Universitat Politècnica de Catalunya
Facultat d'Informàtica de Barcelona
Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona
Facultat de Matemàtiques i Estadística

Degree in Data Science and Engineering

Bachelor's Degree Thesis

# Computer Screenshot Classification for Boosting ADHD Productivity in a VR environment

**Gonzalo Córdova Pou**

Supervised by Silverio Martínez-Fernández
Department of Service and Information System Engineering
Co-directed by David Shepherd (LSU), Juliana Souza (VCU)

June, 2023

## Aknowledgements

I would like to express my heartfelt appreciation to everyone who played a significant role in the success of this research project.

First and foremost, I am deeply grateful to my supervisor, Silverio Martínez-Fernández. His exceptional guidance and unwavering trust have propelled me forward and shaped the outcomes of this research. His attentive dedication as a mentor has been invaluable to me.

I am also immensely grateful to my co-supervisors, David Shepherd and Juliana Souza. David provided me with an exciting opportunity and opened the door to a new country and possibilities. Juliana's extensive knowledge of the subject and attentive personality have been instrumental in the day-to-day progress of this work.

A special thanks goes to the collaborators and colleagues from Virginia Commonwealth University, especially those involved in the Alerta project. Their active participation, dedication, and expertise have enriched this research endeavor in countless ways. I am truly grateful to partners like Miles, Enrique, Boden, and David, among others, for their invaluable contributions.

I would like to extend my appreciation to the GESSI Research Group at Universitat Politècnica de Catalunya (UPC) for their support through the Scholarship for Research Initiation (INIREC). Their investment in this work has been essential, allowing me to pursue this research with passion and dedication.

To all the participants and study subjects who generously volunteered their time and participation, I express my sincere gratitude. Your willingness to be a part of this research has been integral to its success.

I am sincerely grateful to the Data Science  Engineering Degree program at UPC, including the teachers and my classmates, for their support and collaboration. Your contributions and shared knowledge have greatly influenced my journey in this field.

Lastly, I want to express my deepest gratitude to my family members, friends, and loved ones. Your unwavering support, understanding, and encouragement throughout this endeavor have meant the world to me. Your love, patience, and belief in me have been my constant source of strength, and I am eternally grateful for your presence in my life.

## Abstract

Individuals with ADHD face significant challenges in their daily lives due to difficulties with attention, hyperactivity, and impulsivity. These challenges are especially pronounced in the workplace or educational settings, where the ability to sustain attention and manage time effectively is crucial for success. Virtual reality (VR) software has emerged as a promising tool for improving productivity in individuals with ADHD. However, the effectiveness of such software depends on the identification of potential distractions and timely intervention.

The proposed computer screenshot classification approach addresses this need by providing a means for identifying and analyzing potential distractions within VR software. By integrating Convolutional Neural Networks (CNNs), Optical Character Recognition (OCR), and Natural Language Processing (NLP), the proposed approach can accurately classify screenshots and extract features, facilitating the identification of distractions and enabling timely intervention to minimize their impact on productivity.

The implications of this research are significant, as ADHD affects a substantial portion of the population and has a significant impact on productivity and quality of life. By providing a novel approach for studying, detecting, and enhancing productivity, this research has the potential to improve outcomes for individuals with ADHD and increase the efficiency and effectiveness of workplaces and educational settings. Moreover, the proposed approach holds promise for wider applicability to other productivity studies involving computer users, where the classification of screenshots and feature extraction play a crucial role in discerning behavioral patterns.

### Keywords

Convolutional Neural Networks (CNN), Natural Language Processing (NLP), Optical Character Recognition (OCR), Virtual Reality (VR), Attention-Deficit/Hyperactivity Disorder (ADHD), Screenshots, Deep Learning, Machine Learning

## Resumen

Las personas con TDAH se enfrentan a importantes retos en su vida diaria debido a sus dificultades de atención, hiperactividad e impulsividad. Estos retos son especialmente pronunciados en el lugar de trabajo o en entornos educativos, donde la capacidad de mantener la atención y gestionar el tiempo de forma eficaz es crucial para el éxito. El software de realidad virtual (RV) se ha revelado como una herramienta prometedora para mejorar la productividad de las personas con TDAH. Sin embargo, la eficacia de dicho software depende de la identificación de distracciones potenciales y de la intervención oportuna.

El enfoque de clasificación de capturas de pantalla de ordenador propuesto aborda esta necesidad proporcionando un medio para identificar y analizar las distracciones potenciales dentro del software de RV. Mediante la integración de redes neuronales convolucionales (CNN), el reconocimiento óptico de caracteres (OCR) y el procesamiento del lenguaje natural (NLP), el enfoque propuesto puede clasificar con precisión las capturas de pantalla y extraer características, facilitando la identificación de las distracciones y permitiendo una intervención oportuna para minimizar su impacto en la productividad.

Las implicaciones de esta investigación son importantes, ya que el TDAH afecta a una parte sustancial de la población y tiene un impacto significativo en la productividad y la calidad de vida. Al proporcionar un

enfoque novedoso para estudiar, detectar y mejorar la productividad, esta investigación tiene el potencial de mejorar los resultados para las personas con TDAH y aumentar la eficiencia y eficacia de los lugares de trabajo y los entornos educativos. Además, el enfoque propuesto promete una mayor aplicabilidad a otros estudios de productividad en los que participen usuarios de ordenadores, en los que la clasificación de capturas de pantalla y la extracción de características desempeñan un papel crucial a la hora de discernir patrones de comportamiento.

### Palabras clave

Redes neuronales convolucionales, Procesamiento del lenguaje natural, Reconocimiento óptico de caracteres, Realidad virtual, Trastorno por déficit de atención e hiperactividad (TDAH), Capturas de pantalla, Aprendizaje Profundo, Aprendizaje Automático

## Resum

Les persones amb TDAH s'enfronten a reptes importants en la seva vida diària a causa de les dificultats d'atenció, hiperactivitat i impulsivitat. Aquests reptes són especialment pronunciats al lloc de treball o en entorns educatius, on la capacitat de mantenir l'atenció i gestionar el temps de manera eficaç és crucial per a l'èxit. El software de realitat virtual (RV) s'ha revelat com a eina prometedora per millorar la productivitat de les persones amb TDAH. Tanmateix, l'eficàcia del software esmentat depèn de la identificació de distraccions potencials i de la intervenció oportuna.

L'enfocament de classificació de captures de pantalla d'ordinador proposat aborda aquesta necessitat proporcionant un mitjà per identificar i analitzar les distraccions potencials dins del programari de RV. Mitjançant la integració de xarxes neuronals convolucionals (CNN), el reconeixement òptic de caràcters (OCR) i el processament del llenguatge natural (NLP), l'enfocament proposat pot classificar amb precisió les captures de pantalla i extreure'n característiques, facilitant la identificació de les distraccions i permetent una intervenció oportuna per minimitzar-ne l'impacte en la productivitat.

Les implicacions d'aquesta investigació són importants, ja que el TDAH afecta una part substancial de la població i té un impacte significatiu a la productivitat i la qualitat de vida. En proporcionar un enfocament nou per estudiar, detectar i millorar la productivitat, aquesta investigació té el potencial de millorar els resultats per a les persones amb TDAH i augmentar l'eficiència i l'eficàcia dels llocs de treball i els entorns educatius. A més, l'enfocament proposat promet una aplicabilitat més gran a altres estudis de productivitat en què participin usuaris d'ordinadors, en què la classificació de captures de pantalla i l'extracció de característiques tenen un paper crucial a l'hora de discernir patrons de comportament.

### Paraules clau

Xarxes neuronals convolucionals, Processament del llenguatge natural, Reconeixement òptic de caràcters, Realitat virtual, Trastorn per dèficit d'atenció i hiperactivitat (TDAH), Captures de pantalla, Aprenentatge Profund, Aprenentatge Automàtic

# Contents

# 1. Introduction

## 1.1 ADHD

Attention-Deficit/Hyperactivity Disorder (ADHD) is one of the most common neurobehavioral disorders. ADHD has a 9.4% prevalence rate, or 6.1 million children and adolescents and 2.5% of adults in the United States alone [22]. ADHD is characterized by symptoms of inattention, hyperactivity, and impulsivity. Symptoms of hyperactivity/impulsivity often improve, but symptoms of inattention persist into adulthood and are strongly related to impairment [44]. People with ADHD can create a strong concentration in something they are interested in, which is called hyper-focus. Still, in daily activities like studies or work, they need a big effort to keep their attention and not drift out. The usual treatments include medicines and regular neurologists, psychiatrists, and therapist visits.

According to the Diagnostic and Statistical Manual of Mental Disorders, fifth edition (DSM-V) [5] "ADHD is a neurodevelopmental disorder defined by impairing levels of inattention, disorganization, and/or hyperactivity-impulsivity". The first clinical description of the disorder was registered by George Still in 1902 as a defect of moral control [7, 45]. In 1968, the DSM-II presents a "hyperkinetic reaction of childhood" disorder, including in same the definition the problems with attention, organization, and hyperactivity [7, 45, 2]. Later on, in 1980, the DSM-III presented the name "Attention Deficit Disorder (ADD) (with or without hyperactivity)", separating the disorder into two types [45, 4]. After discussions about defining it into two different disorders, the revision DSM-III-R presented the name "Attention deficit-Hyperactivity Disorder (ADHD)", unifying the definition [3].

For decades, ADHD was considered a disorder of childhood that remitted prior to adulthood. However, longitudinal studies following youth with ADHD into adulthood have found that symptoms did not remit, with 60% of adults continuing to meet full diagnostic criteria [8]. More recent data shows that ADHD symptoms of inattention fluctuate across time, such that although 30% may meet criteria for remission at a given point in time, only 9% demonstrate sustained recovery [64]. These and other longitudinal studies also demonstrate that adults with ADHD experience significant impairment across multiple domains of functioning [11, 64].

Impairment is particularly pronounced at work. Adults with ADHD face significantly lower occupational attainment, more job instability, and significantly impaired job performance relative to their peers [33]. These challenges lead to financial difficulties, and adults with ADHD have significantly lower annual income and are more likely to be on public aid [32]. A recent population-based study with over 1 million participants found that individuals with ADHD had a 17 percent lower income, significantly more days of unemployment, and a much higher likelihood of receiving disability pension [38]. Importantly, income was lower and unemployment elevated relative to controls with the same levels of educational attainment. Cost analyses suggest that the human capital value of lost work performance associated with ADHD is $4,336 per worker per year [41], with symptoms of inattention serving as the strongest predictor of work performance [30].

Medication, typically stimulants, is an effective treatment for ADHD symptoms and there is some evidence that it has a positive impact on occupational functioning [33]. However, due to a variety of factors including side effects, more than 50% of youth who take ADHD medications discontinue prior to adulthood [25, 49]. Further, medication does not normalize ADHD symptoms, and adults taking ADHD medications continue to experience significant impairment [28, 33]. Also, after the treatment with medication is interrupted, the good effects achieved stop [34].

Non-medication therapeutic approaches have only started to emerge in the last decade with a recent

review identifying 53 total peer-reviewed studies [31]. Cognitive Behavior Therapy (CBT) delivered by therapists in outpatient settings has the strongest evidence base with several randomized controlled trials. Meta-analyses show that CBT is superior to control and produces significant improvement in self-reported measures of ADHD symptoms [75]. However, therapy for ADHD is costly and resource intensive to implement and is not available in many locations. Also, due to issues of stigma, many adults with ADHD will not pursue treatment in outpatient community settings regardless of availability [29]. Overall, access to evidence-based treatment is low [15, 21] and non-adherence and treatment dropout is high [9, 23].

## 1.2 The Alerta Project

This research project is part of a larger project called "Alerta" carried out at the College of Engineering at Virginia Commonwealth University. The Alerta project aims at using virtual reality (VR) technology to treat students with Attention-Deficit Hyperactivity Disorder (ADHD), with the idea of creating a more favorable environment for the education of such students. This would help these students to concentrate on their educational tasks without tutor supervision.

Alerta is both composed of software and hardware. The software is a three-dimensional environment developed with Unity [69] game engine which consists of a study room with meticulously designed elements, illumination, and a calming atmosphere. The VR environment allows users to customize their study space by adjusting ambient noise. The Alerta project also includes the hardware component, which involves using a Varjo XR-3 headset [51]. Previous related work has shown that the Varjo XR-3 performed significantly better on all foveal vision tests compared to other headsets in the market [40]. This makes it an ideal choice for creating a highly immersive and effective virtual environment for the treatment of ADHD.

Figure 1: Varjo XR-3, headset used for the Alerta project.

The three-dimensional environment is designed to be a mix of virtual and real-world elements. The environment contains a screen of a regular computer inside, which the user will use to work. The Unity program runs on that computer. Additionally, the environment mixes reality to allow the user to see the keyboard, as shown in Figure 3.

While it is true that the TFG is developed as part of a team and a more general project, the specific contributions of this TFG are well-defined and the author's own, with the support and supervision of the VCU team and the UPC supervisor. The challenge has been proposed by the supervisor, and the experiments and solutions are presented by the author with implementation assistance from the team.

Figure 2: Alerta project user during experiment session at VCU.

Figure 3: Alerta project user-view during experiment session at VCU.

# 2. Related work

Screenshots are a popular way of capturing digital information and are frequently used for a variety of purposes. However, the nature of screenshots presents a significant challenge in terms of content classification due to their inherent heterogeneity. The development of automated techniques for screenshot classification and analysis has been of great interest to researchers in recent years. Nonetheless, we have found it to be a rather unexplored domain, especially that of computer screenshots. There is a need for techniques that can automatically classify screenshots based on their contents, enabling the identification of different types of information.

In this section, we will delve into the intersection of Virtual Reality (VR) and Attention-Deficit/Hyperactivity Disorder (ADHD). Furthermore, we will examine in more detail how convolutional neural networks (LeCun et al. [46]) and other deep learning techniques have been applied to computer screenshot classification tasks such as web page and game genre classification and dark propaganda detection. Additionally, we will further explore the classification and analysis of computer screenshots in general, with a focus on the potential of CNNs for this task and see how further research is necessary to explore the classification and analysis of computer screenshots in general.

If we open the scope to other types of screenshots (mainly smartphone screenshots), there is more work that has been done. The analysis of smartphone screenshots is led by The Human Screenome Project presented in [59] by Reeves et al. at the Screenomics Lab at Stanford University.

By surveying these related works, we gain valuable insights into the state-of-the-art methodologies and identify potential directions for advancing the classification and analysis of computer screenshots as well as exploring the untapped potential of VR in alleviating ADHD symptoms.

## 2.1 Virtual Reality and ADHD

Although this work is focused on the models that will be deployed in the Virtual Reality environment, it is important to know how VR has been used to help people with ADHD in different circumstances. Regarding diagnostics, there are several works with good results. The common idea is that VR tests are faster, more accurate, and more interesting than paper-based ones. Pollak, Yehuda, et al. 2009 [54] describes a study comparing the Test of Variables of Attention - Continuous Performance Tests (CPT) paper-based and VR-CPT, showing that the results from the tests are similar. Still, the subjects affirm that the experience with the VR test is much more enjoyable. Rodríguez, Celestino, et al. 2018 [1] claim that their VR approach Aula Nesplora got even better results in differentiating children aged between 6 and 16 years old with ADHD or without it better than a traditional CPT (Test of Variables of Attention; TOVA).

Wiguna, Tjhin, et al. 2020 [72] describe a VR tool prototype that uses deep learning and combines qualitative and quantitative approaches in a digital game to apply VR-CPT in children. Fang, Yantong, Dai Han, and Hong Luo 2019 [27] present a VR CPT for school-aged children with learning problems that objectively evaluated inattention, hyperactivity, and impulsivity. Yeh, Shih-Ching, et al. 2012 [73] present a VR tool that uses a classroom environment to apply auditory and visual tests on attention and cognitive function, such as listening tests, CPT tests, executive tests, and visual memory tests; also, the tool includes elements and sounds in the environment to observe the attention disruption. Lee, Hansey, et al 2017 [47] used a similar environment and eye tracking and added the electroencephalogram (EEG) signal acquisition technologies to the VR-CPT and the Wisconsin card sorting test (WCST) to diagnose and test people with ADHD.

VR is also used for rehabilitation/treatment. Lee, J. M., et al 2001 [48] present an experiment where teenagers with ADHD solve tasks in a VR environment to increase attention and decrease impulsivity. The CPT tests before and after the experiment show considerable positive changes. Rizzo, Albert A., et al. 2000 [60] describe a detailed study with a virtual classroom environment to diagnose and evaluate different groups of ADHD patients, showing the VR's high potential to treat and analyze these groups. Parsons, Thomas D., et al. 2007 [61] and Adams, Rebecca, et al. 2019 [1] also use a virtual classroom, this time to show students' different behavior and response with and without ADHD.

Azadeh Bashiri, Marjan Ghazisaeedi, and Leila Shahmoradi 2017 [10] presented a review of journal papers published between 2000 and 2017 that includes the keywords "ADHD", "Virtual Reality", and "children", focused on rehabilitation. They listed 20 relevant studies from 341 founded works, concluding that VR tools can help create a safe, adequate, and personalized environment to work in. Furthermore, it provides fast feedback and increases students' motivation. It is adequate to apply CPT tests and improve the children's behavior.

Some researchers, such as Cho, Baek Hwan, et al. 2002 [18] and Skalski, Sebastian, et al. 2021 [65], go a step further in the medical aspects of ADHD and use VR to try to stimulate the brain to produce certain substances and reduce inattention by biofeedback.

## 2.2 The 'screenome' approach

In recent years, the field of digital experience has seen the emergence of a new approach called "screenomics," as proposed by Ram et al. [57] and Reeves et al. [58], as part of The Human Screenome Project [59]. This approach involves studying individuals' digital experiences, specifically their unique "screenome", which is the record of experiences on digital devices with screens. Screenomes are made up of sequences of screenshots taken frequently at short intervals of 5 seconds, which provide "the raw material and time-series records needed for fine-grained quantitative and qualitative analysis of fast-changing and highly idiosyncratic digital lives".

To extract maximum intelligence from each screenshot, a custom-designed module wrapped around open-source tools for image and document processing is used to sequence each screenshot. Various techniques for feature extraction are then implemented, including clustering smartphone screenshots with active learning and unsupervised techniques (Chiatti et al. [16]), text extraction techniques applied to smartphone screenshots (Chiatti et al. [17]), and template matching methods for logo detection (Culjak et al. [20]). Graphical features of each screenshot, such as image complexity or image velocity, are quantified in a variety of ways. Text is also quantified in a variety of ways, such as word count or word velocity, with natural language processing methods (NLP) used to interpret the text with respect to sentiment or other language characteristics.

The authors used screenshots for both multiclass and binary classification in their study. To develop meaningful descriptions of the media on the screen, selected subsets of the screenshots were manually labeled with codes that map to specific behavioral taxonomies (e.g., emailing, browsing, shopping) and content categories (e.g., food, health). This approach allowed for multiclass classification of the screenshots, where each screenshot could be labeled with multiple tags or codes that describe the action and/or content depicted in the screenshot.

In addition, the authors used screenshots to classify instances of production or consumption by labeling the screenshots with respect to whether the user was producing or consuming content. This approach allowed for binary classification of the screenshots, where each screenshot was labeled as either an instance of production or consumption. The authors manually labeled 27,000 screenshots to develop ground-truth data that were used to train an extreme gradient boosting model, which accurately classified screenshots as production or consumption with 99.2% accuracy.

The resulting ensemble of text snippets and image data (e.g., number of faces) were then compiled into Unicode text files, one for each screenshot, that were integrated with metadata. The high frequency of screenshot sampling allows for analysis at multiple time scales, from seconds to hours to days to months ([56] Ram & Reeves, 2018).

The Human Screenome Project [59] is a significant initiative aimed at understanding digital experiences through the extraction of features from screenshots. It is important to note that the extraction of features and subsequent analysis is carried out after the use of the device and not during it. However, in this project, we propose an alternative approach where the extraction of features is viewed as a task that can run in parallel to the use of the device.

By integrating the extraction of features as a task that can run alongside the use of the device, we believe that we can gain valuable insights into how the user is interacting with the device and program according to interactions in the software, app, program, or virtual reality environment being used. This approach would allow us to not only better understand how digital experiences are constructed, but also to potentially improve the user's experience in real time.

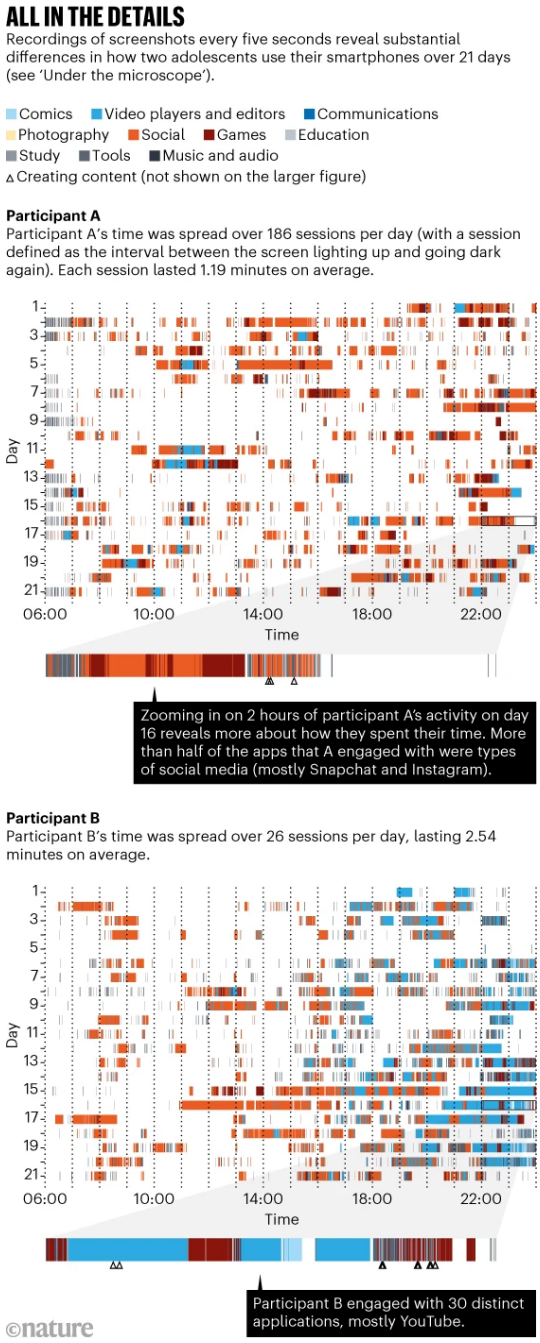


Figure 4: Screenome diagram presented in Nature journal [59].

## 2.3 Computer Vision applied to screenshots

The Human Screenome Project shares a common goal with this project of using images, more precisely screenshots, to extract insights. Deep learning methods, particularly convolutional neural networks, have shown significant success in various visual problems, including object detection, classification, and face recognition. Transfer learning is being used to save time and costs as new CNN architectures are proposed for different problems. In this context, we review some of the related work on screenshot classification, highlighting the need for further exploration and development of effective models for the classification and analysis of computer screenshots.

In a study on laptop screenshots, Sampat et al.[62] used a CNN to classify a dataset into 14 different classes related to the task the user was doing. The authors found that the architecture was not effective for classification. Their work offers important insights into the use of CNNs for screenshot classification and is part of the scarce published work on computer screenshots. Further research is necessary to explore the classification and analysis of computer screenshots in general.

Chayanin Suatap et al.[67] proposed a method for genre classification using typical game images provided in smartphone game stores such as an icon or screenshots. The proposed model based on a convolutional neural network and a soft voting ensemble technique outperformed human testers on average. The accuracy achieved for single icon and single screenshot classification tasks was 40.3% and 46.7% respectively, and the accuracy increased to 55.3% when the screenshots and icon of each game were combined to classify the game genre.

In a research conducted by Hashemi and Hall [36], they explored the application of deep learning methods to identify dark propaganda associated with violent extremist organizations. Using the AlexNet CNN architecture, they achieved an overall generalization accuracy of 86.08% across eight different classes. Instead of using screenshots, they adopted an alternative image-based classification technique by extracting images directly from websites. To build their training dataset, they labeled a collection of 120,000 images obtained from web and social media content.

Aydos et al. [6] introduced a novel approach to web page classification by leveraging the power of multiple neural network outputs. Their model uses a representation of each element through a set of descriptive images, which are then employed for classification purposes. By applying their method to the web page classification problem and utilizing Google Image Search results as descriptive images, they achieved an impressive classification rate of 94.90% on their WebScreenshots dataset, consisting of 20,000 websites categorized into four classes.

Web screenshots have shown to be an explored domain, not only with deep learning techniques, but also with other image processing techniques in the computer vision area. Moreover, in a lot of cases, additional information to the screenshots is used (HTML, URL, files/content extraction...). Typically, these methods involve analyzing the images present on a web page using image processing techniques and classifying the page based on the dominant image class. This particular approach is commonly employed for binary classification tasks, for example, in the detection of pornographic web pages.

Despite the growing interest in using images and screenshots to extract insights, research on computer screenshots is still scarce. While there has been considerable work on the web domain, there is little research on the general use of computers, including desktop and other programs (where is harder to access programmatically) rather than just browsers. The use of convolutional neural networks has shown promise in screenshot classification, with transfer learning being a key component in supervised learning for classification. In addition to deep learning techniques, other image processing techniques, such as Optical Character Recognition, can also be used to extract important features from screenshots. In the

next section, we will explore some of these techniques in detail.

### 2.3.1 Optical Character Recognition

Optical Character Recognition (OCR) is a key technology in the field of computer vision and image processing. Its ability to recognize and extract text from images has numerous applications, including document digitization, data extraction, and image retrieval. OCR has also become an important tool in the field of computer screenshot classification, as it allows for the automated processing of text-based information within screenshots.

However, applying OCR to screenshots presents unique challenges that are not typically encountered in regular OCR applications. Screenshots often contain text that is overlaid on top of complex backgrounds, such as images or other text, which can make it difficult for OCR algorithms to accurately extract the text. In addition, screenshots may contain text in a variety of fonts, sizes, and styles, which can further complicate the OCR process.

Another challenge in applying OCR to screenshots is that screenshots often contain non-text elements, such as icons, logos, and other graphical elements, that can interfere with the OCR process. These non-text elements can be particularly problematic when they overlap with or obscure the text, as they can cause errors in the OCR output.

Given the similarity of the layout between smartphone and computer screenshots, the findings and implications raised by the paper [17] are highly relevant to computer screenshot classification. The paper describes an experimental workflow for text extraction from smartphone screenshots based on OpenCV [14] image-processing and Tesseract OCR modules [66]. The authors identified image pre-processing as a crucial step for obtaining high-quality extracted text, which involved converting images to grayscale and applying binarization methods (Nobuyuki Otsu [50]) to parse grayscale images to their binary (i.e., black and white) counterparts. They also used segmentation to identify rectangular bounding boxes wrapping the textual content of the images, which involved dilating white pixels to create more organic white regions, detecting uniform regions, and drawing a rectangle around the identified area. The segmented regions were passed to the OCR engine, which was based on the Python wrapper for Tesseract. The authors evaluated the extracted text using three metrics: WER, CER, and PER. The published studies, wherein they compared OCR results against ground-truth transcriptions of 2,000 images, show the accuracy of the text extraction procedures at 74% at the individual character level.

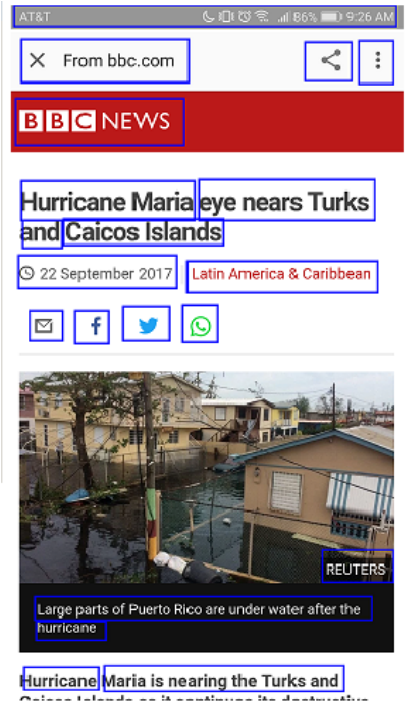


Figure 5: OCR applied to a smartphone screenshot in [17].

The authors also reported that the most prominent errors were associated with mixed icons and text, peculiar fonts, text color blended with background colors, partially overlapping segmented regions leading to duplicated recognition of characters, and human error inherent in the transcription loop. The paper provides valuable insights into

the challenges of applying OCR to smartphone screenshots, which are similar to those encountered in computer screenshot classification. The paper's methodology and results can serve as a basis for further research on OCR applied to computer screenshots, particularly for identifying and addressing the challenges associated with extracting text from screenshots that contain non-textual elements.

## 2.4 Natural Language Processing applied to screenshots

Reading "Natural Language Processing applied to screenshots" may seem counterintuitive due to the fact that people typically associate NLP with textual data, while screenshots are visual. However, the advent of Optical Character Recognition (OCR) techniques has enabled the extraction of text from screenshots with accuracy, as illustrated in section 2.2.1. The extracted text can then be analyzed using NLP techniques, opening up new possibilities for the classification and analysis of computer screenshots.

**Binary Text Classification:** One notable application of NLP to screenshot analysis is binary text classification. By leveraging NLP techniques, it becomes feasible to classify the extracted text into specific categories or labels, such as distinguishing between distracting and non-distracting content. This classification process relies on various NLP methods, including text preprocessing, feature extraction, and machine learning algorithms. Through the integration of NLP, the identification of textual patterns and semantic analysis can contribute to more accurate and nuanced classification results.

**Text Topic Modeling:** Another valuable application of NLP in screenshot analysis is text topic modeling. With the aid of NLP techniques, it becomes possible to discern the main topics or themes within the extracted text from screenshots. Topic modeling algorithms, such as Latent Dirichlet Allocation (LDA [12]), can be employed to automatically identify the underlying topics present in the text. This facilitates a deeper understanding of the content and allows for more advanced analysis, such as studying the distribution of topics within different categories or identifying key areas of focus within the screenshot.

By incorporating NLP into the analysis of screenshots, researchers can harness the power of textual information embedded within the visual data, enabling more comprehensive and insightful interpretations. The combination of OCR and NLP offers novel possibilities for understanding and leveraging the rich textual content available within screenshots.

To gain further insights into the various text classification algorithms used in NLP, a comprehensive survey paper titled "Text Classification Algorithms: A Survey" [43] provides an extensive overview of the existing methodologies and techniques. This survey paper serves as a valuable resource for understanding the state-of-the-art in text classification and can aid in the selection and implementation of appropriate algorithms for the binary text classification task in screenshot analysis.

# 3. Goals of the project

The current intervention inside the Alerta environment is based on logging keyboard and mouse interactions to monitor user behavior. However, this project aims to gain more information about the user's behavior by utilizing the screen to capture screenshots. The goal is to extract valuable insights from these raw data sources, particularly through two proposed tasks. The first task is to develop a binary classification model that can classify the screenshots as either "focused" or "distracted." The second task involves predicting the area of study in which the user is currently engaged while working.

By achieving these goals, the proposed computer screenshot classification approach can provide a valuable means of identifying and analyzing potential distractions within VR software for individuals with ADHD. The ability to accurately classify screenshots as "focused" or "distracted" can enable timely interventions that minimize the impact of distractions on productivity, ultimately improving outcomes for individuals with ADHD and increasing the efficiency and effectiveness of workplaces and educational settings. Additionally, predicting the area of study in which the user is currently engaged while working can help encourage the user to finalize certain tasks before jumping to other tasks, further improving productivity.

The results of this research project have the potential to contribute to a wider range of productivity studies involving computer users, where the classification of screenshots and feature extraction play a crucial role in discerning behavioral patterns. Each session can be registered with a log that can be further analyzed by researchers to study the behavioral patterns of users, especially those with ADHD, during work sessions. The findings of this study can have significant implications for the development of interventions and therapies to improve productivity and quality of life for individuals with ADHD. Overall, the proposed computer screenshot classification approach has the potential to be a valuable tool in improving outcomes for individuals with ADHD and advancing productivity studies in a range of settings.

## 3.1 Requirements

To achieve the goals of this project, several requirements must be met. These include both functional and non-functional requirements.

#### Functional requirements

The main functional requirements that describe the functionality of the system to be developed as well as data requirements are:

**F1. [DATASET] Creation of a sufficient and diverse dataset**: A dataset consisting of at least 10,000 labeled screenshots must be collected from a diverse range of applications, tasks, and user interfaces. The dataset should include a balanced number of both focused and distracted examples, with a minimum of 3,500 examples for each class. The dataset must be diverse enough to ensure that the model can generalize well to different scenarios.

**F2. [DATASET] Data annotation**: The dataset must be annotated to indicate which screenshots are focused and which are distracted. This will be used as the ground truth for training and evaluating the classification model as well as the topic modeling model. At least one supervisor must agree with an accuracy of at least 0.9 of a random sample sufficiently large.

**F3. [MODEL] Development of a computer screenshot classification model**: A binary classification model must be developed that can accurately classify screenshots as either "focused" or "distracted."

A suitable machine learning algorithm must be chosen for the task. This could include popular algorithms such as logistic regression, support vector machines (SVM), or deep learning approaches such as convolutional neural networks (CNNs). Using a deep learning approach, specifically a convolutional neural network (CNN), is highly encouraged.

**F4. [MODEL] Evaluation of the classification model**: To assess the effectiveness of the classification model, it is essential to evaluate its performance using suitable metrics, including accuracy, precision, recall, and F1 score. The model should also be tested on a separate validation dataset to assess its generalization ability. The target values for these metrics should be at least 0.75.

**F5. [MODEL] Development of a topic modeling algorithm**: A topic modeling algorithm must be developed to predict the area of study in which the user is currently engaged while working. This could include popular approaches such as latent Dirichlet allocation (LDA) or non-negative matrix factorization (NMF).

**F6. [MODEL] Implementation and deployment of the model**: The vinary classification model must be implemented and deployed within the Alerta environment for real-time monitoring of user behavior. The model should be able to capture screenshots, classify them as focused or distracted, and predict the user's area of study in real time with a maximum latency of 30 seconds.

### Non-functional requirements

The main non-functional requirements that outline the desired qualities of the system are:

**NF1. Model explainability**: The interpretability of the classification model's decision-making process is a desirable feature that can provide valuable insights into the model's behavior and aid in the development of better models. While it is not a strict requirement for this project, it is highly encouraged to explore methods to improve the interpretability of the model's decisions.

**NF2. Environmental impact**: The tracking of carbon emissions associated with the training of deep learning models is an optional requirement that can be pursued as a task to exceed expectations. Efforts must be made to minimize the environmental impact of the project. (The target carbon footprint for training the models should be less than 50 kg CO2)

**NF3. Ethical considerations**: Ethical considerations must be taken into account throughout the development and deployment of the model. The use of user data must comply with all relevant privacy laws and regulations, and informed consent must be obtained from users before collecting and using their data. Additionally, the model should not be used to discriminate against users or negatively impact their well-being.

### Assumptions

It is important to note that all implementations and experiments in this project are limited to the constraint and controlled domain of the Alerta Project. Therefore, several assumptions may be made throughout the research, such as the assumption that the operating software is the same for all screenshots. These assumptions will be outlined accordingly in each section of the document.

# 4. Proposed solution

This proposed solution presents a novel approach for studying and detecting potential distractions in VR software, using a combination of Convolutional Neural Networks (CNNs) and Optical Character Recognition (OCR) with Natural Language Processing (NLP) techniques. The solution aims to classify screenshots as either "focused" or "distracted" content, using both image and text-based approaches, and subsequently extracting features to identify distractions and study patterns in the "focused" category.

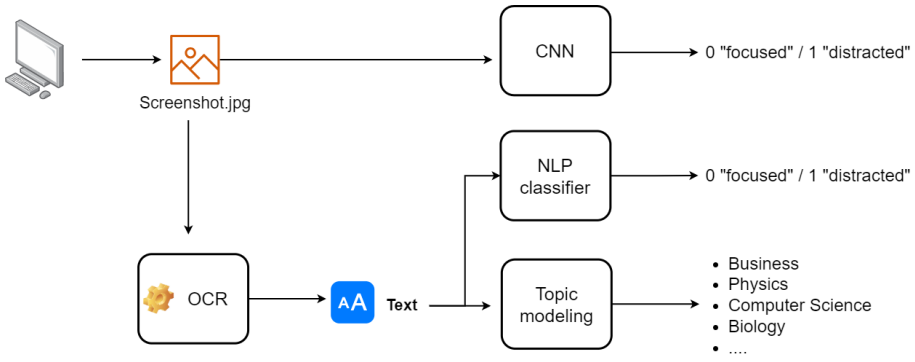


Figure 6: High-level diagram of the proposed solution.

As the diagram outlines, the proposed solution consists of two main components: an image-based approach and a text-based approach. The image-based approach uses a CNN for binary classification to detect if the screenshot contains "focused" or "distracted" content. The text-based approach uses an OCR tool with image preprocessing to extract text from the screenshot. Once the text is extracted, the binary classification is repeated using NLP techniques this time. Additionally, topic modeling is applied to the "focused" category to detect the study area on the screenshot content.

The benefit of approaching the binary classification from two different perspectives is that it provides a more comprehensive and accurate analysis and enables the comparison of the results from both approaches to identify potential discrepancies and further refine the analysis.

The proposed solution is integrated into the Alerta software platform which allows for real-time monitoring and intervention to minimize potential distractions and increase productivity in individuals with ADHD. The proposed solution is designed to be scalable and can be applied to various software applications to enhance productivity and improve outcomes for individuals with ADHD.

Overall, the proposed solution has the potential to make a significant impact on the quality of life and productivity of individuals with ADHD, as well as the efficiency and effectiveness of workplaces and educational settings.

## 4.1 Dataset

The success of machine learning models largely depends on the quality and quantity of the training data used. In this section, we discuss the creation of the dataset used to train our models for the Alerta Project, all with the aim of fulfilling the requirements outlined in section 3.1. We describe the various methods used for data collection, labeling, and augmentation. Additionally, we address the critical issue of data privacy

and the measures taken to ensure the security and confidentiality of our users' data. Finally, we explain the process of generating a text dataset from the image dataset that is suitable for training a natural language processing (NLP) model.

### 4.1.1 Data collection

The source of the data is key to achieving quality and quantity on the training data. Being a critical aspect of training machine learning models, different methods for data collection discussed in this section were contemplated throughout the project, finally deciding to create a dataset from scratch.

**a) Web scrapping**

The use of web scraping for data collection in this scenario was deemed unsuitable due to several factors. Firstly, web scraping involves gathering data from websites that may have varying layouts, resolutions, and aspect ratios, which would result in a dataset with high variability. Given that our analysis focused on screenshots captured from a specific software with a controlled environment and fixed image format, it would be challenging to ensure that the scraped data would be representative of the targeted environment. Moreover, web scraping is known to be a resource-intensive process that could be time-consuming, and it raises legal and ethical concerns (against requirement NF3 specified in section 3.1), including potential copyright infringement, privacy policy infringement, and respecting website terms of service.

To validate these arguments, a series of attempts were made to collect data via web scraping, and multiple challenges affecting data quality were identified. These challenges included the presence of watermarks, cropped screens, webcam superimpositions, and screenshots from other devices such as smartphones and game consoles. Furthermore, additional visual effects such as arrows, boxes, and mouse pointers were also identified, which would negatively impact the dataset's quality and usability for training a deep learning model for the Alerta project.


Figure 7: Examples of challenges affecting data quality (from left to right): Non-computer device screenshot; Watermark and webcam present in image; Cropped image with visual effects (mouse cursor).

**b) Public datasets**

In Section 2 we see how researchers in the field of screenshot classification have generated their own datasets to facilitate their studies. However, the availability of existing datasets is limited, and they often do not adequately represent the specific domain addressed in our project. Some of the datasets focus only on the web domain and exclude other screenshots such as those taken from other programs other than browse engines.

**c) A dataset from scratch**

The Alerta software records the screen of all user sessions and provides a comprehensive collection of screenshots (taken every 10 seconds) for data collection. This software was used to create the dataset from scratch, providing an extensive range of data suitable for the project's needs. Screenshots taken directly from the Alerta software are ideal because they provide an accurate representation of the environment in which the software is used, this being, for example, constant operating system, desktop background, size, and resolution. Moreover, this approach allowed us to control the variability of the dataset and avoid legal and ethical concerns associated with web scraping or public datasets.

On the one hand, getting screenshots from people working or studying was an easy task, as the Alerta project had undergoing experiments requiring users with ADHD to focus on homework or work tasks during sessions of 50 minutes. Each user took 12 sessions. 30 users were recorded in total. This resulted in a vast amount of data with different resolutions, aspect ratios, and content, including a variety of interfaces, menus, and visual elements. On the other hand, getting screenshots from users doing anything rather than work had to be collected with a different protocol, since the ones coming from natural user distractions were not enough to get a balanced dataset.

Therefore, we designed a protocol that consisted of asking users to perform specific tasks while using the Alerta software and recording their sessions. These tasks included browsing the internet, playing games, using social media, and watching videos. We collected an additional 3,000 screenshots using this protocol, resulting in a total dataset of 14,191 screenshots. To ensure a balanced dataset, we randomly sampled 9178 screenshots from the work sessions and 5,013 screenshots from the non-work sessions.

| Dataset | Images | Focused [0] | Distracted [1] | Full-screen | Partial-screen |
|---|---|---|---|---|---|
| dataset1 | 2,219 | 1,482 (66.79 %) | 737 (33.21 %) | 1,174 (52.86 %) | 1,045 (47.14 %) |
| dataset2 | 4,252 | 2,599 (61.15 %) | 1,653 (38.85 %) | 2,652 (62.42 %) | 1,600 (37.58 %) |
| dataset3 | 11,701 | 8,600 (73.55 %) | 3,101 (26.45 %) | 8,190 (69.99 %) | 3,511 (30.01 %) |
| dataset4 | 14,191 | 9,178 (64.67 %) | 5,013 (35.33 %) | 9,879 (69.61 %) | 4,312 (30.39 %) |

Table 1: Datasets created for the project.

To avoid biases when the user is using different window sizes, half-screen or different window displays, a balance between full-screen images or partial-screen images was achieved during the dataset collection and selection process.

Once the screenshots were collected and requirement F1 (section 3.1) was fulfilled, the next step was to label the data.

### 4.1.2 Data labeling

To train a supervised machine learning model, a labeled dataset is required. In our case, the screenshots were manually labeled by experts in the domain to obtain a high-quality dataset. Labeling was done manually

by a team of annotators, who were provided with guidelines and instructions to ensure consistency and accuracy in the labeling process. Each screenshot was labeled as either "focused" or "distracted" based on the task the user was performing during the session. Any ambiguous or unclear screenshots were reviewed by a senior annotator for final labeling. To ensure consistency and accuracy in the labeling process, clear guidelines were provided to the labelers. These guidelines included a description of the different types of activities, examples of screenshots, and instructions on how to handle ambiguous cases. Additionally, a validation process was implemented to ensure the quality of the labels, where a subset of the dataset was labeled by multiple labelers, and the agreement rate was calculated to evaluate the inter-rater reliability.

In addition, images with the "focused" label were also manually segmented in the college course the user was working on. This was known by observing the screenshot or the session video sequence if necessary. The list of courses observed was: "Art", "Astronomy", "Biology", "Business", "Physics", "Chemistry", "Maths and Statistics", "ComputerScience", "Pharmacology", "Psychology", "Literature", "Ethics and Philosophy", "Politics and Society" and "Gender".

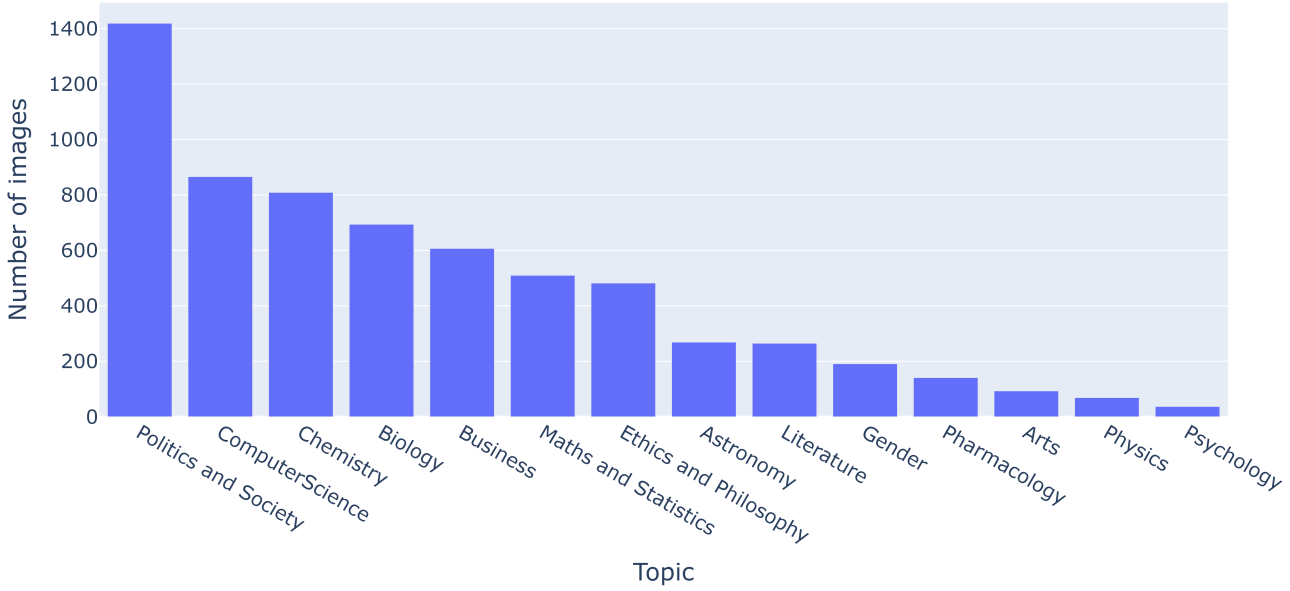


Figure 8: Samples per topic for "focused" screenshots in dataset4.

The labeling process resulted in a dataset of 14,191 labeled screenshots, evenly distributed among the different activity types and in the fulfillment of requirement F2 (section 3.1). This dataset will be used for training and evaluating the machine learning models developed for the Alerta Project.

### 4.1.3 Data augmentation

When designing a Deep Learning model for Computer Vision to be applied to screenshots in a specific software with a controlled environment and fixed image format, there may not be a need to apply common transformations such as cropping and rotation in the preprocessing step, and doing so could actually be detrimental to the model's performance.

This is because in a controlled environment, the screenshots are likely to have consistent layouts and structures (e.g. same screen, same OS...), and the objects of interest will be in the same position and

orientation in each screenshot. As a result, cropping or rotating the images to augment the data may cause the model to lose important information or even introduce noise, which can negatively impact its ability to make accurate predictions.

The particular requirements of the problem were carefully assessed to determine the necessity of any supplementary transformations or techniques. For example, the utilization of data augmentation techniques such as brightness adjustment or color jittering. However, in the absence of such variations in the controlled environment, these potential avenues for data enhancement were deemed unnecessary and were consequently excluded.

### 4.1.4 Data privacy

As specified in requirement NF3 (section 3.1), data privacy is a critical concern when working with any type of sensitive data, including user screenshots. To ensure the privacy and security of the data, several measures were taken throughout the data collection, labeling, and processing pipeline.

Firstly, it's important to note that we had written authorization from all Alerta software users to use the screen recordings for research purposes. Furthermore, no data containing personally identifiable information (PII) has or will be shared with the public. These measures ensure that the privacy of our users is maintained and that we are complying with all relevant data privacy laws and regulations.

Secondly, the labeling process was carried out on secure systems and in local mode, without the use of cloud platforms. This helped prevent any unauthorized access or data breaches.

Finally, access to the labeled dataset was restricted to only authorized personnel who required it for the Alerta Project. Any transfer or sharing of the data was subject to strict guidelines and protocols to prevent any unintended disclosure.

Overall, the Alerta Project team took the necessary precautions to safeguard the privacy and security of the data used in the project. These measures ensured that the data was only accessible to authorized personnel and that any sensitive information was appropriately protected and handled in accordance with industry standards and best practices.

### 4.1.5 Generating a text dataset

Once an image-labeled dataset was obtained, the next step was to generate a corresponding text dataset that would allow for the training of a text-based machine learning model. To accomplish this, the OCR pipeline described in Section X for text extraction was used. However, it was observed that approximately 1% of the samples resulted in blank or incomplete text output.

The text dataset was then preprocessed to remove any irrelevant or sensitive information, such as usernames or URLs, that could potentially compromise user privacy. Additionally, stop words and punctuation marks were removed, and the remaining text was tokenized to prepare it for feature extraction and model training.

It should be noted that the blank or incomplete text samples that were observed in the OCR output were not removed from the test set. These samples were kept to take into account potential errors in the evaluation of the machine learning model when deployed in the Alerta software.

## 4.2 Binary Classifier

In this section, we describe the development of the computer screenshot classification model as part of the process of meeting the requirement outlined in F3 (section 3.1).

### 4.2.1 CNN model

The image-based approach was the first approach for the project. Due to its effectiveness in processing visual data, using a convolutional neural network (CNN) was a logical choice for our project. To do so, we decided to apply transfer learning from similar tasks.

#### Transfer learning: using a pre-trained model

Transfer learning is a widely employed technique used to establish a model that benefits from a favorable initial state by leveraging pre-training on an existing network. This approach proves to be highly effective in enabling the training of deep networks when data availability is limited. The principle underlying transfer learning is rooted in the observation that neurons within the network exhibit varying degrees of generality and specificity. Typically, the lower layers of the network capture more generic features, such as edge filters and color blobs, which possess broad applicability across a wide range of image processing tasks. In contrast, the upper layers of the network specialize in extracting task-specific features, thus enabling enhanced performance on specific tasks.

The approach commonly used in transfer learning involves starting with a pre-trained network that was trained on a base dataset and a base task. To apply this approach to a new target task, we begin by initializing the layers of the target network with those of the base network, followed by fine-tuning the whole target network on the target dataset. According to research, the transferability of knowledge learned in this manner increases as the difference between the target task and the base task decreases.

The typical procedure employed in transfer learning entails the utilization of a pre-trained network that has undergone training on a foundational dataset and task. To adapt this methodology to a novel target task, the first step involves initializing the layers of the target network using those of the base network. Subsequently, the target network is fine-tuned on the target dataset as a whole. Studies indicate that the transferability of acquired knowledge through this process improves as the dissimilarity between the target task and the base task diminishes [74].

#### The ImageNet dataset

The ImageNet dataset [24] has gained significant popularity as a preferred option for pre-training purposes. Studies have demonstrated that utilizing features extracted from a network trained on ImageNet can yield effective general-purpose features for diverse visual tasks, even in the absence of fine-tuning specific to the target problem [63].

Yu et al. in [35] demonstrate that features learned from an ImageNet-trained CNN surpass the performance of existing state-of-the-art alternatives in document image classification and retrieval tasks. This highlights the transferability of the knowledge acquired by CNNs from one domain (ImageNet) to another domain (document images), leading to improved results. In addition, it is worth noting that document images and screenshot images share similarities in terms of visual content and structure. Both types of images often contain text, graphics, and other visual elements arranged in a document-like layout. This

similarity suggests that the transferability of knowledge acquired by CNNs from the ImageNet domain to document images could potentially extend to screenshot images as well.

### Resnet50

Given the findings mentioned above, we were intrigued by the possibility of adapting base models that align with the objectives of our evaluation task. In this regard, we choose the Resnet50 model (Kaiming He et al. [37]) with the default pretrained weights in the PyTorch library [52, 71] trained on the ImageNet dataset containing 1,281,167 training images, 50,000 validation images and 100,000 test images with 1000 object classes [24].

We adapted the network architecture in a simple way. We transformed the class space of the pre-trained model by replacing the last layer (containing 1000 neurons for the ImageNet dataset) with a fully connected layer of 512 neurons (with ReLU activation and dropout) and adding a last layer with two neurons + log-Softmax activation function for binary classification. The use of two neurons instead of one was a decision to facilitate the extension to multiclass classification if desired in the future. In addition, the negative log-likelihood loss was used for training.

### Image transformations

Prior to inputting the images into the network, a transformation process is applied. For compatibility with the resnet50 network, it is required that the input image dimensions are multiples of 32 for height and width, with a channel width of 3. To achieve this, we adopted the PyTorch transforms.Resize(224) function. This resizing technique preserves the aspect ratio of the image, recognizing that altering the proportion is unnecessary. Additionally, it reduces the image size to a standardized format (398x224) since our dataset comprises images of varying resolutions such as 1920x1080 and 3840x2160. This approach aligns with the common practice in CNNs where uniform image sizes are utilized.

In order to facilitate faster training of the model, we made the decision to use smaller images as input (a common practice). Although this may seem counterintuitive, there are practical reasons behind this choice. When the input image is larger, it contains a greater number of pixels that the network needs to process and learn from. This not only increases the memory requirements but also extends the time needed for computations. By resizing the images to a standardized format (398x224), we reduce the overall size of the images while maintaining their aspect ratio. This enables the model to train more efficiently by working with a reduced amount of data. In essence, using smaller images allows the network to learn faster as it needs to process fewer pixels, resulting in improved training speed.

### Training the model

In section 5.1.1, we will present the results obtained from the extensive training of the model with different hyperparameter configurations. The model was trained using a range of learning rates (0.1 to 0.001), batch sizes (32, 64, and 128), optimizers (Adam and SGD), and varying numbers of epochs (up to a maximum of 20). Moreover, experiments retraining the whole architecture and not only the last layers were also made. This thorough experimentation enabled us to gain insights into the impact of these factors on the model's training dynamics, convergence, and overall performance.

The training process was executed on a high-performance GeForce RTX 3090 GPU, leveraging its

computational capabilities for efficient training. Specifically, training the model for 20 epochs on the largest dataset took approximately 6.3 hours, highlighting the substantial computational demands associated with the training process.

By presenting the results in the upcoming section, we aim to analyze and evaluate the performance of the model under different hyperparameter configurations. These findings will contribute to a comprehensive understanding of the model's capabilities and guide future enhancements and refinements.

### 4.2.2 OCR + NLP model

In order to extract text from images, a text extraction approach similar to the one employed in Chiatti et al. (2017) [17] was utilized. As discussed in the literature review section (Section 2.3.1), the aforementioned paper introduced an experimental workflow for extracting text from smartphone screenshots. This workflow relied on the utilization of OpenCV [14] image-processing and Tesseract OCR modules [66].

The implementation was tailored to suit the specific use case of computer screenshots. Notably, an initial cropping step was incorporated into the process. This step involved removing the borders of the image, excluding a 20-pixel margin from each side except for a 60-pixel margin from the bottom. The objective of this step was to eliminate regions without any identifiable text in all the samples. Consequently, this modification resulted in improved performance in terms of text bounding box detection and reduced extraction speed.

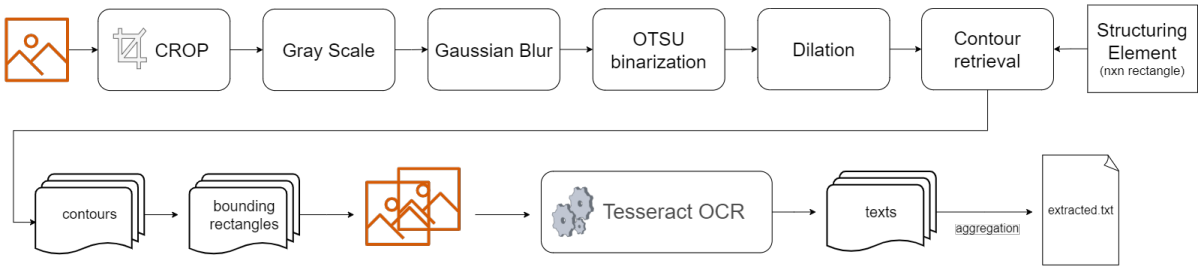


Figure 9: Text extraction process.

Additionally, we made a deviation from the original paper in terms of the kernel size used for dilation. The kernel size directly influenced the size of the rectangles that were detected. By selecting a smaller kernel size, we were able to detect individual words rather than entire sentences. For our purposes, a rectangle with a kernel size of (18, 18) proved to be a suitable option for extracting large portions of text. It is important to note that since the natural language processing (NLP) techniques employed in our project followed a bag-of-words approach, this parameter did not significantly impact the overall results.

The steps performed to convert the image into text are represented in the following diagram. Note that the specific implementation can be found in the code of the project repository (Appendix A.2).

Once we had the screenshot in text format we could apply NLP techniques using the Scikit-Learn Python library [53]:

**Text Cleaning:** Punctuation marks, stopwords, and unnecessary whitespaces were removed from the text. Punctuation marks often do not carry significant meaning for classification tasks, while stopwords are commonly used words (e.g., "the," "is," "and") that add little discriminatory value. By eliminating these

elements, we focused on the more relevant words and phrases within the text. Additionally, the text was lowercased to ensure consistency in the features.

**Conversion to TF-IDF Features:** To represent the collection of cleaned documents, a matrix of TF-IDF features was constructed. The TF-IDF score combines two metrics: Term Frequency (TF) and Inverse Document Frequency (IDF). The TF score captures how frequently a term appears in a specific document, while the IDF score highlights the rarity of a term across the entire corpus. Multiplying these two scores results in the TF-IDF score, which provides a measure of term importance within the document collection.

**Training the Classifier:** Several classifiers were evaluated to identify the most effective model for the binary classification task. A Random Forest Classifier with 1000 estimators was employed to capture complex relationships between features. Additionally, a Multinomial Naive Bayes classifier with standard sklearn parameters was tested. Lastly, Logistic Regression was applied with different values for the inverse of the regularization strength parameter C (0.01, 0.1, 1, 2, 5). These different values of C allowed for the exploration of the impact of regularization on the classifier's performance. Smaller values of C indicate stronger regularization, which can help prevent overfitting.

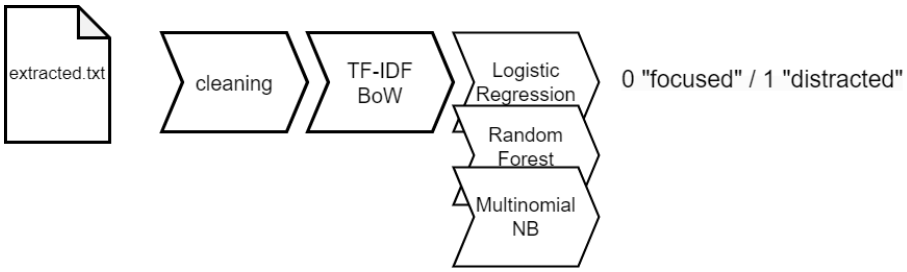


Figure 10: High-level diagram of the NLP classifier pipeline.

In section 5.1.2, we will present the results of the training process, showcasing the performance of the classifiers mentioned above on the binary text classification task.

## 4.3 Topic Modeling

Although the primary focus of this project is on detecting distractions related to the content of the screen, we wanted to showcase an example of how the extracted text can be leveraged for further analysis aiming to meet requirement F5 (section 3.1). In this section, we present a baseline approach for applying topic modeling to the extracted text, enabling us to monitor user activity more effectively. By understanding the topics of the text content, future work will aim to identify patterns and behaviors that could indicate excessive task-switching or lack of focus.

In our approach to classify screenshots converted to text into different courses or areas of study, we use the fastText model for topic modeling. The fastText word embeddings, as described in [13], provide a valuable resource with aligned word vectors for 44 languages. These vectors are pre-trained on extensive datasets, including Common Crawl and Wikipedia. The alignment of the vectors is performed using the RCSLS method outlined in [39]. The fastText models are trained using CBOW with position-weights, utilizing a dimension of 300, character n-grams of length 5, a window size of 5, and 10 negative samples. These embeddings provide similar word representations for words with similar meanings.

In our topic modeling approach, each topic is represented by a set of keywords. These keywords define the topic and serve as the basis for computing their respective embeddings. The topic embedding is obtained by calculating the centroid of the embeddings of the keywords. By utilizing the topic centroid, we can identify the most similar topics to a given text.

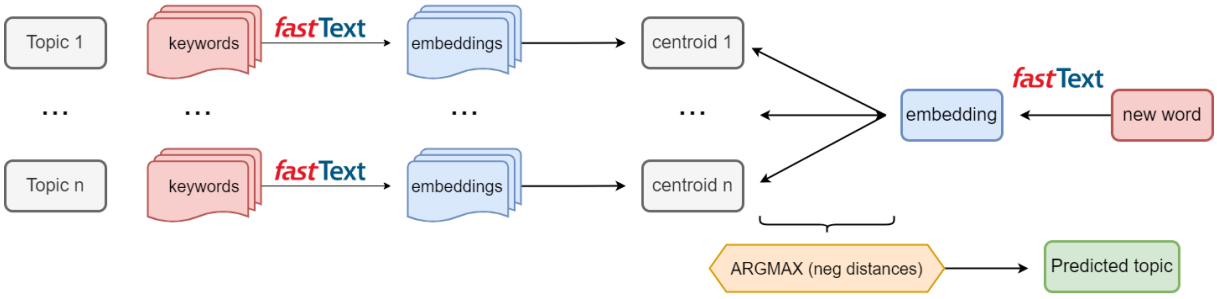


Figure 11: Topic Modeling process applied to a single word.

To initiate the classification process, we load the dataset containing the OCR output of the screenshots. From this dataset, we filter and select only the screenshots categorized as "focused" according to the labeling process explained in 4.1.2. These focused screenshots are further classified into 14 courses or areas of study, such as 'Business' or 'Computer Science', which will be our topics. Adding more topics would only require providing a new list of keywords.

To classify a text (in this case, a screenshot), we begin by performing text cleaning operations, including the removal of punctuation marks, accents, numbers, whitespaces, and the conversion to lowercase. Additionally, we eliminate stopwords to enhance the quality of the text representation. We employ a Bag-of-Words approach by generating a list of word embeddings and computing their centroid. Using this centroid, we calculate the negative Euclidean distance between the topic centroids and the screenshot's centroid. Finally, we apply softmax to obtain a distribution of probabilities across the different courses or areas of study, indicating the likelihood of each classification.

In section 5.1.3, we will present the results of applying the topic modeling approach to classify the extracted text from screenshots into courses or areas of study. We will discuss the performance of the baseline approach and evaluate its effectiveness in identifying relevant topics and capturing user activity patterns.

## 4.4 ML Tracking

In the development of our machine learning project, we adopted the use of the MLflow Python library [55]. MLflow is a comprehensive platform designed to facilitate and streamline the process of machine learning development. It offers functionalities for experiment tracking, packaging code into reproducible runs, and enables sharing and deployment of models. In this section, we highlight the significance of incorporating MLflow and its associated practices in machine learning projects.

**Reproducibility and Experimentation:** MLflow's experiment tracking capabilities played a pivotal role in ensuring reproducibility and facilitating experimentation in our project. By leveraging MLflow, we were able to capture and log crucial information such as parameters, input data, and code versions for each experiment. This meticulous tracking not only allowed us to accurately reproduce experiments but also

enabled seamless collaboration among team members. The ability to compare different approaches, evaluate parameter changes, and analyze the effectiveness of various models significantly expedited our development process.

**Model Performance Monitoring:** Monitoring and comparing model performance is critical in any machine learning project, and MLflow made this process remarkably efficient for us. With MLflow, we could easily track essential metrics (meeting requirement F4 specified in section 3.1), such as accuracy, precision, recall, and F1 score, over time. This comprehensive performance monitoring provided valuable insights into our model's behavior and its performance across different datasets or environments. Identifying potential issues, detecting model degradation, and iteratively improving our models became significantly more manageable and data-driven.

**Hyperparameter Optimization:** Hyperparameter optimization is often a time-consuming and resource-intensive process. MLflow played a vital role in simplifying this aspect of our project. By utilizing MLflow to track hyperparameters and their associated results, we were able to systematically record and evaluate the performance of different hyperparameter configurations. This facilitated the identification of optimal parameter settings that maximized model performance while minimizing unnecessary computation. MLflow's support in this area greatly accelerated our efforts and directed our attention toward the most promising configurations.

**Collaboration and Reproducibility:** Collaboration and reproducibility were central to the success of our project, and MLflow provided us with a centralized platform to achieve these goals. With MLflow, we could effortlessly package our models and associated code into reproducible runs. This allowed us to easily share models with other researchers and deploy them in production environments without the worry of version mismatches or missing dependencies. The ability to collaborate seamlessly and ensure reproducibility significantly accelerated our progress and facilitated knowledge sharing within our team.

**CO2 Emissions Tracking with CodeCarbon:** In order to address requirement NF2 (section 3.1) concerning the environmental impact of our experiments, we gave careful consideration to the CO2 emissions generated throughout the project. We utilized CodeCarbon, a tool specifically designed for tracking CO2 emissions in computing tasks [19], to accurately measure and account for the environmental footprint associated with our computational activities. While the NLP models in our project required minimal computational power, the training of the CNN model involved extensive utilization of an NVIDIA RTX 3090 GPU over prolonged periods. By leveraging CodeCarbon, we precisely quantified the GPU's energy consumption during CNN training, enabling us to calculate the corresponding CO2 emissions. The total emissions resulting from all the CNN experiments amounted to 4.12 kilograms of CO2e, which is approximately equivalent to consuming 153 liters of diesel according to the United States Environmental Protection Agency [26]. This emphasizes the importance of considering the environmental cost when conducting further model training, highlighting our commitment to sustainable machine learning practices and responsible research and development.

By leveraging the comprehensive features provided by MLflow and considering the environmental impact through CodeCarbon, we have been able to establish a solid foundation for our research and enhance the rigor of our methodologies.

## 4.5 Deployment

During the project, we recognized the importance of deploying our machine learning model, the distraction detector, within the Alerta software (as highlighted in section 3.1 by requirement F6). This integration was essential to leverage the potential of the model.

In this section, we discuss the deployment process, focusing on the successful integration of the distraction detector into the Alerta Software. We highlight the technical considerations, challenges overcome, and the benefits of deploying the detector within the VR environment. By combining machine learning with VR technology, we aimed to create an effective solution that addresses distractions and improves productivity for individuals with ADHD.

### 4.5.1 Integration with educational software

For the distraction detector, we made the decision to deploy the binary classifier utilizing a Convolutional Neural Network (CNN) architecture instead of the OCR + NLP approach model. The primary factor influencing this choice was speed. The CNN model runs with an average processing time of approximately 1.2 seconds, significantly faster than the alternative approach, which takes more than 20 seconds on average due to the Optical Character Recognition (OCR) component.

#### Intervention format

During the deployment phase, a crucial consideration was determining an appropriate intervention to be generated when a distraction was detected by the system. After careful deliberation, we concluded that utilizing sound would be an effective option.

To test the intervention mechanism, we initially introduced two different sounds. A lower-pitch sound was played when the distraction detector predicted the "focused" category, indicating that the user was engaged and attentive to the task. Conversely, a higher-pitch sound was triggered when a distraction was detected, alerting the user to refocus their attention.

However, as we moved toward the final deployment, we decided to simplify the intervention by using a single sound exclusively for distraction detection. This sound consisted of a low-pitch tone with a frequency of 200 Hertz and a duration of 1 second.

By employing a distinct sound for distraction detection, we aimed to create an audible cue that would effectively draw the user's attention back to the task at hand. The chosen sound, with its specific pitch and duration, was carefully designed to be noticeable without being overly intrusive, striking a balance between alerting the user and maintaining a seamless user experience within the VR environment.

#### Distraction Detector Loop

During the deployment of the model, we implemented an iterative process that ran continuously while the VR environment was active. To prevent an excessive number of false alarms, we introduced a variable called W (window). Rather than triggering an intervention for every individual screenshot classified as "distracted," we required a consecutive sequence of the last W screenshots to be classified as "distracted" before initiating an intervention.

Additionally, to avoid overwhelming the user with frequent interventions, we implemented a counter that would reset to zero each time an intervention was generated. This prevented repetitive sounds from being played in quick succession. Once an intervention was triggered, no further interventions were generated

until a set duration of time had passed.

To maintain a consistent timing between prediction timesteps, we introduced another variable called INTERVAL. This ensured a constant time interval between each prediction iteration. The loop would sleep until the specified duration for that particular timestep of the algorithm was completed, allowing for a systematic and controlled workflow.

The following pseudocode demonstrates a simplified version of the Distraction Detector Loop triggering the intervention:

**Algorithm 1** Distraction Detector Loop triggering the intervention

```
W ← 3
L ← empty list
while True do
    newScreenshot ← takeScreenshot()
    predicted ← model(newScreenshot)
    L.append(predicted)
    if length(L) > W then
        L.pop(0)                          ▷ We keep the last W predictions
    end if
    if L = [1] * W then
        playSound()
        L ← [0] * W                       ▷ Restart counter
    end if
    sleep()                               ▷ Sleep to achieve a constant interval between screenshots
end while
```

**Note:** [$x$] * $n$ is a list of length $n$ where all elements are equal to $x$.

By incorporating these mechanisms into the deployment process, we aimed to strike a balance between accurately identifying distractions and minimizing the occurrence of false alarms. This approach helped create a smoother and more user-friendly experience within the VR environment, ensuring that interventions were appropriately timed and delivered in a controlled manner.

### Logging the results

In each timestep, the predictions and probabilities were logged into a CSV file for every session, enabling subsequent analysis and visualization of the time series data. This logging feature facilitated a comprehensive examination of the session's progression. While the OCR+NLP model was not employed in the experiments with real users, it would have additionally logged the extracted text and detected topic information.

Additionally, this logged data can be valuable for psychologists, as it provides an opportunity to analyze the behavioral patterns of ADHD students during sessions. By examining the time series data, psychologists can gain insights into the frequency and duration of distractions, the effectiveness of interventions, and the overall engagement levels of ADHD students. This analysis can contribute to a deeper understanding of ADHD behaviors and inform targeted interventions and strategies to enhance productivity and focus in educational and therapeutic settings.

### Human in the loop

The deployment of the model facilitated the development of a "human in the loop" system, where automated predictions were generated and supplemented with human supervision. This approach enabled the rapid expansion of the dataset. By combining the efficiency of automated predictions with human oversight, we were able to accelerate the process of gathering labeled data, improving the model's performance and generalizability.

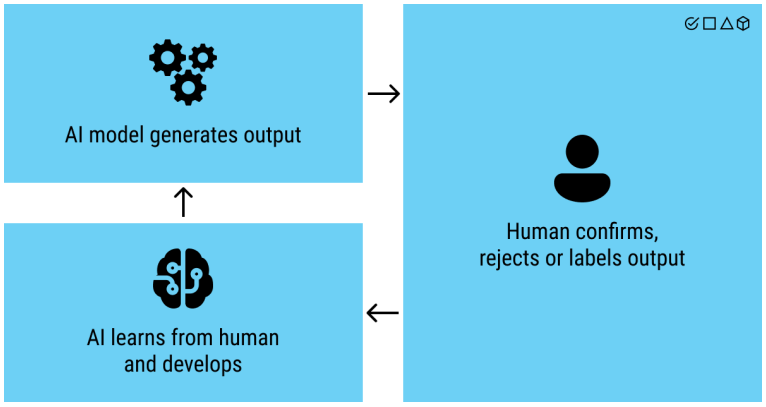


Figure 12: How human-in-the-loop in ML works. Source: Iryna Sydorenko on Label Your Data [68].

## 4.5.2 Experiments with real users

The distraction detection system underwent experimentation involving five participants, four of whom had ADHD, while one did not. Each participant was tasked with utilizing the software during two sessions, each lasting 40 minutes, on separate days.

In the first session, participants were instructed to engage in work-related activities for the entire duration. This session served as a baseline for understanding their typical work behavior and interaction with the environment without any external interventions.

For the second session, participants were first asked as an exercise prior to the session to identify websites or programs that they frequently visited when they became distracted. After 20 minutes of working, participants were explicitly encouraged to visit the websites or programs they had previously identified as distracting.

These experiments aimed to observe the effectiveness of the distraction detection system in real-world scenarios, where participants experienced both focused work and intentional distractions. By capturing their behaviors and responses during these sessions, we sought to evaluate the system's ability to detect and intervene in moments of distraction and provide valuable insights into the impact of the system on ADHD participants' productivity and focus.

### Questionnaire

After each session, participants were asked to complete a questionnaire to gather their subjective feedback and impressions regarding their experience with the distraction detection system. Additionally, participants were interviewed to gain further insights into their perceptions, challenges faced, and suggestions for improvement.

In designing the evaluation metrics for this study, inspiration was drawn from the "Technology Acceptance Model" presented in [70]. The selected metrics aimed to address significant concepts outlined in the paper, including usefulness, quality of output, and intention of use. Adapting the questionnaire to the specific context of this study ensured its relevance and applicability to the evaluation of the intervention tool.

| **Metric** | **Question** | **Scale (1-5)** |
|---|---|---|
| Focus Enhancement (FE) | How much did the VR environment help you stay focused on your task? | 1 (Not at all) to 5 (A great deal) |
| | How much did you feel like the distraction detector helped you improve your concentration and attention? | 1 (Not at all) to 5 (A great deal) |
| Accuracy (A) | How often did you find the distraction detector to be accurate in detecting your distractions? | 1 (Rarely) to 5 (Always) |
| | How often did the distraction detector produce false alarms? | 1 (Always) to 5 (Rarely) |
| Perceived Usefulness (PU) | On your daily work, how frequently do you experience distractions while performing tasks on a computer? | 1 (Rarely) to 5 (Very frequently) |
| | How much do you believe the VR environment and distraction detector are useful tools for managing ADHD symptoms? | 1 (Not at all) to 5 (Extremely) |
| Intention to Use (IU) | How likely would you be to use the VR environment and distraction detector again in the future? | 1 (Not likely at all) to 5 (Very likely) |
| | How likely are you to recommend the VR environment and distraction detector to other people with ADHD? | 1 (Not likely at all) to 5 (Very likely) |
| Potential (P) | How much potential do you think the VR environment and distraction detector have to improve the management of ADHD symptoms? | 1 (Very low potential) to 5 (Very high potential) |
| | How much potential do you think the VR environment and distraction detector have in other contexts beyond education? | 1 (Very low potential) to 5 (Very high potential) |

Table 2: Questionnaire for the evaluation of the distraction detector deployed in VR environment

The questionnaire responses and interview data collected from the participants will be thoroughly analyzed in the subsequent section. This analysis will provide valuable qualitative information, shedding light on the participants' experiences, perceptions of distraction detection interventions, and overall usability of the system. The combination of quantitative and qualitative data will contribute to a comprehensive evaluation of the system's impact and effectiveness in supporting ADHD productivity in a VR environment.

# 5. Results and Evaluation

This section provides an in-depth analysis of the outcomes achieved through the implementation of the various models. Firstly, we will present the results obtained from the different models applied to the specific tasks and datasets, highlighting their performance. Furthermore, we will delve into the evaluation of the distraction detector deployed in the Alerta VR environment, focusing on its real-world usability and user feedback. This evaluation phase will provide valuable insights into the practical application of our solution and its impact on enhancing productivity and managing ADHD symptoms in a VR setting.

## 5.1 Model results

### 5.1.1 CNN Model

The CNN model based on the ResNet50 architecture was evaluated by training with both frozen and unfrozen weights of the pre-trained layers. Considering the availability of a sufficient amount of data, training the entire architecture (unfreezing) resulted in superior performance. Table 3 presents a comparison of the model's performance with frozen and unfrozen weights, using the same architecture and hyperparameters.

| Frozen | Opt | Batch size | LR | Acc | F1 | Precision | Recall |
|---|---|---|---|---|---|---|---|
| False | Adam | 128 | 0.001 | 0.99 | 0.99 | 0.99 | 0.99 |
| True | Adam | 128 | 0.001 | 0.957 | 0.957 | 0.958 | 0.957 |

Table 3: Comparison of CNN model results with frozen and unfrozen weights.

Two different optimizers, Adam and stochastic gradient descent (SGD), were tested. However, SGD exhibited inferior performance on the test dataset compared to Adam. Therefore, Adam was chosen as the preferred optimizer for the CNN model.

Among the tested batch sizes, 128 yielded the highest performance, which was the maximum allowed by the training machine. A batch size of 64 closely followed in performance, with one model achieving exceptional results within this and other batch sizes.

Additionally, the model's performance was evaluated using different learning rates. Table 4 presents the results of the best models. Note that they were trained with varying learning rates (LR).

| Frozen | Opt | Batch size | LR | Acc | F1 | Precision | Recall |
|---|---|---|---|---|---|---|---|
| False | Adam | 128 | 0.01 | 0.951 | 0.951 | 0.955 | 0.951 |
| False | Adam | 128 | 0.001 | 0.99 | 0.99 | 0.99 | 0.99 |
| False | Adam | 128 | 0.0001 | **0.992** | **0.992** | 0.992 | **0.992** |
| False | Adam | 64 | 0.001 | **0.992** | **0.992** | **0.993** | **0.992** |

Table 4: Results of CNN model with different learning rates

In conclusion, the CNN architecture demonstrated excellent performance in detecting screenshots with distracting content. It is important to note that these results are specific to the training set used, and different hyperparameters may yield better performance on other datasets. Nevertheless, the overall success of the approach highlights its potential to accurately classify computer screenshots to identify distractions.

### 5.1.2 OCR + NLP Model

The results of the text binary classifier are presented in Table 5. These results indicate the performance of the different classifiers in terms of accuracy, F1-score, precision, and recall. Overall, the results are very good, indicating that there is valuable information to be learned from the text data and the effectiveness of the classifier in distinguishing between the two classes. However, it is important to note that the performance of this approach is slightly lower compared to the results obtained with the CNN approach.

| Accuracy | F1-score | Precision | Recall | Model |
|---|---|---|---|---|
| 0.962 | 0.944 | 0.938 | 0.95 | Random Forest |
| 0.928 | 0.89 | 0.976 | 0.818 | Multinomial Naive Bayes |
| 0.667 | 0.15 | **1** | 0.081 | Logistic Regression ($C = 0.01$) |
| 0.904 | 0.841 | 0.98 | 0.736 | Logistic Regression ($C = 0.1$) |
| 0.966 | 0.951 | 0.967 | 0.935 | Logistic Regression ($C = 1$) |
| **0.972** | **0.959** | 0.978 | **0.941** | **Logistic Regression ($C = 2$)** |
| 0.968 | 0.953 | 0.965 | **0.941** | Logistic Regression ($C = 5$) |

Table 5: Results of the different Text Binary Classifiers.

Among the different models evaluated, Logistic Regression with a regularization strength parameter $C = 2$ achieved the best overall performance. It obtained an accuracy of 0.972, an F1-score of 0.959, precision of 0.978, and recall of 0.941. These metrics indicate the model's ability to correctly classify the text instances and strike a balance between precision and recall.

On the other hand, the model with $C = 0.01$ exhibited very strong regularization, leading to saturated precision of 1 and a significant decrease in recall to 0.081. This indicates that the model became overly conservative in making positive predictions, resulting in a high precision but a low ability to identify true positive cases.

In comparison, the Multinomial Naive Bayes model achieved an accuracy of 0.928, an F1-score of 0.89, precision of 0.976, and recall of 0.818. Random Forest achieved an accuracy of 0.962, an F1-score of 0.944, precision of 0.938, and recall of 0.95. Overall, the logistic regression classifiers showed better results.

In summary, the Text Binary Classifier demonstrated promising results, with the Logistic Regression model (specifically with $C = 2$) achieving the highest performance. These results provide valuable insights into the effectiveness of the classifier and its potential application in classifying text data in the research project.

#### Explainability

Explainability is of crucial importance for our text binary classifier model, given that it is a key requirement (NF1) mentioned in Section 3.1. The ability to provide clear explanations for the model's predictions allows us to understand the underlying factors driving its decisions. By uncovering the key features and words that contribute to the classification outcomes, we gain valuable insights into the reasoning behind the model's outputs. This interpretability not only helps build trust and confidence in the model's performance but also enables us to identify and address any biases or shortcomings.

The observed results, with words like 'game', 'play', or 'shop' being influential for the "distracted" class, and 'file', 'lecture', or 'course' being influential for the "focused" class, align with our expectations. These findings demonstrate that the model has learned meaningful associations between certain words and the respective classes.

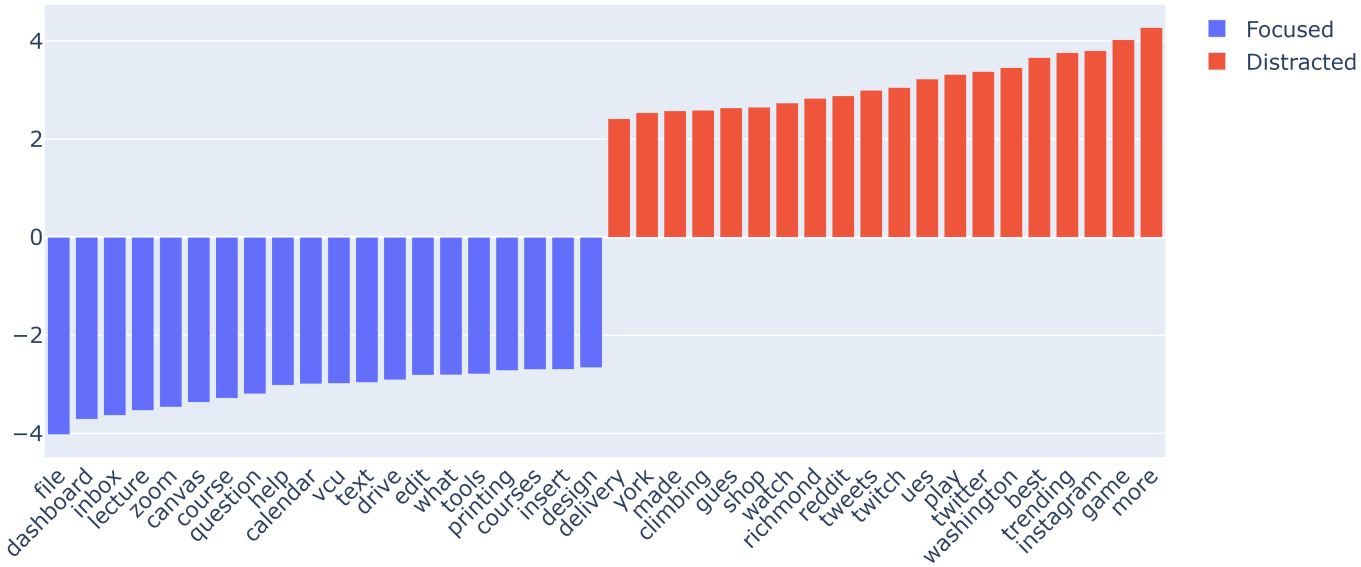


Figure 13: Top-20 Words with most Feature Importance per class.

This information can be beneficial in several ways. Firstly, it helps in detecting biases or potential sources of bias in the model. For example, the presence of words like "vcu" (Virginia Commonwealth University) as an influential word for the "focused" class and "richmond" for the "distracted" class suggests a potential bias towards certain educational institutions or locations. Being aware of these biases allows us to address them by either modifying the model, modifying the dataset, or incorporating additional considerations during the decision-making process.

Additionally, the identified influential words can guide the selection of stopwords or domain-specific terms that can be excluded from the text analysis. For instance, if we observe that certain words are consistently associated with a particular class but do not provide meaningful information for classification, we can consider adding them as stopwords to improve the model's performance and eliminate noise in the data.

Overall, these interpretability results not only help us gain insights into the model's decision-making process but also allow us to detect biases, refine the model, and make informed decisions regarding the inclusion or exclusion of specific terms or stopwords, ultimately enhancing the accuracy and fairness of the text binary classifier.

### 5.1.3 Topic Modeling

The results of the topic modeling approach for classifying the extracted text from screenshots into courses or areas of study are as follows. We evaluated the performance using the top-k accuracy metric, which measures the number of times the correct label is among the top k labels predicted, ranked by predicted scores.

In our experiments, we compared the top-k accuracy of our approach with random guessing as a baseline. Our approach clearly outperformed random guessing, indicating that it captured some meaningful patterns in the data. However, it is important to note that the results were not highly accurate.

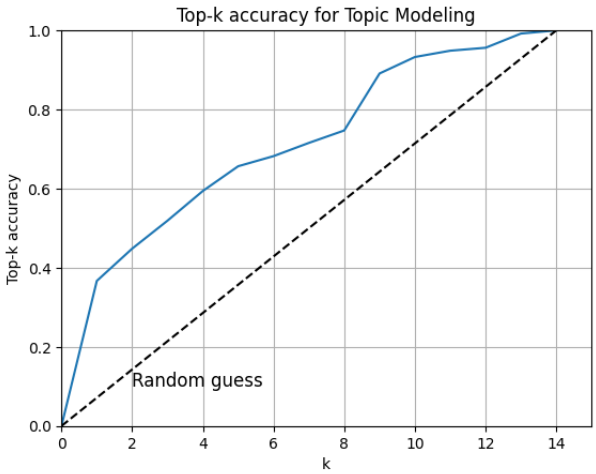


Figure 14: Top-k accuracy of our Topic Modeling approach.

We obtained a top-1 accuracy of 36.7%, which means that the correct label was predicted as the top choice in 36.7% of the cases. Additionally, we achieved a top-4 accuracy of 60%, indicating that the correct label was included in the top four predicted labels in 60% of the cases.

While our approach was able to provide some level of accuracy, the occasional incorrect predictions could be attributed to the fact that some topics share similar vocabulary, making them challenging to differentiate. Additionally, the current model's low complexity may limit its ability to capture more nuanced relationships between topics and the extracted text. These results highlight the need for further refinement and improvement in our topic modeling approach.

## 5.2 Evaluation with real users

The evaluation of the distraction detector deployed with the VR environment for ADHD students yielded overall very positive results. Participants provided feedback on various metrics, including focus enhancement, accuracy, perceived usefulness, intention to use, and future potential.

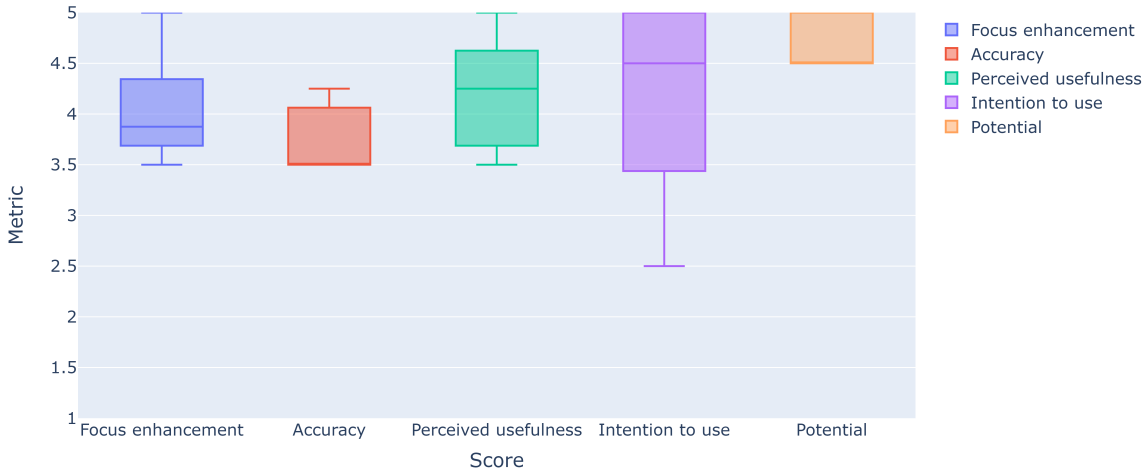


Figure 15: Boxplot with the deployment evaluation metrics.

The average scores for these metrics ranged from 3.5 to 5.0 on a scale of 1 to 5. The mean scores were consistently high, with an average of 4.17. The metric with the highest score and consensus was potential, which received an average score of 4.5. These results indicate that the VR environment and distraction detector have the potential to be valuable tools for ADHD students. In the following subsections, we will analyze each metric in detail, highlighting the main quotes from participant interviews. The transcription of all the main quotes from the interview during the experiments can be found in Appendix A.1.

### 5.2.1 Focus Enhancement

Average Score: 4.05
Notable quotes:

- P2: *"It was good. I definitely felt really focused. I think it helps a lot because you can't really get distracted by anything else."*
- P1: *"The combination of them [the VR environment and the distraction detector] is what's so important. I mean, you have the visual part, where I'm not looking at my phone all the time doing whatever and looking around the room, but there's also the distraction thing where it does play a sound if I try to play a game [...] It's a nice combination of blocking out distractions and then the virtual distractions that are left are mitigated by the distraction detector."*

The evaluation of the "Focus Enhancement" metric received an average score of 4.05, indicating a positive perception among the participants. The immersive VR environment and the distraction detector were reported to be effective in promoting focus and minimizing distractions. Participant 1 expressed that being in the VR world eliminated visual and auditory distractions, allowing them to solely focus on

their work. They also mentioned that the detector helped them regain focus when they momentarily deviated from their tasks. Another participant, Participant 2, highlighted the benefit of blocking out random distractions and the reinforcement of maintaining focus as a habit. The isolation provided by the VR environment was appreciated by Participant 3, who found it helpful in concentration. Participant 4, in addition to affirming that the environment and detector facilitated their ability to focus, acknowledged that the detector's intermittent beeping served as a helpful reminder to stay on task. Overall, the participants recognized the combined effectiveness of the VR environment and the distraction detector in enhancing their focus and minimizing distractions, resulting in a positive evaluation of this metric.

### 5.2.2 Accuracy of the detector

Average Score: 3.75
Notable quotes:

- P1: *"[What kind of distractions did the detector identify correctly?] It was great with videos. For example like random YouTube or Reddit whenever there was a video it would ping like immediately. But other things like text not so much. Like whenever I'd read a Reddit story you know like paragraphs and paragraphs it would think it's focused."*
- P3: *""It has not generated many false alarms."*
- P4: *""If it didn't have false positives it would help a lot."*

The evaluation of the "Accuracy" metric received an average score of 3.75, which is the lowest among all the metrics assessed. Participants provided mixed feedback regarding the accuracy of the distraction detector and its impact on their focus. While Participant 1 expressed satisfaction with the detector's accuracy, stating that it correctly identified distractions like videos, they also noted that it missed distractions related to reading longer texts. Participant 2 had a similar experience, where the detector initially performed well but had more false alarms toward the end of the session. They mentioned instances when the detector triggered false alarms while they were reading code on GitHub. On the other hand, Participant 3 reported minimal false alarms and acknowledged that the detector missed detecting audio distractions from background YouTube videos. Participant 4 faced frequent false alarms despite not engaging in non-work-related activities. However, they found the detector effective for most distractions, except for a black-and-white comic book in PDF format. Overall, the participants' feedback reflects a mixed evaluation of the "Focus Enhancement" metric, with concerns raised about the accuracy and handling of false alarms in specific scenarios, contributing to the lower average score.

The participants' feedback on the distraction detection system indicates that it has generally performed well in terms of false positives. They reported that the number of false alarms during the sessions was relatively low, which is a positive aspect of the system. While the occurrence of false positives in the distraction detection system may be relatively low, their impact on the overall user experience is significant. Participants expressed frustration whenever a wrong intervention occurred. Constant and unnecessary interruptions can disrupt workflow, decrease productivity, and cause frustration for users, even if they occur occasionally. Therefore, it is essential to continue refining the deployment of the detection algorithm to ensure that interventions are triggered only when genuinely needed. Furthermore, some participants expressed the desire for a way to "justify" their work activities to the system, as it currently does not have a mechanism to recognize focused work and may continue beeping until a change in program or display occurs. By striking a balance between accurate detection and avoiding unnecessary interruptions, the system can effectively support users in maintaining focus and achieving their tasks efficiently.

### 5.2.3 Perceived usefulness

Average Score: 4.2
Notable quotes:

- P4: *"I think it's very helpful to help me manage ADHD symptoms. I don't think it's perfect though. You can't use it for more than an hour I think. It is uncomfortable."*
- P1: *""So that was 10 out of 10. It was saying to get back to work when I wasn't necessarily working. So I would say overall five. I mean, it was incredibly helpful."*

The evaluation of the "Perceived Usefulness" metric received an average score of 4.2, indicating a positive reception among participants. They expressed how the combination of the VR environment and distraction detector was instrumental in managing ADHD symptoms and improving work behavior. Participant 1 highlighted the transformative effect on their productivity and described the distraction detector as an improvement, praising its ability to block out distractions and mitigate virtual distractions. Participant 2 emphasized the system's effectiveness in preventing random distractions and acknowledged how it stopped them from using non-work-related websites like Reddit. They also recognized the potential of the system for tasks requiring intense focus and believed it could benefit others in managing ADHD symptoms. Participant 3 found the distraction detector subtly helpful and suggested its application in educational settings. However, they mentioned a decrease in appeal as one gets older and raised concerns about privacy. Participant 4 regarded the system as helpful for managing ADHD symptoms but noted the discomfort of prolonged VR use. Despite this, they acknowledged the great potential of the tools and recommended others to try them. Overall, the participants' feedback indicates that the VR environment and distraction detector are perceived as useful in managing ADHD symptoms, with recommendations for improvement and consideration of various user needs.

### 5.2.4 Intention to use

Average Score: 4.15
Notable quotes:

- P1: *"I'm 100% recommending it to other people."*
- P2: *"Like I know for me, honestly, if I had to study for an exam and it was like the night before, I would actually probably put this thing on. [...] I would use both the VR with the detector."*
- P4: *"If I had it at home, I think it would be cool, but wouldn't use it. At least the VR. The distraction detector I would use for sure. If it didn't beep too much, of course."*

The evaluation of the "Intention to Use" metric yielded an average score of 4.15, indicating a positive inclination among participants to utilize the VR environment and its associated tools. While there was some variance in the responses, participants expressed a strong desire to incorporate the system into their daily work and study routines. Participant 1 expressed enthusiasm for using the VR environment for various academic activities such as watching lectures and doing homework. They highly recommended the system to others and emphasized its potential for widespread use. Participant 2 also expressed a high likelihood of using the system, particularly when intense concentration was required. Although they mentioned the need for improved hardware, such as more comfortable glasses, they still expressed a positive inclination to

recommend the system. Participant 3 highlighted the importance of separating the distraction detection aspect from the VR headset, suggesting that a standalone distraction detector would be more practical and efficient. They mentioned potential challenges associated with prolonged VR headset use but expressed a desire to use the system more if it were more comfortable. Participant 4 demonstrated a preference for the distraction detector over the VR headset and expressed a definite intention to use the detector in the future while expressing less enthusiasm for the VR component. Overall, participants showed a positive intention to use the system, with suggestions for improvements in comfort and flexibility, indicating the potential for increased adoption and usage among ADHD students.

### 5.2.5 Potential

Average Score: 4.7
Notable quotes:

- P2: *"I can definitely see people using it in the future if they have it at home or if they go to a library and want to study. I think the potential is really high. At least for me, it helped me manage my ADHD symptoms a lot."*

The evaluation of the "Potential" metric received a highly positive response, with participants giving it an average score of 4.7. They recognized the system's potential for customization, improved hardware, and wider applicability beyond education. Participants highlighted its usefulness for managing ADHD symptoms, creating focused work environments, and supporting children's concentration. The system's potential integration with parental controls and blocking of distracting content were also noted. Participants expressed confidence in its potential for various settings, including office work post-pandemic. However, addressing false positives was suggested as an area for improvement. Overall, participants were impressed with the system's potential and provided valuable insights for further enhancement.

### 5.2.6 Comfort

Notable quotes:

- P2: *"I would suggest maybe get a lighter headset. That's the biggest thing for me. Like for me, the headset is way too heavy. [...] I got tired from it."*

The evaluation of comfort, despite not being an original metric, received significant feedback from participants. They highlighted several aspects related to comfort and provided valuable suggestions for improvement. Participants expressed concerns about the weight and discomfort of the VR headset, noting that it can cause fatigue and pain during prolonged use. However, they also acknowledged the benefits of the isolated environment for minimizing distractions. Feedback on the distraction detector sound was mixed, with some participants finding it helpful and not overly annoying, while others suggested improving its feedback or combining it with visual cues. Overall, participants emphasized the need for a lighter and more comfortable headset to enhance the overall experience. Their feedback provides important insights for refining the VR environment for ADHD students, ensuring both functionality and user comfort.

# 6. Conclusions

In conclusion, this research project has successfully accomplished its predetermined objectives and requirements as detailed in Section 3.1. A comprehensive and diverse dataset was meticulously constructed (F1), which, through an arduous labeling process (F2), facilitated the training of a highly effective computer screenshot classification model (F3) that demonstrated its efficacy following rigorous evaluation (F4). Furthermore, a topic modeling algorithm was implemented to classify the screenshots according to their respective areas of study (F5). The system was seamlessly integrated and subjected to rigorous testing by real users within the virtual reality environment (F6).

Throughout the project, significant attention was dedicated to addressing crucial aspects outlined as non-functional requirements: the explainability of the NLP model (NF1), the environmental impact associated with CNN training (NF2), and the ethical considerations pertaining to data collection and utilization (NF3).

Notably, this research has convincingly demonstrated that the detection of non-work or non-study related content can be achieved with a remarkable level of accuracy by analyzing screenshots. More precisely, through the training of convolutional neural networks or text extraction and subsequent training of natural language processing models. One noteworthy advantage of this technology is its ability to operate solely on system screenshots, making it applicable to any computer environment, whether it is a virtual reality setting or not.

In addition, it is worth noting that the code developed for this research project is publicly available in the GitHub repository referenced in Appendix A.2. Researchers and practitioners interested in exploring or building upon our work can access the codebase, enabling them to delve deeper into the technical aspects of our implementation. Furthermore, for those seeking to directly test and utilize the Distraction Detector (CNN binary classification model) or the Topic Modeling model, we have provided a Hugging Face Space, which is also referenced in Appendix A.3. This space offers a convenient platform for users to interact with and evaluate the models, fostering a collaborative environment for future experimentation and advancements in the field. It is important to acknowledge that the training, testing, and deployment experiments of the models were conducted within the constraints and controlled domain of the Alerta project, as described in Section 3.1. Therefore, it is crucial to recognize that the models may have inherent biases (e.g. location, operating system, desktop background) due to the specific context in which they were developed.

The satisfaction and positive feedback received from real subjects, including individuals with ADHD, during the testing phase have been exceptionally encouraging. Users have acknowledged the value of the deployed model and its role as a preventive measure against distractions. Although opportunities for improvement exist, which will be discussed in the subsequent section, subjects have recognized the vast potential offered by both virtual reality and the model itself.

Ultimately, this work aspires to inspire professionals in the technology and psychology sectors to build upon its findings. The challenge of aiding individuals, particularly those with ADHD, in enhancing their concentration and boosting productivity during study or work endeavors is an exhilarating pursuit with the potential for remarkable contributions to society.

# 7. Future work

In this section, we discuss several avenues for future research and improvements that can enhance the effectiveness and applicability of the project results.

Firstly, the VR environment could be enhanced in terms of its capabilities and user experience. Improvements in hardware, such as lighter headsets and sharper image quality, would contribute to a more immersive and comfortable experience for students. These enhancements would facilitate prolonged focus and engagement during educational activities.

The topic modeling approach presented in this project can be further enhanced. While we have established a baseline and demonstrated the feasibility of the method, there is room for improvement. Future research should aim to explore advanced topic modeling techniques and compare their performance against the established baseline.

This work has demonstrated that both image and text features play a crucial role in monitoring user experience from screenshots. Therefore, the next logical step is to investigate the integration of these modalities in a multimodal model. A potential direction is to train an end-to-end architecture, inspired by existing work ([42]), capable of understanding both text and image layouts. Developing such a model specific to our task would be highly interesting and could lead to improved performance and speed when recognizing text.

One line of future work could be mixing other inputs (e.g., keyboard use, mouse clicks, eye tracking) already used by Alerta (intentionally excluded from this project) with the findings of this research. By incorporating additional data sources, a more accurate distraction detector could be deployed, leveraging the combined power of multiple modalities. Integrating keyboard use, mouse clicks, and eye tracking data, along with the analyzed image and text features from the screen, could provide a comprehensive understanding of user engagement and distraction levels. This approach would enable a more holistic and nuanced assessment of the user's experience, leading to improved accuracy and effectiveness in identifying and addressing distractions.

Furthermore, time is an important factor in the deployment of the model in Alerta and the creation of suitable interventions. It would be valuable to incorporate the time dimension directly into the model architecture, rather than solely relying on software deployment timestamps. This can be achieved by extending the current architecture or designing a new architecture from scratch, specifically tailored to incorporate temporal information.

Moreover, conducting a more in-depth analysis of the results and session reports by psychologists specializing in ADHD would provide valuable insights for the current and future work. Their expertise and perspectives can shed light on the interpretation of the model's outputs and guide further improvements.

By pursuing these avenues of research, we can refine and expand upon the existing framework, ultimately enhancing its effectiveness in boosting productivity and mitigating ADHD symptoms through the use of VR technology.

# References


[1] Rebecca Adams, Paul Finn, Elisabeth Moes, Kathleen Flannery, and Albert "Skip" Rizzo. Distractibility in attention/deficit/hyperactivity disorder (adhd): The virtual reality classroom. *Child neuropsychology*, 15(2):120–135, 2009.

[2] American Psychiatric Association et al. Diagnostic and statistical manual of mental disorders second edition—dsm-ii. *American Psychiatric Association: Washington, DC, USA*, 1968.

[3] American Psychiatric Association et al. Diagnostic and statistical manual of mental disorders third edition text revision—dsm-iii-r. *American Psychiatric Association: Washington, DC, USA*, 1980.

[4] American Psychiatric Association et al. Diagnostic and statistical manual of mental disorders third edition—dsm-iii. *American Psychiatric Association: Washington, DC, USA*, 1980.

[5] American Psychiatric Association et al. Diagnostic and statistical manual of mental disorders fifth edition text revision—dsm-5-tr. *American Psychiatric Association: Washington, DC, USA*, 2022.

[6] Fahri Aydos, A Murat Özbayoğlu, Yahya Şirin, and M Fatih Demirci. Web page classification with google image search results. *arXiv preprint arXiv:2006.00226*, 2020.

[7] Russell A Barkley. *ADHD and the nature of self-control*. Guilford press, 1997.

[8] Russell A Barkley, Mariellen Fischer, Lori Smallish, and Kenneth Fletcher. The persistence of attention-deficit/hyperactivity disorder into young adulthood as a function of reporting source and definition of disorder. *Journal of abnormal psychology*, 111(2):279, 2002.

[9] Geoffrey Baruch, Ioanna Vrouva, and Pasco Fearon. A follow-up study of characteristics of young people that dropout and continue psychotherapy: Service implications for a clinic in the community. *Child and Adolescent Mental Health*, 14(2):69–75, 2009.

[10] Azadeh Bashiri, Marjan Ghazisaeedi, and Leila Shahmoradi. The opportunities of virtual reality in the rehabilitation of children with attention deficit hyperactivity disorder: a literature review. *Korean journal of pediatrics*, 60(11):337, 2017.

[11] Joseph Biederman, Stephen V Faraone, Thomas J Spencer, Eric Mick, Michael C Monuteaux, and Megan Aleardi. Functional impairments in adults with self-reports of diagnosed adhd: A controlled study of 1001 adults in the community. *The Journal of clinical psychiatry*, 67(4):7488, 2006.

[12] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[13] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[14] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[15] Anil Chacko, Brian T Wymbs, Estrella Rajwan, Frances Wymbs, and Nicole Feirsen. Characteristics of parents of children with adhd who never attend, drop out, and complete behavioral parent training. *Journal of Child and Family Studies*, 26(3):950–960, 2017.

[16] Agnese Chiatti, Dolzodmaa Davaasuren, Nilam Ram, Prasenjit Mitra, Byron Reeves, and Thomas Robinson. Guess what's on my screen? clustering smartphone screenshots with active learning. *arXiv preprint arXiv:1901.02701*, 2019.

[17] Agnese Chiatti, Xiao Yang, Miriam Brinberg, Mu Jung Cho, Anupriya Gagneja, Nilam Ram, Byron Reeves, and C Lee Giles. Text extraction from smartphone screenshots to archive in situ media behavior. In *Proceedings of the Knowledge Capture Conference*, pages 1–4, 2017.

[18] Baek Hwan Cho, Jong-Min Lee, JH Ku, Dong Pyo Jang, JS Kim, In-Young Kim, Jang-Han Lee, and Sun I Kim. Attention enhancement system using virtual reality and eeg biofeedback. In *Proceedings IEEE Virtual Reality 2002*, pages 156–163. IEEE, 2002.

[19] CodeCarbon.io. Codecarbon python library.

[20] Ivan Culjak, David Abram, Tomislav Pribanic, Hrvoje Dzapo, and Mario Cifrek. A brief introduction to opencv. In *2012 proceedings of the 35th international convention MIPRO*, pages 1725–1730. IEEE, 2012.

[21] Victoria Dahl, Amrita Ramakrishnan, Angela Page Spears, Annlady Jorge, Janice Lu, Nina Abraham Bigio, and Anil Chacko. Psychoeducation interventions for parents and teachers of children and adolescents with adhd: a systematic review of the literature. *Journal of Developmental and Physical Disabilities*, 32(2):257–292, 2020.

[22] Melissa L Danielson, Rebecca H Bitsko, Reem M Ghandour, Joseph R Holbrook, Michael D Kogan, and Stephen J Blumberg. Prevalence of parent-reported adhd diagnosis and associated treatment among us children and adolescents, 2016. *Journal of Clinical Child & Adolescent Psychology*, 47(2):199–212, 2018.

[23] Anna M De Haan, Albert E Boon, Joop TVM De Jong, Machteld Hoeve, and Robert RJM Vermeiren. A meta-analytic review on treatment dropout in child and adolescent outpatient mental health care. *Clinical psychology review*, 33(5):698–711, 2013.

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[25] Dan Edvinsson and Lisa Ekselius. Six-year outcome in subjects diagnosed with attention-deficit/hyperactivity disorder as adults. *European Archives of Psychiatry and Clinical Neuroscience*, 268(4):337–347, 2018.

[26] Environmental Protection Agency. Greenhouse Gas Equivalencies Calculator. `https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator`. Accessed: June 11, 2023.

[27] Yantong Fang, Dai Han, and Hong Luo. A virtual reality application for assessment for attention deficit hyperactivity disorder in school-aged children. *Neuropsychiatric disease and treatment*, pages 1517–1523, 2019.

[28] Stephen V Faraone, Stephen J Glatt, Oscar G Bukstein, Frank A Lopez, L Eugene Arnold, and Robert L Findling. Effects of once-daily oral and transdermal methylphenidate on sleep behavior of children with adhd. *Journal of Attention Disorders*, 12(4):308–315, 2009.

[29] Jason M Fogler, David Burke, James Lynch, William J Barbaresi, and Eugenia Chan. Topical review: transitional services for teens and young adults with attention - deficit hyperactivity disorder: a process map and proposed model to overcoming barriers to care. *Journal of Pediatric Psychology*, 42(10):1108–1113, 2017.

[30] Anselm Fuermaier, Lara Tucha, Marah Butzbach, Matthias Weisbrod, Steffen Aschenbrenner, and Oliver Tucha. Adhd at the workplace: Adhd symptoms, diagnostic status, and work-related functioning. *Journal of Neural Transmission*, 128(7):1021–1031, 2021.

[31] Tim Fullen, Sarah L Jones, Lisa Marie Emerson, and Marios Adamou. Psychological treatments in adult adhd: a systematic review. *Journal of Psychopathology and Behavioral Assessment*, 42(3):500–518, 2020.

[32] Chanelle T Gordon and Gregory A Fabiano. The transition of youth with adhd into the workforce: Review and future directions. *Clinical child and family psychology review*, 22(3):316–347, 2019.

[33] Chanelle T Gordon, Gregory A Fabiano, Nicole K Schatz, Karen Hulme, and Rebecca K Vujnovic. Parenting stress during late adolescence in mothers of individuals with adhd with and without odd. *Journal of Child and Family Studies*, 30(12):2966–2979, 2021.

[34] MTA Cooperative Group et al. A 14-month randomized clinical trial of treatment strategies for attention-deficit/hyperactivity disorder. *Archives of general psychiatry*, 56(12):1073–1086, 1999.

[35] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE, 2015.

[36] Mahdi Hashemi and Margeret Hall. Detecting and classifying online dark visual propaganda. *Image and Vision Computing*, 89:95–105, 2019.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[38] Andreas Jangmo, Ralf Kuja-Halkola, Ana Pérez-Vigil, Catarina Almqvist, Cynthia M Bulik, Brian D'Onofrio, Paul Lichtenstein, Ewa Ahnemark, Tamara Werner-Kiechle, and Henrik Larsson. Attention-deficit/hyperactivity disorder and occupational outcomes: The role of educational attainment, comorbid developmental disorders, and intellectual disability. *PloS one*, 16(3):e0247724, 2021.

[39] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. *arXiv preprint arXiv:1804.07745*, 2018.

[40] Elizabeth Kappler, Rosemarie Figueroa Jacinto, and Steve Arndt. Evaluation of visual acuity and perceptual field of view using the varjo xr-3 headset in a virtual environment. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 66, pages 2193–2197. SAGE Publications Sage CA: Los Angeles, CA, 2022.

[41] RC Kessler, M Lane, PE Stang, and DL Van Brunt. The prevalence and workplace costs of adult attention deficit hyperactivity disorder in a large manufacturing firm. *Psychological medicine*, 39(1):137–147, 2009.

[42] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. *arXiv preprint arXiv:2111.15664*, 2021.

[43] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

[44] Joshua M. Langberg, L. Eugene Arnold, Amanda M. Flowers, Jeffery N. Epstein, Mekibib Altaye, Stephen P. Hinshaw, James M. Swanson, Ronald Kotkin, Stephen Simpson, Brooke S. G. Molina, Peter S. Jensen, Howard Abikoff, William E. Pelham Jr., Benedetto Vitiello, Karen C. Wells, and Lily Hechtman. Parent-reported homework problems in the mta study: Evidence for sustained improvement with behavioral treatment. *Journal of Clinical Child & Adolescent Psychology*, 39(2):220–233, 2010. PMID: 20390813.

[45] Klaus W Lange, Susanne Reichl, Katharina M Lange, Lara Tucha, and Oliver Tucha. The history of attention deficit hyperactivity disorder. *ADHD Attention Deficit and Hyperactivity Disorders*, 2:241–255, 2010.

[46] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

[47] Hansey Lee, Youdi Li, Shih-Ching Yeh, Yanyan Huang, Zhengyu Wu, and Zanli Du. Adhd assessment and testing system design based on virtual reality. In *2017 2nd International Conference on Information Technology (INCIT)*, pages 1–5. IEEE, 2017.

[48] JM Lee, BH Cho, JH Ku, JS Kim, JH Lee, IY Kim, and SI Kim. A study on the system for treatment of adhd using virtual reality. In *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 4, pages 3754–3757. IEEE, 2001.

[49] Suzanne McCarthy, Lynda Wilton, Macey L Murray, Paul Hodgkins, Philip Asherson, and Ian CK Wong. The epidemiology of pharmacologically treated attention deficit hyperactivity disorder (adhd) in children, adolescents and adults in uk primary care. *BMC pediatrics*, 12(1):1–11, 2012.

[50] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[51] Varjo Technologies Oy. Varjo xr-3 - the industry's highest resolution xr headset. Varjo, November 1 2021.

[52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[54] Yehuda Pollak, Patricia L Weiss, Albert A Rizzo, Merav Weizer, Liron Shriki, Ruth S Shalev, and Varda Gross-Tsur. The utility of a continuous performance test embedded in virtual reality in measuring adhd-related deficits. *Journal of Developmental & Behavioral Pediatrics*, 30(1):2–6, 2009.

[55] MLflow Project. Mlflow python library.

[56] N Ram and B Reeves. Time sampling in bornstein mh, arterberry me, fingerman kl, & lansford je (eds.), encyclopedia of lifespan human development (pp. 2247–2258), 2018.

[57] Nilam Ram, Xiao Yang, Mu-Jung Cho, Miriam Brinberg, Fiona Muirhead, Byron Reeves, and Thomas N Robinson. Screenomics: A new approach for observing and studying individuals' digital lives. *Journal of Adolescent Research*, 35(1):16–50, 2020.

[58] Byron Reeves, Nilam Ram, Thomas N Robinson, James J Cummings, C Lee Giles, Jennifer Pan, Agnese Chiatti, MJ Cho, Katie Roehrick, Xiao Yang, et al. Screenomics: A framework to capture and analyze personal life experiences and the ways that technology shapes them. *Human–Computer Interaction*, 36(2):150–201, 2021.

[59] Byron Reeves, Thomas Robinson, and Nilam Ram. Time for the human screenome project. *Nature*, 577(7790):314–317, 2020.

[60] Albert A Rizzo, J Galen Buckwalter, Todd Bowerly, Cheryl Van Der Zaag, Lorie Humphrey, Ulrich Neumann, Clint Chua, Chris Kyriakakis, Andre Van Rooyen, and D Sisemore. The virtual classroom: a virtual reality environment for the assessment and rehabilitation of attention deficits. *CyberPsychology & Behavior*, 3(3):483–499, 2000.

[61] Celestino Rodríguez, Débora Areces, Trinidad García, Marisol Cueli, and Paloma González-Castro. Comparison between two continuous performance tests for identifying adhd: Traditional vs. virtual reality. *International journal of clinical and health psychology*, 18(3):254–263, 2018.

[62] Anand Sampat and Avery Haskell. Cnn for task classification using computer screenshots for integration into dynamic calendar/task management systems, 2015.

[63] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[64] Margaret H Sibley, L Eugene Arnold, James M Swanson, Lily T Hechtman, Traci M Kennedy, Elizabeth Owens, Brooke SG Molina, Peter S Jensen, Stephen P Hinshaw, and Arunima Roy. Variable patterns of remission from adhd in the multimodal treatment study of adhd. *American Journal of Psychiatry*, pages appi–ajp, 2021.

[65] Sebastian Skalski, Karol Konaszewski, Grzegorz Pochwatko, Robert Balas, and Janusz Surzykiewicz. Effects of hemoencephalographic biofeedback with virtual reality on selected aspects of attention in children with adhd. *International Journal of Psychophysiology*, 170:59–66, 2021.

[66] Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE, 2007.

[67] Chayanin Suatap and Karn Patanukhom. Game genre classification from icon and screenshot images using convolutional neural networks. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*, pages 51–58, 2019.

[68] Iryna Sydorenko. Human-in-the-loop in machine learning: A handful of arguments in favor. *Label Your Data*, 2021. Consulted on June 10, 2023.

[69] Unity Technologies. Unity. https://unity.com, Accessed 2023.

[70] Viswanath Venkatesh and Hillol Bala. Technology acceptance model 3 and a research agenda on interventions. *Decision sciences*, 39(2):273–315, 2008.

[71] Vasilis Vryniotis. How to train state-of-the-art models using torchvision's latest primitives, November 18 2021.

[72] Tjhin Wiguna, Ngurah Agung Wigantara, Raden Irawati Ismail, Fransiska Kaligis, Kusuma Minayati, Raymond Bahana, and Bayu Dirgantoro. A four-step method for the development of an adhd-vr digital game diagnostic tool prototype for children using a dl model. *Frontiers in Psychiatry*, 11:829, 2020.

[73] Shih-Ching Yeh, Chia-Fen Tsai, Yao-Chung Fan, Pin-Chun Liu, and Albert Rizzo. An innovative adhd assessment system using virtual reality. In *2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences*, pages 78–83. IEEE, 2012.

[74] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[75] Zoe Young, Nima Moghaddam, and Anna Tickle. The efficacy of cognitive behavioral therapy for adults with adhd: A systematic review and meta-analysis of randomized controlled trials. *Journal of Attention Disorders*, 24(6):875–888, 2020.

# A. Appendix

## A.1 Experiment interviews transcription

We present the interview transcription of the main quotes from the experiments conducted as part of our research project. These interviews provide valuable insights and firsthand accounts from individuals who participated in our study, shedding light on their perspectives, experiences, and opinions.

### A.1.1 Participant 1

#### Session 1

- " I was watching a lecture video that was posted to YouTube, so I was quite impressed that it never said I was distracted while on YouTube. When I first started this session, it alerted when I was on Canvas trying to find a lecture to watch, but it stopped as soon as I started watching the lecture and typing and actually doing work. So I would say it functioned very well, helped me stay focused."
- "[How did it help to stay focused on your tasks?] So I'm not looking around all the time like I normally am because I'm in this environment where I don't have visual distractions, I don't have auditory distractions like people walking around usually distracts me, and in VR I can't see anything else, I can't hear anything else, I'm just in the world with my work. [...] I mean it's incredibly helpful for me."
- "When I wasn't specifically working on my stuff, like when I was looking for a video, it did kind of make me want to find a video faster and get to work. So I would say it helped me focus and get working again if I had stopped."
- "[Were you distracted at any point during the session?] No, I would say I wasn't. I mean I was just looking for a video in the very beginning, which is when it was beeping."
- "[So how often did the distraction produce false alarms?] Not at all. I wasn't necessarily distracted when it went off, but I wasn't doing work. I would say it's pretty accurate."
- "[Was the sound of the distraction detector disturbing too much?] No, I actually think it was great. I had the volume fairly low anyways, but with a lower volume, it was perfect. With the sound level, the frequency of it, I actually really liked the sound of it for getting back to work."
- "On my daily work I get distracted all the time."
- "[How much would you say on a scale from one to five that these tools (the environment and the distraction detector) are useful for managing ADHD symptoms?] Personally, it's a five. I mean, it completely changes my entire work behavior. Even just with the VR, the distraction detector is for me like an improvement. [...] So that was 10 out of 10. It was saying to get back to work when I wasn't necessarily working. So I would say overall five. I mean, it was incredibly helpful."
- "When the alarm is just a sound, it only comes up when I'm distracted and I'm not worried about it the rest of the time because I can't see it."

- " If I could, I would love to use it all the time. I mean, all the time when I do schoolwork would be wonderful. I would love to watch lectures on it. I'd like to do homework in it."
- "I'm 100% recommending it to other people."
- "I think there are a couple of things you could fine-tune to make it more optimal. A more customizable environment would be actually really interesting and nice because that would solve each person's issues."
- "I definitely think it's something that not only could but should be used in the future."

**Session 2**

- "It didn't beep a lot while I was working. It beeped a little after I came back from the distracted stuff to watch my YouTube lecture again. It beeped for a little bit and then I made the screen a little smaller and it didn't it didn't beep anymore."
- "It did go off a few times. Like I said while watching the YouTube video. But it wasn't distracting or anything. It was still fairly quiet so that like I would notice it but it wasn't something that pulled me out of whatever I was thinking about. It was just there."
- "Most of the time when it went off it was accurate. I didn't think it was going off as frequently as it should have been. I think it should be a little bit quicker to kind of get my attention a little more. It wasn't really that impactful every ten seconds."
- "It was pretty easy to tune out [the sound]. Especially when I had the lecture still playing and the lecture was louder than the sound."
- "[What kind of distractions did the detector identify correctly?] It was great with videos. For example, on random YouTube or Reddit whenever there was a video it would ping like immediately. But other things like text not so much. Like whenever I'd read a Reddit story you know like paragraphs and paragraphs it would think it's focused."
- "The combination of them [the VR environment and the distraction detector] is what's so important. I mean you have the visual like I'm not looking at my phone all the time doing whatever, looking around the room but there's also the distraction thing where it does play a sound if I go like try to play a game of Sudoku which I do all the time when I'm working because I don't know, I'm weird like that. But yeah, it's a nice combination of blocking out distractions, and then the virtual distractions that are left are mitigated by the distraction detector."
- "If the headset improves significantly, I think it would be super, super like applicable in everything. But you know, you've got the cable, it's heavy, it's a little uncomfortable, I can imagine it's super pricey. So after all of those things kind of lessen as issues, I think yeah, I think it could definitely be super useful."
- "I don't think it will change the world, but I think, well, not like revolutionize the world, but I think it will definitely change the way we work in the future."
- "There is one thing, it's only looking at the screen. So like if I start looking around like this, I'm not focused on whatever I'm doing, but it doesn't pick up on that, which we did have, I think, in the other one, kind of. But yeah, that's the only thing. Like, I could look down at my watch to see a notification if I wanted to. And it wouldn't tell me I'm doing anything wrong."

### A.1.2 Participant 2

#### Session 1

- "The session was good in the beginning because the algorithm was working but then at the end, it started buzzing me even if I was doing work."
- "I think it helps a lot (the environment and detector) because I don't have random distractions. A lot of the times I will be sitting at my desk and I'll just pick something up even if it's not my phone. So the fact that it blocks that out is really helpful."
- "At one point I was just like, I'm just going to go check out Reddit because that's what I would do at home. But after a while, it played the sound and I was like, okay, okay. It actually stops you from using it. So yeah, I actually did get distracted and it stopped that."
- "So for me, it would beep if I was on GitHub, but I'm not distracted when I'm on GitHub. And so that was pretty bad. But when it did work, it worked really well."
- "For the first maybe 20 minutes, it rarely had false alarms, but then the last 20 minutes, it had a lot. It would detect 100% of the distractions, but for false alarms, it was quite often too."
- "[So you didn't have any distractions that didn't catch?] No, none."
- "On my daily work I experience distractions literally every 10 minutes. Every 10 minutes I'll open up a new site or I'll be like, oh look at this, while I'm working. It makes a huge difference."
- "The only thing is, at the end, I don't know if it's the VR glasses or maybe I've just been paying attention for so long, but I get fatigued. I get tired."
- "The biggest downside is the weight of the headset, but once you're in there and you're working, you can't get distracted. That's actually a big thing, yeah. It helps a lot."
- "I don't think the alarm is too annoying actually. The only thing is, maybe it doesn't give you enough feedback that you are distracted. Maybe it should be more loud or something."
- "I would definitely be very likely to use it, especially if I had to work on something that required extreme attention."
- "The hardware would have to be a bit better. It would have to be glasses you just put on and stuff like that. But still, I'd definitely recommend it."
- "I think only using the detector (without the VR) wouldn't be as useful, because one of the things with the VR environment is that you can't really turn your head and play with physical objects, for me at least."
- "I can definitely see people using it in the future if they have it at home or if they go to a library and want to study. I think the potential is really high. At least for me, it helped me manage my ADHD symptoms a lot."
- "In workspaces, I could see it if you have a white-collar job. I used to be an accountant. I could see people using it there if the environment around them is too distracting."
- "Really the only thing I would improve is the false alarms when it beeps. That for me got pretty annoying at the end. Lighter headsets would also be really nice."

**Session 2**

- "It was good. I definitely felt like really focused."
- "I think it helps a lot because you can't really get distracted on anything else."
- "I feel like at first it was when I was being distracted, it was always accurate. Like if I was distracted, it would always be accurate, but it would sometimes think that I was distracted when I wasn't."
- "When false alarms went off, usually I would be like reading through code on GitHub."
- "It got like, if I would go onto social media, like Reddit, it would get it. It also got like, I like doing like shopping websites. So like it got those too. I didn't try YouTube because honestly, after trying the first two, like I just couldn't get distracted. I think it got them all."
- "Doing my daily work like, literally, I get distracted like all the time."
- "Like I know for me, honestly, if I didn't like study for an exam and it was like the night before, I would actually probably put this thing on. [...] I would use both the VR with the detector."
- "I know that one of the biggest things with ADHD people is that it's hard to make habits. So like when I'm on the computer, it might be hard to make the habit of staying focused on something. And like this kind of helps like reinforce it."
- "I think that, yeah, beyond education there is a lot of potential too. You have people who are like trying to find places to like work outside of home. And so like there's no real place you can just kind of like show up and work and, you know, like be able to focus. Like if you go to a Starbucks, like I can't focus."
- "[about the alarm sound] I think that it's, I think that it's the right amount because it's not terrible on the ears. But like it's just bad enough to where like, I remember when you told me to go on something that would distract me, I did for like the first minute. And then I was like, I can't keep listening to this. I'm just going to go back to work. [...] I like the sound." [Would you prefer a light?] I think I would kind of like a bit of both, like a sound to get like the alert and then like maybe the screen to tell me to get back to work or something."
- "I would suggest maybe getting a lighter headset. That's the biggest thing for me. Like for me, the headset is like, it's way too heavy. [...] Like I get like, I get tired from it, how heavy it is."

### A.1.3 Participant 3

**Session 1**

- "The VR environment has helped me because it's a closed environment. There are no distractions. For example, my cell phone rang quite a few times, but my cell phone wasn't there. It was there, but I couldn't see it."
- "The distraction detector has helped me stay focused. I think it's good that it's not a very intrusive signal because otherwise, it throws you off your concentration."

- "It hasn't helped me that much [the distraction detector] because I haven't lost my concentration that much during the session. In fact, it helps the fact that you know that if you lose focus it's going to let you know."

- "It has not generated many false alarms."

- "I find the environment quite useful [for managing distractions]. Maybe not for people our age [21-24]. And not as something constant, but for children, who find it harder to concentrate. It's quite an interactive way for them to concentrate on something. The signal that they are losing concentration is quite subtle, so to speak, the beeping. It's better than a teacher yelling at them or telling them to simply concentrate."

- "I think that for older people it should be much easier to use. If you didn't need the headset I would be more likely to use it."

- "I think its appeal decreases a bit as you get older because it's more childish and you don't get into the situation as much."

- "I don't know to what extent you can do that, but, if you can separate the environment from the distraction detection itself. That is, you don't have to have the glasses. If you had the distraction detection part separately, with that, you would already be more efficient. [...] I would use it more that way. I could extend my session more in time because the headset is heavy. I don't know how feasible it is for a person to wear the headset for a prolonged period of time."

- "Using the environment at work seems just as interesting to me, but I would see it as more intrusive. The fact that they control 100 percent of the work you do... I find it to be useful, but do I think it's good? that would be another thing. In order to increase the user's productivity it would be very useful, but privacy would have to be a priority."

- "I would consider adding ambient classroom or office noise. It would be a good line in the future to be able to adapt the sound to the environment you want to generate."

**Session 2**

- "I noticed that if I'm constantly distracted it beeps every 10 seconds. If you are constantly on YouTube, for example. On Spotify it has also beeped. On news sites as well. [...] Whenever there was a distraction, it always started beeping."

- "False alarms have been rare. On one of the pages I have worked on it was beeping. Also when opening a PDF."

- "The VR environment, since it is more isolating, I think it helps with concentration."

- "One thing it didn't detect is if, for example, I'm working but I had a YouTube video in the background, it didn't detect it. I mean the sound."

- "If it were much more comfortable I would use it in my day-to-day life."

- "It seems to me a system that can very easily be added to parental controls that exist right now on phones and computers. As soon as a kid enters YouTube, apart from the fact that it can beep, I don't know, for example, for two minutes, and as soon as the kid doesn't leave YouTube, the page is blocked and that's it. Or for sensitive content as well."

- "I would have liked to tell the system that the 'Clickup' page was work. But of course, if you put a mute button, it's very easy to hit it every time it beeps."

### A.1.4 Participant 4

#### Session 1

- "The session was good and I was able to focus. The environment and the detector helped me."

- "The headset was a little heavy and blurry at some point. It's a little uncomfortable to use but the idea of an isolated environment it's really good."

- "I never opened anything that was not about work, but the distraction detector still beeped a lot of times, [...] 5 or 6 times during the session."

- "Something very interesting happened. I was working with two windows, so it started to beep when I opened a second. So I just changed the size of the windows and it stopped."

- "Generally I don't open games or social media on my computer. I do this on my phone. [...] So when I'm using the computer, I get distracted changing tasks, switching all the time or opening too many things."

- "I think it's very helpful to help me manage ADHD symptoms. I don't think it's perfect though. You can't use it for more than an hour I think. It is uncomfortable."

- "When it started to beep I asked myself if I was changing tasks because that was important for the detector. I wasn't getting distracted but it was a reminder. Made me ask myself if that was necessary."

- "If I had it at home, I think it would be cool, but wouldn't use it. At least the VR. The distraction detector I would use it for sure. If it didn't beep too much, of course."

- "I would recommend everyone to try it."

- "I think these tools have great potential if some things get better. Great potential. I really think so."

- "I think people going back to office work after the lockdown (COVID) and who miss the silence of working from home would really find it useful."

- "You just need to stop it from beeping when I'm working."

- "I like the low pitch of the sound."

**Session 2**

- "Was not a good session. I think it's because I was wearing the glasses so the headset didn't fit well so in the end I was with my head in pain and my face in pain. [...] The detector thing was nice. Yeah, and I had a good time trying it."
- "Like with respect to the environment, it was uncomfortable. The attention is good because it blocks everything that's happening in the surroundings and it's the idea that I'm using something only to work so I'm supposed to be working. So the headset works like a reminder that I should be working too. So it's possible to relate to the environment and the headset to what I should be doing."
- "I didn't see the cabin this time. It disappears for me. I was just looking at the computer. I forgot about the environment so it's good."
- "The environment can help me focus on my distraction, too. Not when the distraction detector is on. [...] When I stopped working to read a comic book it helped me to be very focused on the comic book. So when you asked me to go back to work, I didn't want to because I was reading the comic book. So I was focused on my distraction."
- "Sometimes as I was working and the detector went off. [...] So when that happened I was like 'Oh this thing exists'. And then it was good to make me think about what I'm doing. That was good."
- "If it didn't have false positives it would help a lot."
- "When I was distracted it was really good. It only didn't get one thing. A comic book in PDF format. Maybe because it was in black and white. It was really good with strong colors I think. I opened Netflix, Amazon shopping, a blog... it was really good with those."
- "Generally, I get distracted by my cell phone. With the computer, I get distracted changing tasks."
- "Using the VR glasses is bad."
- "I would definitely use the detector in the future, the VR with the headset not so much."
- "If I really wanted to get distracted, I could live with the sound beeping."

### A.1.5 Participant 5

Participant 5 expressed a preference not to be recorded during the interviews. Therefore, the information provided by this participant was solely captured through the questionnaire.

## A.2 GitHub code repository

The code repository associated with this research project is publicly available at the following link: `https://github.com/gonzalo-cordova-pou/MLADHD`. Researchers and practitioners interested in accessing and utilizing the codebase can visit this link for comprehensive documentation and implementation details. In the event of any issues encountered while accessing the repository, the author can be contacted via email at `gonzalo.cordova@estudiantat.upc.edu` or `gonzalocp6@gmail.com` for prompt assistance and support.

## A.3 Hugging Face spaces

For those seeking to directly test and utilize the Distraction Detector (CNN binary classification model) or the Topic Modeling model, we have provided two Hugging Face Spaces:

- **Topic Modeling - Topic DeTextor**
  Link: `https://huggingface.co/spaces/mladhd/TopicDeTextor`
- **CNN Binary Classifier - Distraction Detector**
  Link: `https://huggingface.co/spaces/mladhd/DistractionDetector`

In the event of any issues encountered while accessing the space, the author can be contacted via email at `gonzalo.cordova@estudiantat.upc.edu` or `gonzalocp6@gmail.com` for prompt assistance and support.