



Eye Tracking and Wellness: The Quest for Unobtrusive Biomarkers for Designing Smart Neuro Information Systems

by

Javad Norouzi Nia

A Dissertation

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Doctor of Philosophy

in

Business Administration (Information Technology)

August 2022

Approved:

Prof. Soussan Djamassbi, PhD Committee Chair, School of Business, WPI

Prof. Diane Strong, PhD Committee Member, School of Business, WPI

Prof. Randy Paffenroth, PhD Committee Member, Data Science, WPI

Table of Contents

1. INTRODUCTION.....	1
2. BACKGROUND.....	4
2.1. Eye Tracking Chronic Pain	5
2.1.1. Participants' Characteristics.....	7
2.1.2. Stimulus and Task	7
2.1.3. Analysis Method.....	9
2.1.4. Operationalization and Quantification of Attentional Bias.....	9
2.2. Limitation of Previous Studies	30
3. Methodology	32
3.1. Stimulus Design	32
3.2. Eye-tracking Apparatus.....	33
3.3. Data collection Process.....	33
3.4. Eye Movement Metrics.....	34
3.5. ML Methods and Settings	42
3.6. Model Evaluation	43
3.6.1. Confusion Matrix.....	43
3.6.2. Accuracy.....	44
3.6.3. F1-Score	44

3.7.	Model Selection.....	45
3.8.	Pre-Processing Data.....	45
3.8.1.	Creating AOIs	45
3.8.2.	Data Cleaning	46
3.8.3.	Adding new variables.....	46
3.9.	Preparing Data for Machine Learning Algorithms.....	47
3.9.1.	Splitting Data to Train and Test Sets	47
3.9.2.	Balancing Data	48
3.9.3.	Un-arranged data.....	48
3.9.4.	Re-arranging the Data.....	49
3.9.5.	Addressing Missing Values.....	50
3.10.	Proposed iterative process for developing ETML to detect chronic pain.....	50
3.10.1.	ETML Development – Step 1: Collecting Initial Eye-Tracking Dataset.....	52
3.10.2.	ETML Development – Step 2: Developing a Proof of Concept	53
3.10.3.	ETML Development – Step 3: Validating the Proof of Concept (the Best Model of Step 2).....	64
3.10.4.	ETML Development – Step 4: Refining the Proof of Concept (the Best Model of Step 3)	68
3.10.5.	ETML Development – Step 5: Validating the Refined Proof of Concept (From Step 4).....	73
3.10.6.	ETML Development – Step 6: Refining the Proof of Concept (With Step 3 Merged Set).....	75
4.	Discussion	81
5.	References	87

6. Appendix I..... 93

7. Appendix II..... 98

Abstract

Human needs are being increasingly addressed by information technologies. As a consequence, market competition has shifted toward developing innovative user experiences. Neuro information systems (NeuroIS) that detect user needs automatically can play a major role in addressing the continual demand for innovative user experiences in today's digital economy. By using sensors that can collect physiological measurements from users, NeuroIS can provide a continuous stream of valuable objective data for detecting user needs in various problem domains.

One problem domain that can be addressed by NeuroIS is chronic pain. Chronic pain is a major public health problem that impacts 1 out of 5 American adults. Assessment of chronic pain is often achieved via self-reported scales. Research indicates that identifying measures that can objectively assess chronic pain can improve its effective treatment. Grounded in pain and eye tracking literature, I developed a theory-based eye tracking machine learning (ETML) engine as a proof of concept. Grounded in prior research, I develop a set of visual stimuli and an extensive set of eye tracking features that can be used to detect a person's chronic pain status from the person's eye movements. Grounded in user-centered design framework, I also propose and test an iterative process for developing such a NeuroIS (ETML engine) over time as more data becomes available. The results of my project show that the visual stimuli and the eye tracking features that I have developed for designing the ETML can indeed help to build a reliable NeuroIS for detecting chronic pain status. The results of my project also suggest that my proposed iterative process is likely to produce a robust ETML that can predict chronic pain status with 80% accuracy or more.

Acknowledgment

My gratitude goes out to Professor Strong and Paffenroth for their support. The lessons I learned from you over the last few years will definitely help me in my next steps. In addition to teaching and transferring experience and best practices over the past 4 years, I would like to thank Professor Djamasbi for her patience and for providing me with the chance to learn.

Additionally, I would like to thank my parents who have supported me for decades, and last but not least, I would like to thank Shabnam, my wife, who has patiently supported me throughout all my difficult moments.

1. INTRODUCTION

Today's digital economy is driven by continuous market demand for innovation. This market need can be met by creating smart adaptive devices that provide useful services with excellent user experiences (Djamasbi and Strong 2019). Neuro information system (NeuroIS) research can address this pressing market need thanks to advances in technology that make it possible to use sensors to collect physiological measures without burdening users. NeuroIS research aims to use physiological measures to detect changes in user experiences and/or behaviors, so that systems can be designed to respond in real time to user needs (Fehrenbacher and Djamasbi 2017; Shojaeizadeh et al. 2019). Because vision is the dominant sense of human being (Pocock 1981) research in this area recently received more attention and was used in NeuroIS research. To collect data about this dominant sense, the use of modern eye trackers has become more popular. These eye-trackers can unobtrusively collect accurate gaze data without requiring additional gears such as chin rests. As eye movements reveal a great deal about a person's attention to objects in visual environments, and that they can be collected unobtrusively, eye tracking has become a gold standard for investigating user experience and behavior objectively (Alrefaei et al. 2022; Djamasbi 2014; Norouzi Nia et al. 2021; Shojaeizadeh et al. 2019).

One problem domain that can benefit from developing eye-tracking enabled NeuroIS is chronic pain. Pain is defined as “a distressing experience associated with actual or potential tissue damage with sensory, emotional, cognitive, and social components” (Williams and Craig 2016). A pain experience that is rated as 4 or higher on a 0-10 low to high scale and lasts more than 3 months is called chronic pain (Alrefaei et al. 2022). Chronic pain is considered as a major public health problem that afflicts about %20 of American adults (Yong et al. 2022). Chronic pain is often assessed with self-reported measures. While understanding pain from a patient point of view is important, it lacks the objectivity that is often needed for effective treatment and management of chronic pain (Alrefaei et al. 2022). Developing a NeuroIS that can detect chronic pain status from eye

movements can serve as a first step toward providing objective measures for chronic pain.

Regarding the importance of chronic pain, the National Health Survey introduced a pain module in 2019. The module estimated that 50.2 million (1 out of 5) American adults experienced chronic pain. This group missed 10.3 working days compared to non-chronic pain people who, on average, only missed 2.8 days. This difference between the missed days of the two groups is significant ($p < 0.001$). Based on the survey, the impact of chronic pain on the economy is notable. The US annual lost wages are estimated to be \$79.9 billion, and the lost productivity can account for \$300 billion. In contrast, the estimated yearly US expenditure on medical expenses lost productivity, and disability programs comprise 560 billion dollars. The estimates shed light on the magnitude of the topic and demonstrate the importance of improving treatment of chronic pain, for example, by providing objective measures for identifying and treating chronic pain.

In this dissertation, I address this gap in knowledge by developing an eye tracking machine learning ETML engine that can detect chronic pain from the objective measure of eye movements as a proof of concept. In addition to building the proof of concept, I propose and test an iterative process to continue refining the initial proof of concept. When designing new products, minimum viable products (MVPs) are tested iteratively with smaller sets of data until a success threshold is achieved (Tullis and Albert 2013). By gathering insight iteratively from smaller datasets, such an iterative process facilitates efficiency. Similarly, building an ETML proof of concept with a relatively smaller set of data, and validating and refining it as needed (iteratively) with new sets of data allow us to build reliable engines over time efficiently and cost-effectively.

Research has shown that eye movements can provide insight into task load (Fehrenbacher and Djamasbi 2017) and can be used to detect cognitive load automatically (Shojaeizadeh et al. 2019). As pain impacts cognition, and eye movements can reliably measure cognitive load, it is reasonable to assume that eye movements can also measure chronic pain. I explore this possibility through developing an eye-tracking machine learning (ETML) engine, as a proof of concept that is grounded in eye tracking

and pain literature, that predicts chronic pain status. My proposed iterative process for developing such an ETML, requires iterative validation with new sets of data. I set the threshold for validation success in my project to 80% accuracy. For this accuracy threshold and given the complexity of the problem domain (e.g., predicting chronic pain status from eye movements), it is reasonable to expect that the development process must go through several iterations before it can pass the success criteria. As expected, the initial proof of concept in my project did not achieve the 80% validation accuracy threshold that was set as the success criteria for developing a finished product. Hence, I followed my proposed iterative process, and as the last step in my project, I used the entire dataset to refine the ETML. The refined ETML developed based on the combined datasets supported the initial theoretically based proof of concept in this project; it showed that the eye tracking metrics and visual stimuli used in the project were effective in predicting chronic pain. Hence the refined ETML, based on my proposed iterative process, warrants further development. While refining the ETML with the combined datasets was the last step in my current project, the results obtained provide support for the theoretical soundness of the ETML as well as the visual stimuli, and task used to collect eye movements. The results from iterative refinement of the ETML, suggests that the proposed iterative process is likely to results in a robust ETML for predicting chronic pain.

In the following sections, I explain the theoretical background for the design of the ETML. This was achieved by conducting a thorough review of literature that used eye tracking to investigate the impact of chronic pain on attention to visual stimuli. Following the review of literature, in the method section I describe the method that I used to construct the ETML that can detect chronic pain status by expanding the task paradigm and eye movement metrics used in the reviewed studies. I also explain the method that I used to validate the ETML proof concept. Then I explain my proposed process for developing such ETML iteratively with newly collected sensor data over time. Finally, I explain how I used my proposed iterative process to develop a theoretically based refined proof of concept for predicting chronic pain-status using only eye movements.

The results of my project show that the task (reading) and visual stimuli (text passages) used to develop the ETML provided a rich context for investigating the impact of pain on visual behavior. Hence, a major contribution of my project is extending the task paradigm in chronic pain literature. The current chronic pain literature focuses mostly on measuring the initial stages of attention (e.g., using metrics such as first fixations, or duration of first visit). Extending the task paradigm into reading allows us to investigate the impact of chronic pain more effectively in later stages of attention, such as attention maintenance (e.g., using saccadic eye movements to measure a change in attention and pupillometry to measure the intensity of effort). I extended the currently used eye metrics to a rich set of eye movement features, which were not previously used in the chronic pain literature. This set is suitable for measuring both initial and later stages of attention. In this regard, my project majorly contributes to chronic pain and eye tracking literature.

The results of my project have also major contribution for NeuroIS research. The results showing that the eye movement metrics I used to develop the ETML were reliable predictors of chronic pain status, establish eye movements as reliable biomarkers of pain. The methodology and iterative process I proposed for developing the ETML proof of concept lays the groundwork for developing NeuroIS that can predict chronic pain objectively, automatically, and unobtrusively. According to a recent user-centered framework (Djamasbi and Strong 2019), to address the challenge of continual market demand for innovation, we must develop smart engines that can adjust and modify as more data becomes available. The iterative process that I propose in this project for designing an ETML iteratively over time can serve as an initial framework for developing such smart NeuroIS systems that can address the market demand for innovation.

2. BACKGROUND

The ETML engine development in my project was carried out using data collected through an eye tracking study that required participants to read four different short texts (each about 100 words). Eye movement metrics used to develop the ETML as well as the design

of visual stimuli and the task in the study were based on a systematic review of literature involving chronic pain and eye tracking. This review, which examines eye movement metrics, visual stimuli, and task paradigms used to compare viewing behavior of people with and without chronic pain, are provided in this section.

2.1. Eye Tracking Chronic Pain

I began my research by looking into prior chronic pain studies that used eye tracking.

The amount of chronic pain research that has used eye tracking methodology is somewhat limited as attested by a review article published in 2020 (Chan, Suen, Jackson, et al. 2020). An overview of the research that used eye tracking data to study pain has been published in a 2020 review paper. The systematic review conducted by this article resulted in 24 papers. Out of the 24 papers, 9 of them dealt with chronic pain. To update the findings of the 2020 review regarding chronic pain to include papers from 2020-present, I conducted a systematic review using the same keyword combinations used in Chan's review paper to maintain consistency throughout the study. The search was adjusted to return related papers in the three databases, as there were differences in the logic order of the three databases. I used the same databases except one, Web of Science. I used Scopus instead of Web of Science because it contains a larger set of journals. There are 21,950 journals in Scopus in comparison to 13,100 journals on the Web of Science according to(Iowa State University 2022). A further fact to consider is that about 99.11% of journals indexed by Web of Science are also indexed by Scopus. The keywords used for each database can be found in Table 1.

Table 1: Databases and Keyword	
ProQuest	TI,AB,IF(pain) AND TI,AB,IF(eye PRE/0 track* OR gaze AND behavio*r OR ems OR eye PRE/0 movement* OR fixation) AND TI,AB,IF(attention* PRE/0 bias* OR selective PRE/1 attention OR vigilance OR hypervigilance OR avoidance OR maintenance OR disengagement)
PubMed	((((pain[TW]) AND (("eye-track*" [TW] OR "eye track*" [TW] OR "gaze behavior" [TW] OR EMs [TW] OR "eye movement*" [TW] OR fixation [TW]))) AND

	(2020/5/1:3000/12/12[pdat])) AND ("attention* bias*" [Text Word] OR "selective attention" [Text Word] OR vigilance [Text Word] OR hypervigilance [Text Word] OR avoidance [Text Word] OR maintenance [Text Word] OR disengagement [Text Word]))
Scopus	TITLE-ABS-KEY ((pain) AND (eye- AND tracking OR eye AND tracking OR gaze AND behavio*r OR ems OR eye AND movement* OR fixation) AND (attention OR bias* OR selective AND attention* OR vigilance OR hypervigilance OR avoidance OR maintenance OR disengagement))

My systematic review resulted in 96 papers, after removing duplication and making sure they were about chronic pain 9 papers were left. following the criteria of the 2020 review paper, I updated the search to include any related papers published after May 2nd, 2020, which resulted in finding 18, 59, and 19 papers in ProQuest, PubMed, and Scopus databases, respectively. The repetitive papers were removed from the review. Due to the fact that the purpose of this study is to differentiate chronic pain from healthy people, I excluded any research that did not include chronic pain and healthy participants, did not use eye-tracking data, or was a review of other studies. My systematic review resulted in 96 papers, after removing duplication and making sure they were about chronic pain 9 papers were left. This process resulted in a total of 18 papers (see Table 2).

During the process of searching for relevant original papers, I came across of another review paper (Jones et al. 2021) that covered papers up to 2021. My updated review included all the relevant papers that were listed in the 2021 review and expanded that review to include all relevant papers to present.

I report the findings of the comprehensive review in the following fashion. I begin by discussing participants' characteristics in the reviewed papers. I then summarize the task/stimulus they used and continue by discussing the analysis methods that were used in the reviewed studies. Next, I explain the type of analysis used in these studies. Then I explain how eye-tracking was operationalized in the context of chronic pain and provide a summary of the analysis of the 18 relevant papers reviewed for my project in Table 2.

2.1.1. Participants' Characteristics

Most studies conducted their investigations by comparing the viewing behavior of those who suffered from chronic pain with those who were pain free. There were a few researchers who, in addition to recruiting both chronic pain and healthy groups, have also divided the participants into subgroups and matched the number of participants in each group in the study. For instance, (ten Brink et al. 2021) focused on one type of chronic pain, Complex Regional Pain Syndrome (CRPS), but in addition to recruiting participants with CRPS along with healthy participants, the same numbers of patients with chronic pain but with different kinds of pain were also recruited, to create a pain control group. (Giel et al. 2018), for example, created a control group with depressive symptoms matching the chronic-pain (CP) group, or (Chan et al. 2022) investigated the reactions of participants in four groups of young CP and pain-free (PF), as well as old CP and PF. Here, for comparing the eye movement and other characteristics of chronic pain patients with those of healthy participants, only studies that recruited chronic pain patients and healthy participants were included.

2.1.2. Stimulus and Task

There has been a trend to use a dot-probe task as one of the most commonly used tasks in pain studies (Chan, Suen, Jackson, et al. 2020), and it has even become popular among eye tracking research that is focused on chronic pain. Researchers used to use the dot-probe method, where the participants were presented with either pain or non-pain words or faces, and they were then asked to respond to a dot that appeared after the stated words or faces appeared to compare their response to pain or non-pain. Due to the limited amount of information about a user based on reaction time alone, researchers have started to use eye trackers to gather information beyond what the initial attention to a dot in the dot probe will reveal about the user.

Amongst those researchers, (Fashler and Katz 2014; Yang et al. 2013) used neutral words and pain words as stimuli for dot-probe tests with an eye tracker, while (Fashler and Katz 2016; Franklin et al. 2019) used scene images (neutrals, injuries) and (Mazidi et al. 2021) used faces (happy, sad, neutral).

Another common task used in these studies is free viewing, during which participants are asked to explore one or more images, faces, or activities that appeared in random locations or participants were told where to look for them. In some studies, emotional faces such as happy, sad, painful, and angry photos have been presented to the participant, such as in (BlaisdaleJones et al. 2021; Chan et al. 2022; Chan, Suen, Hsiao, et al. 2020; Giel et al. 2018; Lioffi et al. 2014; Priebe et al. 2021; Soltani et al. 2022). There have also been studies in which scene images such as painful daily activities as well as neutral daily activities were used (Mahmoodi-Aghdam et al. 2017; Shiro et al. 2021). As an example of a painful activity, (Shiro et al. 2021) use a video where a person touches the hand of another person. Considering that participants in this study had chronic pain in one hand, touching the same hand of the person in the video could trigger the pain for the participant and result in different eye movements, which were recorded and analyzed.

Researchers also used visual search to investigate the impact of pain on cognition. Hence, for example, researchers have used face images (pain, angry, happy, neutral) (Schoth et al. 2015) or another shape such as a diamond (Koenig et al. 2021) images to investigate participant's ability to distinguish between the target and distractors. (Soltani et al. 2020) used Flanker task, a visual filtering task.

The paper by (ten Brink et al. 2021) implemented a multi-task design involving dot-probe, free viewing, and visual search, as well as free-viewing.

While Dot probe is the most popular task used in the reviewed studies it has an inherent limitation. When single words are used as a stimulus, it is possible that participants interpret the same word differently. For example, one participant might perceive "sharp" as a neutral word given that it recalls a sharp knife, while another might view it as a pain

word because it recalls sharp pain. This ambiguity then can affect the results obtained from eye movements. A different stimulus with richer context is likely to reduce ambiguity and by doing so remove the noise caused by possible different interpretation of the stimulus.

2.1.3. Analysis Method

Regarding the analysis method, 16 of 18 relevant papers used only basic statistical methods, while 2 papers used machine learning as well as basic statistical methods.

None of the papers which used basic statistical methods were able to find significant differences in viewing behavior between the groups. The two papers (Chan et al. 2022; Chan, Suen, Hsiao, et al. 2020) that used machine learning methods used Hidden Markov models to cluster the eye movements of participants to nose-centric versus eye-centric in one paper and explorative versus focused in the other paper. Then they used basic statistics to compare viewing behavior (e.g., nose vs. eye centric or explorative vs. focused) between the chronic pain and pain-free groups. Yet, they could not find a significant difference between eye movements of the two groups.

Overall, neither of the approaches found a significant difference between the eye movements of chronic pain and healthy participants.

2.1.4. Operationalization and Quantification of Attentional Bias

Eye tracking analysis starts by identifying areas of interest (AOI) in visual stimuli. A defined AOI is used by the investigator to study viewing behavior related to that specific region of stimuli. Participants will not be able to see the AOI, as it is only visible to the researchers.

Eye movement variables used in the reviewed studies are related to fixation, visit, and only one study used pupillometry metrics. Fixations refer to relatively slow eye movements, during which a visual stimulus can be processed and understood. Hence, fixations are considered as reliable indication of attention.

Visit refers to the period of time between the first fixation to an AOI, and the last one before a different AOI is looked at. A visit can include a single fixation or a series of fixations, and a saccade (high velocity eye movement between fixations) or more.

Pupillometry metrics, which measure pupillary responses to stimuli, are considered reliable indication of cognitive activity (Shojaeizadeh et al. 2019). Only one out of 18 papers used pupillometry metrics to study chronic pain.

None of the studies used saccadic eye movements, which refer to rapid eye movements that change the focus of attention from one focal point (fixation) to another.

In the following paragraphs, I will explain the eye movement variables that were used in the reviewed studies. I start by discussing the fixation metrics followed by visit metrics. For a fair comparison, some researchers calculated Cohen's *d*. Whenever this value is reported by researchers, it is referred to herein as "*d*".

First fixation proportion:

The percentage of the number of fixations on an AOI during the first visit divided by the total number of fixations on that AOI during all visits.

A study by (Mahmoodi-Aghdam et al. 2017) found that all participants had significantly more first fixations ($d = 2.00$) for painful activity than neutral images, whereas a study by (Giel et al. 2018) found that all participants had more first fixations ($d = 1.44$) for happy faces than neutral faces. However, these differences were not significant between groups. The proportion of first fixations was not found to differ significantly across studies

carried out by (ten Brink et al. 2021; Chan et al. 2022; Liossi et al. 2014; Mazidi et al. 2021; Schoth et al. 2015; Soltani et al. 2020; Yang et al. 2013).

Time To First Fixation (First Fixation Latency):

The time between the onset of the stimulus and the first fixation on a particular AOI.

(BlaisdaleJones et al. 2021; Schoth et al. 2015) have both found that all participants had a shorter first fixation latency for pained faces and happy faces compared to neutral faces, but there were no significant differences between the groups. (Mahmoodi-Aghdam et al. 2017), on the other hand, reported shorter first fixation latency between pain-free groups for images of neutral versus painful activity ($d = 0.70$). Additionally, (Franklin et al. 2019) reported that the chronic-pain group had shorter first fixation latency ($d = 0.90$) for painful activity images than neutral images, and the chronic-pain group also had shorter first fixation latency ($d = 0.70$) on painful activity images than the pain free group. As a contrast,(Mazidi et al. 2021; Yang et al. 2013) did not find any significant differences between groups in regard to their time to first fixation.

First Fixation Duration:

The duration of the first fixation.

A study by (Yang et al. 2013) reported that the chronic-pain group had a shorter first fixation duration on health-catastrophe words than the pain-free group when compared with the chronic-pain group. According to (Mahmoodi-Aghdam et al. 2017), people with higher current-week pain severity had shorter latency for first fixation of painful activity images ($d = 1.38$) when compared to those with lower pain severity. On the other hand, (Koenig et al. 2021; Liossi et al. 2014; Mazidi et al. 2021) found no significant differences in the duration of the first fixation between the two groups.

Average Fixation Duration:

Average duration of single fixations.

(Lioffi et al. 2014) reported that participants' average fixations on happy faces were longer than those on pained faces ($d = 0.27$ for all participants). On the other hand, (Franklin et al. 2019) found that the average duration of fixations for neutral images was longer in the chronic-pain group than for images that contained painful activity ($d = 0.80$). A longer average fixation duration ($d = 0.90$) was also reported in the chronic-pain group compared to the pain-free group on images showing painful activity. The average duration of fixation, however, did not appear to differ significantly between the two studies conducted by (Fashler and Katz 2014, 2016).

Total Fixation Count:

Total frequency of fixations

(Fashler and Katz 2014) reported that the chronic-pain group had more total fixations on sensory-pain words than the pain-free group ($d = 0.48$). The studies conducted by (Fashler and Katz 2014; Mahmoodi-Aghdam et al. 2017), respectively, also found that all participants had higher total fixations on pain stimuli as compared to neutral stimuli ($d = 1.54$) and ($d = 2.73$). (BlaisdaleJones et al. 2021) also found that there was a significant difference in the number of fixations made on pain stimuli by all participants, but no significant difference between groups. (Franklin et al. 2019), however, found only the chronic-pain group had more fixations ($d = 1.57$) for images containing painful activity than those with neutral activity. In spite of these findings, (BlaisdaleJones et al. 2021; Chan, Suen, Hsiao, et al. 2020; Mazidi et al. 2021; Shiro et al. 2021) did not find any discernible difference between the groups.

Total Fixation Duration:

Total duration of fixations on a specific AOI.

Visit duration includes the amount of time for both fixation and saccades (rapid eye movements between fixations). In addition to recording the visit duration of the stimulus over its whole visible duration, some researchers recorded it over shorter periods. In dot-probe, for example, if the words were visible for 2000 ms, the visits during 0-500, 500-1000, 1000-1500, and 1500-2000 ms are separately recorded.

As reported by (Fashler and Katz 2014), CP participants experienced a longer period of time during their 1000-2000 ms visit for sensory pain compared to neutral words ($d = 0.76$). In contrast, (Fashler and Katz 2016) noted that all participants had longer total fixation durations for neutral images during 0-500 ms ($d = 0.77$) and for injury images during 500-1000 ms ($d = 0.53$), and CP group had a longer total fixation duration for injury images during 1000-2000 ms.

As for faces, (Giel et al. 2018) found that participants had a longer total fixation duration ($d = 1.69$) for happy faces than for neutral faces, while (Mazidi et al. 2021) found that the duration of the visit for happy faces was longer than the duration for pained faces during 1000–1500 ms ($d = 0.31$).

In (BlaisdaleJones et al. 2021), there were significant differences in visit duration between all participants subjected to a pain stimulus, whereas no significant differences were identified between groups on pain stimulus. Furthermore, (ten Brink et al. 2021; Chan et al. 2022; Mahmoodi-Aghdam et al. 2017; Yang et al. 2013) have not found any remarkable differences in the duration of total fixations.

Total Visit Count:

Frequency of visits of an AOI.

A greater number of total visits ($d = 1.11$) were reported in (Fashler and Katz 2014) for sensory pain words than for neutral words by everyone in the study. In addition, (Lioffi et al. 2014) found that the number of visits to neutral faces was higher than that to angry faces ($d = 0.26$) and pained faces ($d = 0.38$), as well as to neutral faces compared to angry faces ($d = 0.45$) and pained faces ($d = 0.57$). As per (Fashler and Katz 2016) similar results were obtained with more participants visiting for injuries ($d = 1.66$) than for neutral images. The findings of (Koenig et al. 2021), however, showed a higher number of distractors visiting in higher threat conditions, but not a significant difference between the two groups.

Average Visit Duration:

Average time spent on an AOI.

There was a longer visit duration for sensory pain than neutral words ($d = 0.87$) in a study by (Fashler and Katz 2014) among chronic-pain patients. A more recent study by (Fashler and Katz 2016) found that the average visit duration for injury images ($d = 0.74$) was longer than the average visit duration for neutral images among all participants.

In my review, I discussed popular variables used in chronic pain studies. Despite this, there are a few researchers who use other eye-tracking variables. For instance, only a team of researchers used the average duration of a visit in their analyses. Using none of those variables, the researchers were able to separate the chronic pain group from the healthy group.

Moreover, there were limited variables in the relevant papers due to the limitations of the analysis method, basic statistics. 16 out of these 18 papers used basic statistical methods with no more than 6 variables. Yet, both papers that used machine learning methods that

could handle large sets of variables used no more than five variables. Overall, in previous research, very few variables are used to distinguish chronic pain from healthy people.

Additionally, there have been eye tracking studies which, in order to understand user characteristics such as cognitive loads and develop automatic tools for detecting user status, used other variables such as saccadic or pupillometry features (Shojaeizadeh et al. 2019). As pain is a complex phenomenon that can have a significant impact on cognitive function, I use a wide range of eye tracking variables that measure attention (fixation) change in attention (saccade) and cognitive effort (pupillometry) in order to develop my tool.

A summary of 18 related papers is provided in Table 2:

Table 2: Summary of Literature Review

	Participants	Task	Research Question/ Hypotheses	Eye Movement Variables	Key Findings
1	(Yang et al. 2013) CP and PF groups with high and low fear of pain subgroups (24 CP, 24 PF)	Dot-Probe (2000 ms) Words (sensory-pain, health catastrophe, neutral)	<ol style="list-style-type: none"> Participants with higher fear of pain were expected to display more fixation and shorter time to first fixation on pain/health catastrophe words than neutral alternatives Effects of Fear of pain and pain status on subsequent biases in attention maintenance (first fixation and total fixation durations were explored) 	<ol style="list-style-type: none"> First fixation proportion First fixation latency First fixation duration Total fixation duration 	<ol style="list-style-type: none"> CP had a shorter first fixation duration on health-catastrophe words than the PF group No significant finding for first fixation latency and total fixation duration or first fixation proportion
2	(Fashler and Katz 2014) 51 CP, 62 PF	Dot-Probe (2000 ms) (sensory-pain, neutral)	<p>CP participants, compared to PF, will:</p> <ol style="list-style-type: none"> Fixate more to sensory pain-related words 	<ol style="list-style-type: none"> Total fixation count Total visit count Average fixation duration 	<ol style="list-style-type: none"> CP group had more total fixations on sensory pain words than PF group ($d = 0.48$)

			<ol style="list-style-type: none"> 2. Exhibit a different pattern of sustained attention to sensory pain-related and neutral words 3. Show an attentional bias toward sensory pain-related words at different stages of visual attentional processing 	<ol style="list-style-type: none"> 4. Average visit duration 5. Total fixation duration during 0–500, 500–1000, and 1000–2000 ms 	<ol style="list-style-type: none"> 2. CP participants had longer visits ($d = 0.87$) and longer total fixation duration during 1000–2000 ms ($d = 0.76$) for sensory pain than neutral words 3. All participants had more total fixations ($d = 1.54$), more total visits ($d = 1.11$), and longer total fixation duration ($d = 0.54$–1.26) for sensory-pain than neutral words 4. No significant finding for average fixation duration
3	(Lioffi et al. 2014) 23 CP, 23 PF	Free viewing (4000 ms) Face images (pain, angry, happy, neutral)	CP individuals, compared to PFs, would demonstrate: <ol style="list-style-type: none"> 1. A higher proportion of initial fixations on pain expressions 2. Longer first fixation duration 3. More visits to pain expressions. 	<ol style="list-style-type: none"> 1. First fixation proportion 2. Total visit count 3. First fixation duration 4. Average fixation duration 	<ol style="list-style-type: none"> 1. CP had more first fixations on pain than neutral faces ($d=1.21$) 2. CP had more first fixations on pained faces than PF group ($d=0.79$) 3. All participants had more visits to happy compared to angry ($d = 0.45$) and pained

			<p>4. The current investigation also explored the specificity of bias when negatively-valenced, positively-valenced, and neutral target pictures were presented simultaneously.</p>		<p>faces ($d = 0.57$), and 4to neutral compared to angry ($d = 0.26$) and pained faces ($d = 0.38$)</p> <p>4. All participants had longer fixations for happy than pained faces ($d = 0.27$)</p> <p>5. No significant finding for first fixation duration</p>
4	(Schoth et al. 2015) 23 CP, 24 PF	Visual search, Face images (pain, angry, happy, neutral)	<p>CP individuals relative to PFs would show:</p> <ol style="list-style-type: none"> 1. A significantly higher proportion of initial fixations to target pain expressions 2. Significantly shorter time to first fixation on pain expressions 	<ol style="list-style-type: none"> 1. First fixation proportion 2. First fixation latency 	<ol style="list-style-type: none"> 1. CP group had more first fixations on pain than neutral faces ($d = 1.01$) 2. CP had more first fixations on pained faces than PF ($d=0.93$) 3. All participants had shorter first fixation latency for pained ($d = 1.12$) and happy faces ($d = 0.53$) than neutral faces

5	(Fashler and Katz 2016) 51 CP, 62 PF	Dot-probe (2000 ms) Scene images (injury, natural)	<p>CP individuals would show:</p> <ol style="list-style-type: none"> 1. More fixations on injury pictures 2. Higher fixation duration on injury pictures 3. Longer visit of injury pictures in later phases of attention than PFs. 	<ol style="list-style-type: none"> 1. Total fixation count 2. Total visit count 3. Average fixation duration 4. Average visit duration 5. Total visit duration during 0–500, 500–1000, and 1000–2000 ms 	<ol style="list-style-type: none"> 1. All participants had more fixations ($d = 1.74$), more visit ($d = 1.66$) and longer average visit 2. Duration ($d = 0.74$) for injury than neutral images 3. All participants had a longer total fixation duration for neutral images during 0–500 ms ($d = 0.77$) and for injury images during 500–1000 ms ($d = 0.53$) 4. CP group had a longer total fixation duration for injury images during 1000–2000 ms than the PF group ($d = 0.48$) 5. No significant finding for average fixation duration
6	(Mahmoodi-Aghdam et al. 2017) 20 CP, 18 PF	Free viewing (1000 ms) Scene	<ol style="list-style-type: none"> 1. CP patients show an engagement bias (i.e., the initial orientation of attention to pain- 	<ol style="list-style-type: none"> 1. First fixation proportion 2. First fixation latency 	<ol style="list-style-type: none"> 1. All participants had more first fixations ($d = 2.00$), more total fixations ($d = 2.73$) and shorter first visit duration ($d =$

		images (painful and neutral daily activity)	<p>provoking activity pictures than neutral ones).</p> <p>2. Regarding sustained attention, CP patients, relative to PFs, have difficulty disengaging from pictures of pain-provoking activities or disengaging from those pictures faster than PFs.</p>	<p>3. Total fixation count</p> <p>4. First fixation duration</p> <p>5. First visit duration</p> <p>6. Total fixation duration</p>	<p>0.94) for painful activity than neutral images</p> <p>2. PF group had shorter first fixation latency for neutral than painful activity images ($d = 0.70$)</p> <p>3. People with higher current week pain severity had shorter first fixation latency for painful activity images ($d = 1.38$)</p>
7	(Giel et al. 2018) 17 CP, 17 control group with depressive symptoms matched to the CP group, 17 PF	Free viewing (3000 ms) Face images (happy, sad, neutral)	<p>1. PFs show an early and maintained attentional bias for emotional faces.</p> <p>2. CPs and DCs (depressive symptoms) show facilitated orienting to and longer maintenance on negative emotions (sad faces) while less</p>	<p>1. First fixation proportion</p> <p>2. Total fixation duration.</p>	<p>1. All participants had more first fixations ($d = 1.44$) and longer total fixation duration ($d = 1.69$) for happy than neutral faces</p>

			orienting to and maintaining positive emotions (happy faces).		
8	(Franklin et al. 2019) 18 CP, 17 PF	Dot-probe (500 ms) Scene images (painful and neutral daily activity)	Chronic back pain participants compared to PFs would have: <ol style="list-style-type: none"> 1. A significantly higher percentage of fixations to threatening stimuli 2. A longer average fixation duration to threatening images 3. Exhibit a faster reaction time to threatening images in the dot-probe task. 	<ol style="list-style-type: none"> 1. First fixation latency 2. Total fixation count 3. Average fixation duration 	<ol style="list-style-type: none"> 1. CP group had a shorter first fixation latency ($d = 0.90$), more fixations ($d = 1.57$) and longer 2. Average fixation duration ($d = 0.80$) for painful activity than neutral images 3. CP group had a shorter first fixation latency ($d = 0.70$) and longer average fixation duration 4. ($d = 0.90$) on painful activity images than PF group

9	(Mazidi et al. 2021) 28 CP, 29 PF	Dot-probe (1500 ms) Face images (pain, happy, neutral)	<ol style="list-style-type: none"> 1. CP people would demonstrate increased vigilance towards pain faces compared to PF people. 2. Attentional control would moderate the relationship between catastrophizing and increased attention to pain faces. 	<ol style="list-style-type: none"> 1. First fixation proportion 2. First fixation latency 3. Total fixation count 4. First fixation duration 5. Total fixation duration 6. Total fixation duration during 0–500, 500–1000, and 1000–1500 ms 	<ol style="list-style-type: none"> 1. All participants had a longer total fixation duration for happy than pained faces during 1000–1500 ms ($d = 0.31$) 2. No significant finding for first fixation proportion, first fixation latency, total fixation count, and first fixation duration
---	-----------------------------------	---	---	---	--

10	(Priebe et al. 2021) 20 CP, 20 PF	Free-viewing (2000 ms) Face images (happy, angry, pain)	1. Pain patients would show difficulties disengaging from pain faces, with group differences becoming increasingly evident during later stages of processing	1. Probability of first fixation 2. Fixation duration over 0-500ms, 500-1000ms, 1000-1500ms and 1500-2000 ms epochs 3. The difference between the fixation time for emotional/painful faces and the fixation time for neutral faces	1. All participants had significantly lower first fixation probabilities on pain faces than anger ones ($d = 0.65$) 2. Similar fixation duration between groups 3. Preference for pain faces in some epochs, but no significant difference between groups
11	(Chan, Suen, Hsiao, et al. 2020) 32 PF, 31 CP	Free-viewing (500 ms) Face images (with doctors', patients',	1. People with chronic pain may be either more nose-centered or more eye-centered than healthy controls for neutral faces with different identity labels	1. Fixation location 2. Fixation duration 3. Fixation number Eye movement is used to create eye-centered vs. Nose-centered and	1. The CP group endorsed more negative interpretations for ambiguous scenarios but did not differ in attentional processing of faces with different identities.

		and healthy labels)	2. Correlations between participants' interpretation styles and eye movements without specifying a direction	holistic vs. Analytics patterns	2. No difference in eye movements between people with and without chronic pain
12	(BlaisdaleJones et al. 2021) 74 CP, 66 PF	Free viewing, with face images: (happy, sad, angry, and pain) and words (sensory pain and neutral words), Recognition task (interpretation bias), Flanker task	<ol style="list-style-type: none"> 1. More pronounced cognitive biases towards pain and poorer attentional in CP participants 2. Attentional biases, interpretation biases, and attentional control would be significantly associated with each other 3. Cognitive biases would be significantly associated with pain-related outcomes. 	<ol style="list-style-type: none"> 1. Probability of first fixation 2. Time to first fixation 3. First fixation duration 4. Visit duration 5. Number of fixations within the AOI. 	Although significant differences in duration of first fixation, visit duration, time to first fixation, and the number of fixations between all participant on pain stimulus was found, no significant differences in groups were identified

		(attentional control)			
13	(Koenig et al. 2021) 25 CP, 25 PF	Visual search (finding the diamond)	<ol style="list-style-type: none"> 1. If chronic pain patients fail to establish the CS- as a safety signal, both the CS+ and the CS- should elicit comparable levels of emotional arousal (indexed by pupil dilation) compared to healthy controls. 2. The failure to discriminate between the CS should lead to increased attention to the CS- which in healthy controls exhibits shorter fixation dwell time than the CS+ 3. If deficits in differential learning in CP patients are based on 	<ol style="list-style-type: none"> 1. Pupil dilation 2. Fixation probability on the diamond over a 5-sec interval 3. Total fixation duration 4. Number of visits of distractors 5. Duration of first fixation (if on distractor) 	<ol style="list-style-type: none"> 1. Higher dilation in higher threat, but not significant between groups. 2. No significant difference in fixation probability and total fixation duration 3. Higher visit of distractors in higher threat conditions, but no significant difference between groups 4. No significant difference between groups in the duration of first fixations

			<p>overgeneralization of fear, stimuli that are irrelevant for predicting shock should attract more attention in CP patients than in healthy controls</p>		
14	<p>(ten Brink et al. 2021)</p> <p>40 CP Complex Regional Pain Syndrome (CRPS), 40 pain controls, 40 PF</p>	<p>Free viewing, visual search, temporal order judgment, and dot-probe</p>	<ol style="list-style-type: none"> 1. People with CRPS would show a visuospatial attention bias away from the affected side that was larger for, or only evident in, conditions that were more likely to recruit body representation 2. There would be an interaction between any visuospatial attention bias and the location of the body-part stimulus (i.e., on the affected or 	<ol style="list-style-type: none"> 1. Proportion of first fixations 2. Visit duration of the affected side/total visit duration 3. Number of first fixations on the affected side/total number of first fixations 4. Average latency of first fixations on the unaffected side/average 	<ol style="list-style-type: none"> 1. There is no evidence for a body-related or general visuospatial attention bias in people with CRPS. However, there are indications that a body-related visuospatial attention bias might be present in some people with CRPS.

			unaffected side of the screen)	latency of all first fixations	
15	(Soltani et al. 2020) 102 CP, 53 PF	Flanker visual filtering task	<ol style="list-style-type: none"> 1. The nature of attentional bias in youth with chronic pain vs. A pain-free control group 2. The moderating effect of attentional control on attentional bias. 	<ol style="list-style-type: none"> 1. Probability of first fixation 2. Mean total fixation time 	<ol style="list-style-type: none"> 1. No significant difference between groups in the probability of first fixation 2. No significant difference between groups in mean total fixation time
16	(Soltani et al. 2022) 125 CP, 52 PF	Free viewing, with face images: (neutral, and low, moderate, and high pain)	<ol style="list-style-type: none"> 1. Higher levels of anxiety sensitivity, pain catastrophizing, and fear of pain would be associated with a greater attentional bias for pain expressions. 2. Greater attentional biases for pain expressions would be associated with worse clinical outcomes 	<ol style="list-style-type: none"> 1. First fixation (bias) proportion 2. Total fixation bias 	<ol style="list-style-type: none"> 1. All participants exhibited first fixation bias and total fixation bias for pain facial expressions, regardless of chronic pain status 2. Pain catastrophizing, anxiety sensitivity, and fear of pain were not related to attentional bias in youth with chronic pain 3. Other than the first fixation bias and pain intensity at follow-up, the rest of the attentional bias variables

					were not correlated with clinical outcomes (pain and mental). Higher first fixation bias for pain faces was associated with lower self-reported pain intensity at follow-up
17	(Chan et al. 2022) 32 young PF, 31 young CP, 31 old PF, 32 old CP	Free-viewing (500 ms) Face images (with doctors', patients', and healthy labels	<ol style="list-style-type: none"> 1. People with chronic pain would endorse more injury-/illness-related interpretations for ambiguous scenarios and may be more vigilant toward injury scene images than pain-free controls. 2. There is an age difference in interpretive and attentional processing. 3. Interpretation and attentional biases are 	<ol style="list-style-type: none"> 1. Proportion of fixations on pain AOI 2. Total duration of fixations within AOIs 3. First fixation proportions on the pain-related AOIs 4. Duration of visit of pain AOIs 5. Sequence of fixation locations in each trial 	<ol style="list-style-type: none"> 1. CP endorsed more negative interpretations for injury-/illness-related scenarios than PF, but the two groups did not differ in their eye movements on injury scenes 2. CP participants, regardless of age group, had a more negative interpretation bias for injury-/illness-related scenarios than controls 3. Distinct roles of interpretation and attention in chronic pain

			correlated and may together predict later pain functioning		
18	(Shiro et al. 2021) 8 CP, 8 PF	Free viewing (a clip with neutral and one with pain-related content)	<ol style="list-style-type: none"> 1. Whether CP patients have a visual attentional bias toward the bodies of others 2. Whether CP patients have a visual attentional bias when a stranger touches the patient's hand 3. Relation between attentional behavior and clinical symptoms 	<ol style="list-style-type: none"> 1. Fixation duration 2. Fixation count 	<ol style="list-style-type: none"> 1. No significant difference between groups in fixation duration or fixation count 2. No significant correlation between clinical symptoms and attentional patterns

2.2. Limitation of Previous Studies

In early pain research, Dot-probe were used to examine participants' initial attention to dots appearing after words or faces. Initial attention did not provide all the information about visual attention to the content; however, technologies for tracking sustained attention to stimuli were unavailable. After eye tracking emerged and was used in this topic of research, researchers were able to collect more information about eye movements.

Along with initial orientation, which is typically measured through first fixation proportion and first fixation latency, eye tracking also allowed for continuous attention to be measured. Some researchers, such as (Chan, Suen, Jackson, et al. 2020), have labeled continued attention as attentional engagement, measured by the number of fixations and visits, and attentional maintenance (i.e., the first fixation/visit duration, the average fixation/visit duration, and the total gaze duration).

According to previous research (Chan, Suen, Jackson, et al. 2020), attentional engagement and maintenance could be better indicators for studying pain than initial orientation, which was not accessible before eye trackers. In order to capture this valuable information to study chronic pain and healthy people's attentional biases, I use an eye tracker.

Aside from technological limitations, most researchers used dot-probes or other stimuli with short tasks. The typical duration of dot-probes is 2 seconds, which does not provide much opportunity for collecting continuation of attention, but initial attention. I propose asking participants to read short passages to overcome this limitation. This new stimulus (text passages) provides a suitable context for studying attentional engagement and maintenance more fully, hence, affords the opportunity to obtain more information on these influential indicators.

Furthermore, as discussed previously, the stimulus in dot-probe tasks, for example, could be richer if the ambiguity of different interpretations of a single word was reduced.

Through passages instead of a single word, the participant can gain a greater understanding of the context. I addressed the limitation of the stimulus' richness by replacing words with passages.

Despite the fact that eye trackers also enabled researchers to collect more information about eye movements, not many eye movement variables were used by researchers on this topic. I used a large list of variables that are used in other eye tracking studies to investigate attention (visit, fixation, and saccadic metrics), and cognitive effort (pupillometry) to take advantage of this overlooked opportunity. For instance, saccades can reveal changes in attention which is critical to study attentional biases. Also, pupillary responses can reveal information about cognitive effort when processing (attending to) information (Shojaeizadeh et al. 2019). However, both these series of variables are not utilized in previous chronic pain eye-tracking studies. I also summarized the researchers' analysis methods in my literature review. A majority of studies used basic statistical methods, which have limitations for decoding complex patterns such as attentional bias and are less powerful than machine learning models for dealing with multiple variables. This might explain why previous researchers have used a short list of variables. In order to analyze complex patterns of eye movements, I use a number of machine learning methods.

In order to create an accurate ETML model for identifying chronic pain from healthy participants, I propose replacing the stimulus to facilitate richer context and longer exposure to afford us opportunity to more extensively study attention. I also propose a rich list of eye-tracking variables that include fixation, saccade, and pupillometry metrics. And propose utilizing machine learning methods that can operate the proposed rich set of eye-tracking variables.

My proposal also involves an iterative approach to create a model that is both accurate and reliable across a variety of sample pools.

3. METHODOLOGY

In this section I explain the methodology for implementing my proposal.

3.1. Stimulus Design

To address the limitation of previous studies and to increase the exposure time to stimuli, and provide richer context, I used four textual passages as visual stimuli. To develop the visual stimuli a set of online articles were reviewed, from which 9 articles were selected. These articles were reviewed by me and three other PhD students to select the final 4 articles for the study. These articles were modified to have similar length, approximately 100 words each. Hence the visual stimulus included a total of four passages, containing a total of 19 sentences, and 410 words. The four text passages were presented to participants in a random order.

The topic for 2 of the 4 text passages was pain (headache and neckache); the other two passages had a neutral topic (bees and furniture). This is because prior pain research suggests that that people with chronic pain exhibit attentional bias (attention and/or avoidance) towards pain-related stimuli. These pain studies often include both pain-related and neutral stimuli to study the impact of pain on attention. Additionally, because pain affects cognition, I created two levels of task load by manipulating the reading difficulty of the text passages (Shojaeizadeh et al. 2017). Two of the passages (one in each topic) was designed to be harder to read (14th, and 16th grade reading level) and 2 were designed to be relatively easy to read (7th grade reading level) (see appendix 1 for text passages and their respective reading level). The reading difficulty of the passages were determined by an online tool available at (“Readability Formulas” n.d.)

The visual stimuli was formatted to match the optimal accuracy of the eye tracking device. The optimal accuracy of the eye tracker used in my project (Tobii spectrum 600Hz) was 0.4 degrees (Tobii Technology Inc. 2017). Based on that, the line spacing was adjusted

to about 1.5 to consider this accuracy and to minimize the probability of overlapping attention to different lines of the text. The conservative space of 0.5 degrees translated to 20 pixels or 0.5 cm. Hence, I used 16-point size Arial font for the passages.

The passages and their respective reading level are available at appendix 1.

3.2. Eye-tracking Apparatus

To collect eye movements, I used an eye-tracker called Tobii Pro Spectrum that captured 600 samples per second. I used Tobii Pro lab 1.162.32461 software to use this device. Using an I-VT filter with a threshold of 30 degrees/second, a minimum fixation time of 100 milliseconds, and gap fill-in, fixation, saccade, and unclassified gazes were labeled.

Previous research (Nuske et al. 2015) reported a change in pupillometry data at different luminosity levels, and behavior change at different temperature settings (Ramsey, Jerry D; Burford, Charles L; Beshir, Mohamed Youssef; Jensen 1983). To minimize such effects, the experiment was conducted in a room with controlled temperature and light.

A Tobii Spectrum IPS built-in monitor with 92 pixels per inch and a 16:9 ratio was used (DisplayDB 2016). The monitor provided images with a resolution of 1920 x 1080.

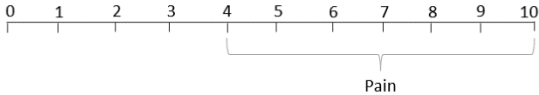
3.3. Data collection Process

The data for my project was collected via an IRB approved eye-tracking study at the User Experience and Decision Making (UXDM) laboratory at WPI. Eye tracking data was collected individually, one participant per study session. Participants were recruited from WPI community (due to COVID, no outside visitors were allowed on campus). At the beginning of each data collection session, I explained the study to participants and requested their consent. Once the participant consented to collect the accurate gaze data,

they went through a calibration process. Tobii Pro lab, the software I used for the data collection, includes five calibration points and then 4 validation points. The results of calibration were shown afterward. The calibration process took on average about a minute. For some participants, redoing the calibration process to get acceptable results was done.

Once I received accurate calibration results, the participant was asked to read 4 short text passages, each about 100 words. Participants' eye movements were recorded when they were reading the text. After completing the task, each participant was asked to self-identify, based on the definition of chronic pain as someone who suffers from chronic pain, someone who is pain-free, or someone with an "in-between" experience (Figure 1). This information was used to categorize eye movement datasets into chronic-pain and pain-free groups. The experiment finished by thanking participants for their participation. Each participant received a \$20 Amazon gift card as a token of our appreciation.

Daily chronic pain is defined as experiencing pain with a rating greater than or equal to 4 on a scale of 0-10, for 3 months or more



Based on this definition, would you say that right now:

- You suffer from daily chronic pain
- You are free of daily chronic pain
- You are somewhere in-between

Figure 1 - Self-reported health status question

3.4. Eye Movement Metrics

All variables and their definitions are presented in Table 3. Additionally, variables used in the literature review are marked. The variables I calculated for this study are also marked.

Table 3: Eye Movement Metrics

	Variable	Category	Definition	Used in previous chronic pain eye tracking research
1	Total duration of whole fixations	Fixation	The total duration of the fixations inside an AOI during an interval (excluding partial fixations).*	
2	Average duration of whole fixations	Fixation	The average duration of the fixations inside an AOI during an interval (excluding partial fixations).*	
3	Minimum duration of whole fixations	Fixation	The duration of the shortest fixation inside an AOI during an interval (excluding partial fixations).*	
4	Maximum duration of whole fixations	Fixation	The duration of the longest fixation inside an AOI-during an interval (excluding partial fixations).*	
5	Number of whole fixations	Fixation	The number of fixations occurring in an AOI during an interval (excluding partial fixations).*	
6	Time to first whole fixation	Fixation	The time to the first fixation inside an AOI during an interval (excluding partial fixations).*	
7	Duration of first whole fixation	Fixation	The duration of the first fixation inside an AOI during an interval (excluding partial fixations). *	

8	First-pass first fixation duration	Fixation	The duration of the first fixation during first-pass inside an AOI during an interval. *	
9	First-pass duration	Fixation	The total duration of the fixations during first-pass inside an AOI during an interval. *	
10	Go-past duration	Fixation	The total duration of the fixations from first fixation in this area of interest until a fixation occurs in an area of interest progressive to this one, during an interval. *	
11	First pass regression	Fixation	Indicates whether the reader exits the AOI with a regression (1) or reads on progressively (0) during an interval. *	
12	Regression-path duration	Fixation	The total duration of the fixations from first fixation in this area of interest until a fixation occurs in an AOI progressive to this one, including fixations in regressive AOIs, during an interval. *	
13	Re-reading duration	Fixation	Regression path duration excluding first pass fixations during an interval. *	
14	Average duration of fixations	Fixation	The average duration of the fixations inside an AOI during an interval.	X
15	Standard deviation of Duration of fixations	Fixation	The standard deviation of duration of the fixations inside an AOI during an interval.	
16	Average fixation inner density	Fixation	The average inner density of fixations inside an AOI during an interval.	
17	Standard deviation of fixation inner density	Fixation	The standard deviation of inner density of fixations inside an AOI during an interval.	
18	Number of fixations	Fixation	The number of fixations occurring in an AOI during an interval.	X

19	Total duration of fixations	Fixation	The total duration of the fixations inside an AOI during an interval.	X
20	Minimum duration of fixations	Fixation	The duration of the shortest fixation inside an AOI during an interval.	
21	Maximum duration of fixations	Fixation	The duration of the longest fixation inside an AOI-during an interval.	
22	Time to first fixation	Fixation	The time to the first fixation inside an AOI during an interval.	X
23	Duration of first fixation	Fixation	The duration of the first fixation inside an AOI during an interval.	X
24	Number of fixations during first visit	Fixation	The number of fixations during the first visit of AOI.	X
25	Total duration of fixations/total duration of visits	Fixation	Normalizing duration of fixation: Ratio of duration of fixations to duration of visits	
26	Number of fixations/Number of fixations on sentence	Fixation	Normalizing the number of fixations on sentence: The number of fixations divided by the number of fixations on corresponding sentence.	
27	Number of fixations/Number of fixations on passage	Fixation	Normalizing the number of fixations on passage: The number of fixations divided by the number of fixations on corresponding passage.	X
28	Number of fixations/Number of fixations on all passages	Fixation	Normalizing the number of fixations on passage: The number of fixations divided by the number of fixations on all passages.	

29	Normalized duration of fixations on sentence	Fixation	Normalizing duration of fixation: Ratio of duration of fixations to duration of fixation on corresponding sentence	
30	Normalized duration of fixations on passage	Fixation	Normalizing duration of fixation: Ratio of duration of fixations to duration of fixation on corresponding passage	
31	Normalized duration of fixations on all passages	Fixation	Normalizing duration of fixation: Ratio of duration of fixations to duration of fixation on all passages	
32	Normalized number of fixations in first visit	Fixation	Normalizing number of fixations: Ratio of number of fixations in the first visit to number of fixations on all passages	
33	Average size of pupil during fixations	Pupillometry	The average size of pupil during fixations inside an AOI during an interval.	
34	Standard deviation of average size of pupil during fixations	Pupillometry	The standard deviation of size of pupil during fixations inside an AOI during an interval.	
35	Median size of pupil during fixations	Pupillometry	The median size of pupil during fixations inside an AOI during an interval.	
36	Standard deviation of median size of pupil during fixations	Pupillometry	The standard deviation of median of pupil size during fixations inside an AOI during an interval.	
37	Average pupil dilation during fixations	Pupillometry	The average change of size of pupil during fixations compared to baseline	
38	Standard deviation of pupil dilation during fixations	Pupillometry	The standard deviation of change of size of pupil during fixations compared to baseline	

39	Average pupil size	Pupillometry	The average size of pupil among all gazes of the participant during the visit of whole stimuli	X
40	Median of pupil size	Pupillometry	The median size of pupil among all gazes of the participant during the visit of whole stimuli	
41	Average pupil dilation	Pupillometry	The average change of size of pupil during whole experiment compared to baseline	
42	Number of saccades in AOI	Saccade	The number of saccades occurring in an AOI during an interval. *	
43	Time to entry saccade	Saccade	The duration until the start of the first saccade that ends in an AOI during an interval. *	
44	Time to exit saccade	Saccade	The duration until the start of the first saccade that exits an AOI during an interval. *	
45	Peak velocity of entry saccade	Saccade	The peak velocity of the first saccade that ends in an AOI during an interval. *	
46	Peak velocity of exit saccade	Saccade	The peak velocity of the first saccade that exits an AOI during an interval. *	
47	Number of saccades	Saccade	The number of saccades occurring during an interval. *	
48	Average peak velocity of saccades	Saccade	The average peak velocity of all saccades in this interval. *	
49	Minimum peak velocity of saccades	Saccade	The peak velocity of the saccade with the lowest peak velocity in this interval. *	
50	Maximum peak velocity of saccades	Saccade	The peak velocity of the saccade with the highest peak velocity in this interval. *	

51	Standard deviation of peak velocity of saccades	Saccade	The standard deviation of all peak velocities of the saccades in this interval. *	
52	Average amplitude of saccades	Saccade	The average amplitude of all saccades in this interval. *	
53	Minimum amplitude of saccades	Saccade	The amplitude of the saccade with the lowest amplitude in this interval. *	
54	Maximum amplitude of saccades	Saccade	The amplitude of the saccade with the highest amplitude in this interval. *	
55	Total amplitude of saccades	Saccade	The total amplitude of all saccades in this interval. *	
56	Time to first saccade	Saccade	The time to the first saccade during an interval. *	
57	Direction of first saccade	Saccade	The direction of the first saccade in the interval. *	
58	Peak velocity of first saccade	Saccade	The peak velocity of the first saccade in the interval. *	
59	Average velocity of first saccade	Saccade	The average velocity of the first saccade in the interval. *	
60	Amplitude of first saccade	Saccade	The amplitude of the first saccade in the interval. *	
61	Total duration of Glances	Visit	The total duration of the Glances inside an AOI during an interval. *	
62	Average duration of Glances	Visit	The average duration of the Glances inside an AOI during an interval. *	

63	Minimum duration of Glances	Visit	The duration of the shortest Glance inside an AOI during an interval. *	
64	Maximum duration of Glances	Visit	The duration of the longest Glance inside an AOI during an interval. *	
65	Number of Glances	Visit	The number of Glances occurring in an AOI during an interval. *	
66	Time to first Glance	Visit	Time in milliseconds to the first Glance inside an AOI during an interval. *	
67	Duration of first Glance	Visit	The duration of the first Glance inside an AOI during an interval. *	
68	Total duration of visits	Visit	The total duration of the visits inside an AOI during an interval.	X
69	Duration of first visit	Visit	The duration of the first visit inside an AOI during an interval.	
70	Minimum duration of visits	Visit	The duration of the shortest visit inside an AOI during an interval.	
71	Maximum duration of visits	Visit	The duration of the longest visit inside an AOI during an interval.	
72	Average duration of visits	Visit	The average duration of the visits inside an AOI during an interval.	X
73	Duration of next visits	Visit	The duration of visits after the first visit inside an AOI during an interval.	X
74	Number of visits	Visit	The number of visits occurring in an AOI during an interval.	X
75	Duration of interval	Visit	The duration of an interval. *	
76	Start of interval	Visit	The start time of an interval. *	
77	Native language of participant		1 if English, 0 if not. *	
Eye-tracker provided variables are marked by *.				

3.5. ML Methods and Settings

Various data inputs were determined for this project (passage, sentence, word, combined passage-sentence-word, pain passages, neutral passages, difficult passages, easy passages). Each data input was run with a set of commonly used algorithms, each was tested with different settings and the best model was selected by cross-validation. The algorithms include:

1. Random forest with 200 estimators
2. A neural network called Multi-layer Perceptron classifier with maximum iteration of 50,000
3. Logistic Regression (LR) with maximum iteration of 250,000
4. Linear Support Vector Classifier (SVC)

All algorithms were used with three variable settings, including:

1. Using all variables
2. Using principal component analysis (PCA) to reduce variables and only use 3 new ones
3. Using principal component analysis (PCA) to reduce variables and use as many new variables as needed to cover 90% of the variation of the data

All algorithms are used with two settings for the number of observations which include:

1. Actual observations
2. Equal observations in both classes by oversampling the minority class

The combination of these algorithms and settings were used to train and test 24 models on each input.

3.6. Model Evaluation

3.6.1. Confusion Matrix

Confusion matrices show the number of correct and incorrect predictions for each group. A schematic confusion matrix is shown in the Table 4.

Table 4: Schematic confusion matrix		Predicted as negative	Predicted as positive
True condition	condition: negative	TN	FP
	condition: positive	FN	TP

The abbreviations in the above table are explained below.

TP: True positive: Positive observations that are correctly predicted to be positive. In this analyses TP refers to pain-free participants who are predicted as pain-free.

FP: False positive: Positive observations that are incorrectly predicted as negative. Here, FP shows the chronic pain subjects who are labeled as pain free.

FN: False negative: Negative observations that are incorrectly predicted as positive. In this study, FN shows the pain-free participants who are classified as chronic pain.

TN: True negative: Negative observations that are correctly predicted as negative. Here, TN points to chronic pain participants who are correctly classified as chronic pain.

In this research, chronic pain can be considered as a negative condition, and pain-free as a positive one.

Based on the confusion matrix, some indicators are calculated and compared below.

3.6.2. Accuracy

A measure of how accurate the predictions are calculated as follows:

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

Although accuracy is a great indicator of model performance, it can be misleading because it aggregates the results of both groups. It is therefore beneficial to compare models also based on their specificity and sensitivity. Here are how these terms are calculated:

$$\text{Specificity, True negative rate} = \frac{TN}{TN + FP}$$

In the equation above, TN represents correctly predicted chronic pain cases, and (TN+FP) represents all negative (chronic pain) subjects. The ratio shows the percentage of correct predictions among chronic pain subjects.

$$\text{Sensitivity, Recall, or True positive rate} = \frac{TP}{TP + FN}$$

Where TP is the number of correctly predicted pain-free cases, and (TP+FN) is the number of all positive (pain-free) subjects. In this ratio, the share of correct predictions among pain-free participants is shown.

3.6.3. F1-Score

The F1 score is a measure of accuracy that takes precision and recall into account. "Precision" is the ratio of true positive predictions to all observations that are predicted as positive, as shown in equation1. The number of positive samples is considered, rather than positive sample accuracy alone; therefore, it is a fairer evaluation.

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

The recall is defined as the number of true positive predictions divided by the number of true samples. F1 is calculated as follows.

$$F1 = \frac{2 * \textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}}$$

F1 score ranges from 0 to 1, with 1 being the best result and 0 being the worst. The higher the F1 score, the better the result.

3.7. Model Selection

After developing all models, the best model was validated. The procedure is explained below.

As explained in section 3.5, for each input, 24 models were developed. Out of these 24 models, the model with the highest accuracy on testing data was selected. This model then was compared to models with different inputs and the model which overall reached the highest accuracy was selected for validation.

3.8. Pre-Processing Data

3.8.1. Creating AOIs

Areas of Interests (AOIs) are researcher-defined areas in which the gaze data can be captured and reported. To study the difference in attention to the content, I defined AOIs for each passage, sentence, and word. To maintain consistency in defining AOIs, the height of AOIs at the sentence and word level was kept the same and equal to 83 pixels. Obviously, the length of AOIs depended on the length of the word or sentence. The height

of the AOIs were calculated based on the accuracy of eye-tracker in degrees and monitor pixel density. Based on these AOIs, I extracted data for each passage, sentence, and word.

To distinguish repetitive AOIs, which happened only on words such as a, the, is, a naming format including sentence and word number was used. I also created an AOI around each passage to provide the most aggregated data. It worth noting here that sentences were not necessarily in one line and the sentence-AOIs covered the sentence regardless of being in one or more lines of text.

3.8.2. Data Cleaning

The collected data needs to meet some standards to qualify for being used in the analysis. Some of these standards are enforced during the experiment, during calibration, for example. After collecting the data, participants' data will be examined to meet the minimum of 80% gaze sample. If a participant's gaze samples were below 80%, it means the eye-tracker was incapable of capturing their eye movements for more than 80% of the experiment, which can be caused by numerous blinks or by looking away from the screen. As there is no ground truth for determining whether participants are looking away or blinking, participants with a sample rate less than 80% are removed from analysis, based on previous research (Varzгани et al. 2021).

3.8.3. Adding new variables

In addition to the variables provided by the eye-tracker software, some additional variables had to be calculated before using the collected data as input to the algorithm. To incorporate all potential variables in a model that can distinguish chronic pain participants from healthy participants, I calculated these additional variables from gaze level data, the rawest data the eye tracker can provide. The variables have previously

been used in chronic pain research or in eye-tracking research. As an example, previous studies (Shojaeizadeh et al. 2019) have found that pupillometry data can function as a cognitive process indicator, or saccadic variables have been widely used in previous studies with reading stimuli (Rayner et al. 2006).

3.9. Preparing Data for Machine Learning Algorithms

Data processing was performed using Python 3.8.8, Pandas 1.2.4, and NumPy 1.19.2. Multiple machine learning models were trained and tested using Scikit-learn 0.24.1 package.

3.9.1. Splitting Data to Train and Test Sets

Data was divided into training and testing sets. Using the training set, the models were trained, and then tested with the testing set. To ensure fair training and testing, I used a random selection model in which some observations are randomly selected for testing and the others are randomly selected for training. I used k-fold cross-validation to accomplish this. By using this method, the data is split into multiple chunks, specified by k, with one chunk used for testing and the rest for training. Each time the algorithm runs, a different chunk is used for testing, the rest for training, and so forth until all chunks have been used. Figure 1 is a schematic representation of how k-fold cross-validation works.

This study used k-fold cross-validation with k of 5.

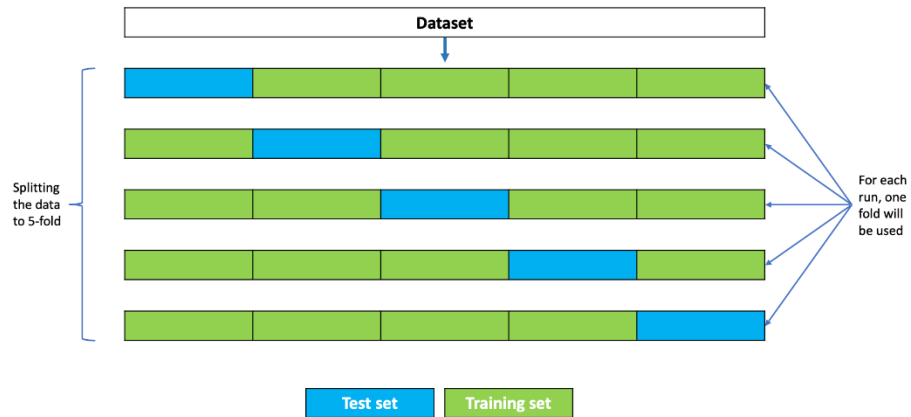


Figure 2 - Schematic View of K-Fold Cross Validation

3.9.2. Balancing Data

After pre-processing the collected data of this step, 28 participants' data were qualified to be used in the analyses, including 19 pain-free participants and 9 chronic pain subjects (these groups were created based on self-identification, using the survey question displayed in Figure 1). To prevent bias in the trained model, I needed to balance the data before being used in the algorithm. However, I pursued a strategy to ensure the legitimate manipulation of the data only impacts the training set, and not the testing set. This strategy helps reduce the effect of manipulation and limits its impacts on the evaluations of the algorithms. To achieve this, after splitting the data into train and test sets, the training set was oversampled. For oversampling, I used "RandomOverSampler" which bootstraps the minority group of observation to add them and create a balanced dataset.

3.9.3. Un-arranged data

The collected data, regardless of word, sentence, or passage level, provides insight into some of these AOIs. The information includes eye movement variables used in previous

chronic pain research or other previous eye-tracking research examining attention, cognitive effort and/or load.

No matter which AOI is used, each AOI in the dataset is provided in one row, and many of these rows are randomly selected for training and testing. In order to prevent data snooping and prevent the model from predicting based only on the participant ID, the participant ID was removed from all rows.

In this dataset, the model will be trained and tested using many observations but only one AOI's data at a time. This means that the model does not learn about the connections between all AOIs (e.g., sentences or words) in a passage, for example, and the fact that many of those AOIs were read by one participant. This limits the richness of input to the model. Due to the low richness of data, it was expected that the model using un-arranged data would have low accuracy.

3.9.4. Re-arranging the Data

To train the model with richer data, all participant data was rearranged into one row. This method provided significantly richer data on the eye movements of participants when looking at different AOIs. One disadvantage of this strategy is the lower number of observations which would be equal to the number of participants and remarkably less than training with un-arranged data.

I rearranged the data in one row by combining the AOI name with the variable name and then displacing the value corresponding to that variable name.

3.9.5. Addressing Missing Values

Scikit Learn, the machine learning package we used, does not work with data with missing values or NaNs (Not a number). A number of methods are available in the literature to resolve this issue, each with its own advantages and disadvantages. Removing missing values is the simplest yet safest method, but it comes with the disadvantage of losing valuable data. Data variables with at least one missing value were removed from the analysis using this method. Despite the loss of some eye-movement features, I still used this strategy and removed all columns with missing values because this simple method minimizes the risk of adding unwanted changes.

Nonetheless, missing values cannot be removed from datasets that are used to validate existing models constructed from datasets with different variables due to another limitation of Scikit Learn. In this situation, I must predict the missing values and fill them in. A detailed explanation of how this is done will be provided in the corresponding section.

3.10. Proposed iterative process for developing ETML to detect chronic pain

In this section, I propose an iterative process for creating a robust ETML model. While it is expected that in near future eye movements can be collected remotely for studies (Alrefaei et al. 2022), currently collecting eye movement data sets are done through laboratory experiments via individual sessions, which is inherently a lengthy process. Rather than waiting until a large pool of participants' data is recorded, grounded in the "test and refine" cycles in user-centered design, I propose an iterative process in which an ETML is developed iteratively. First, a proof of concept is developed based on a relatively small pool of data. This ETML is then validated via a new set of data collected from a new set of participants. If the validation is successful (e.g., reaches a minimum desired accuracy threshold) with the newly collected dataset then process stops otherwise the model is refined by using the combined datasets and validated via another

set of newly collected datasets. Such a process guarantees that the validation data is not available when the model is being developed. It also allows to build the model efficiently and cost effectively (collect additional dataset as needed).

This iterative process starts by collecting eye-tracking data to create an initial dataset (step 1). Then, using this dataset multiple models is developed using a few popular machine learning algorithms (in my project, I use 4 different algorithms). The model with the highest accuracy will be chosen for validation (step 2). For validating the best model, which is identified in step 2, a new set of eye-movement data is collected (using the same study design and process) (step 3). If the validation process in step 3 does not reach the desired threshold (in our case, 80% accuracy with sensitivity and specificity above 50%), the collected data in steps 1 and 3 would be merged. A portion of the merged dataset will be kept aside for validation, and using the rest, multiple models, similar to step 2, will be created (refined model). Following that, the best model among the new models will be chosen for validation (step 4). Next, step 4's best model will be validated using the portion of the merged dataset, which was set aside in step 4 (step 5). If the model still serves as a good proof of concept (sensitivity and specificity above 50%) then, in step 6, all data in the merged dataset in step 3 will be used to train and test a new model, and the best model will be chosen to be validated with a new set of data collected from a different set of participants (go to step 3). This iterative process continues until the validated model satisfies the requirements, or the accuracy does not improve for two consecutive iterations.

The following provides an overview of the steps completed in my project; each step is explained in more detail in the following sections:

Step 1: Collected an initial set of data from 28 participants (9 chronic pain, 19 pain free)

Step 2: Developed 24 models for each input, selected the best of each, and the best of all 24 bests

Step 3: Collected a new set data (27 participants, including 21 chronic pain and 6 pain free); validated the best of all models in step 2 with this set.

Step 4: Validation in last step failed (did not reach 80% accuracy), hence I merged the datasets, split the merged dataset into train/test/validation, and developed new models with the train/test datasets.

Step 5: Validated the models developed in step 4 with the validation dataset that was kept aside.

Step 6: Validation showed that the proof concept still worked well, hence all parts of datasets (train, test, and validation) were merged (n=55 participants) develop (test and train) new models. The results were satisfactory for continuing the process. Future studies are needed to collect a new dataset for validating the final model developed in my project and continue the iterations until a robust model (with 80% accuracy or better) is developed.

3.10.1. ETML Development – Step 1: Collecting Initial Eye-Tracking Dataset

Here, I explain how I collected participants' eye movements.

For this step, 41 participants from the WPI community attended and consented to the experiment. The data collection took place during Spring 2021. Among them, two participants were unable to calibrate, and one participant's data was not collected properly due to a technical problem with the eye-tracker. It is worth noting that calibration accuracy was not satisfactory for a participant with Lasik surgery and a participant who wears 3-focal glasses. Thus, their data was removed from the project.

1 of the 38 these participants' gaze samples was below 80%, whose data, according to section 3.8.2 was removed. During data collection, the first few participants were closely monitored to ensure high-quality data was collected and to identify improvement opportunities in the experiment design. Following these observations and participants' feedback, some task instructions were revised, and passages were modified to ensure

providing a suitable reading experience. For example, words or phrases which were considered too cumbersome were replaced by more common words and the length of articles was shortened from two paragraphs to one (about 100 words each). This resulted in the final analysis not using the data from the first 7 participants, which were collected before the final adjustments. In addition, 2 participants reported as in-between health status. Due to the study's focus on identifying attentional biases of people with and without chronic pain, the models excluded people whose self-identified condition was in-between.

As a result of all data cleaning steps, I had 28 participants' data (Age mean: 24.2, Age Standard deviation: 3.2) ready to be analyzed, 19 of whom were pain-free and 9 of whom were chronic pain sufferers.

3.10.2. ETML Development – Step 2: Developing a Proof of Concept

After cleaning the data and preprocessing it using the steps discussed in the method section, data was ready to be used as input for the selected machine learning algorithms. Using different inputs, algorithms, and settings, multiple models were developed. Then, as discussed in section 3.7, the best model for each input was selected. Below the best model for each input along with the size of training dataset, and other model settings are presented. It is important to remember that the testing set is different from the validation set, which will be collected in the next step, Hence the untouched validation set is yet unavailable to this analysis. This study uses k-fold ($k=5$) to create a testing set every time it runs. The eye tracking variables used for this part are provided in Table 3.

This first set of models was developed based on a single AOI, a passage. Hence, the models were trained and tested using 112 observations (28 participants X 4 passages) and 67 variables (Table 5). This set of models uses samples with high-view data at the passage level. Furthermore, each time the models use one passage's data with no clue about the participant or other passages which are read by the same participant. This

means that the accuracy of this set of models due to the limited access to very little information about the participant, only a sample, is likely to be unreliable. In other words, the best model is overfitted to the available data and expected not to be able to predict unseen data well. With this argument, even if this set of models provides the highest accuracy among all models, it would not be selected as the best model for validation. However, I will evaluate its performance as an exploratory analysis at the end.

Table 5: Model Set 1 – Passage level with 112 observations (28 participants X 4 passages) and 67 variables

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Passage level	Random Forest Classifier (RF)	Yes	No	0.771	0.531	0.886	0.84

The input for the next set of models was still at passage level, however, this time the rearranged data was used so that the models are trained and tested using based on not just a single passage, but all passages read by each participant.

It is worth reminding here that the re-arranged data has also unintended positive consequences. Because after transposing columns with missing values will be removed, there is a chance that multiple missing values will be arranged in one column and hence reduce the number of removed columns. This means the re-arranged data may include more variables than un-arranged dataset. As a result, the re-arranged data can provide significantly more information for the models to be trained with. Moreover, when the information for four passages read by the same participant is collapsed into a single row, naturally the number of variables (columns) in re-arranged datasets would become higher than their corresponding un-arranged sets (e.g., passage, sentence, and word level variables must all be presented as unique variables), which is why this set of models is built by 28 observations with 126 variables. Here this set of models has access to the

data of all four passages read by each participant compared to one passage's data in model set 1. The set of models' accuracy is lower than the unreasonably high accuracy of model set 1, but reasonable for the provided information. Results are shown in Table 6.

Table 6: Model Set 2 – Passage level (re-arranged) 28 observations and 126 variables

Input	algorithm	oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 score
Re-arranged passage level	Support Vector Classifier (SVC)	Yes	3	0.575	0.7	0.523	0.625

The sentence-level data were used to train a set of models similar to model set 1. The results of testing this set, which used un-arranged, and separate sentences, without a clue as to which sentences each participant read, are shown in Table 7. A total of 532 sentences and 9 variables were used in the training and testing of this set of models (it is possible to observe 19 sentences per participant). More missing data in sentences compared to passages led to more missing data in columns, resulting in more removed columns to prepare the data to be used by Scikit Learn.

Table 7: Model Set 3 – Sentence level, 532 observations (28 participants X 19 sentences) with 9 variables

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Sentence level	Logistic Regression (LR)	Yes	No	0.573	0.686	0.519	0.622

The next set of models uses rearranged sentences read by one participant, which created one row for each participant, providing the data of 19 sentences read by each participant.

The data includes all variables for each sentence. The results are shown in table 8. This set of models included 28 observations and 639 variables.

Table 8: Model Set 4 – Sentence level with 28 observations and 639 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged sentence level	Logistic Regression (LR)	No	No	0.672	0.53	0.755	0.758

Using the same settings of sets of models 1 and 3, a set of models was trained and tested using all variables at the word-level, i.e., the models made decisions based on seeing eye movement data for one single word. Columns with missing values were removed, and no affiliation of words belonging to a sentence or passage was visible to the set of models. This set of models uses a remarkably higher number of observations of 11480 with 10 parameters to be trained and tested.

Table 9: Model Set 5 – Word level with 11480 observations (28 participants X 410 words) and 10 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 score
Word level	MLP Classifier - Neural Network (NN)	Yes	No	0.535	0.671	0.471	0.579

The word-level data were rearranged to create a dataset containing 28 observations and 4344 features from all world level AOIs. Thus, each reader's eye movements on all words were provided in this dataset to this set of models. Table 10 shows this set of models and its results.

Table 10: Model Set 6 – Word level with 28 observations and 4344 variables

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged word level	Random Forest Classifier (RF)	Yes	3	0.613	0.6	0.616	0.682

Next, I developed a set of models with combined passage, sentence, and word level input. To train and test these models, data from all passages, sentences, and words seen by all participants was used to provide more information about participants' eye movements. A total of 12124 observations with 8 variables were used for these models.

As discussed before, increasing the number of samples, especially of different types, like passages, sentences, and words for this set of models, may result in appearing missing values in more columns. The result can be fewer columns remaining when columns with missing values are removed. Moreover, although the data of each word in a sentence and the data of the sentence itself may appear the same, that may not be true for all variables. The summation of the number of fixations on each word of a sentence, for instance, is equal to the number of fixations on that sentence. The number of visits to a sentence cannot, however, be calculated from the number of visits to its words. This set of models' results are shown in Table 11.

Table 11: Model Set 7 – Passage, sentence, and word level with 12124 observations and 8 variables

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Passage & sentence & word level	Random Forest Classifier (RF)	Yes	No	0.554	0.559	0.551	0.626

Finally, I created the most information-rich set of models by combining all levels' eye movements for each participant. It is generally expected that the more information-rich a model, the better its accuracy and generalizability.

To provide all samples viewed by each participant, the data was rearranged. Rearranging the data provides the models with information about different AOIs, enabling it to learn more about each participant and not just stick to one sample. Therefore, it is expected that these models will outperform previous models due to more available details.

However, because Scikit Learn cannot work with missing data, adding more samples to the set can result in fewer variables after removing the missing data. For instance, model set 6 uses only words and has 4344 variables as input, while this set of models includes passage, sentence, and word level data and has 3658 variables. Despite this, a broader range of available samples should still enhance the generalizability and reliability of models like this set.

This set of models used 28 observations with 3658 variables. As shown in Table 12, this set yields the following results.

Table 12: Model Set 8 – Passage, sentence, and word level with 28 observations and 3658 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level	Support Vector Classifier (SVC)	Yes	No	0.703	0.76	0.672	0.734

My next four sets of models examine how pain/neutral or easy/difficult stimulus can affect eye movements and can be used to distinguish chronic pain participants from healthy participants.

Because pain literature suggests that chronic pain influences attention to pain-related stimuli, I developed a set of models that included only pain passages. In other words, a set of models using only pain stimuli's data was trained and tested in order to determine if there are possible differences in attention to pain stimuli. Using rearranged data, this set of models used 28 observations and 1797 variables. The performance of this set is shown in Table 13.

Table 13: Model Set 9 – Passage, sentence, and word level for pain stimuli with 28 observations and 1797 variables

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Pain Passages only	Support Vector Classifier (SVC)	Yes	No	0.735	0.78	0.720	0.775

In order to better understand the role of the pain stimuli in differentiating eye movements between the CP and PF participants, a set of models was trained with the same settings as model set 9 but using non-pain (neutral) passages. This set of models uses 28 observations with 1862 features. This set's performance is shown in Table 14.

Table 14: Model Set 10 – Passage, sentence, and word level for neutral stimuli with 28 observations and 1862 variables

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Neutral Passages only	Support Vector Classifier (SVC)	Yes	No	0.696	0.79	0.647	0.718

Next, I looked at a set of models that used input based on text difficulty. Because pain affects cognition, and cognitive load is reflected in eye movements (Mina’s paper), the level of text difficulty might be helpful in distinguishing pain status. The first set of models in this series used only difficult passages for input. This set of models used 28 observations and 1910 variables, and its performance is shown in Table 15.

Table 15: Model Set 11 – Passage, sentence, and word level for difficult stimuli with 28 observations and 1910 variables

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Difficult Passages only	Support Vector Classifier (SVC)	Yes	No	0.723	0.78	0.698	0.757

Similar to model set 11, a set of models that only used easy passages' eye movements was trained and tested. This set of models used 28 observations and 1749 variables, and its performance and settings are shown in table 16.

Table 16: Model Set 12 – Passage, sentence, and word level for easy stimuli with 28 observations and 1749 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Easy Passages only	Support Vector Classifier (SVC)	Yes	No	0.684	0.8	0.635	0.719

For each set of models 9, 10, 11, and 12, two passages (pain vs. neutral, difficult vs. easy) were used as input, resulting in some interesting findings. I narrowed down the input of the next four set of models to one type of passage including pain difficult, pain easy, neutral difficult, and neutral easy). This analysis gives us information about the nature of attention influenced by pain-related and difficulty level separately. Hence it helps us to see which type of stimuli was best in predicting health status. Furthermore, we investigate whether a reliable model can be trained using data from only a single passage. The results of these analyses are displayed in Tables 17 to 20.

Table 17: Model Set 13 – Passage, sentence, and word level for neutral easy stimuli with 28 observations and 911 variables

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Neutral Easy Passage only	Support Vector Classifier (SVC)	Yes	No	0.703	0.88	0.625	0.71

Table 18: Model Set 14 – Passage, sentence, and word level for neutral difficult stimuli with 28 observations and 952 variables

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Neutral Difficult Passage only	Random Forest Classifier (RF)	No	3	0.683	0.61	0.728	0.742

Table 19: Model Set 15 – Passage, sentence, and word level for pain difficult stimuli with 28 observations and 959 variables

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Pain Difficult Passage only	Support Vector Classifier (SVC)	Yes	No	0.753	0.76	0.743	0.788

Table 20: Model Set 16 – Passage, sentence, and word level for pain easy stimuli with 28 observations and 939 variables

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Pain Easy Passage only	Support Vector Classifier (SVC)	Yes	No	0.713	0.77	0.683	0.744

Among sets of models 13 to 16 that were trained using only one passage, the best mode of the model set 15, which used pain difficult passage, provided the highest accuracy. Prior research shows that people attend to pain stimuli differently than neutral stimuli

(Franklin et al. 2019; Soltani et al. 2022) which is why model sets 13 and 14 are developed to investigate the impact of pain/non-pain stimulus on predicting the health status. Additionally, based on (Shojaeizadeh et al. 2019) the difficulty level of passage can impact the eye movement of the two groups by altering the cognitive load of the reader. Thus, these sets of models have high potential to analyze influential factors on eye movement and differentiate the two groups. The other three sets of models resulted in a slightly lower accuracy, which still is important because each set of models uses limited data of one passage.

The results of the models reported in this section show that it is possible to develop an eye-tracking ML model to differentiate chronic pain from pain-free participants. In other words, the metrics and stimuli that were used in the project were suitable for creating the proof of concept. Hence, we continue the project by validating the best-obtained model in this section with a new dataset that is collected in a separate eye-tracking experiment.

As discussed before, the models with un-arranged data use very limited data of one sample, which could be a passage, sentence or word. Thus, they are not reliable. Among the developed models with rearranged data, which provides the data of a few to many samples of each participant, model set 15, SVC algorithm with oversampling and without using principal component analysis, showed the highest accuracy.

3.10.3. ETML Development – Step 3: Validating the Proof of Concept (the Best Model of Step 2)

This best model of step 2 was validated using a new dataset collected during Spring 2021. Since the data was collected in a separate step (hence unavailable at the time the model was being developed in step 2), it can provide an excellent way to validate the model's reliability.

The validation data was similarly prepared. It started by removing participants who self-identified as in-betweens, resulting in a database of 27 participants (Age mean: 30.3, Age Standard deviation: 15.6), and then rearranged the data. Then I needed to address the missing values as discussed in section 3.6.5 because our ML package cannot use columns with missing data. The package also cannot use inputs with variables that are not the same as the trained model. So, to begin addressing the package limitations, the columns in step 3 were trimmed so that they match those in step 1. In other words, some details about step 3 participants were ignored. As an example, some participants in step 1 did not see the word "the". Therefore, it was removed from step 1, but it was not removed from step 3 since it had been viewed by all participants of this step. But due to Scikit Learn's limitations, during validation this additional information was removed from step 3, and the model predicted participants based on variables included in the models which developed based on step 1 data. As a result of trimming the validation set to match the training/testing set (step 1' data), we obtained a dataset with 27 observations and 3658 variables. Out of 3658 variables, 254 had missing values, so Scikit Learn could not use them.

In order to fill in the missing values, I used a random forest model. First, the target of this dissertation, health status, was removed from both steps' datasets in order to protect the model against data snooping and prevent the model from using this variable in predicting the missing values. Next, the step 3 set has been labeled as X for columns without missing data and Y for columns with NANs. On the basis of these two sets, step 1 data were also split into X and Y. After that, a random forest (Rodriguez-Galiano et al. 2012) using step 1's X and Y was trained and applied to the step 3 data to impute the missing values.

Once imputing procedure was completed, and the trimmed step 3 set was free of missing values, the dataset was ready for validation of the best model of model set 15. The result of validating this model using this prepared dataset is shown in Table 21.

Table 21: Validating Step 1 – Model Set 15

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Validating model 15 (pain difficult)	Support Vector Classifier (SVC)	Yes	No	0.381	0.957	0.217	0.353

As this prediction is not as accurate as flipping a coin with a 50% chance in both classes, the model is not ready for implementation. With the small training/testing set of step 1 with 28 observations, this poor performance was predictable. The differences in variation between step 1 and 3 sets, considering the high variance in small sets, could explain the poor validation results.

Despite the fact that the selected model from model set 15 provided the highest accuracy among models which used rearranged data, this model uses only one passage and offers limited context. In contrast, the best of model set 8 provides the most contextual information as it provides passages, sentences, and words for all four passages. Since this range of data can help create a more solid model, I also validate it. The result of validating this model is shown in Table 22.

Table 22: Validating Step 1 – Model 8

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	Score
Validating model 8 (all passages)	Support Vector Classifier (SVC)	Yes	No	0.175	1.0	0.096	0.175

Also, to compare the validity of models which used limited context of one passage, model sets of 13 to 16 were validated. For the sake of comprehensiveness, the result of model set 15 is repeated here. The results are shown in Table ...

Table 23: Validating Step 1 - Models Sets 13, 14, 15, 16							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	Score
Validating model 13 (neutral easy)	Support Vector Classifier (SVC)	Yes	No	0.296	1.0	0.094	0.172
Validating model 14 (neutral difficult)	Random Forest Classifier (RF)	No	3	0.68	0.183	0.822	0.8
Validating model 15 (Pain difficult)	Support Vector Classifier (SVC)	Yes	No	0.381	0.957	0.217	0.353
Validating model 16 (pain easy)	Support Vector Classifier (SVC)	Yes	No	0.253	0.993	0.041	0.079

In step 1, model set 1 had the least amount of information but performed the best. The high accuracy of this model on the testing set is likely due to overfitting. Here, I evaluated my argument that this model is unreliable, and its validation results were as poor as other models, shown in Table 24. Due to the limited data available to this model and similar model sets, it is impossible to develop a robust model. Therefore, in each iteration, I focus only on the performance of model 8 to 16, which has the potential to produce solid models.

Table 24: Validating of Step 1 – Model Set 1

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Validating model 1 (all passages)	Random Forest Classifier (RF)	Yes	No	0.745	0.027	0.95	0.853

Overall, none of the models provided accurate predictions in both groups, which means none of the models is reliable. The results show that the high accuracy of these models is achieved by overfitting the step 1 data and not learning from eye movements of the two groups.

To create a more reliable model, I follow the iterative process that was discussed in section 3.10.

3.10.4. ETML Development – Step 4: Refining the Proof of Concept (the Best Model of Step 3)

Accordingly, the two datasets of steps 1 and 3 were merged based on the iterative process of validation, described in section 3.10. Next, 20% of the new larger dataset was randomly selected and kept aside for validation. The models were trained and tested with the remaining 80% of the data. Then similar models of 8 to 16 of step 1 using k-fold of 5 with the new training/testing set were trained and tested. For each setting, the best model based on testing accuracy is selected and presented. The names of corresponding models were kept the same for the sake of simplicity in comparing models of step 1 and 3.

The result of redoing model set 8 with the new dataset is shown in Table 25.

Table 25: Model Set 8 – Passage, sentence, and word level for all passages with 44 observations and 3405 variables

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level	Logistic Regression (LR)	No	No	0.719	0.520	0.794	0.799

The result of redoing model set 9 with the new dataset is shown in Table 26.

Table 26: Model Set 9 – Passage, sentence, and word level for pain passages with 44 observations and 1664 variables

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Pain Passages only	Logistic Regression (LR)	Yes	No	0.649	0.547	0.699	0.734

The result of redoing model set 10 with the new dataset is shown in Table 27.

Table 27: Model Set 10 – Passage, sentence, and word level for neutral stimuli with 44 observations and 1742 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Neutral Passages only	Support Vector Classifier (SVC)	Yes	No	0.557	0.808	0.462	0.603

The result of redoing model set 11 with the new dataset is shown in Table 28.

Table 28: Model Set 11 – Passage, sentence, and word level for difficult stimuli with 44 observations and 1755 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Difficult Passages only	Support Vector Classifier (SVC)	Yes	No	0.589	0.793	0.514	0.631

The result of redoing model set 12 with the new dataset is shown in Table 29.

Table 29: Model Set 12 – Passage, sentence, and word level for easy stimuli with 44 observations and 1651 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Easy Passages only	MLP Classifier - Neural Network (NN)	No	No	0.652	0.475	0.719	0.75

The result of redoing model set 13 with the new dataset is shown in Table 30.

Table 30: Model Set 13 – Passage, sentence, and word level for neutral easy stimuli with 44 observations and 872 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Neutral Easy Passage only	Support Vector Classifier (SVC)	Yes	No	0.548	0.85	0.434	0.583

The result of redoing model set 14 with the new dataset is shown in Table 31.

Table 31: Model Set 14 – Passage, sentence, and word level for neutral difficult stimuli with 44 observations and 871 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Neutral Difficult Passage only	Logistic Regression (LR)	Yes	3	0.520	0.57	0.503	0.573

The result of redoing model set 15 with the new dataset is shown in Table 32.

Table 32: Model Set 15 – Passage, sentence, and word level for pain difficult stimuli with 44 observations and 885 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Pain Difficult Passage only	Support Vector Classifier (SVC)	Yes	No	0.639	0.793	0.580	0.688

The result of redoing model set 16 with the new dataset is shown in Table 33.

Table 33: Model Set 16 – Passage, sentence, and word level for pain easy stimuli with 44 observations and 780 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Pain Easy Passage only	Support Vector Classifier (SVC)	Yes	No	0.53	0.783	0.434	0.573

Model set 8 provided the highest overall accuracy. Also, among the 4 model sets which only used a single passage, the model set 15, which used pain difficult passage again, provided the highest accuracy. In the next step, I will validate the best model.

3.10.5. ETML Development – Step 5: Validating the Refined Proof of Concept (From Step 4)

After merging the two steps' data sets, model set 8 provided the best results. The best model of model set 8 is a Logistic Regression (LR) without PCA and oversampling, with an overall accuracy of 71.9%. The result of validating this model with the unseen data is shown in Table 34.

Table 34: Validating the Refined ETML - Model Set 8

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Validating model 8 (all passages)	Logistic Regression (LR)	No	No	0.598	0.633	0.585	0.679

Using a larger dataset (44 observations), the new model can distinguish the two groups with about 60% accuracy. Because the performance of the model on specificity and sensitivity is above 50%, merging the datasets and adding the number of samples for training the model helped create a more reliable model. Following the iterative process that was discussed in section 3.10 can likely result in a reliable and more accurate model to separate the groups.

Following the previous discussion on validating the models which only used one passage, Table 35 shows the results of these models' validation.

Table 35: Validating the Refined ETML - Models Sets 14, 15

Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Validating model 14 (neutral difficult)	Logistic Regression (LR)	Yes	3	0.595	0.64	0.578	0.674
Validating model 15 (Pain difficult)	Support Vector Classifier (SVC)	Yes	No	0.385	0.72	0.26	0.381

The models sets 14 and 15 provided greater than 50% accuracy, specificity, and sensitivity on the test set when only one passage was used for training.

Among models that used one passage for training, only model sets 14 and 15 exceeded 50% accuracy, specificity, and sensitivity on the test set. Among these two model sets, only model set 14, which used only neutral difficult passage, performed well in validation. Despite the fact that the best model of this model sets' performance is almost as good as the best model of model set 8, which used all passages since the available samples for this model set are much fewer than model set 8, we should keep an eye on its performance going forward. There is a possibility that in future iterations, we will find the high performance of this model set, likely as a result of differentiating the cognitive loads that this passage creates for the participants. It is possible, however, that this finding will weaken when sample sizes are larger. Regardless, further iterations can reveal more insights.

With the additional samples, validation of the best model of model set 15, which performed well in previous iterations, does not indicate that this model can differentiate between groups. Similarly, more samples can change this finding.

3.10.6. ETML Development – Step 6: Refining the Proof of Concept (With Step 3 Merged Set)

At step 5, we validated the best model from step 4 and found that the improved results met the requirement for a minimum viable product. In other words, the validation of the ETML model, which was trained with more samples, suggests that continuing this iterative process will lead to a stronger model. Accordingly, as explained in section ..., I now redo model sets 8 to 16 with the merged set created in step 3. Our next step is to validate the new models by returning to step 3.

Here are the models that were trained with the whole merged dataset.

The result of redoing model set 8 with the new dataset is shown in Table 36.

Table 36: Model Set 8 – Passage, sentence, and word level for all passages with 55 observations and 3405 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level	Logistic Regression (LR)	Yes	3	0.591	0.540	0.61	0.668

The result of redoing model set 9 with the new dataset is shown in Table 37.

Table 37: Model Set 9 – Passage, sentence, and word level for pain passages with 55 observations and 1664 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Pain Passages only	Logistic Regression (LR)	Yes	No	0.651	0.513	0.703	0.740

The result of redoing model set 10 with the new dataset is shown in Table 38.

Table 38: Model Set 10 – Passage, sentence, and word level for neutral stimuli with 55 observations and 1742 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Neutral Passages only	Logistic Regression (LR)	Yes	3	0.585	0.533	0.605	0.671

The result of redoing model set 11 with the new dataset is shown in Table 39.

Table 39: Model Set 11 – Passage, sentence, and word level for difficult stimuli with 55 observations and 1755 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Difficult Passages only	Support Vector Classifier (SVC)	Yes	No	0.549	0.793	0.458	0.596

The result of redoing model set 12 with the new dataset is shown in Table 40.

Table 40: Model Set 12– Passage, sentence, and word level for easy stimuli with 55 observations and 1651 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Easy Passages only	Logistic Regression (LR)	Yes	3	0.615	0.48	0.665	0.715

The result of redoing model set 13 with the new dataset is shown in Table 41.

Table 41: Model Set 13 – Passage, sentence, and word level for neutral easy stimuli with 55 observations and 872 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Neutral Easy Passage only	Logistic Regression (LR)	Yes	3	0.636	0.526	0.678	0.721

The result of redoing model set 14 with the new dataset is shown in Table 42.

Table 42: Model Set 14 – Passage, sentence, and word level for neutral difficult stimuli with 55 observations and 871 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Neutral Difficult Passage only	Logistic Regression (LR)	Yes	3	0.547	0.594	0.53	0.613

The result of redoing model set 15 with the new dataset is shown in Table 43.

Table 43: Model Set 15 – Passage, sentence, and word level for pain difficult stimuli with 55 observations and 885 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Pain Difficult Passage only	Logistic Regression (LR)	Yes	3	0.583	0.533	0.603	0.667

The result of redoing model set 16 with the new dataset is shown in Table 44.

Table 44: Model Set 16 – Passage, sentence, and word level for pain easy stimuli with 55 observations and 780 variables							
Input	Algorithm	Oversampling	PCA	Overall Accuracy	Specificity	Sensitivity	F1 Score
Re-arranged passage & sentence & word level - Pain Easy Passage only	Logistic Regression (LR)	Yes	3	0.538	0.5	0.552	0.635

The accuracy of the best model of model set 8, the model set with most samples, is less than previous models, and now the best overall model is used only pain passages, resulting in 65% accuracy.

An interesting finding is that all models which use only one passage provide higher than 50% accuracy, specificity, and sensitivity.

Another interesting finding is with this dataset, almost all best models are logistic regression and use PCA. This phenomenon might be explained by the fact that PCA helps to reduce dimensions. In a feature rich model such as the one in this project, is helpful to reduce the dimension of the data.

Next step in the process requires validating the new best models with a new set of data collected from a new set of people (step 3). While this step is beyond the scope of my current project, all the steps completed in this project suggest that continuing the process is likely to lead to a robust model.

4. DISCUSSION

The objective of this project was twofold 1) develop an ETML as a minimum viable proof concept for predicting chronic pain and 2) propose an iterative approach to continue developing the ETML into a robust predictive model that can detect chronic pain people from eye movements automatically with 80% accuracy.

To achieve this goal, I started by reviewing the existing studies that used eye-tracking methodology to study the impact of chronic pain on attention to visual stimuli. The review of literature resulted in a small set (18) of relevant papers; there has been relatively little work in this area. Chronic pain studies typically examine the impact of chronic pain on visual attention to detect bias toward pain-related stimuli. This objective is often achieved by capturing participants' reactions to pairs of pain-related and neutral visual stimuli (e.g., words and/or images) for a fixed short period of time (e.g., 2 to 5 seconds). In pain studies that use eye tracking methodology, attention to stimuli is typically captured by fixations and visits metrics (Table 2 that shows the summary of literature review). Because fixations reveal the maintenance of one's gaze on a specific object or stimuli and visits include a collection of one's consecutive fixations on an object or stimuli, they provide excellent eye tracking metrics for such studies, particularly if the focus is on measuring initial engagement (e.g., time to first fixation or duration of first visit). However, measuring later stages of attention (i.e., attentional engagement and maintenance), which seem to provide more consistent evidence for detecting attentional bias in such stimuli presentation tasks (Chan, Suen, Jackson, et al. 2020) can benefit from stimuli that is richer in context and exposure time. Stimuli that is rich in context, naturally require longer processing time, which is likely to provide more opportunities for detecting differences in all stages of attention, particularly for detecting differences in attentional engagement and maintenance. Hence, in this project, I extend the task paradigm that is used in prior eye tracking chronic pain literature. Rather than comparing attention to simple pain-related and neutral words or images, I examined viewing behavior when people read four short (about 100 words) text passages, two of which covered pain-related topics (headache and backache) and two neutral topics (bees' communications and furniture). The

extension of the task paradigm also allowed me to extend the set of eye tracking metrics used in prior chronic pain studies. For example, saccadic variables were not used in prior chronic pain research. Saccadic variables, however, which reflect a change of focus, are very important to assess attention maintenance.

Because chronic pain impacts cognition (Phelps et al. 2021), the viewing behavior of people with chronic pain is likely to be affected by task load. To test this possibility, I created two levels of task loads (easy and difficult) by manipulating the reading level of the stimuli (2 text passages were simplified to be at 7th grade reading level; 2 were more difficult at 14th and 16th reading grade levels). This is yet another extension to the existing task paradigm in chronic pain literature.

Because prior research shows that pupillometry serves as a great metric for automatic detection of cognitive load, I included a host of pupillary metrics that can reliably detect cognitive load (e.g., Shojaeizadeh et al. 2019).

Using the above discussed extended task paradigm and extended list of eye movement metrics, I developed a feature-rich ETML (76 eye tracking features) from the eye movements of 28 people (9 with chronic pain and 19 pain free) that read 4 text passages (about 400 words), randomly presented to them, at their own pace (section 3.1). The results showed that the ETML which used the richest set of related data (rearranged data, see section 3.9.4) reached a promising level of accuracy (model 8, 70.0% overall accuracy with good specificity and sensitivity).

Comparatively, model 1, with an unarranged dataset, reached 77.1% accuracy with only passage-level eye movements with no other information on the passage (word, sentence) or other passages that the same participant read. Among all passages, this model received the least input, and it is likely that overfitting led to such high accuracy. While models 8 and 15 were support vector machines with oversampling and without dimension reduction (PCA), model 1 used the random forest algorithm with the same settings. Although the high accuracy achieved by model 1 (the least information-rich model, which

used only passage level eye movement data for only one passage) was not deemed reliable, this model was still validated on an exploratory basis.

The above results showed that the selected feature set that was obtained after removing the missing values were suitable for predicting chronic pain. The results also showed that it might be possible to reach relatively high level of accuracy with a model that takes eye movements at passage, sentence, and word levels for the difficult pain-related passages (model 15, with 75.3% accuracy). This result is important because it indicated we might be able to build a reliable ETML by asking users to read only one short paragraph (4-6 sentences, about 100 words) rather than all four passages. These encouraging results, which were obtained with a relatively small number of participants (n=28), warranted further iterative development. Hence, I continued to refine my investigations by following the iterative model that I proposed in the methodology section of this dissertation, for developing an ETML engine for predicting chronic pain.

The next step in the process (step 3) was to collect a new set of eye tracking data using the same stimuli and task paradigm as before. I collected eye movement data for 27 new participants. The new dataset, which was roughly the size of the original dataset, had the same inherent skewness; there were more pain-free participants in the pool than people who suffered from chronic pain (6 chronic pain and 21 pain-free people).

I used the newly collected dataset to validate the ETML that was developed in step 2 (model 8). Also, the validation of the two other models (model 1 and 15) was also explored. None of the three models maintained a specificity or sensitivity above 50% while maintaining relatively acceptable accuracy. This low validation accuracy of the models was expected because the proof of concept was developed (trained and tested) in step 2 with collected data from a small number of participants (n=28). Hence, as specified in the step 4 of my proposed process, I merged all data from both collected datasets (n=55) and divided the data randomly into two chunks: 80% of the data was used for training and testing new models and 20% were kept aside for validating the models. This allowed me to see how the models perform when the two sets of data were uniformly distributed.

Overall, the best model in this step was model 8, which covered passage, sentence, and word-level input for all 4 passages. The best results were obtained by a logistic regression without oversampling or PCA, resulting in an accuracy of 71.9%. Among the models which used passage, sentence, and world level input for a single passage, models 14 and 15, with a SVC algorithm with oversampling but no PCA, and logistic regression with oversampling and a PCA with 3 components, provided 52% and 63.9% accuracy, respectively. These promising results provide support that continuing the process is likely to result in developing a robust ETML that can predict chronic pain with a minimum of 80% accuracy. Hence, these “good” models were validated using the portion of the dataset set aside for validation. Both models 8 and 14 reached approximately 60% accuracy in validation, but model 15's was not acceptable. Improved validation results for models 8 and 14 encourage us to continue the iterative process until high validation accuracy is reached.

Because step 5 result was encouraging, as instructed in my proposed iterative process, I used all the data to test and train new ETMLs to be validated with a newly collected set of data (go back to step 3) in a future study.

The results of this step (6) were also encouraging. Almost all models' specificity and sensitivity were above 50% (except model 11 and 12) and some reached as high as 65% accuracy. The best model was model 9, which used two pain passages as input. The best performance on this input was achieved by logistic regression with oversampling and without PCA, resulting in 65.1% accuracy. Among models which used only one passage for input, the neutral difficult (54.7% accuracy) was the poorest performer, and the best performance (highest accuracy) was achieved by model 13, which used the neutral easy passage, leading to 63.6% accuracy. Developing a model using all samples of the merged set resulted in a slight decrease in the accuracy of the models that were developed with the smaller datasets (n=28).

Another interesting finding of this iteration is the trend of using dimension reduction between best models. Upon adding the number of samples for training, the best models started to use PCA. In step 6, almost all the best models used PCA. Because adding

samples may introduce complex eye movements, the model likely will use PCA to reduce the dimension of data to manage the number of variables while keeping the performance high. Additionally, the results seem to show a trend for winning models among the best models. In step 2, SVC performed best in almost all inputs. Later, some best models used logistic regression, and in step 6, all of the best models used logistic regression. The different algorithms or dimension reduction settings used among the best models in later iterations likely can explain the slight reduction of accuracy, and with more samples, it is possible to see a convergence of algorithms and settings between the best models.

Given the small sample sizes used to develop these initial models, the fluctuations in accuracy of the models and variation in best performing algorithms is not surprising. The fact that a reasonably good model can be built in each iteration provides support that building an ETML for predicting chronic pain is likely to be successful, following the suggested iterative process. As more and more data are collected in each iteration, more differentiating eye movement nuances are captured and incorporated in the models. Meanwhile, the iterative process allows us to monitor how the fluctuation in accuracy stabilizes over time and which algorithms and settings continue to outperform others.

The results of this research have important implications. By extending the task paradigm as well as using more eye movement metrics, my research contributes to chronic pain literature that investigates the impact of chronic pain on attentional bias. By developing an ETML proof of concept, this study contributes to NeuroIS literature. As more and more consumer-grade eye tracking devices become available, NeuroIS for supporting clinical decision-making becomes more affordable and hence feasible to use in clinical settings and perhaps in a not so distant future, at home, during remote clinical visits (Alrefaei et al. 2022). According to a recent user-centered framework for product development (Djamasbi and Strong 2019), there is a need for developing smart machine learning engines to address the continual market demand for user experience-driven innovations. My proposed iterative process, which allows for collecting sensor data over time, presents an initial step towards developing such smart NeuroIS systems over time as more sensor data becomes available. Finally, the developed proof of concept in this study provides

evidence for appropriateness of using eye movement data as an objective biomarker for chronic pain.

Similar to any project in its formative stages, my dissertation is not without limitations that should be addressed in future iterations. The ETML in this research was built with 4 text passages (2 pain-related and 3 neutral passages) with limited topics. For example, the pain-related text covered headache and backache. Expanding the pain stimuli to cover a wider variety of chronic pain topics may improve the accuracy of the models and even help to build models that can distinguish between people with different types of chronic pain. Previous eye-tracking studies suggest that including relevant images can improve attention to text (Norouzi Nia et al. 2021). Hence, including images in textual passages may improve the effectiveness of the visual stimuli in detecting attentional biases between pain-free and chronic pain groups.

Eye tracking data was collected primarily from WPI community members for my project. Collecting eye movement data from a wider population with different educational backgrounds is necessary to build a robust model. None of the participants in this research suffered from acute chronic pain. Thus, collecting eye movements from people with mild, moderate, and acute chronic pain experience will improve the robustness of the models.

5. REFERENCES

- Alrefaei, D., Sankar, G., Nia, J. N., Djamasbi, S., and Strong, D. (n.d.). *Examining the Impact of Chronic Pain on Information Processing Behavior: An Exploratory Eye-Tracking Study*. (https://doi.org/10.1007/978-3-031-05457-0_1).
- BlaisdaleJones, E., Sharpe, L., Todd, J., MacDougall, H., Nicholas, M., and Colagiuri, B. 2021. "Examining Attentional Biases, Interpretation Biases, and Attentional Control in People with and without Chronic Pain," *Pain* (162:7). (<https://doi.org/10.1097/j.pain.0000000000002212>).
- ten Brink, A. F., Halicka, M., Vittersø, A. D., Keogh, E., and Bultitude, J. H. 2021. "Ignoring Space around a Painful Limb? No Evidence for a Body-Related Visuospatial Attention Bias in Complex Regional Pain Syndrome," *Cortex* (136). (<https://doi.org/10.1016/j.cortex.2020.12.007>).
- Chan, F. H. F., Suen, H., Chan, A. B., Hsiao, J. H., and Barry, T. J. 2022. "The Effects of Attentional and Interpretation Biases on Later Pain Outcomes among Younger and Older Adults: A Prospective Study," *European Journal of Pain (United Kingdom)* (26:1). (<https://doi.org/10.1002/ejp.1853>).
- Chan, F. H. F., Suen, H., Hsiao, J. H., Chan, A. B., and Barry, T. J. 2020. "Interpretation Biases and Visual Attention in the Processing of Ambiguous Information in Chronic Pain," *European Journal of Pain (United Kingdom)* (24:7). (<https://doi.org/10.1002/ejp.1565>).
- Chan, F. H. F., Suen, H., Jackson, T., Vlaeyen, J. W. S., and Barry, T. J. 2020. "Pain-Related Attentional Processes: A Systematic Review of Eye-Tracking Research," *Clinical Psychology Review*. (<https://doi.org/10.1016/j.cpr.2020.101884>).
- DisplayDB. 2016. "EIZO FlexScan EV2451 Specifications."

- Djamasbi, S. 2014. "Eye Tracking and Web Experience," *AIS Transactions on Human-Computer Interaction* (6:2), pp. 37–54. (<https://doi.org/10.17705/1thci.00060>).
- Djamasbi, S., and Strong, D. (n.d.). *User Experience-Driven Innovation—Theory and Practice: Introduction to Special Issue*. (<https://doi.org/10.17705/1thci.00120>).
- Fashler, S. R., and Katz, J. 2014. "More than Meets the Eye: Visual Attention Biases in Individuals Reporting Chronic Pain," *Journal of Pain Research* (7). (<https://doi.org/10.2147/JPR.S67431>).
- Fashler, S. R., and Katz, J. 2016. "Keeping an Eye on Pain: Investigating Visual Attention Biases in Individuals with Chronic Pain Using Eye-Tracking Methodology," *Journal of Pain Research* (9), pp. 551–561. (<https://doi.org/10.2147/JPR.S104268>).
- Fehrenbacher, D. D., and Djamasbi, S. 2017. "Information Systems and Task Demand: An Exploratory Pupillometry Study of Computerized Decision Making," *Decision Support Systems* (97), Elsevier B.V., pp. 1–11. (<https://doi.org/10.1016/j.dss.2017.02.007>).
- Franklin, Z. C., Holmes, P. S., and Fowler, N. E. 2019. "Eye Gaze Markers Indicate Visual Attention to Threatening Images in Individuals with Chronic Back Pain," *Journal of Clinical Medicine* (8:1). (<https://doi.org/10.3390/jcm8010031>).
- Giel, K. E., Paganini, S., Schank, I., Enck, P., Zipfel, S., and Junne, F. 2018. "Processing of Emotional Faces in Patients with Chronic Pain Disorder: An Eye-Tracking Study," *Frontiers in Psychiatry* (9:MAR). (<https://doi.org/10.3389/fpsy.2018.00063>).
- Iowa State University. 2022. "No Title." (<https://instr.iastate.libguides.com/c.php?g=901522&p=6492159>).
- Jones, E. B., Sharpe, L., Andrews, S., Colagiuri, B., Dudeney, J., Fox, E., Heathcote, L. C., Lau, J. Y. F., Todd, J., van Damme, S., van Ryckeghem, D. M. L., and Vervoort, T. 2021. "The Time Course of Attentional Biases in Pain: A Meta-Analysis of Eye-Tracking Studies," *Pain* (162:3). (<https://doi.org/10.1097/j.pain.0000000000002083>).

- Koenig, S., Körfer, K., Lachnit, H., and Glombiewski, J. A. 2021. "An Attentional Perspective on Differential Fear Conditioning in Chronic Pain: The Informational Value of Safety Cues.," *Behaviour Research and Therapy* (144). (<https://doi.org/10.1016/j.brat.2021.103917>).
- Liossi, C., Schoth, D. E., Godwin, H. J., and Liversedge, S. P. 2014. "Using Eye Movements to Investigate Selective Attention in Chronic Daily Headache," *Pain* (155:3). (<https://doi.org/10.1016/j.pain.2013.11.014>).
- Mahmoodi-Aghdam, M., Dehghani, M., Ahmadi, M., Banaraki, A. K., and Khatibi, A. 2017. "Chronic Pain and Selective Attention to Pain Arousing Daily Activity Pictures: Evidence from an Eye Tracking Study," *Basic and Clinical Neuroscience* (8:6). (<https://doi.org/10.29252/nirp.bcn.8.6.467>).
- Mazidi, M., Dehghani, M., Sharpe, L., Dolatshahi, B., Ranjbar, S., and Khatibi, A. 2021. "Time Course of Attentional Bias to Painful Facial Expressions and the Moderating Role of Attentional Control: An Eye-Tracking Study," *British Journal of Pain* (15:1). (<https://doi.org/10.1177/2049463719866877>).
- Norouzi Nia, J., Varzгани, F., Djamasbi, S., and Tulu, B. 2021. "Visual Hierarchy and Communication Effectiveness in Medical Decision Tools for Surrogate-Decision-Makers of Critically Ill Traumatic Brain Injury Patients," *HCI Interaction Conference*.
- Nuske, H. J., Vivanti, G., and Dissanayake, C. 2015. "No Evidence of Emotional Dysregulation or Aversion to Mutual Gaze in Preschoolers with Autism Spectrum Disorder: An Eye-Tracking Pupillometry Study," *Journal of Autism and Developmental Disorders* (45:11). (<https://doi.org/10.1007/s10803-015-2479-5>).
- Phelps, C. E., Navratilova, E., and Porreca, F. 2021. "Cognition in the Chronic Pain Experience: Preclinical Insights," *Trends in Cognitive Sciences*. (<https://doi.org/10.1016/j.tics.2021.01.001>).

- Pocock, D. C. D. 1981. "Sight and Knowledge.," *Transactions, Institute of British Geographers* (6:4). (<https://doi.org/10.2307/621875>).
- Priebe, J. A., Horn-Hofmann, C., Wolf, D., Wolff, S., Heesen, M., Knippenberg-Bigge, K., Lang, P., and Lautenbacher, S. 2021. "Attentional Processing of Pain Faces and Other Emotional Faces in Chronic Pain—an Eye-Tracking Study," *PLoS ONE* (16:5 May). (<https://doi.org/10.1371/journal.pone.0252398>).
- Ramsey, Jerry D; Burford, Charles L; Beshir, Mohamed Youssef; Jensen, R. C. 1983. "Effects of Workplace Thermal Conditions on Safe Work Behavior," *Journal of Safety Research* (14(3)), pp. 105–114. ([https://doi.org/10.1016/0022-4375\(83\)90021-X](https://doi.org/10.1016/0022-4375(83)90021-X)).
- Rayner, K., Chace, K. H., Slattery, T. J., and Ashby, J. 2006. "Eye Movements as Reflections of Comprehension Processes in Reading," *Scientific Studies of Reading*. (https://doi.org/10.1207/s1532799xssr1003_3).
- "Readability Formulas." (n.d.). (<https://readabilityformulas.com/>).
- Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sanchez, J. P. 2012. "An Assessment of the Effectiveness of a Random Forest Classifier for Land-Cover Classification," *ISPRS Journal of Photogrammetry and Remote Sensing* (67:1). (<https://doi.org/10.1016/j.isprsjprs.2011.11.002>).
- Schoth, D. E., Godwin, H. J., Liversedge, S. P., and Lioffi, C. 2015. "Eye Movements during Visual Search for Emotional Faces in Individuals with Chronic Headache," *European Journal of Pain (United Kingdom)* (19:5). (<https://doi.org/10.1002/ejp.595>).
- Shiro, Y., Nagai, S., Hayashi, K., Aono, S., Nishihara, M., and Ushida, T. 2021. "Changes in Visual Attentional Behavior in Complex Regional Pain Syndrome: A Preliminary Study," *PLoS ONE* (16:2 February). (<https://doi.org/10.1371/journal.pone.0247064>).
- Shojaeizadeh, M., Djamasbi, S., Chen, P., and Rochford, J. 2017. "Text Simplification and Pupillometry: An Exploratory Study," in *Lecture Notes in Computer Science*

(Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 10285). (https://doi.org/10.1007/978-3-319-58625-0_5).

Shojaeizadeh, M., Djamasbi, S., Paffenroth, R. C., and Trapp, A. C. 2019. "Detecting Task Demand via an Eye Tracking Machine Learning System," *Decision Support Systems* (116:June), Elsevier, pp. 91–101. (<https://doi.org/10.1016/j.dss.2018.10.012>).

Soltani, S., van Ryckeghem, D. M. L., Vervoort, T., Heathcote, L. C., Yeates, K. O., Sears, C., and Noel, M. 2022. "Clinical Relevance of Attentional Biases in Pediatric Chronic Pain: An Eye-Tracking Study," *Pain* (163:2). (<https://doi.org/10.1097/j.pain.0000000000002346>).

Soltani, S., van Ryckeghem, D. M. L., Vervoort, T., Heathcote, L. C., Yeates, K., Sears, C., and Noel, M. 2020. "Attentional Biases in Pediatric Chronic Pain: An Eye-Tracking Study Assessing the Nature of the Bias and Its Relation to Attentional Control," *Pain* (161:10). (<https://doi.org/10.1097/j.pain.0000000000001916>).

Strong, S. 2019. "User Experience-Driven Innovation in Smart and Connected Worlds," *AIS Transactions on Human-Computer Interaction* (11:4), p. 215. (<https://doi.org/10.17705/1thci.00121>).

Tobii Technology Inc. 2017. "Tobii Pro Spectrum Product Description." (<https://www.tobii.com/siteassets/tobii-pro/product-descriptions/tobii-pro-spectrum-product-description.pdf?v=2.0%0Ahttps://drive.google.com/file/d/1tWGKNvsHRWv-HOT68TkEnsRvDR9LEpV0/view?usp=sharing>).

Tullis, T., and Albert, B. 2013. "Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics: Second Edition," *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics: Second Edition*. (<https://doi.org/10.1016/C2011-0-00016-9>).

- Varzgani, F., Nia, J. N., Alrefaei, D., and Shojaeizadeh, M. (n.d.). "Effects of Text Simplification on Reading Behavior of Older and Younger Users," *HCI Interaction Conference*, pp. 1–13.
- Williams, A. C. D. C., and Craig, K. D. 2016. "Updating the Definition of Pain," *Pain*. (<https://doi.org/10.1097/j.pain.0000000000000613>).
- Yang, Z., Jackson, T., and Chen, H. 2013. "Effects of Chronic Pain and Pain-Related Fear on Orienting and Maintenance of Attention: An Eye Movement Study," *Journal of Pain* (14:10). (<https://doi.org/10.1016/j.jpain.2013.04.017>).
- Yong, R. J., Mullins, P. M., and Bhattacharyya, N. 2022. "Prevalence of Chronic Pain among Adults in the United States," *Pain* (163:2). (<https://doi.org/10.1097/j.pain.0000000000002291>).

6. APPENDIX I

Bees may have tiny brains, but they are surprisingly intelligent. Bees can learn from their environment to gain a reward, and then teach other bees to do the same. When experienced bees, those who know how to solve a problem, are put in a hive with naïve bees, they somehow communicate the solution to their inexperienced peers. Bees can learn to solve relatively complex problems particularly when the reward is sugar. They can learn to slide or lift a cap and then move a ball to access sugar. Once bees learn something new, they can figure out how to improve it.

Figure 3 - Neutral easy passage (Difficulty level: 7th grade)

Despite attempts to build heightened obsolescence into products to shorten their life span, some experts argue that furniture retailers and manufacturers are likely to benefit from style changes that are pursued more gradually. It is primarily at the lower and younger end of the market that a fashion-oriented strategy has met with success, perhaps most visibly in the case of large furniture retailers. The multinational furniture retailers' emphasis on disposability is evident in their advertisement campaigns which suggest that "Even furniture can go bad". Such advertisement campaigns invite comparison between furniture and other commodities with a shorter life span, such as food and flowers.

Figure 4 - Neutral difficult Passage - Difficulty level: 14th grade

Stabbing headaches, or "ice pick headaches," are short, stabbing, extremely intense headaches that generally last only seconds. People with new or never-evaluated stabbing headache should be evaluated to make sure that they do not have a different primary headache disorder that can mimic primary stabbing headache. Other primary headache disorders that mimic primary stabbing headache include short-lasting unilateral neuralgiform headache attacks with conjunctival injection and tearing (SUNCT), and short-lasting unilateral neuralgiform headache attacks with cranial autonomic symptoms (SUNA). Primary stabbing headache has been previously called ice-pick pains; jabs and jolts; needle-in-the-eye syndrome; ophthalmodynia periodica; and sharp short-lived head pain.

Figure 5 - Pain difficult Passage - Difficulty level: 16th grade

A common cause of a stiff neck is poor posture while working, eating, and sleeping. A compromised posture, such as looking down at your smartphone for a long time, can make your neck ache — a problem that has been dubbed "text neck" or "tech neck". If your neck bothers you, you should also pay attention to your pillow. A pillow that bends your neck forward or to one side will only make your neck pain worse. The kind of pillow you choose and how you sleep on it makes a big difference in how much pain you feel when you wake up in the morning.

Figure 6 - Pain easy passage - Difficulty level: 7th grade

Despite attempts to build heightened obsolescence into products to shorten their life span, some experts argue that furniture retailers and manufacturers are likely to benefit from style changes that are pursued more gradually. It is primarily at the lower and younger end of the market that a fashion-oriented strategy has met with success, perhaps most visibly in the case of large furniture retailers. The multinational furniture retailers' emphasis on disposability is evident in their advertisement campaigns which suggest that "Even furniture can go bad". Such advertisement campaigns invite comparison between furniture and other commodities with a shorter life span, such as food and flowers.

Figure 7 – Passage-level AOI (AOI at the size of the page)

Despite attempts to build heightened obsolescence into products to shorten their life span, some experts argue that furniture retailers and manufacturers are likely to benefit from style changes that are pursued more gradually. It is primarily at the lower and younger end of the market that a fashion-oriented strategy has met with success, perhaps most visibly in the case of large furniture retailers. The multinational furniture retailers' emphasis on disposability is evident in their advertisement campaigns which suggest that "Even furniture can go bad". Such advertisement campaigns invite comparison between furniture and other commodities with a shorter life span, such as food and flowers.

Figure 8 - Sentence-level AOI



Figure 9 - Word-level AOI

7. APPENDIX II

Ph.D. defense presentation slide deck



Ph.D. Dissertation Defense

Eye Tracking and Wellness: The Quest for Unobtrusive Biomarkers for Designing Smart Neuro Information Systems

JAVAD NOROUZI NIA

Dissertation
Committee:

Prof. Soussan Djamasbi (chair)

Prof. Diane Strong

Prof. Randy Paffenroth

Today's Agenda

- Digital Economy and Smart Neuro Information Systems
- Problem Domain
- Research goal
- Background
- Methodology
- Results
- Conclusion





Neuro Information Systems

Today's market needs for quick response to users' needs require adaptive devices. Using NeuroIS, we can build such adaptive devices.

Neuro Information Systems (NeuroIS) research refers to research that uses neuroscience knowledge, tools and physiological measures to design intelligent information systems that can detect user needs automatically.^{1,2}

1. <http://www.neurois.org/what-is-neurois/>

2. Shojaeizadeh, M., Djamasi, S., Paffenroth, R. C., & Trapp, A. C. (2019). Detecting task demand via an Eye Tracking Machine Learning System. *Decision Support Systems*, 116, 91–101.

Smart NeuroIS for Detecting Chronic Pain

✓ Problem Domain

- Chronic pain is pain that persists or recurs for more than 3 months¹
- It impacted 50.2 million² American adults and the US pays \$560 billion³ on medical expenses, lost productivity and disability programs, annually.

✓ Eye Tracking

- Vision is a dominant sense⁴
- Modern eye-tracking devices allow to capture gaze unobtrusively

1. Treede, R. D., Rief, W., Barke, A., Aziz, Q., Bennett, M. I., Benoliel, R., Cohen, M., Evers, S., Finnerup, N. B., First, M. B., Giamberardino, M. A., Kaasa, S., Korwisi, B., Kosek, E., Lavand'homme, P., Nicholas, M., Perrot, S., Scholz, J., Schug, S., Smith, B. H., ... Wang, S. J. (2019). Chronic pain as a symptom or a disease: the IASP Classification of Chronic Pain for the International Classification of Diseases (ICD-11). *Pain*, 160(1), 19–27.

2. Yong RJ, Mullins PM, Bhattacharyya N. Prevalence of chronic pain among adults in the United States. *Pain*. 2021 Apr 2. Epub ahead of print.

3. https://www.washingtonpost.com/national/health-science/the-big-number-50-million-adults-experience-chronic-pain/2018/10/19/30831828-d2e0-11e8-83d6-291fced2ab1_story.html

4. Pocock, D. C. D. (1981). Sight and Knowledge. *Transactions of the Institute of British Geographers*, 6(4), 385–393.

Goals & Contribution



◆ Research Goal

- Develop a proof of concept for an eye tracking machine learning (ETML) System that can detect chronic pain automatically and exclusively from eye movement
 - Develop and test an iterative process for data collection and model development following user-centered design methodology
-

◆ Research Contribution

- NeuroIS (designing smart systems)
- Eye-tracking (identifying and expanding set of metrics for attentional biases)
- Health and wellness (objective biomarker to diagnose chronic pain or assess treatment efficacy)

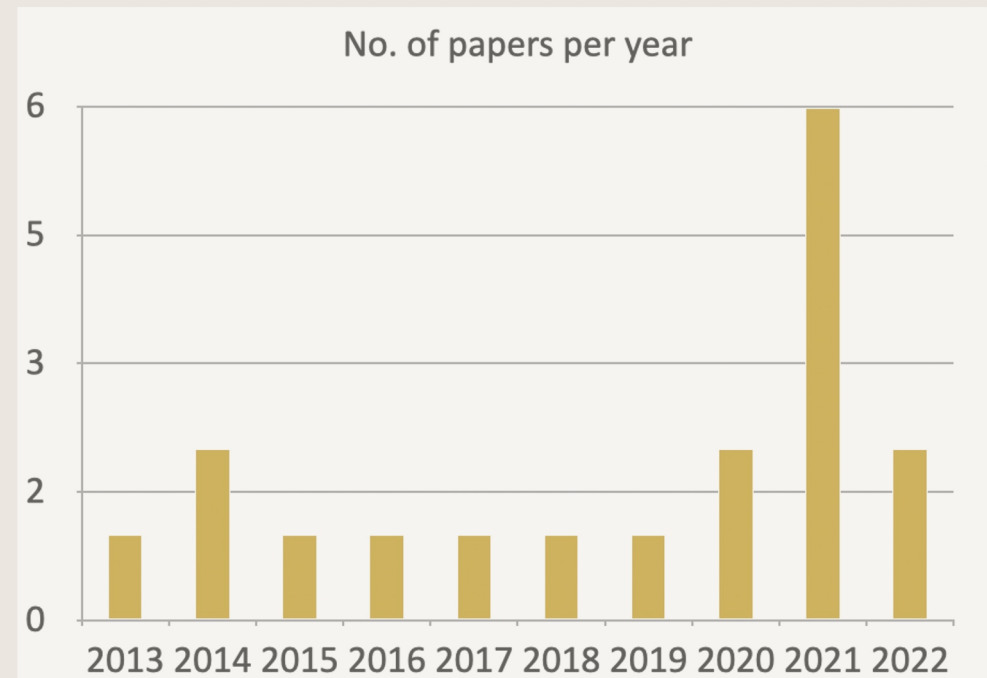


Background

Background

- May 2020 review paper of eye tracking pain studies (9 chronic pain – eye-tracking papers)¹
- May 2020 – present: 9 newer studies

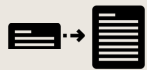
Total of **18 papers** used eye tracking to study chronic pain



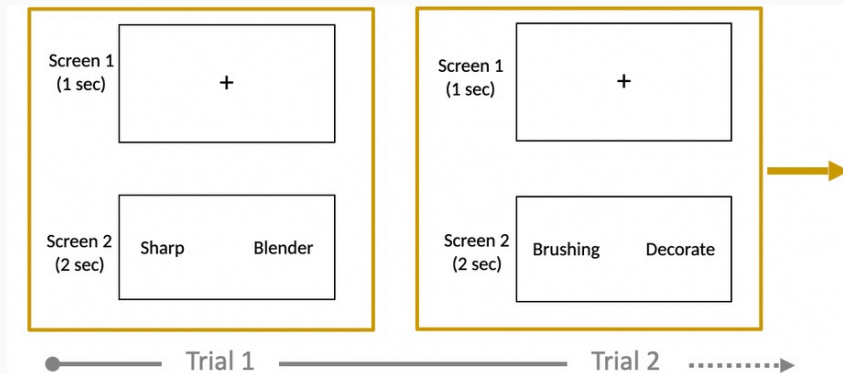
¹ Chan, F., Suen, H., Jackson, T., Vlaeyen, J., & Barry, T. J. (2020). Pain-related attentional processes: A systematic review of eye-tracking research. *Clinical psychology review*, 80, 101884. <https://doi.org/10.1016/j.cpr.2020.101884>

Limitation of Prior Research

Limited context



Short exposure time



A common cause of a stiff neck is poor posture while working, eating, and sleeping. A compromised posture, such as looking down at your smartphone for a long time, can make your neck ache - a problem that has been dubbed "text neck" or "tech neck. If your neck bothers you, you should also pay attention to your pillow. A pillow that bends your neck forward or to one side will only make your neck pain worse. The kind of pillow you choose and how you sleep on it makes a big difference in how much pain you feel when you wake up in the morning.

Limitation of Prior Research

Limited context



Short exposure time



Small set of eye-tracking variables



Analysis method



Experiment Design



✓ Process

- Designing and testing the visual stimuli with 10 participants
- Designing and developing ETML iteratively with 55 different participants

✓ Experimental Session

- Consent
- Calibration
- Text Passages
- Debriefing

✓ Equipment

- 600 Hz Tobii Pro Spectrum
- Tobii Pro Lab software 1.162
- IVT filter, with 100 ms for fixation threshold
- 30 degree/s for saccade threshold



Each participant views



4 Passages



19 sentences



410 words

Pain easy passage

A common cause of a stiff neck is poor posture while working, eating, and sleeping. A compromised posture, such as looking down at your smartphone for a long time, can make your neck ache — a problem that has been dubbed “text neck” or “tech neck”. If your neck bothers you, you should also pay attention to your pillow. A pillow that bends your neck forward or to one side will only make your neck pain worse. The kind of pillow you choose and how you sleep on it makes a big difference in how much pain you feel when you wake up in the morning.

Neutral easy passage

Bees may have tiny brains, but they are surprisingly intelligent. Bees can learn from their environment to gain a reward, and then teach other bees to do the same. When experienced bees, those who know how to solve a problem, are put in a hive with naïve bees, they somehow communicate the solution to their inexperienced peers. Bees can learn to solve relatively complex problems particularly when the reward is sugar. They can learn to slide or lift a cap and then move a ball to access sugar. Once bees learn something new, they can figure out how to improve it.

Pain difficult passage

Stabbing headaches, or “ice pick headaches,” are short, stabbing, extremely intense headaches that generally last only seconds. People with new or never-evaluated stabbing headache should be evaluated to make sure that they do not have a different primary headache disorder that can mimic primary stabbing headache. Other primary headache disorders that mimic primary stabbing headache include short-lasting unilateral neuralgiform headache attacks with conjunctival injection and tearing (SUNCT), and short-lasting unilateral neuralgiform headache attacks with cranial autonomic symptoms (SUNA). Primary stabbing headache has been previously called ice-pick pains; jabs and jolts; needle-in-the-eye syndrome; ophthalmodynia periodica; and sharp short-lived head pain.

Neutral difficult passage

Despite attempts to build heightened obsolescence into products to shorten their life span, some experts argue that furniture retailers and manufacturers are likely to benefit from style changes that are pursued more gradually. It is primarily at the lower and younger end of the market that a fashion-oriented strategy has met with success, perhaps most visibly in the case of large furniture retailers. The multinational furniture retailers’ emphasis on disposability is evident in their advertisement campaigns which suggest that “Even furniture can go bad”. Such advertisement campaigns invite comparison between furniture and other commodities with a shorter life span, such as food and flowers.

Measuring Attentional Biases

Category of eye movement metrics used in this research

	Fixation	Visit	Saccade	Pupillometry
Used variables in the systematic review	7	4	0	1
Used variables in my proposed ETML	32	16	19	9

Preparing Datasets for My Proposed Theory-based ML

- ✓ Split the data to train and test sets (k-fold=5), with 10 repeat
- ✓ Balancing the data
- ✓ Addressing missing values (NANs)
- ✓ Arranging the data

	Passage 1 - Fixation No.	Passage 1 - Fixation Duration	Passage 2 - Fixation No.	Passage 2 - Fixation Duration	...
Participant 1					
Participant 2					
Participant 3					
Participant 4					

Evaluation of Model

Evaluation and selection of models based on performance of testing set on:

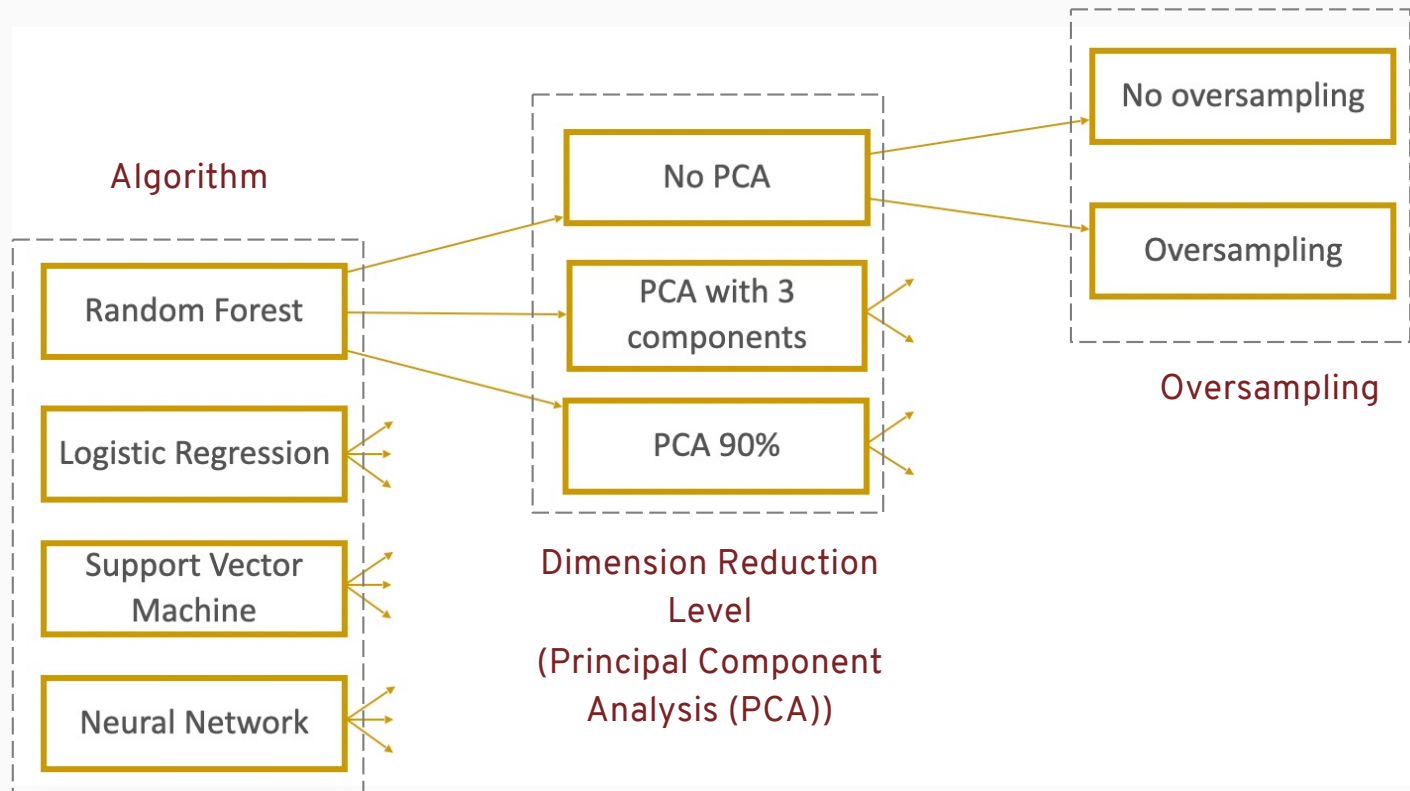
✓ Accuracy

✓ Specificity

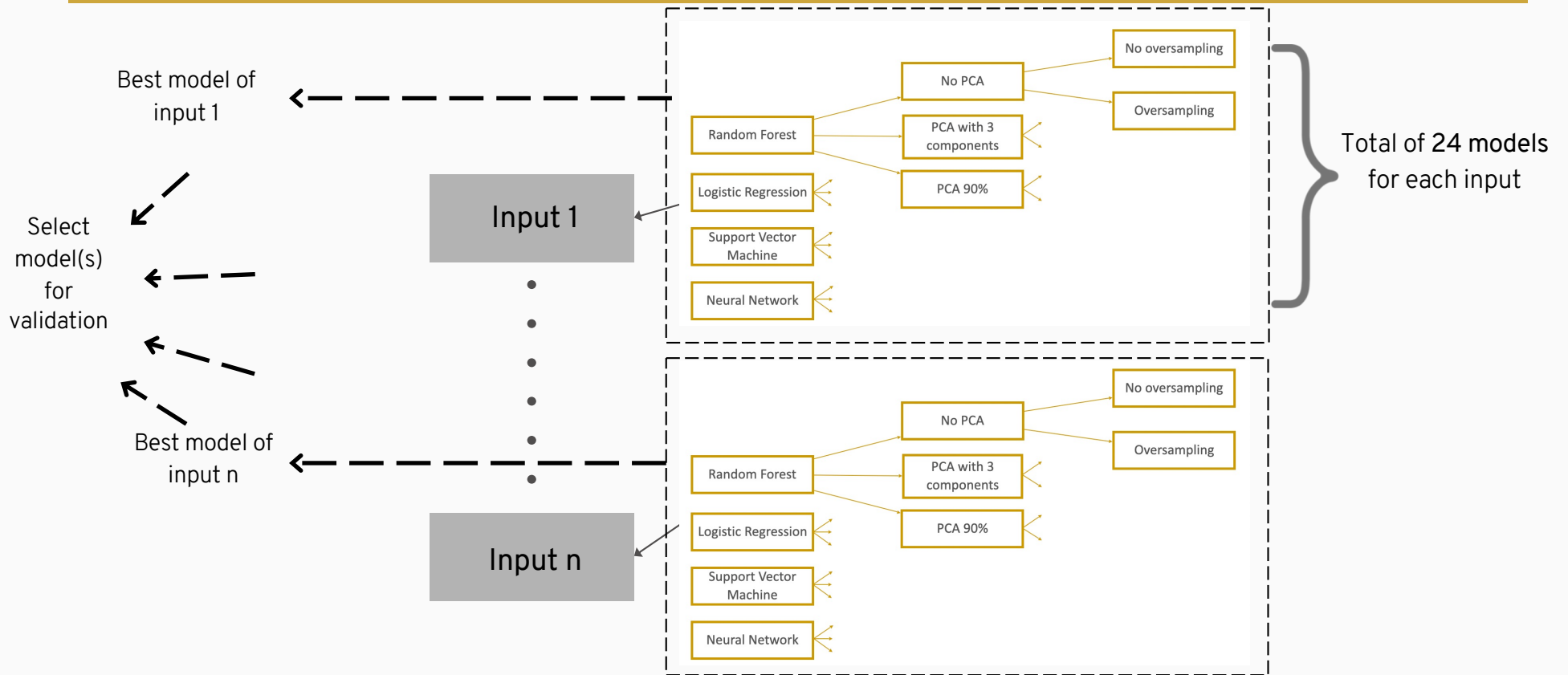
✓ Sensitivity

✓ F1 score

ML Models and Settings



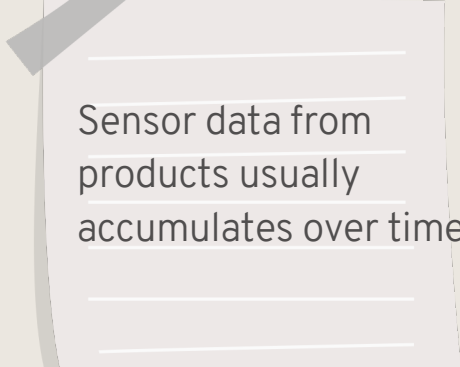
Model Selection




Model Set Inputs

	Model Set #	Data Level	Arrangement	Stimuli
Exploratory	1	Passage-level	Unarranged	All passages
	2	Passage-level	Re-arranged	All passages
	3	Sentence-level	Unarranged	All passages
	4	Sentence-level	Re-arranged	All passages
	5	Word-level	Unarranged	All passages
	6	Word-level	Re-arranged	All passages
	7	Passage, sentence, word-level	Unarranged	All passages
Theory-based	8	Passage, sentence, word-level	Re-arranged	All passages
	9	Passage, sentence, word-level	Re-arranged	Pain Passages only
	10	Passage, sentence, word-level	Re-arranged	Neutral Passages only
	11	Passage, sentence, word-level	Re-arranged	Difficult Passages only
	12	Passage, sentence, word-level	Re-arranged	Easy Passages only
	13	Passage, sentence, word-level	Re-arranged	Neutral Easy Passage only
	14	Passage, sentence, word-level	Re-arranged	Neutral Difficult Passage only
	15	Passage, sentence, word-level	Re-arranged	Pain Difficult Passage only
	16	Passage, sentence, word-level	Re-arranged	Pain Easy Passage only

Proposed Iterative Process



Sensor data from products usually accumulates over time

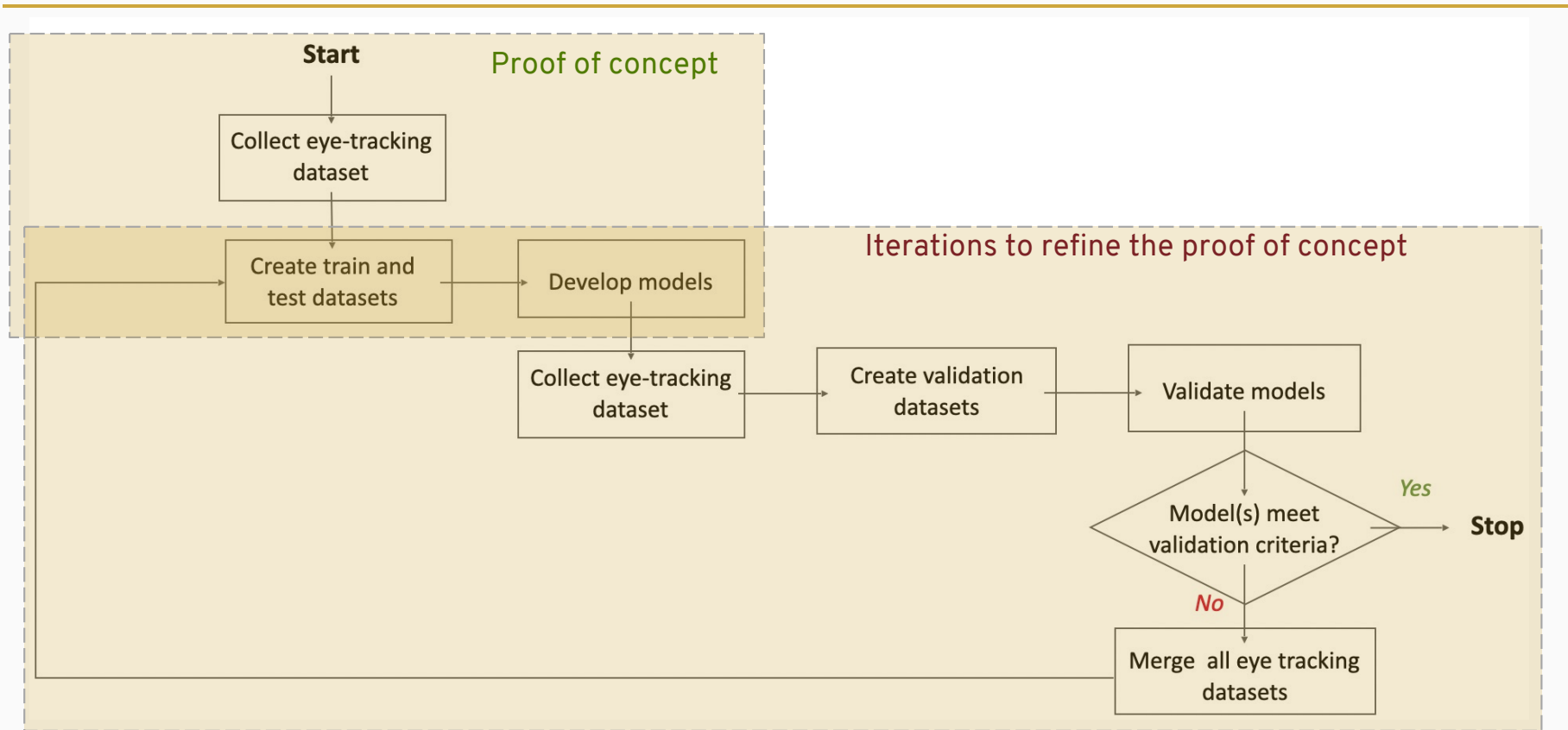


Learning while developing a product

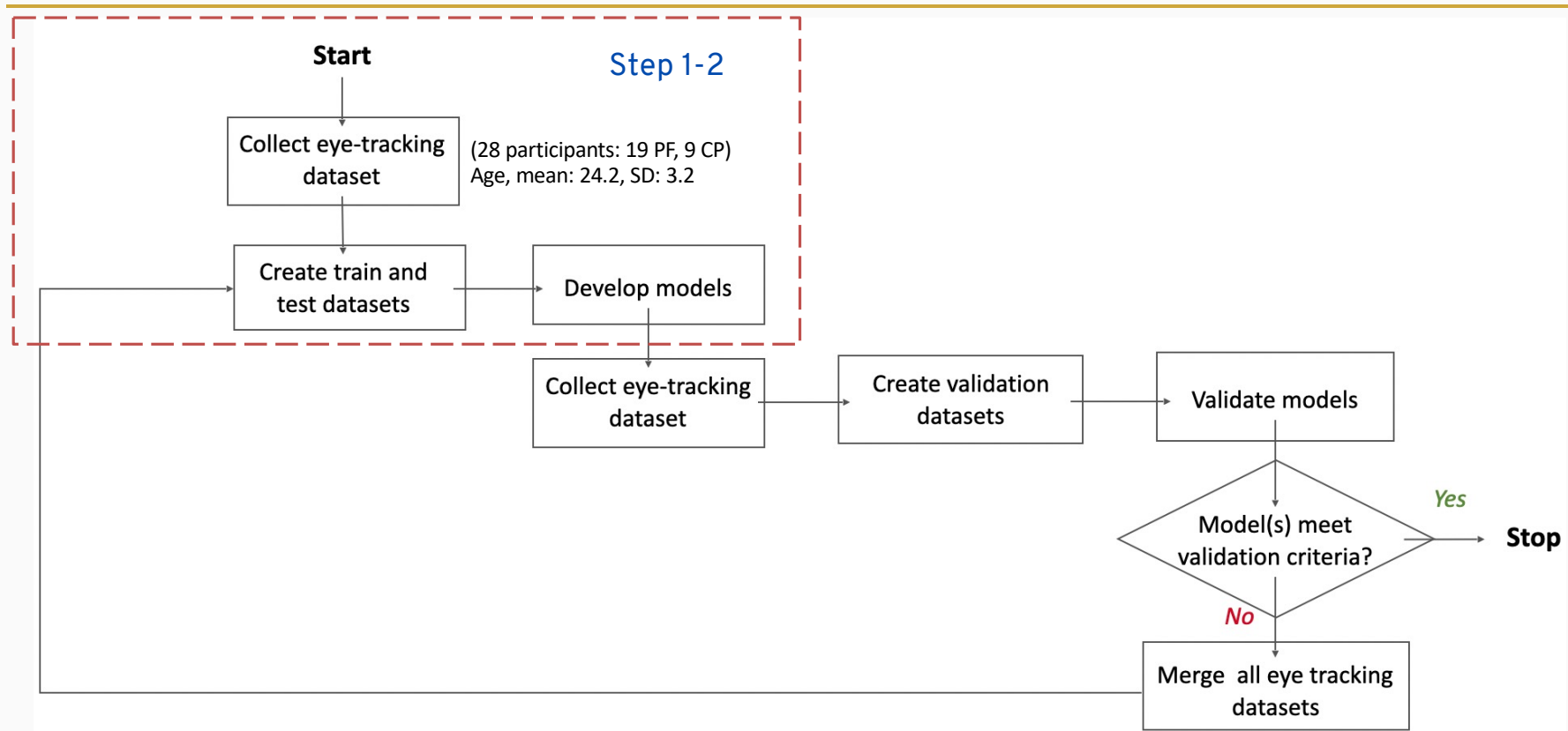
✓ Iterative process based on UXDI¹ model for product development

1. Djamasbi, S., & Strong, D. (2019). User Experience-driven Innovation—Theory and Practice: Introduction to Special Issue. *AIS Transactions on Human-Computer Interaction*, 11(4), 208-214. <https://doi.org/10.17705/1thci.00120>

Proposed Iterative Process



Implemented Iterative Process



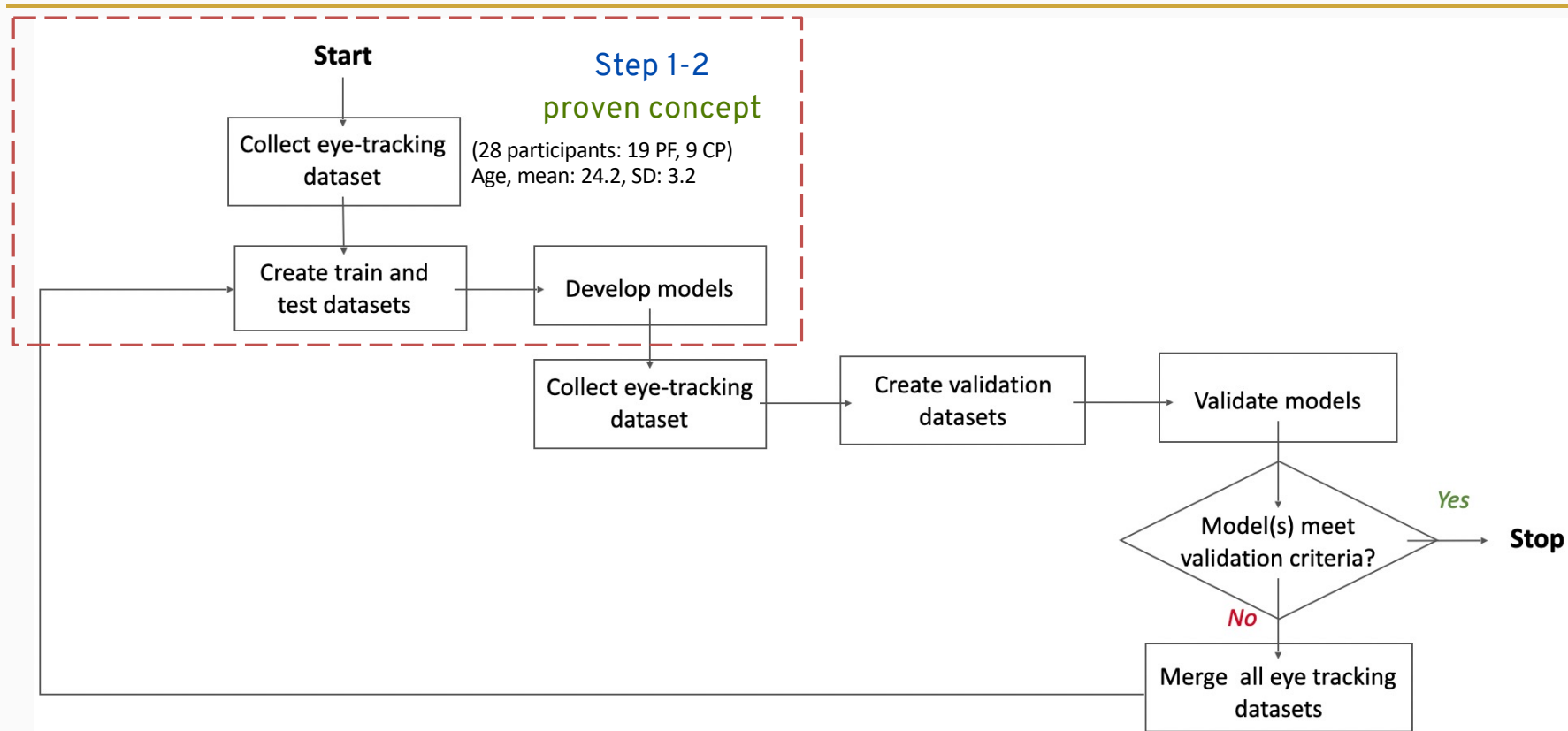
Step 2 Results

Develop Models with Step 1 Train/Test data

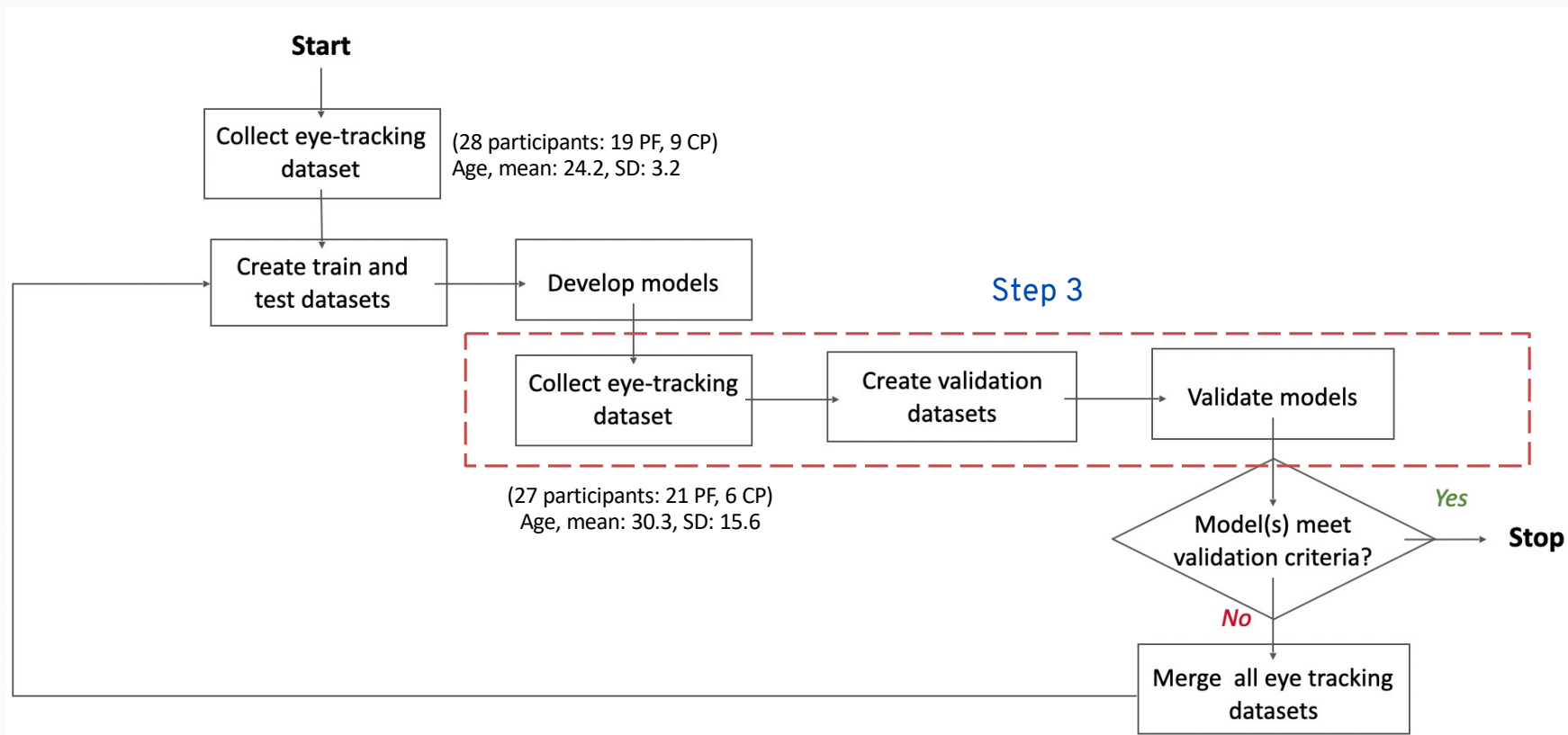
Model Set #	Input	Algorithm	Oversampling	PCA	Accuracy	Specificity	Sensitivity	F1 score	# of Observations	# of Variables
8	PSW (R) – All passages	SVC	Yes	No	0.703	0.76	0.672	0.734	28	3658
9	PSW (R) – Pain passages only	SVC	Yes	No	0.735	0.78	0.720	0.775		1797
10	PSW (R) – Neutral passages only	SVC	Yes	No	0.696	0.79	0.647	0.718		1862
11	PSW (R) – Difficult passages only	SVC	Yes	No	0.723	0.78	0.698	0.757		1910
12	PSW (R) – Easy passages only	SVC	Yes	No	0.684	0.8	0.635	0.719		1749
13	PSW (R) – Neutral easy passage only	SVC	Yes	No	0.703	0.88	0.625	0.71		911
14	PSW (R) – Neutral difficult passage only	RF	No	3	0.683	0.61	0.728	0.742		952
15	PSW (R) – Pain difficult passage only	SVC	Yes	No	0.753	0.76	0.743	0.788		959
16	PSW (R) – Pain easy passage only	SVC	Yes	No	0.713	0.77	0.683	0.744		939

PSW: passage, sentence, word-level data; (R): re-arranged

Implemented Iterative Process



Implemented Iterative Process

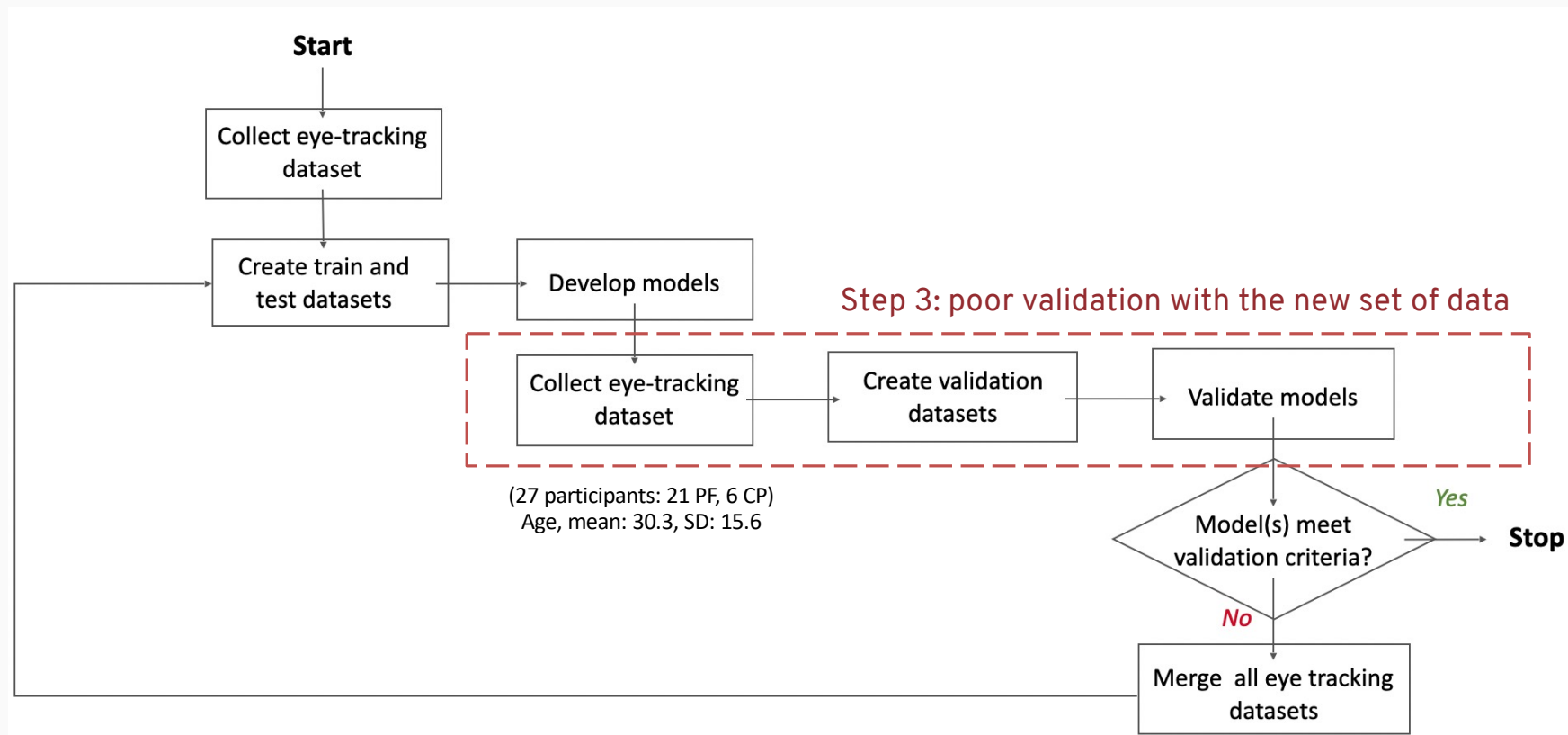


Step 3 - Validating step 2 models with a new dataset

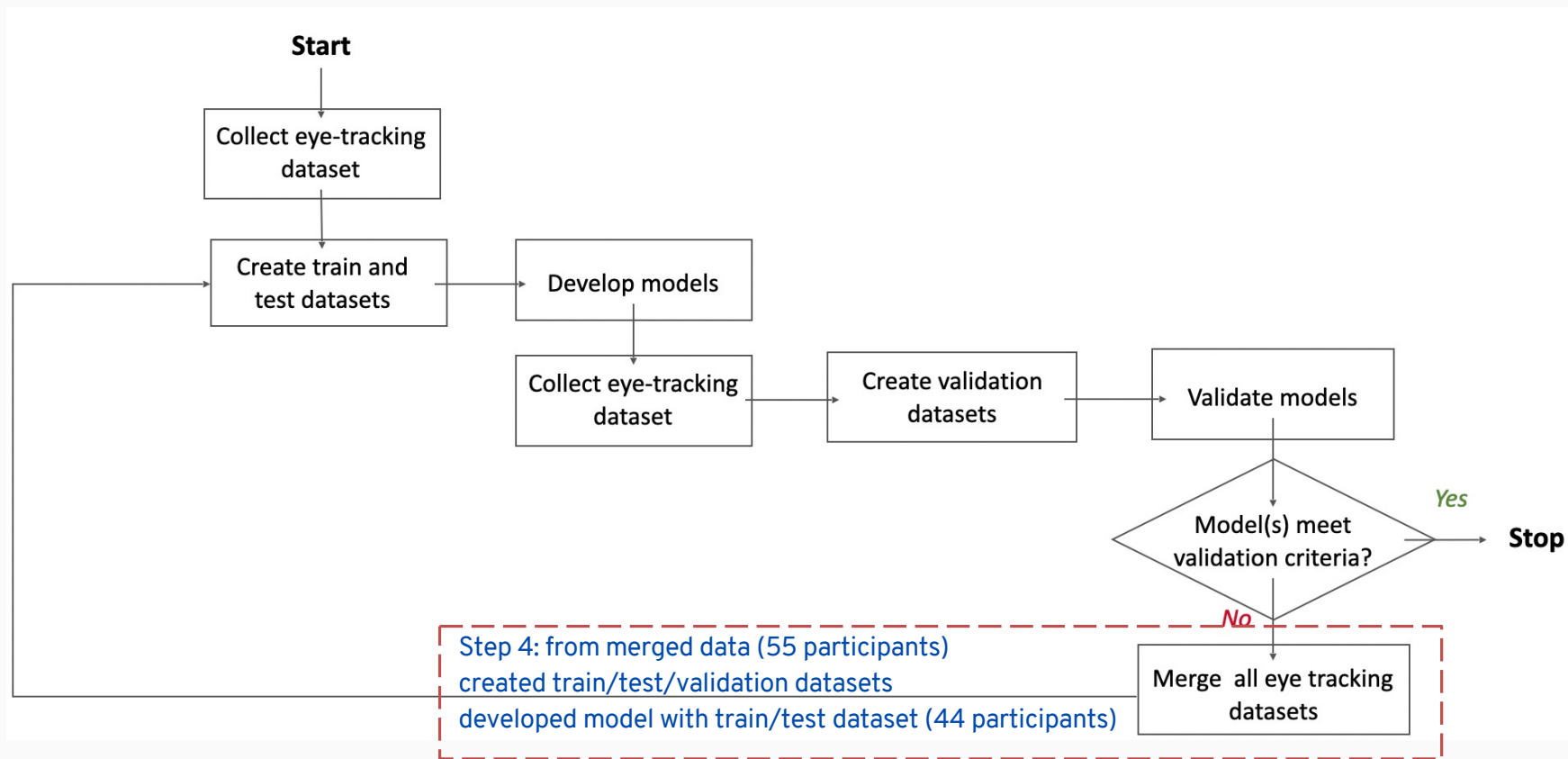
Model Set #	Input	Algorithm	Oversampling	PCA	Accuracy	Specificity	Sensitivity	F1 score
8	PSW (R) – All passages	SVC	Yes	No	0.175	1.0	0.096 X	0.175
9	PSW (R) – Pain passages only	SVC	Yes	No	0.364	1.0	0.182 X	0.308
10	PSW (R) – Neutral passages only	SVC	Yes	No	0.267	1.0	0.058 X	0.11
11	PSW (R) – Difficult passages only	SVC	Yes	No	0.344	1.0	0.157 X	0.272
12	PSW (R) – Easy passages only	SVC	Yes	No	0.279	1.0	0.073 X	0.137
13	PSW (R) – Neutral easy passage only	SVC	Yes	No	0.296	1.0	0.094 X	0.172
14	PSW (R) – Neutral difficult passage only	RF	No	3	0.68	0.183 X	0.822	0.8
15	PSW (R) – Pain difficult passage only	SVC	Yes	No	0.381	0.957	0.217 X	0.353
16	PSW (R) – Pain easy passage only	SVC	Yes	No	0.253	0.993	0.041 X	0.079

Not acceptable validation results

Implemented Iterative Process



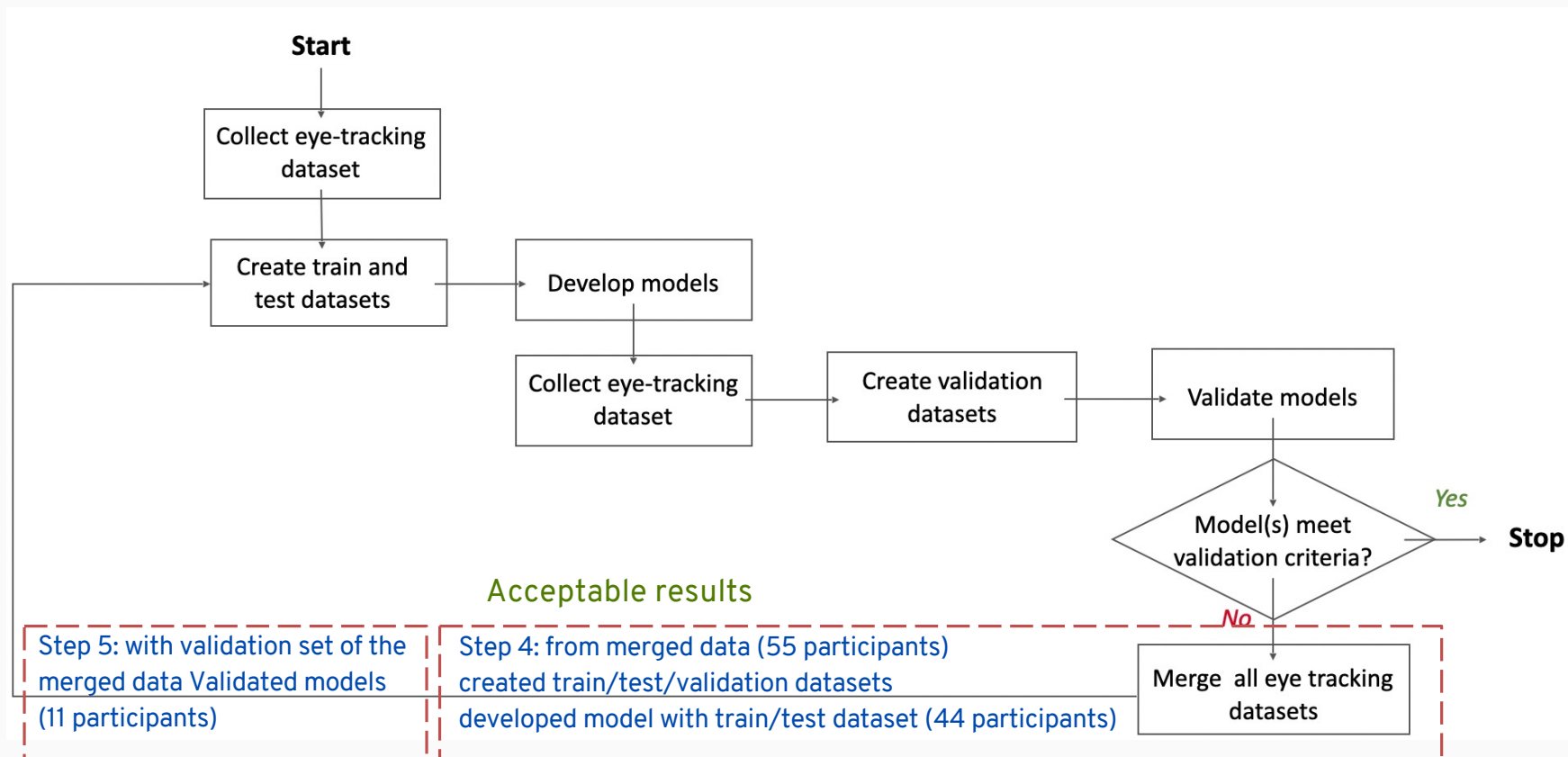
Implemented Iterative Process



Step 4 Results – Develop Models with Step 3 Train/Test data

Model Set #	Input	Algorithm	Oversampling	PCA	Accuracy	Specificity	Sensitivity	F1 score	# of Observations	# of Variables
8	PSW (R) – All passages	LR	No	No	0.719	0.520	0.794	0.799	44	3405
9	PSW (R) – Pain passages only	LR	Yes	No	0.649	0.547	0.699	0.734		1664
10	PSW (R) – Neutral passages only	SVC	Yes	No	0.557	0.808	0.462	0.603		1742
11	PSW (R) – Difficult passages only	SVC	Yes	No	0.589	0.793	0.514	0.631		1755
12	PSW (R) – Easy passages only	NN	No	No	0.652	0.475 X	0.719	0.75		1651
13	PSW (R) – Neutral easy passage only	SVC	Yes	No	0.548	0.85	0.434 X	0.583		872
14	PSW (R) – Neutral difficult passage only	LR	Yes	3	0.520	0.57	0.503	0.573		871
15	PSW (R) – Pain difficult passage only	SVC	Yes	No	0.639	0.793	0.580	0.688		885
16	PSW (R) – Pain easy passage only	SVC	Yes	No	0.53	0.783	0.434 X	0.573		780

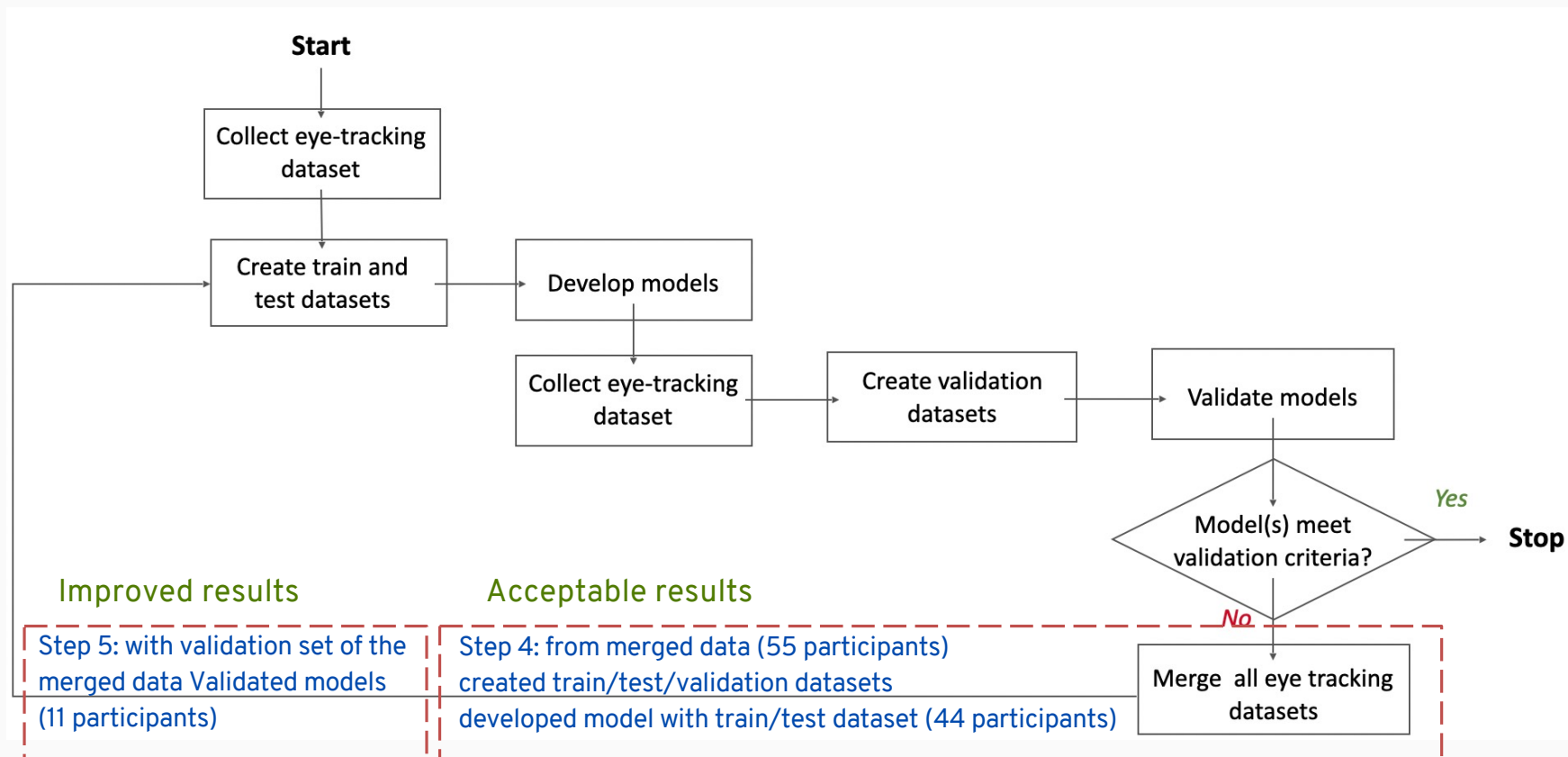
Implemented Iterative Process



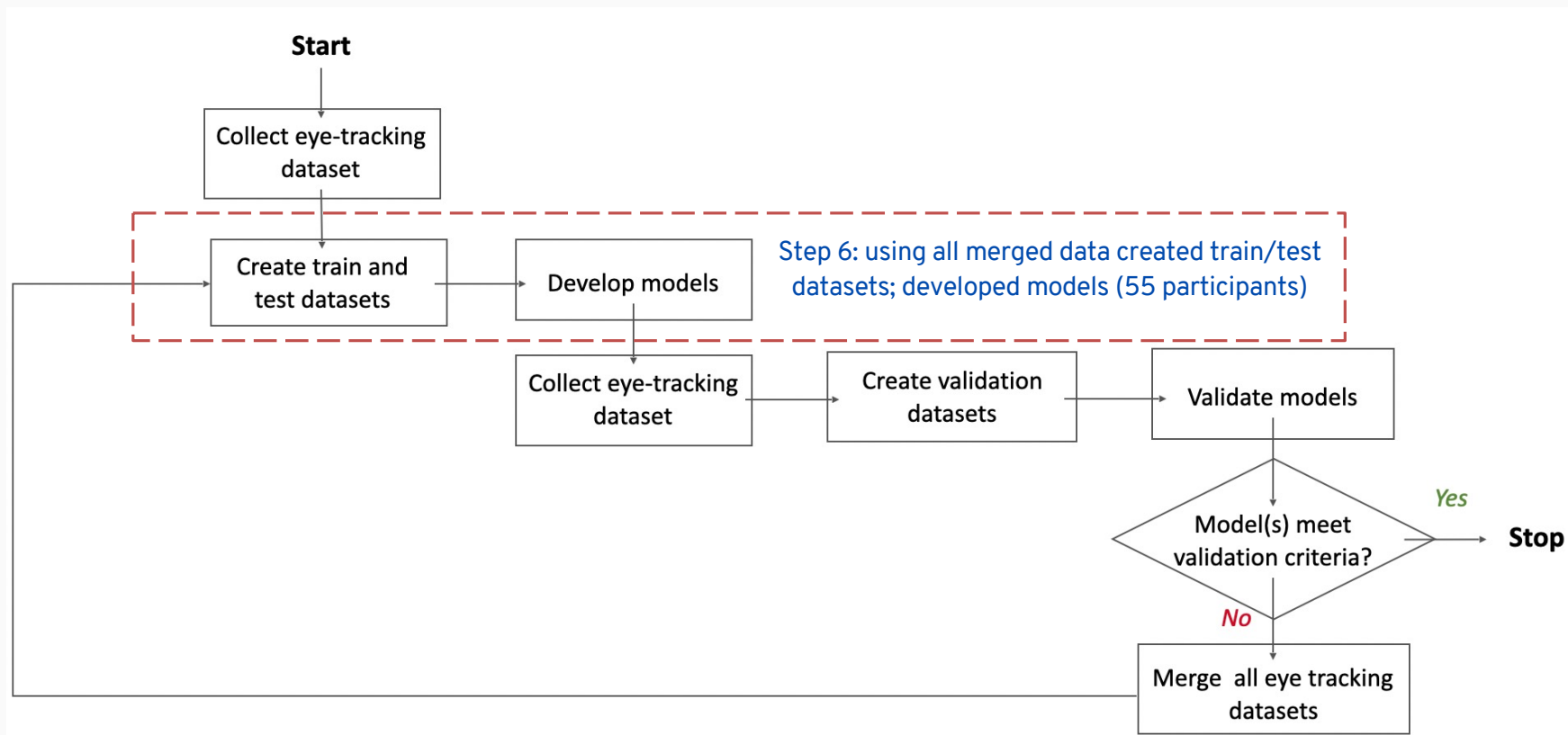
Step 5 - Validating step 4 models

Model Set #	Input	Algorithm	Oversampling	PCA	Accuracy	Specificity	Sensitivity	F1 score
8	PSW (R) – All passages	LR	No	No	0.598	0.633	0.585	0.679
9	PSW (R) – Pain passages only	LR	Yes	No	0.685	0.553	0.735	0.773
10	PSW (R) – Neutral passages only	SVC	Yes	No	0.371	0.993	0.138 X	0.241
11	PSW (R) – Difficult passages only	SVC	Yes	No	0.404	0.82	0.248 X	0.376
14	PSW (R) – Neutral difficult passage only	LR	Yes	3	0.595	0.64	0.578	0.674
15	PSW (R) – Pain difficult passage only	SVC	Yes	No	0.385	0.72	0.26 X	0.381

Implemented Iterative Process



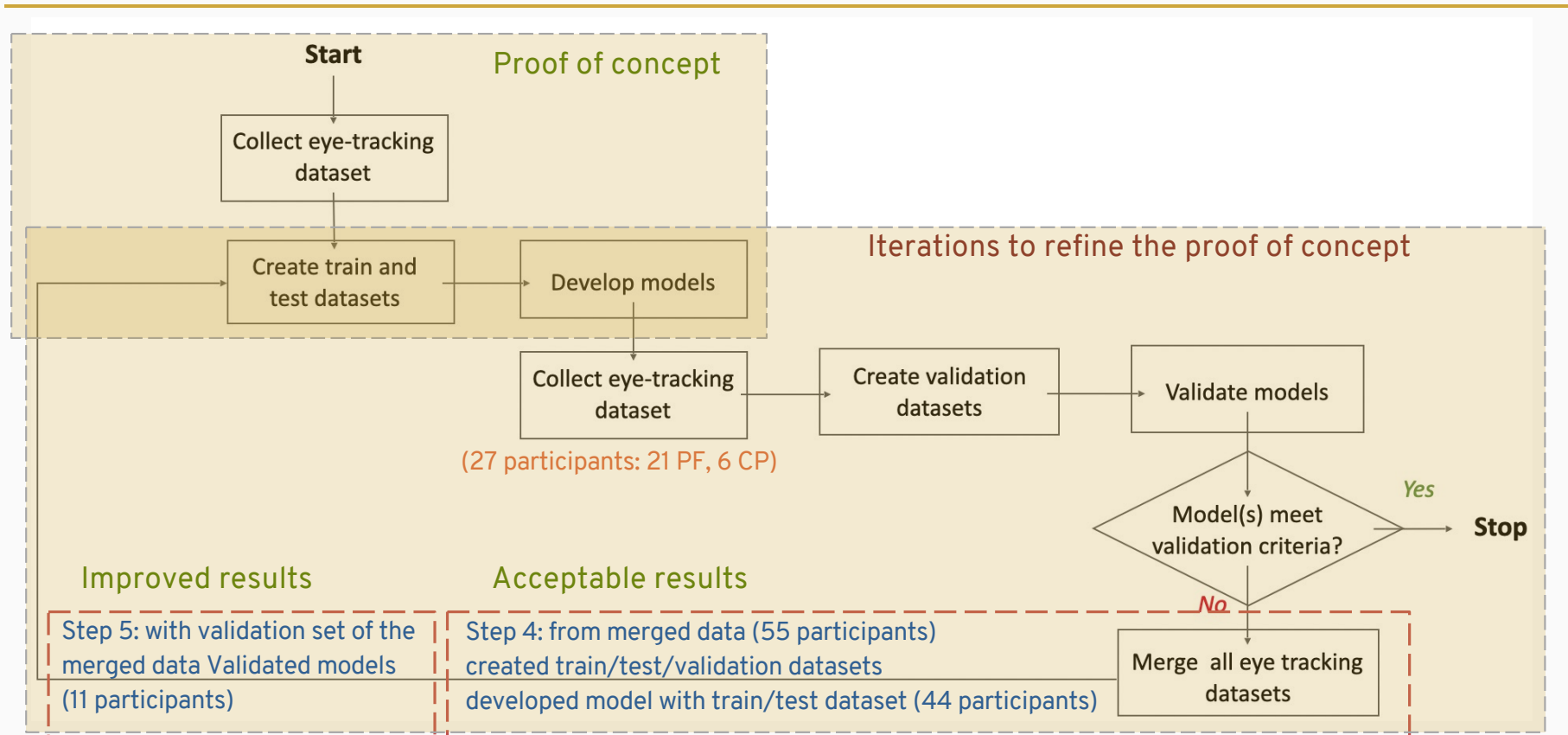
Implemented Iterative Process



Step 6 Results – Develop Models with Step 3 All Data

Model Set #	Input	Algorithm	Oversampling	PCA	Accuracy	Specificity	Sensitivity	F1 score	# of Observations	# of Variables
8	PSW (R) – All passages	LR	Yes	3	0.591	0.540	0.61	0.668	55	3405
9	PSW (R) – Pain passages only	LR	Yes	No	0.651	0.513	0.703	0.740	55	1664
10	PSW (R) – Neutral passages only	LR	Yes	3	0.585	0.533	0.605	0.671	55	1742
11	PSW (R) – Difficult passages only	SVC	Yes	No	0.549	0.793	0.458	0.596	55	1755
12	PSW (R) – Easy passages only	LR	Yes	3	0.615	0.48 X	0.665	0.715	55	1651
13	PSW (R) – Neutral easy passage only	LR	Yes	3	0.636	0.526	0.678	0.721	55	872
14	PSW (R) – Neutral difficult passage only	LR	Yes	3	0.547	0.594	0.53	0.613	55	871
15	PSW (R) – Pain difficult passage only	LR	Yes	3	0.583	0.533	0.603	0.667	55	885
16	PSW (R) – Pain easy passage only	LR	Yes	3	0.538	0.5	0.552	0.635	55	780

Implemented Iterative Process



Conclusion

Conclusion

- ✓ Proof of concept for an ETML model to differentiate chronic pain and pain-free people with validated models on step 5

Model Set #	Input	Algorithm	Oversampling	PCA	Accuracy	Specificity	Sensitivity	F1 score
8	PSW (R) – All passages	LR	No	No	0.598	0.633	0.585	0.679
9	PSW (R) – Pain passages only	LR	Yes	No	0.685	0.553	0.735	0.773
14	PSW (R) – Neutral difficult passage only	LR	Yes	3	0.595	0.64	0.578	0.674

- ✓ Evidence for working iterative process

Limitation

- ✓ Text passages
- ✓ Universal label for all Chronic pains
- ✓ Initial datasets came from WPI community (COVID)

Future Research

- ✓ Adding images to stimuli
- ✓ Model eye movements of people with different types of pain
- ✓ Wider range of participants



Thank You

for attending my presentation!

JAVAD NOROUZI NIA

School of Business
Worcester Polytechnic Institute