Yes/No evaluation of only male participants

|                | <b>MLA</b> | <b>avg P</b> | <b>avg R</b> | <b>acc</b> |
|----------------|------------|--------------|--------------|------------|
| <b>SVM</b>     | empiric    | 0.56         | 0.55         | 0.60       |
|                | HGS        | 0.57         | 0.57         | 0.58       |
| <b>DT</b>      | empiric    | 0.50         | 0.50         | 0.47       |
|                | HGS        | 0.43         | 0.42         | 0.44       |
| <b>kNN</b>     | empiric    | 0.51         | 0.51         | 0.51       |
|                | HGS        | 0.51         | 0.51         | 0.51       |
| <b>k-means</b> |            | 0.44         | 0.45         | 0.40       |
| <b>MLP</b>     |            | 0.55         | 0.55         | 0.57       |

Table 7.13 presents the average P and R as well as the accuracy, of the results for the Yes/No classification of data from only female participants. The kNN with empirically chosen parameters reaches an accuracy of 0.6. The kNN and DT with parameters determined by HGS, as well as the SVMs, performed similarly. The accuracy of the other classifiers is closer to 0.5, while the P and R are equal to the accuracy.

**Table 7.13:** Average Precision, Recall, and Accuracy for Yes/No evaluation of only female participants

|                | <b>MLA</b> | <b>avg P</b> | <b>avg R</b> | <b>acc</b> |
|----------------|------------|--------------|--------------|------------|
| <b>SVM</b>     | empiric    | 0.57         | 0.57         | 0.57       |
|                | HGS        | 0.59         | 0.58         | 0.58       |
| <b>DT</b>      | empiric    | 0.56         | 0.55         | 0.56       |
|                | HGS        | 0.43         | 0.43         | 0.43       |
| <b>kNN</b>     | empiric    | 0.60         | 0.60         | 0.60       |
|                | HGS        | 0.57         | 0.57         | 0.57       |
| <b>k-means</b> |            | 0.44         | 0.45         | 0.40       |
| <b>MLP</b>     |            | 0.50         | 0.50         | 0.50       |

The SVM is considered one of the better classifiers for both genders. However, the kNN and MLP show promising results for one gender. In contrast to the first classification with all genders, the highest result was better. Across tests with all labels, the SVM reaches also a 0.60 accuracy for data from male participants. while the kNN can reach the highest accuracy of 0.58 for female participants. Gender-specific classification can therefore reach better results than non-gendered classification.

The Agree/Disagree classification is random for male participants. None of the classifiers perform significantly better than an accuracy of 0.5. Although the P and R are about 0.6 for the kNN, as the classes are not balanced, this classifier might be better suited. For females, the accuracy of the kNN is approximately 0.6. The MLP has a similar accuracy. The evaluation set for females is nearly balanced.

When considering both labels together, the accuracy for males is 0.60 for the SVM. The kNN with empirically determined parameters achieved the best results for females, with an accuracy of 0.58. The SVM determined by HGS has similar results.

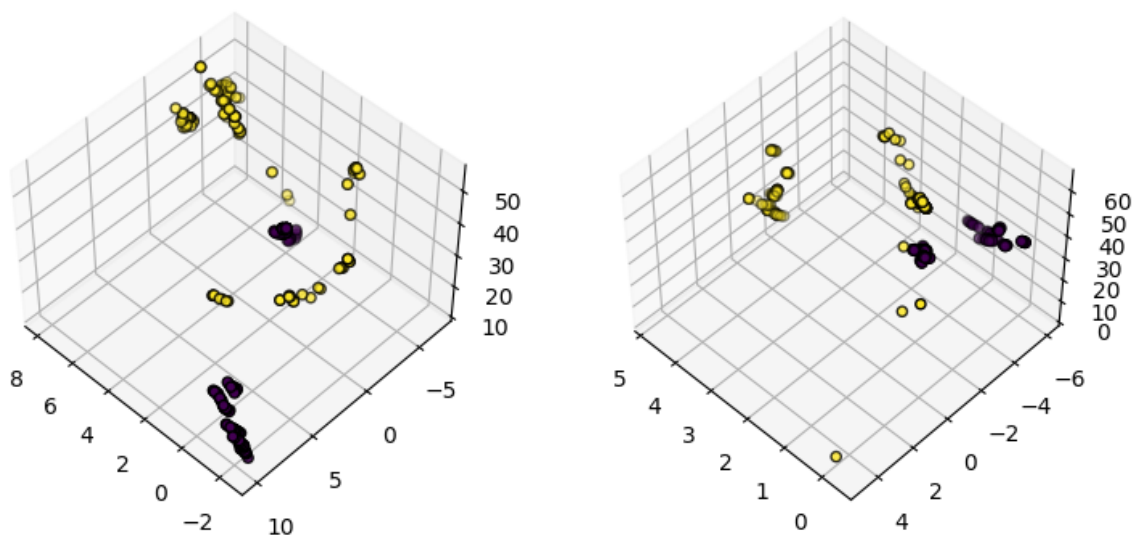
## 7.6 Discussion

At the current state, the results show that the success of classifying the data, according to our set aims varies.

### 7.6.1 Discussion of Gender Classification

The most successful outcome for gender classification were achieved with a DT. However, it is important to note, that DTs are likely to overfit when more features than samples are given [42]. We did not set a maximal depth for the tree, even then only seven features were considered for the classification.

Figure 7.3 shows the data samples for the training (Figure 7.3a) and evaluation set (Figure 7.3b) in the feature space, with colors denoting the different genders. Each figure shows four distinct groups, with the axes indicating the values of the same three features. For the clustering algorithm, no two clear groups are visible. While the different positions of the



(a) Samples for gender classification in the feature space for the training data

(b) Samples for gender classification in the feature space for the evaluation data

**Figure 7.3:** Samples for gender classification in the feature space

groups lead to a low accuracy for the SVM. The kNN has a better classification as the groups are still close together, although as Figure 7.1b shows, about 67% of the sample are classified as 'Male'. Consequently, the R is higher than for the 'Female' class. More data could also help evaluate the robustness of the classifiers [111]. With more data from female participants, less data would need to be discarded to balance the data. Balanced data is important for DT because bias towards the dominant class could occur otherwise.

As explained in Section 3.4, different genders exhibit distinct eye behaviors like higher blink patterns [82] or a different baseline for the pupil size [31]. Note that the gender classification is based on data from only four people, even if training samples are mixed up. To decide whether gender classification over more people is still possible additional data has to be collected. This classification shows promise for further studies.

Currently, only data from eye measurements was used. We have not specifically used suggested features like eyelid movement [82] or fixation points [83, 84]. This preliminary study did not focus on providing visual stimulation. Should we decide to use fixation points, experiments using visual aids have to be designed. Considering that our goal is to study the data potentially collected by a VR headset, additional potential data sources should be considered. Height can be an indicator for gender [112], which could be determined by the position of the headset in the room.

We analyzed the eye movement and gaze position through the collected x- and y-coordinates. However, the importance of these features is questionable given our experiment layout, which primarily used audio stimuli and reduced visual stimuli. During the sound experiments, the participants were encouraged to focus on the green dot presented on the screen. Therefore, the natural moving patterns were not present in the data. When presenting the text for questions and statements, the recorded gaze follows the text, with the eyes making

left-right movements. This pattern is similar for all participants. Further analysis would be required to investigate differences in specific patterns. To exploit eye movement further, future experiments should focus more on visual stimuli.

Specific eyelid movements were not analyzed, as we focused on the data automatically recorded by the eye tracker. Although the eye tracker records videos of the eye that could be analyzed for eyelid movement, this thesis only focuses on the pupil. Therefore, analysis of eyelid movements is a consideration for future work. Furthermore, while we considered blink count, by examining the automatically extracted blinks, we only counted the blinks present during the specific samples, to prevent data leakage between samples. Using average blinks per minute could be a valuable feature for gender classification as Sforza *et al.* [82] showed. Although our calculations cannot denote average blink rates, a small difference between female and male participants can be observed. This is also illustrated in the blink rate distribution, as displayed in Figure 7.2. However, we collected data from a limited number of people so currently, we can not confirm its value as a feature. It could be considered when more data is available.

In this work, we classified gender, but it is unclear whether the analyzed pupil responses are related to gender, so learned behavior, or the biological sex, represented by biological attributes [113]. Currently, we are unable to analyze differences as none were reported by the participants. Other factors that could contribute to gender-divergent reactions are hormones, which would require further research. This could be relevant should transgender individuals take part in future studies. Further research could examine whether the reactions of transgender individuals are more similar to their original sex, the sex they transitioned into, or if they differ from both. Furthermore, if hormones have an influence it could be explored whether the use of hormone-based birth control or pregnancy hormones affect the pupil.

In conclusion, the classification of gender data was successful. Despite the limited visual stimulation, which was a major focus in other studies [83, 84], we reached 75% accuracy. For further studies, we have several options to expand the input data, as VR environments aim to emulate realistic surroundings. The visual-focused programming encourages natural movements, that are continuously recorded. So the collected eye movements when using a VR headset could be used to derive the user's gender.

### 7.6.2 Discussion of Caffeine Detection

The evaluation data is unbalanced, with only one-third of the samples belonging to a caffeine consumer. This means that the evaluation is focused on identifying the caffeine consumer. Due to the unbalanced data, accuracy is not the preferred performance metric. Instead, P and R are better indicators of the classifier performance [51]. The lower P indicates, that while the samples of the caffeine consumer are likely to be identified, samples of other participants will be wrongly identified as caffeine consumers. But as the R is higher we are likely to identify whether caffeine was consumed without missing many samples that also indicate caffeine consumption. Generally, having both a high P as well as a high R would be optimal [51]. It shows that the classifier is precise.

The accuracy of nearly 90% achieved for the detection of caffeine consumption needs to be confirmed on a larger sample size, as MLP reached this. Generally, neural networks require a larger number of samples. On the other hand, the SVM can produce satisfactory results on smaller sample sizes, reaching an accuracy of nearly 70%. As in this dataset, only one caffeine consumer needs to be identified, it is, questionable whether the classifiers will hold in further studies with a larger participant group.

Caffeine is considered a drug, belonging to the methylxanthine class of stimulants, with a psychoactive effect [114]. Studies have shown, that caffeine can increase pupil size [91] and slow the pupillary unrest [92]. We included the mean pupil diameter in the feature extraction, therefore it is already used in the classification.

Both studies show that the effects grow over time. The reduction of the pupillary unrest peaks after 75 minutes [92]. Abokyi *et al.* [91] also demonstrate an increase in pupil size for 90 minutes after caffeine consumption. In our study, we did not specify the time since the last caffeine intake. This should be considered in future user studies. Furthermore, excessive caffeine consumption can prolong its effects [92]. The participants in this preliminary study were primarily students of RIT. All the participants regularly consume caffeinated beverages. Consequently, it is unclear if some measurements were still affected. To obtain clearer results, a test group, that rarely consumes caffeine should be included. Alternatively, participants who consume caffeine should abstain from it before participating in the experiments.

While designing the experiments, participant comfort was highly valued. Consequently, we chose to use the lowest illumination, resulting in the least noisy data. Due to the low illumination, pupils reach their largest dilation. Any stimuli affecting the pupil can either cause a constriction or dilation of the pupil. Keeping the pupil at the maximum dilation allows for only one reaction direction. Therefore, choosing the lowest illumination might have led to information being lost. Additionally, observed noise may provide more distinct pupil reaction patterns to a given stimulus.

Before conducting further user studies, we will test whether better classification results can be achieved by conducting experiments in a brighter environment. This will also allow for more variability should more visual stimuli be employed. According to Bradley *et al.* [11], the pupil constricts during the illumination influx when presenting pictures, which is expected. They found that this effect can be mitigated by allowing the pupil to adjust to the same mean illumination before displaying a picture.

It is unclear if an increase in pupil dilation can be detected due to caffeine consumption, as it is assumed that the pupils are already close to maximal dilation. Additionally, Abokyi *et al.* [91] conducted their data collection in a dimly lit room, in difference to the completely dark room we are using. Pupillary unrest was avoided in the feature extraction, so it cannot be used to detect caffeine consumption. The sample only included data between the audio onset and one second after it ended. Pupillary unrest arises from the recurring correction of the pupil light reflex [31]. We tried to avoid this natural behavior by limiting the amount of data belonging to one sample. The sample should mainly show the excitation phase. To analyze the pupillary unrest we would need to examine data during which no stimuli are present. The 30 seconds before the first onset, which allows the pupil to adjust to the lighting condition, is

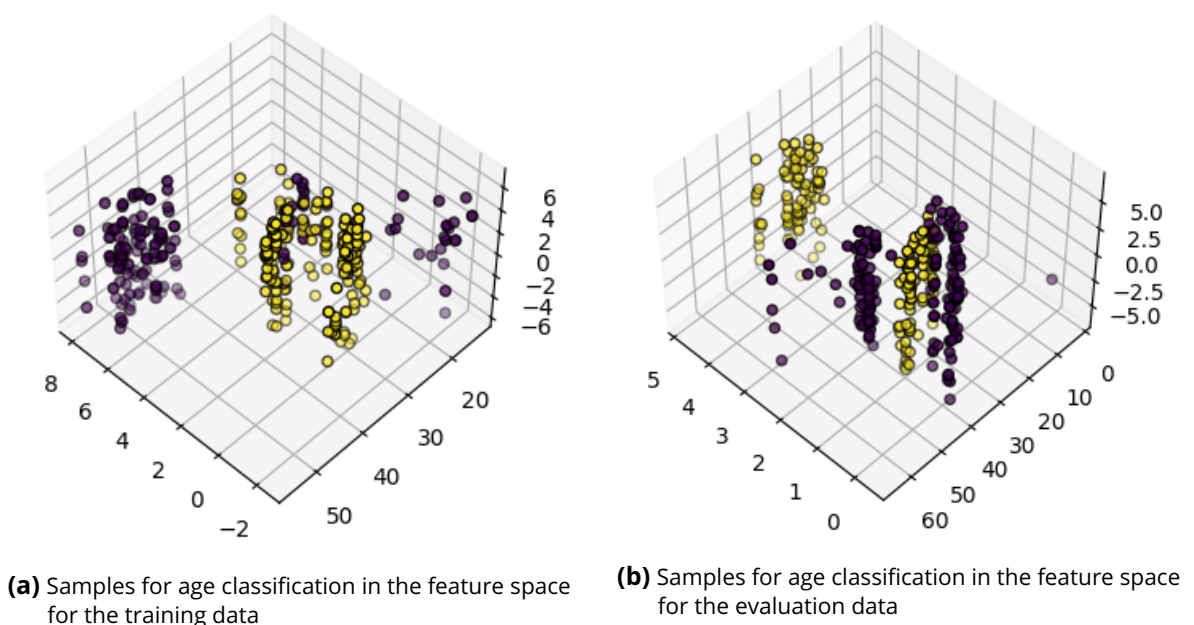
not suitable for this analysis. The example in Figure 4.2 shows a continuous increase in pupil size. To accurately analyze this, a specific time frame during which the recording is made must be provided.

For detecting addiction related to other drugs, saccadic eye movements, in the presence of drug-related stimuli, were used [16]. To emulate this study, we could analyze eye movements while displaying caffeine-related stimuli, like a picture of coffee, and compare them to eye movements while other stimuli are present. Since we were able to detect the caffeine consumer in the evaluation set with an accuracy of nearly 90%, conducting this additional experiment seems neglectable. We will keep it in mind should further studies fail to validate our ability to detect caffeine consumption. Before that, we will attempt to utilize the other suggested features.

In summary, this experiment shows that the consumption of drugs can be detected through eye data, that might be collected with a VR headset. It is not clear if any specific stimuli are needed, for example, stimuli that increase cognitive processes, to detect differences in pupil behavior. If not, any collected eye data could be used to derive this personal information, which could then be used to adjust advertisements to personal preferences.

### 7.6.3 Discussion of Age Classification

The training set and the evaluation set both include samples from two participants belonging to the 'Older' class, resulting in an unbalanced training set and a balanced evaluation set. To balance the training set, over 60% of the samples in the 'Younger' class had to be removed during the training, while all samples from the 'Older' class were used in all training rounds. This produces a similar classification as for gender, but the classifiers behaved differently. This could be the result of the ANOVA feature selection. As here the data came from different individuals, different features separate the samples most effectively.



**Figure 7.4:** Samples for age classification in the feature space

Figure 7.4 shows the data samples of the training set (Figure 7.4a) and the evaluation set (Figure 7.4b) in the feature space. The colors represent the age classes. Although the classes are not entirely mixed, the clusters are not as distinct as those for the genders. This might be the reason for the lower performance of the DT. Depending on the specific balancing of samples during the training, this distinction can increase or decrease, which leads to the volatile performance of the DT. The kNN algorithm is more robust because even in the mixed sample groups, the majority of the neighboring samples are similar. Consequently, the performance of the kNN is similar for both age and gender classification.

For the SVM the positions of the classes in the feature space do not seem to change between training and evaluation as much as they do for gender classification. However, due to the blend of the classes, the R for the specific classes is very different. Many of the 'Younger' class samples are classified as 'Older'.

Currently, only data from two individuals per class are analyzed for training and evaluation. This limited sample size may result in classification performance that is not applicable to larger populations. While the difference between the training and evaluation sets suggests that generalization may not be possible, the robustness of the kNN algorithm with an accuracy of 61%, indicates otherwise. However, the detection has less certainty than that of caffeine consumption or gender. To obtain clearer results, additional data has to be collected. To achieve this, efforts should be made to balance the age classes more effectively.

Erbilek *et al.* [7] achieved only an accuracy of 52% with the kNN. The result for the SVM is similar to ours. In their study, they achieved the best results by combining different classifiers. This approach could be employed to increase the quality of our classification results, in future work.

In previous studies, the age gap between age groups was larger than in our study. Sforza *et al.* [82] selected two distinct groups: one with an age range between 20-30 years, and the other with people above 50 years of age. Alexandridis [31] also seems to suggest that the differences become apparent with a significant age gap. It is important to note that the age range in our study is relatively narrow compared to other studies [7, 71, 82, 115]. In our study, the age range was divided into two classes, which are considered one class by Sforza *et al.* [82]. Therefore, in future work, it is essential to ensure that a more diverse age range is represented in the people participating in the study, as it could potentially improve the classification results. Erbilek *et al.* [7] did a three-class classification, which could be considered when a broader age range exists. Using more specific groups provides more specific information about a person.

It is worth noting that blink rates not only differ between genders but also change with age [82]. As previously mentioned in relation to gender classification, the blink rates were not analyzed in a way that the average blink rate could be used as a classification feature. This could be added in future work, especially if the data is kept separated by people because in mixed samples it could lead to data leakage. However, blinks offer more usable features as Zhu *et al.* [52] shows, such as frequency or time between blinks.

For further improvement, new features could be considered. An aspect that has not yet been analyzed is spontaneous eyelid movements. According to Sforza *et al.* [82], the velocity of eyelid opening and closing significantly changes with age, with movements slowing by more than 70% in the older age group. Moreover, not all eyelid movements lead to a full closure. Older people fully close their eyelids less frequently than young people. These behavioral patterns could be utilized to more accurately distinguish between age groups.

Additionally, it may be beneficial to consider a wider range of features beyond just pupil data, such as the area around the eye. Dehshibi and Bastanfard [116] used wrinkles as a feature for age classification, which could be considered in future work. With the addition of facial expression trackers [3], Dibeklioglu *et al.* [115] could become more relevant, as they detected age through smiles. Other potential features that could be explored are eyelid movement [82] or pupil oscillation [31], which could also be used to detect caffeine consumption.

In summary, different age groups were detected, but the accuracy could be improved. Widening the age range as well as adding new features could give the means to achieve this goal. With more sensors added to the VR headset age detection will become more precise. Increasing the potential to exploit the collected data to extract PII.

When comparing the suggested features, it becomes apparent that the same feature can be used to detect various information. Blink rates can be used to classify age and gender. While pupil oscillation helps in the detection of age and caffeine consumption. Gingras *et al.* [72] describe that women have larger pupils than men, while a large pupil can also be a sign of caffeine consumption.

#### **7.6.4 Discussion of Decision Detection**

To classify the made decisions kNN and SVM have the best performance overall, achieving approximately an accuracy of 0.6. When examining the results of the gendered datasets, it becomes apparent that the kNN performs better for female participants and the SVM performs better for male participants independent of which label was classified.

The data for evaluation is unbalanced, although the discrepancies between the labels are stronger in the gendered datasets. This shows for male participants when the P and R are higher than the accuracy. Considering the unbalanced data accuracy might not be the best-suited performance measurement.

In their study of pupil size changes during the decision-making process, Gee *et al.* [13] did not attempt to classify the data. However, they were able to detect a difference between negative and positive responses. Their experiments were based on determining whether a signal was present. A visual analysis did not show consistent differences between two given options. Further analysis in this could be considered.

The accuracy increases when using the gendered datasets. The data seems to not generalize well over multiple individuals. So a test for more specific subsets might further improve the classification. Possible subsets could be created according to demographic aspects such



as whether glasses were worn or the age of the individual. Another possibility is to classify the data of one individual, or according to the type of questions that were asked. However, we would have to examine the number of samples we would need for the classification, for example by studying the learning curve depending on the number of samples [117], to test the feasibility of the classification on such limited datasets.

Not all situations that require decision-making have two distinct answers. As a person might be indecisive about some questions [14]. We tried to mitigate this in our experiments by selecting simple questions, to which most people know the answer to. For the statements we provided the option to have 'No Opinion'. Generally, the classification for the statements was worse than for questions. This could be the result of indecisiveness. To address this problem we should focus on using binary questions in further experiments.

Rosner *et al.* [14] employed a more visual approach leading to complex combinations of fixations between the possible answers. In comparison, we did not study fixation due to the absence of visuals related to decision-making. By focusing on audio stimuli, we limit the possible features that could be collected for the visuals, simplifying the process. However, this may result in a loss of information. For further examination of decision-making, we could consider including visual aids in future work.

Dealing with human data can be complex. Regarding our experiment, there are several challenges that we can only attempt to mitigate. For example, we must ensure that participants remain focused on the task and do not become distracted by each question, which could lead to them thinking more deeply about the topic. Another example is individual differences in reaction time, which could be an incentive to specify the time frame, in which pupil responses occur, for each person individually. Additionally, since we included a visual representation of the questions and statements, their reading speed could surpass the audio speed, potentially making specifying the time frame for the answer difficult.

The time frame that is considered for each sample, is one major factor that influences the quality of the analyzed data. This is especially important for decision-making, as unrelated data might be collected if the participant has not yet processed the prompt or has already moved on to different thoughts. Additionally, the stimuli are presented as text in the current layout, which could lead to discrepancies in the expected response time. Currently, it is assumed that participants answer after the audio has ended. However, if a participant reads the question faster than its audio presentation, they may answer earlier, and therefore not all following data is necessarily related to the decision-making process. Furthermore, even if the participant consciously answers after the audio, the decision-making process could have been reflected in the pupil size earlier. Libet [118] discovered that the brainwaves reflect a decision before a person becomes aware of it and can act on it. It is possible that the pupil size, which cannot be consciously controlled, changes simultaneously to the brainwaves before the individual is aware of their answer. In conclusion, the pupil's reaction might already be over by the time a person makes the decision. However, these differences are within one second [118]. Therefore, we have the necessary data within our sample. Nevertheless, there could be more noise in the sample than we initially expected. To improve the quality of our

data, we should discard the text representation to avoid any deviation between the reading speed and the audio. Furthermore, more research should be done to better specify the time frame of the sample, so less unrelated data is analyzed.

At the current state, it is not possible to detect a clear decision, only which decision might be more likely. Therefore, personal information cannot be derived from asking a single question. Furthermore, the user would need an outside prompt to consider their decision. It is unclear if the user would automatically reflect on their decision when hearing a question. Additional research is needed to determine the feasibility of detecting a Yes/No decision through eye data.

### 7.6.5 Evaluation of the Classification Models

In terms of gender classification, the DT algorithm has the highest average accuracy, although it is not robust and is dependent on the training data. On the other hand, kNN has a lower accuracy but is more stable. In both cases, the algorithm with empirically determined parameters performed better. When it comes to caffeine consumption, MLP showed the highest accuracy on the evaluation set with 0.87, followed by SVM and kNN. The kNN classifier proves to be the most stable and accurate for age classification, with the highest average accuracy. The second-best accuracy was achieved by SVM and the DT, but the DT is volatile. For Yes/No questions, the DT and the SVM classifiers performed best. However, for Agree/Disagree and over all labels, the classifiers are close to random. In the follow-up experiment, where the classification was done for a subset based on gender, the SVM showed the best results for males, while for females SVM and kNN have similar good results. The kNN and SVM also prove to yield the best accuracy for the gender-based classification across all labels.

Overall, the best classifier was kNN. Although it did not yield the best results for each classification problem, the results were more robust than other classifiers. SVM is another classifier that has shown good results. This is consistent with prior research on classifying pupil data [5, 52]. Among the other classifiers, many were found to be volatile. Zhu *et al.* [5] discovered that the kNN requires the least samples to achieve good results. For the classification of caffeine consumption, the MLP consistently shows high results.

Although, the kNN yields the best results, Amor and Liu [110] suggest using a radius-based classifier. This classifier considers all neighbors within a certain radius for prediction instead of defining a definite number of neighbors. However, it loses effectiveness in high-dimensional spaces. The results of MLP and SVM could be improved through data scaling [45]. Currently, only the individual samples are normalized. The purpose was to prevent data leakage. One way to improve the data is to scale the overall training and evaluation data separately. Another option would be to change the scoring function of the MLP. For binary classification, a logistic activation function is recommended, especially if no hidden layers are used [44]. In the current model, the default number of hidden layers is 100. Therefore, alternative functions such as swish or ReLU could be used [44].

To improve the classification results, tuning the parameters could be considered. In all the experiments the parameters for the MLAs were kept the same. The HGS did not find that a different set of parameters would increase the quality of the results. However, other parameters for the individual test might change the result. To refine the HGS, different scoring metrics could be employed. Currently, the data of the training data is unbalanced and through balancing data is lost. The scoring metric for the HGS is by default accuracy [119], which is not preferable for unbalanced data [51]. By changing the scoring metric the full training set could be used during the HGS. The P or R might be better alternatives for the scoring metrics. Another metric that could be applied is the F1-Score, which calculates the harmonic mean of P and R [51]. Additionally, a different search algorithm could be implemented. For HGS specific values are set for a parameter and the combinations are tested. Randomized parameter optimization samples parameters from a given distribution [119]. Therefore, more parameter combinations are tested.

It became apparent during the decision-making classification that the data does not generalize over all people. We succeeded in improving the classification by using gender-specific subsets. The utilization of eye data as biometrics during authentication confirms that the eye reactions can be unique [5, 52]. Therefore, a pre-trained model could be employed, trained on data from different individuals, and then refitted for smaller subsets. The most specific data would be to detect the decisions of a single person. However, this would imply that a significant amount of data would be required to be collected from a single individual. By using pre-trained models data of a single person could be analyzed. As a generalized model could be trained and then specified. This is a topic for future work.

Overall, it appears that insufficient data was collected to make a definitive determination about the classifiers. For this preliminary study, we were able to collect 1020 usable samples. While we can use the complete dataset for caffeine consumption, gender, and age classification, decision-making, in particular, has a smaller number of samples. It is necessary to ensure that the collected data is better balanced, to minimize data loss, so more can be utilized for analysis. It is advised to have multiple independent samples for each parameter [120]. Therefore, the number of samples depends on the complexity of the MLA. The MLP, in particular, can have a large number of parameters, if each weight is considered to be a parameter, and several thousand samples might be necessary. As the number of data samples we can collect is limited, due to the effort it takes, MLP might not be the optimal algorithm for our chosen classification problem. The kNN has fewer parameters, so according to this suggestion it may be better suited for small datasets. Another guideline is to use at least 10% more samples than input features [120]. We use an ANOVA filter to select 50 features, which means at least 60 samples are needed for all classification problems. In our study, this premise is fulfilled except for the follow-up experiment. In comparison, other studies collected at least 2000 samples for analysis [5, 13, 72]. If we take this as a benchmark the number of collected samples would need to be doubled.

The influence of prescription was not analyzed. It is important to note that 75% of the participants normally wear glasses. As the eye-tracking device is worn like glasses they interfere with each other. We tested the possibility of wearing both, but the glasses interfered with the tracking of the pupil and no usable data could be collected. It remains to be seen whether the pupil near reflex [32] had a negative influence on the collected data. The experiments

that would probably be the most affected are the decision-making tests. The sound experiments have no changing visuals, therefore the influence should be minimal. During the other experiments, the question was visually displayed, and attempting to focus on the text could have triggered the reflex.

## 8 Conclusion and Future Work

This work examines the feasibility of extracting PII through eye-related data. We assessed the possibility by focusing on four different types of information that we attempted to detect.

### *Is the extraction of age possible?*

We were able to distinguish between the two set age groups. A better classification might be possible by widening the age range and incorporating new features besides the pupil measurements. To get a more precise age classification, a multiclass classification could be used.

### *Is there a difference in eye behavior between the genders?*

We were able to classify the data into two genders. Further testing is needed to ensure generalizability to a larger population. It should be examined whether there is a difference in eye behavior between genders with added visual stimulation, including fixation points as features.

### *Does the eye data show whether caffeine was consumed?*

The evaluation data allowed us to identify caffeine consumers. It should be determined whether the eye data can indicate caffeine consumption. Allowing participants more leeway for natural eye movement could improve classification.

### *Is it possible to infer binary decisions made by the user from eye data?*

At the current state, a certain detection of the decision a participant makes is not possible. Our findings indicate that the implemented classifiers perform better for binary Yes/No decisions than for more complex ones, including three possible answers or different levels of certainty. It also became apparent, that the classification of more specified data, such as gender-specific data, is better. As detecting decisions can be a tool to receive valuable information, further research in this area is warranted.

In conclusion, our study design was able to detect private information in three out of the four tasks we aimed to test. This indicates that the privacy concerns for VR devices are valid. In Adams *et al.* [65], the interviewed people also raised concerns about the data that sensors in VR devices could collect, particularly the outside cameras or the microphone. With the inclusion of eye-tracking in VR headsets more data becomes available for exploitation.

Before conducting further studies, some changes have to be made. Specifically, we need to re-analyze which brightness level should be used during the experiments and attempt to ensure that the collected data is better balanced, so less data has to be discarded during the training process. The length and setting of our experiments appear to be adequate. The site, where we conduct our experiments could be better shielded from outside interference, which could be achieved by relocating to a more remote space. Additionally, distractions through external light could also be negated through relocation.

In future work, we need to adjust our decision-detection methods. Detecting decisions could be used to derive opinions on political topics, personal preferences, or information about living situations. As we can not detect certain information from one question, repeating the question could increase the chance of calculating the correct decision. Consequently, we decided to split one question into twelve sub-questions from which we can derive the answer. The answer would be divided in a fashion that if the main question would be answered one way half the answers would be 'Yes'. While for the complementary answer, the answers to the subquestion would be inverted.

To improve decision-making, we aim to explore the inclusion of more visual stimuli. To reduce indecisiveness when confronted with an unfamiliar topic, we propose presenting a short, informative video on the topic before asking questions. This will give participants the opportunity to consider their standpoint beforehand. Afterward, we would ask questions of a similar makeup to those without video. If standpoint detection is successful, we could consider the feasibility of detecting the standpoint while watching the video.

Further work for this research involves the reevaluation of the MLA. This requires verifying the parameter by implementing new performance metrics and testing a different grid search algorithm. Additionally, new classifiers such as Nearest Neighbors based on radius should be tested. The influence of pre-training models for smaller subsets also needs to be examined.

Additional PII could be derived from the eye data. One potential idea is to detect the native language [80]. Ito *et al.* [80] used fixation on words in the more familiar language to discern the native language. In the current experiment layout, questions are presented as text. By displaying the questions in different languages, we could gather information on the participant's language proficiency based on which language they focus on. This study did not explore the different nationalities reported by the participants due to their diversity, as most are only represented by one individual.

Besides new experiment ideas, new features could be explored. This study's purpose is to examine the feasibility of extracting PII of data collected by a VR device. Head movement, which can be calculated by the position change within the room, can indicate preferences [75]. New sensors are being developed for VR. Facial trackers might provide information about preferences and emotions or help make conclusions about age as demonstrated by Dibeklioglu *et al.* [115]. With the addition of new sensors, many more possibilities to collect data will arise. This poses a threat to users' privacy as it allows for new ways to determine PII without their knowledge.

## Bibliography

- [1] *Vive Wrist Tracker*, HTC America, Inc., 2024. [Online]. Available: <https://business.vive.com/eu/product/vive-wrist-tracker/>, Accessed: Jan. 11, 2024.
- [2] *Top 5 Virtual Reality Trends of 2024: The Future of VR*, Program Ace, 2023. [Online]. Available: <https://program-ace.com/blog/virtual-reality-trends/>, Accessed: Jan. 11, 2024.
- [3] *VIVE Focus 3 Facial Tracker*, en, <https://business.vive.com/eu/product/vive-focus-3-facial-tracker/>, HTC America, Inc., Accessed: Jan. 11, 2024.
- [4] *Eye Tracking on VR (Virtual Reality) headsets*, Pimax Inc. [Online]. Available: <https://pimax.com/de/eye-tracking-in-pimax-crystal-vr-headset/>, Accessed: Jan. 11, 2024.
- [5] H. Zhu, M. Xiao, D. Sherman, and M. Li, "SoundLock: A Novel User Authentication Scheme for VR Devices Using Auditory-Pupillary Response", 2023. DOI: 10.14722/ndss.2023.24298.
- [6] V. Nair, G. M. Garrido, D. Song, and J. O'Brien, "Exploring the Privacy Risks of Adversarial VR Game Design", *Proceedings on Privacy Enhancing Technologies*, vol. 2023, no. 4, pp. 238–256, Oct. 2023. DOI: 10.56553/popets-2023-0108. [Online]. Available: <https://doi.org/10.56553%2Fpopets-2023-0108>.
- [7] M. Erbilek, M. Fairhurst, and M. C. D. C. Abreu, "Age prediction from iris biometrics", in *5th International Conference on Imaging for Crime Detection and Prevention (ICDP 2013)*, 2013, pp. 1–5. DOI: 10.1049/ic.2013.0258.
- [8] B. L. Esther Xiu Wen Wu and S. Magnussen, "Through the eyes of the own-race bias: Eye-tracking and pupillometry during face recognition", *Social Neuroscience*, vol. 7, no. 2, pp. 202–216, 2012. DOI: 10.1080/17470919.2011.596946.
- [9] W. Toivo and C. Scheepers, "Pupillary responses to affective words in bilinguals' first versus second language", *PLoS one*, vol. 14, no. 4, e0210450, 2019. DOI: 10.1371/journal.pone.0210450.
- [10] J. Huijding, B. Mayer, E. H. W. Koster, and P. Muris, "To look or not to look: An eye movement study of hypervigilance during change detection in high and low spider fearful students", *Emotion (Washington, D.C.)*, vol. 11, no. 3, pp. 666–674, 2011. DOI: 10.1037/a0022996.
- [11] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation", *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2008. DOI: 10.1111/j.1469-8986.2008.00654.x.
- [12] A. Babiker, I. Faye, K. Prehn, and A. Malik, "Machine Learning to Differentiate Between Positive and Negative Emotions Using Pupil Diameter", *Frontiers in psychology*, vol. 6, p. 1921, 2015. DOI: 10.3389/fpsyg.2015.01921.
- [13] J. W. de Gee, T. Knapen, and T. H. Donner, "Decision-related pupil dilation reflects upcoming choice and individual bias", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 5, E618–25, 2014. DOI: 10.1073/pnas.1317557111.

- [14] A. Rosner *et al.*, "Ambivalence in decision making: An eye tracking study", *Cognitive Psychology*, vol. 134, p. 101-164, 2022.
- [15] J. Pettiford, R. V. Kozink, A. M. Lutz, S. H. Kollins, J. E. Rose, and F. J. McClernon, "Increases in impulsivity following smoking abstinence are related to baseline nicotine intake and boredom susceptibility", *Addictive behaviors*, vol. 32, no. 10, pp. 2351-2357, 2007. DOI: 10.1016/j.addbeh.2007.02.004.
- [16] N. R. Dias *et al.*, "Anti-saccade error rates as a measure of attentional bias in cocaine dependent subjects", *Behavioural Brain Research*, vol. 292, pp. 493-499, 2015. DOI: 10.1016/j.bbr.2015.07.006.
- [17] D. Bittner, I. Wieseler, H. Wilhelm, M. Riepe, and N. Mueller, "Repetitive Pupil Light Reflex: Potential Marker in Alzheimer's Disease?", *Journal of Alzheimer's disease : JAD*, vol. 42, Jul. 2014. DOI: 10.3233/JAD-140969.
- [18] E. Giza, D. Fotiou, S. Bostantjopoulou, Z. Katsarou, and A. Karlovasitou, "Pupil light reflex in Parkinson's disease: evaluation with pupillometry", *The International journal of neuroscience*, vol. 121, no. 1, pp. 37-43, 2011. DOI: 10.3109/00207454.2010.526730.
- [19] *Personally Identifiable Information (PII)*, U.S. DEPARTMENT OF LABOR. [Online]. Available: <https://www.dol.gov/general/ppii>, Accessed: Jan. 11, 2024.
- [20] X. Li, W. Yi, H.-L. Chi, X. Wang, and A. P. Chan, "A critical review of virtual and augmented reality (VR/AR) applications in construction safety", *Automation in Construction*, vol. 86, pp. 150-162, 2018. DOI: <https://doi.org/10.1016/j.autcon.2017.11.003>.
- [21] S. Benford, C. Greenhalgh, G. Reynard, C. Brown, and B. Koleva, "Understanding and Constructing Shared Spaces with Mixed-Reality Boundaries", *ACM Trans. Comput.-Hum. Interact.*, vol. 5, no. 3, pp. 185-223, Sep. 1998. DOI: 10.1145/292834.292836.
- [22] P. Milgram and H. Colquhoun, "A Taxonomy of Real and Virtual World Display Integration", vol. 1, pp. 1-26, Jan. 2001. DOI: 10.1007/978-3-642-87512-0\_1.
- [23] X. Wang, M. Truijens, L. Hou, Y. Wang, and Y. Zhou, "Integrating Augmented Reality with Building Information Modeling: Onsite construction process controlling for liquefied natural gas industry", *Automation in Construction*, vol. 40, pp. 96-105, 2014. DOI: <https://doi.org/10.1016/j.autcon.2013.12.003>.
- [24] M. F. Bear, B. W. Connors, and M. A. Paradiso, "Das auge", in *Neurowissenschaften: Ein grundlegendes Lehrbuch für Biologie, Medizin und Psychologie*, A. K. Engel, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2018, pp. 309-347, ISBN: 978-3-662-57263-4.
- [25] PDQ Pediatric Treatment Editorial Board, "Retinoblastoma Treatment (PDQ®): Patient Version", *PDQ Cancer Information Summaries [Internet]*, Oct. 2021, Available: [https://www.ncbi.nlm.nih.gov/books/NBK65754/figure/CDR0000258033\\_\\_141/](https://www.ncbi.nlm.nih.gov/books/NBK65754/figure/CDR0000258033__141/), Accessed: 23 Nov. 2023.
- [26] F. Grehn, "Anatomie, Physiologie und Pathophysiologie des Auges", in *Augenheilkunde*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2019, pp. 5-21, ISBN: 978-3-662-59154-3.



- [27] W. Kahle and M. Frotscher, "The Eye", English, in *Color Atlas of Human Anatomy: Vol. 3 Nervous System and Sensory Organs* (Color Atlas of Human Anatomy, Vol. 3: Nervous System and Sensory Organs), Color Atlas of Human Anatomy, Vol. 3: Nervous System and Sensory Organs. Germany: Thieme Medical Publishers, Incorporated, 2015, pp. 341–364, ISBN: 978-3-135-33507-0.
- [28] *Diagram of Extraocular Muscles of the Right Eye*, Encyclopædia Britannica. [Online]. Available: <https://www.britannica.com/science/ophthalmoplegia>, Accessed: Nov. 23, 2023.
- [29] D. Purves *et al.*, "Types of Eye Movements and Their Functions", in *Neuroscience. 2nd edition*, Sinauer Associates, 2001. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK10991/>, Accessed: Feb. 22, 2024.
- [30] F. Grehn, "Pupille", in *Augenheilkunde*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2019, pp. 229–240, ISBN: 978-3-662-59154-3.
- [31] E. Alexandridis, "Die normale Pupille", in *Die Pupille: Physiologie - Untersuchung - Pathologie*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1982, pp. 1–22, ISBN: 978-3-662-00496-8.
- [32] S. Mathôt, "Pupillometry: Psychology, Physiology, and Function", *Journal of Cognition*, vol. 1, no. 1, p. 16, Feb. 2018. DOI: 10.5334/joc.18.
- [33] E. Alexandridis, "Pathologische Pupille", in *Die Pupille: Physiologie — Untersuchung — Pathologie*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1982, pp. 37–77, ISBN: 978-3-662-00496-8.
- [34] W. Leydhecker, "Die Pupille", in *Augenheilkunde*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1990, pp. 92–96, ISBN: 978-3-662-21656-9.
- [35] Z.-H. Zhou, "Introduction", in *Machine Learning*, Z.-H. Zhou, Ed., Singapore: Springer, 2021, pp. 1–24, ISBN: 978-9-811-51967-3.
- [36] Z.-H. Zhou, "Clustering", in *Machine Learning*, Z.-H. Zhou, Ed., Singapore: Springer, 2021, pp. 211–240, ISBN: 978-9-811-51967-3.
- [37] Z.-H. Zhou, "Dimensionality Reduction and Metric Learning", in *Machine Learning*, Z.-H. Zhou, Ed., Singapore: Springer, 2021, pp. 241–264, ISBN: 978-9-811-51967-3.
- [38] A. Amor and L. Liu, *Clustering*, scikit-learn. [Online]. Available: <https://scikit-learn.org/stable/modules/clustering.html>, Accessed: Feb. 07, 2024.
- [39] Z.-H. Zhou, "Support Vector Machine", in *Machine Learning*, Z.-H. Zhou, Ed., Singapore: Springer, 2021, pp. 129–153, ISBN: 978-9-811-51967-3.
- [40] A. Amor and L. Liu, *Support Vector Machines*, scikit-learn. [Online]. Available: <https://scikit-learn.org/stable/modules/svm.html>, Accessed: Feb. 07, 2024.
- [41] Z.-H. Zhou, "Decision Trees", in *Machine Learning*, Z.-H. Zhou, Ed., Singapore: Springer, 2021, pp. 79–102, ISBN: 978-9-811-51967-3.
- [42] A. Amor and L. Liu, *Decision Trees*, scikit-learn. [Online]. Available: <https://scikit-learn.org/stable/modules/tree.html>, Accessed: Feb. 07, 2024.
- [43] Z.-H. Zhou, "Neural Networks", in *Machine Learning*, Z.-H. Zhou, Ed., Singapore: Springer, 2021, pp. 103–128, ISBN: 978-9-811-51967-3.

- [44] P. Baheti, *Activation Functions in Neural Networks - 12 Types & Use Cases*, 2021. [Online]. Available: <https://www.v7labs.com/blog/neural-networks-activation-functions>, Accessed: Feb. 07, 2024.
- [45] A. Amor and L. Liu, *Neural network models (supervised)*, scikit-learn. [Online]. Available: [https://scikit-learn.org/stable/modules/neural\\_networks\\_supervised.html](https://scikit-learn.org/stable/modules/neural_networks_supervised.html), Accessed: Feb. 07, 2024.
- [46] Z.-H. Zhou, "Feature Selection and Sparse Learning", in *Machine Learning*, Z.-H. Zhou, Ed., Singapore: Springer, 2021, pp. 265–285, ISBN: 978-9-811-51967-3.
- [47] J. Brownlee, *How to Choose a Feature Selection Method For Machine Learning*, Aug. 2020. [Online]. Available: <https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/>, Accessed: Jan. 22, 2024.
- [48] H. Ding, P.-M. Feng, W. Chen, and H. Lin, "Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis", *Molecular BioSystems*, vol. 10, no. 8, pp. 2229–2235, 2014.
- [49] Z.-H. Zhou, "Model Selection and Evaluation", in *Machine Learning*, Z.-H. Zhou, Ed., Singapore: Springer, 2021, pp. 25–55, ISBN: 978-9-811-51967-3.
- [50] A. Amor and L. Liu, *Cross-validation: evaluating estimator performance*, scikit-learn. [Online]. Available: [https://scikit-learn.org/stable/modules/cross\\_validation.html](https://scikit-learn.org/stable/modules/cross_validation.html), Accessed: Feb. 07, 2024.
- [51] D. Shah, *Top Performance Metrics in Machine Learning: A Comprehensive Guide*, 2023. [Online]. Available: <https://www.v7labs.com/blog/performance-metrics-in-machine-learning>, Accessed: Feb. 07, 2024.
- [52] H. Zhu, W. Jin, M. Xiao, S. Murali, and M. Li, "BlinKey: A Two-Factor User Authentication Method for Virtual Reality Devices", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, pp. 1–29, Dec. 2020. DOI: 10.1145/3432217.
- [53] S. R. K. Gopal, D. Shukla, J. D. Wheelock, and N. Saxena, "Hidden Reality: Caution, Your Hand Gesture Inputs in the Immersive Virtual World are Visible to All!", in *32nd USENIX Security Symposium (USENIX Security 23)*, Anaheim, CA: USENIX Association, Aug. 2023, pp. 859–876, ISBN: 978-1-939133-37-3. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity23/presentation/gopal>.
- [54] C. Slocum, Y. Zhang, N. Abu-Ghazaleh, and J. Chen, "Going through the motions: AR/VR keylogging from user head motions", in *32nd USENIX Security Symposium (USENIX Security 23)*, Anaheim, CA: USENIX Association, Aug. 2023, pp. 159–174, ISBN: 978-1-939133-37-3. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity23/presentation/slocum>.
- [55] S. Stephenson, B. Pal, S. Fan, E. Fernandes, Y. Zhao, and R. Chatterjee, "SoK: Authentication in Augmented and Virtual Reality", in *2022 IEEE Symposium on Security and Privacy (SP)*, 2022, pp. 267–284. DOI: 10.1109/SP46214.2022.9833742.
- [56] Z. Yu, H.-N. Liang, C. Fleming, and K. L. Man, "An exploration of usable authentication mechanisms for virtual reality systems", in *2016 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, 2016, pp. 458–460. DOI: 10.1109/APCCAS.2016.7804002.

- [57] Y. Gao, W. Wang, V. V. Phoha, W. Sun, and Z. Jin, "EarEcho: Using Ear Canal Echo for Wearable Authentication", *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 3, Sep. 2019. DOI: 10.1145/3351239.
- [58] F. Lin, K. W. Cho, C. Song, W. Xu, and Z. Jin, "Brain Password: A Secure and Truly Cancelable Brain Biometrics for Smart Headwear", in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '18, Munich, Germany: Association for Computing Machinery, 2018, pp. 296–309. DOI: 10.1145/3210240.3210344.
- [59] S. Li, A. Ashok, Y. Zhang, C. Xu, J. Lindqvist, and M. Gruteser, "Whose move is it anyway? Authenticating smart wearable devices using unique head movement patterns", in *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2016, pp. 1–9. DOI: 10.1109/PERCOM.2016.7456514.
- [60] K. Pfeuffer, M. Geiger, S. Prange, L. Mecke, D. Buschek, and F. Alt, "Behavioural Biometrics in VR: Identifying People from Body Motion and Relations in Virtual Reality", Apr. 2019, pp. 1–12. DOI: 10.1145/3290605.3300340.
- [61] F. Mathis, H. I. Fawaz, and M. Khamis, "Knowledge-driven Biometric Authentication in Virtual Reality", in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '20, USA: Association for Computing Machinery, 2020, pp. 1–10. DOI: 10.1145/3334480.3382799.
- [62] A. Giaretta, "Security and Privacy in Virtual Reality—A Literature Survey", *arXiv preprint arXiv:2205.00208*, 2022.
- [63] Y. Zhang, C. Slocum, J. Chen, and N. Abu-Ghazaleh, "It's all in your head(set): Side-channel attacks on AR/VR systems", in *32nd USENIX Security Symposium (USENIX Security 23)*, Anaheim, CA: USENIX Association, Aug. 2023, pp. 3979–3996, ISBN: 978-1-939133-37-3.
- [64] P. Casey, I. Baggili, and A. Yarramreddy, "Immersive Virtual Reality Attacks and the Human Joystick", *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 2, pp. 550–562, 2021. DOI: 10.1109/TDSC.2019.2907942.
- [65] D. Adams, A. Bah, C. Barwulor, N. Musaby, K. Pitkin, and E. M. Redmiles, "Ethics Emerging: the Story of Privacy and Security Perceptions in Virtual Reality", in *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, Baltimore, MD: USENIX Association, Aug. 2018, pp. 427–442, ISBN: 978-1-939133-10-6.
- [66] A. Gallardo *et al.*, "Speculative Privacy Concerns About AR Glasses Data Collection", *Proceedings on Privacy Enhancing Technologies*, vol. 4, pp. 416–435, 2023.
- [67] C. Nwaneri, "Ready lawyer one: Legal Issues in the Innovation of Virtual Reality", *Harvard journal of law & technology*, vol. 30, no. 2, p. 601, 2017.
- [68] L. E. Buck and B. Bodenheimer, "Privacy and Personal Space: Addressing Interactions and Interaction Data as a Privacy Concern", in *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2021, pp. 399–400. DOI: 10.1109/VRW52623.2021.00086.

- [69] J. von Willich, M. Funk, F. Müller, K. Marky, J. Riemann, and M. Mühlhäuser, "You Invaded my Tracking Space! Using Augmented Virtuality for Spotting Passersby in Room-Scale Virtual Reality", in *Proceedings of the 2019 on Designing Interactive Systems Conference*, ser. DIS '19, San Diego, CA, USA: Association for Computing Machinery, 2019, pp. 487–496. DOI: 10.1145/3322276.3322334.
- [70] I. Jarin, Y. Duan, R. Trimananda, H. Cui, S. Elmalaki, and A. Markopoulou, "BehaVR: User Identification Based on VR Sensor Data", *arXiv preprint arXiv:2308.07304*, 2023.
- [71] J. L. Kröger, O. H.-M. Lutz, and F. Müller, "What Does Your Gaze Reveal About You? On the Privacy Implications of Eye Tracking", *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers 14*, pp. 226–241, 2020.
- [72] B. Gingras, M. M. Marin, E. Puig-Waldmüller, and W. T. Fitch, "The Eye is Listening: Music-Induced Arousal and Individual Differences Predict Pupillary Responses", *Frontiers in Human Neuroscience*, vol. 9, 2015. DOI: 10.3389/fnhum.2015.00619.
- [73] A. A. Zekveld, T. Koelewijn, and S. E. Kramer, "The Pupil Dilation Response to Auditory Stimuli: Current State of Knowledge", *Trends in Hearing*, vol. 22, pp. 1–25, 2018. DOI: 10.1177/2331216518777174.
- [74] R. J. Snowden, K. R. O'Farrell, D. Burley, J. T. Erichsen, N. V. Newton, and N. S. Gray, "The pupil's response to affective pictures: Role of image duration, habituation, and viewing mode", *Psychophysiology*, vol. 53, no. 8, pp. 1217–1223, 2016.
- [75] M. Behnke, N. Bianchi-Berthouze, and L. D. Kaczmarek, "Head movement differs for positive and negative emotions in video recordings of sitting individuals", *Scientific reports*, vol. 11, no. 1, p. 7405, 2021.
- [76] S. Berkovsky *et al.*, "Detecting Personality Traits Using Eye-Tracking Data", in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Apr. 2019, pp. 1–12. DOI: 10.1145/3290605.3300451.
- [77] S. Hoppe, T. Loetscher, S. Morey, and A. Bulling, "Eye Movements During Everyday Behavior Predict Personality Traits", *Frontiers in Human Neuroscience*, vol. 12, p. 105, Apr. 2018. DOI: 10.3389/fnhum.2018.00105.
- [78] S. D. Goldinger, Y. He, and M. H. Papesh, "Deficits in cross-race face learning: insights from eye movements and pupillometry", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 35, no. 5, p. 1105, 2009.
- [79] D. Green, Q. Li, J. J. Lockman, and G. Gredebäck, "Culture influences action understanding in infancy: Prediction of actions performed with chopsticks and spoons in Chinese and Swedish infants", *Child development*, vol. 87, no. 3, pp. 736–746, 2016.
- [80] A. Ito, M. Pickering, and M. Corley, "Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study", *Journal of Memory and Language*, vol. 98, Sep. 2017. DOI: 10.1016/j.jml.2017.09.002.
- [81] K. Kunze, H. Kawaichi, K. Yoshimura, and K. Kise, "Towards inferring language expertise using eye tracking", Apr. 2013, pp. 217–222. DOI: 10.1145/2468356.2468396.

- [82] C. Sforza, M. Rango, D. Galante, N. Bresolin, and V. F. Ferrario, "Spontaneous blinking in healthy persons: an optoelectronic study of eyelid motion", *Ophthalmic and Physiological Optics*, vol. 28, no. 4, pp. 345–353, 2008.
- [83] F. J. Mercer Moss, R. Baddeley, and N. Canagarajah, "Eye Movements to Natural Images as a Function of Sex and Personality", *PloS one*, vol. 7, no. 11, e47870, 2012.
- [84] Y. M. Hwang and K. C. Lee, "Using an Eye Tracking Approach to Explore Gender Differences in Visual Attention and Shopping Attitudes in an Online Shopping Environment", *International Journal of Human-Computer Interaction*, vol. 34, no. 1, pp. 15–24, 2018.
- [85] B. Cheval, E. Grob, J. Chanal, P. Ghisletta, F. Bianchi-Demicheli, and R. Radel, "Homophobia Is Related to a Low Interest in Sexuality in General: An Analysis of Pupillometric Evoked Responses", *The Journal of Sexual Medicine*, vol. 13, no. 10, pp. 1539–1545, 2016. DOI: <https://doi.org/10.1016/j.jsxm.2016.07.013>.
- [86] T. Godet and G. Niveau, "Eye tracking and child sexual offenders: A systematic review", *Forensic Sciences Research*, vol. 6, no. 2, pp. 133–140, 2021. DOI: [10.1080/20961790.2021.1940737](https://doi.org/10.1080/20961790.2021.1940737).
- [87] G. Rieger and R. C. Savin-Williams, "The eyes have it: Sex and sexual orientation differences in pupil dilation patterns", *PloS one*, vol. 7, no. 8, e40256, 2012.
- [88] S. Jain *et al.*, "Pupillary unrest correlates with arousal symptoms and motor signs in Parkinson disease", *Movement disorders*, vol. 26, no. 7, pp. 1344–1347, 2011.
- [89] D. F. Fotiou *et al.*, "Pupil reaction to light in Alzheimer's disease: evaluation of pupil size changes and mobility", *Aging clinical and experimental research*, vol. 19, pp. 364–371, 2007.
- [90] A. L. Bertrand, J. B. S. Garcia, E. B. Viera, A. M. Santos, and R. H. Bertrand, "Pupillometry: The Influence of Gender and Anxiety on the Pain Response", *Pain Physician*, vol. 16, no. 3, E257–E266, May 2013. DOI: [10.36076/ppj.2013/16/E257](https://doi.org/10.36076/ppj.2013/16/E257).
- [91] S. Abokyi, J. Owusu-Mensah, and K. Osei, "Caffeine intake is associated with pupil dilation and enhanced accommodation", *Eye*, vol. 31, no. 4, pp. 615–619, 2017.
- [92] B. Wilhelm, G. Stuibler, H. Lüdtke, and H. Wilhelm, "The effect of caffeine on spontaneous pupillary oscillations", *Ophthalmic and Physiological Optics*, vol. 34, no. 1, pp. 73–81, 2014. DOI: <https://doi.org/10.1111/opo.12094>.
- [93] *Pupil Core*, Pupil Labs, 2023. [Online]. Available: <https://pupil-labs.com/products/core/>, Accessed: Oct. 10, 2023.
- [94] M.-K. Jeon and J.-W. Oh, "Study on listening to white noise of nursing college students and improvement of concentration.", *Medico-Legal Update*, vol. 19, no. 1, 2019.
- [95] *Core - Pupil Player - Pupil Labs Docs*, Pupil Labs, 2023. [Online]. Available: <https://docs.pupil-labs.com/core/software/pupil-player/>, Accessed: Jan. 18, 2023.
- [96] M. Oliva and A. Anikin, "Pupil dilation reflects the time course of emotion recognition in human vocalizations", *Scientific Reports*, vol. 8, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:256947365>.
- [97] A. Baird, E. Parada-Cabaleiro, C. Fraser, S. Hantke, and B. Schuller, "The Perceived Emotion of Isolated Synthetic Audio: The EmoSynth Dataset and Results", in *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, ser. AM'18, Wrexham, United Kingdom: ACM, 2018, 7:1–7:8. DOI: [10.1145/3243274.3243277](https://doi.org/10.1145/3243274.3243277).

- [98] J. T. Martin and M. Spitschan, *Pyplr (version 1.0.2)*, version v1.0.2, Aug. 2021. DOI: 10.5281/zenodo.5749833.
- [99] *Normalisierung*, Google LCC, 2022. [Online]. Available: <https://developers.google.com/machine-learning/data-prep/transform/normalization?hl=de>, Accessed: Jan. 18, 2024.
- [100] T. Bäckström *et al.*, *Introduction to Speech Processing*, 2nd ed. 2022. DOI: 10.5281/zenodo.6821775.
- [101] P. Antoniadis, *Introduction to Curve Fitting*, 2023. [Online]. Available: <https://www.baeldung.com/cs/curve-fitting>, Accessed: Jan. 18, 2024.
- [102] L. Buitinck *et al.*, "API design for machine learning software: Experiences from the scikit-learn project", in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.
- [103] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [104] A. Amor and L. Liu, *sklearn.model\_selection.HalvingGridSearchCV*, scikit-learn. [Online]. Available: [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.HalvingGridSearchCV.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.HalvingGridSearchCV.html), Accessed: Feb. 07, 2024.
- [105] A. Amor and L. Liu, *sklearn.tree.DecisionTreeClassifier*, scikit-learn. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>, Accessed: Feb. 07, 2024.
- [106] A. Amor and L. Liu, *sklearn.cluster.KMeans*, scikit-learn. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>, Accessed: Feb. 07, 2024.
- [107] A. Amor and L. Liu, *sklearn.svm.SVC*, scikit-learn. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>, Accessed: Feb. 07, 2024.
- [108] A. Amor and L. Liu, *sklearn.neighbors.KNeighborsClassifier*, scikit-learn. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>, Accessed: Feb. 07, 2024.
- [109] A. Amor and L. Liu, *sklearn.neural\_network.MLPClassifier*, scikit-learn. [Online]. Available: [https://scikit-learn.org/stable/modules/generated/sklearn.neural\\_network.MLPClassifier.html](https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html), Accessed: Feb. 07, 2024.
- [110] A. Amor and L. Liu, *Nearest Neighbors*, scikit-learn. [Online]. Available: <https://scikit-learn.org/stable/modules/neighbors.html>, Accessed: Feb. 07, 2024.
- [111] S. Ray, *8 Ways to Improve Accuracy of Machine Learning Models*. [Online]. Available: <https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/>, Accessed: Feb. 07, 2024.
- [112] *Durchschnittsgrößen von Mann und Frau*, Laenderdaten.info. [Online]. Available: <https://www.laenderdaten.info/durchschnittliche-koerpergroessen.php>, Accessed: Feb. 07, 2024.
- [113] C. I. o. H. R. Government of Canada, *What is gender? What is sex? - CIHR*, Last Modified: 2023-05-08, Jan. 2014. [Online]. Available: <https://cihr-irsc.gc.ca/e/48642.html>, Accessed: Feb. 22, 2024.

- [114] J. Evans, J. R. Richards, and A. S. Battisti, "Caffeine", eng, in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2024. [Online]. Available: <http://www.ncbi.nlm.nih.gov/books/NBK519490/>, Accessed: 07 Feb. 2024.
- [115] H. Dibeklioglu, T. Gevers, A. A. Salah, and R. Valenti, "A smile can reveal your age: Enabling facial dynamics in age estimation", in *Proceedings of the 20th ACM International Conference on Multimedia*, ser. MM '12, Nara, Japan: Association for Computing Machinery, 2012, pp. 209–218. DOI: 10.1145/2393347.2393382.
- [116] M. M. Dehshibi and A. Bastanfard, "A new algorithm for age recognition from facial images", *Signal Processing*, vol. 90, no. 8, pp. 2431–2444, 2010. DOI: <https://doi.org/10.1016/j.sigpro.2010.02.015>.
- [117] A. Amor and L. Liu, *sklearn.model\_selection.LearningCurveDisplay*, scikit-learn. [Online]. Available: [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.LearningCurveDisplay.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.LearningCurveDisplay.html), Accessed: Feb. 22, 2024.
- [118] B. Libet, "Do we have free will?", *Journal of consciousness studies*, vol. 6, no. 8-9, pp. 47–57, 1999.
- [119] A. Amor and L. Liu, *Tuning the hyper-parameters of an estimator*, scikit-learn. [Online]. Available: [https://scikit-learn.org/stable/modules/grid\\_search.html](https://scikit-learn.org/stable/modules/grid_search.html), Accessed: Feb. 07, 2024.
- [120] J. Brownlee, *How Much Training Data is Required for Machine Learning?*, Jul. 2017. [Online]. Available: <https://machinelearningmastery.com/much-training-data-required-machine-learning/>, Accessed: 07 Feb. 2024.

# Appendix A: Survey Questionair

## A.1 Demographic Survey

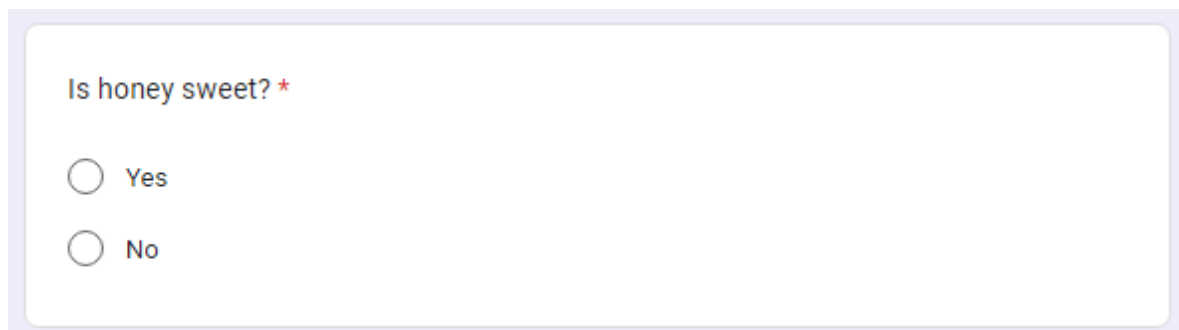
This is the list of Questions asked for the survey about the participants' demographic data:

1. What is your Biological Gender (Sex)?
2. What is your Declared Gender?
3. What is your Weight (kg)?
4. What is your Height (cm)?
5. How old are you?
6. What is your Nationality?
7. Did you consume caffeine today?
8. What is your Native Language?
9. What other Languages do you Speak?
10. Do you wear Glasses?
11. If yes, what kind of lens do you have?
  - a) Single Vision Lenses
  - b) Bifocal Lenses
  - c) Multifocal Lenses
  - d) none of these
12. Are you Far- or Nearsighted?

## A.2 Ground Truth Survey [Digital Appendix]

We used Google Forms to create a survey, in which every question and statement was presented with the possible answers. Figure A.1 shows the visuals for an example question.

Due to the extensive length of the full survey, which includes more questions than could be asked of a single participant, the survey will not be printed in the thesis. Please refer, to the digital Appendix for the complete list in the document: GroundTruthSurvey.pdf.



The image shows a screenshot of a Google Form question. The question text is "Is honey sweet? \*" in a dark grey font. Below the question, there are two radio button options: "Yes" and "No". Both radio buttons are currently unselected. The entire form area is enclosed in a light purple border.

**Figure A.1:** Visuals of an example question from the Ground Truth Survey



## Appendix B: Study Material [Digital Appendix]

The stimuli used in the experiments were presented in a video. Each experiment has an accompanying video in an 'mp4' file, which can be found in the folder 'Experiment\_Videos' in the Digital Appendix.

Table B.1 shows the videos that were created during the experiment design, with their file name. They are sorted according to the phase in which the experiments were conducted.

Table B.2 lists all experiments created for the preliminary study, with their file name. It is important to note that more videos were designed than were later used in the study, due to the set time limit.

The experiment name consists of the Type of experiment denoted by the letter (A, B, C). The first number shows the revision of the experiment video, while the last number symbolizes the number it had during the design phase.

**Table B.1:** Overview for the videos used in the design process

|  | <b>Experiment</b> | <b>Description</b>                    | <b>File Name</b>                  |
|--|-------------------|---------------------------------------|-----------------------------------|
| <b>Protocol 1</b>                          | <b>EX001</b>      | Time for stabilizing the pupil        | EX001_PauseLength.mp4             |
|  | <b>EX002</b>      | Illumination through background color | EX002_GrayScale.mp4               |
|  | <b>EX003</b>      | comfortable color for focus point     | EX003_ColorFixiation.mp4          |
| <b>Follow-up experiments to Protocol 1</b> | <b>EX004</b>      | Black background, Dark green point    | EX004_BlackDarkGreen.mp4          |
|  | <b>EX004</b>      | Black background, Vivid green point   | EX004_BlackVividGreen.mp4         |
|  | <b>EX004</b>      | Gray background, Dark green point     | EX004_GrayDarkGreen.mp4           |
|  | <b>EX004</b>      | Gray background, Vivid green point    | EX004_GrayVividGreen.mp4          |
| <b>Protocol 2</b>                          | <b>EX005</b>      | Questions                             | EX005_Questions.mp4               |
|  | <b>EX006</b>      | Statements                            | EX006_Statements.mp4              |
|  | <b>EX007</b>      | Controversial questions               | EX007_ControversialQuestions.mp4  |
|  | <b>EX008</b>      | Controversial statements              | EX008_ControversialStatements.mp4 |

Table B.2: Overview for the videos created for the preliminary study

| Type       | Experiment | Description                | File Name                         |
|------------|------------|----------------------------|-----------------------------------|
| Sound      | EXA24      | Sounds                     | EXA24_Sounds1.mp4                 |
|            | EXA44      | Frequencies                | EXA44_Frequencies.mp4             |
| Questions  | EXB45      | General knowledge          | EXB45_Questions4.mp4              |
|            | EXB55      | General knowledge          | EXB55_Questions5.mp4              |
|            | EXB65      | General knowledge          | EXB65_Questions6.mp4              |
|            | EXB75      | Debated, non-controversial | EXB75_DebateQuestions.mp4         |
|            | EXB85      | Personal                   | EXB85_PersonalQuestions.mp4       |
|            | EXB17      | Controversial              | EXB17_ControversialQuestions.mp4  |
| Statements | EXC46      | General knowledge          | EXC46_Statements4.mp4             |
|            | EXC56      | General knowledge          | EXC56_Statements5.mp4             |
|            | EXC66      | General knowledge          | EXC66_Statements6.mp4             |
|            | EXC76      | Debated, non-controversial | EXC76_DebateStatements.mp4        |
|            | EXC86      | Personal                   | EXC86_PersonalStatements.mp4      |
|            | EXC08      | Controversial statements   | EXC08_ControversialStatements.mp4 |

## Statutory Declaration in Lieu of an Oath

I – Alexandra Lengert – do hereby declare in lieu of an oath that I have composed the presented work independently on my own and without any other resources than the ones given.

All thoughts taken directly or indirectly from external sources are correctly acknowledged.

This work has neither been previously submitted to another authority nor has it been published yet.

Mittweida, 24. February 2024

Location, Date

Alexandra Lengert, B.Sc.