The Eyes Have It : Characteristics of Deep Reading Activities on Desktop and Mobile Devices

XIUGE CHEN, The University of Melbourne, Australia
NAMRATA SRIVASTAVA, University of Melbourne, and Monash University, Australia
RAJIV JAIN, Adobe Research, United States
JENNIFER HEALEY, Adobe Research, United States
TILMAN DINGLER, University of Melbourne, Australia

Deep reading fosters text comprehension, memory, and critical thinking. But with Mobile interfaces and online incentives around attention capture, concerns have grown about deep reading activities being replaced by shallow skimming and sifting through information. Traditionally, reading quality is assessed using comprehension tests, which require readers to explicitly answer a set of carefully composed questions. To quantify and understand reading behaviour in natural settings and at scale, however, implicit measures are needed of deep versus shallow reading across Desktop and Mobile devices. In this paper, we present an approach to systematically induce and detect deep and shallow reading patterns using eye movement and interaction data. Based on a user study with 36 participants, we created models that detect deep reading on both devices with up to 0.82 AUC. We present the characteristics of deep reading and discuss how our models can be used to monitor long-term changes in reading behaviours.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; *Human computer interaction (HCI)*.

Additional Key Words and Phrases: Reading mode classification, Digital Devices, Eye tracking, Gaze features

1 INTRODUCTION

Over the past five thousand years humans have developed the ability to read and to read deeply, spending years developing an array of sophisticated cognitive processes that include inferential and deductive reasoning, analogical skills, critical analysis, reflection, and insight [49]. These deep reading skills have followed the development of widespread literacy and the ability to produce written documents in astounding quantities, especially since the Industrial Revolution. Today, more than 86% of the world's population is considered literate [1] and this literacy has changed how we can rapidly communicate complex ideas across communities, countries and nations. While reading has never been a static discipline and reading habits have evolved over time (for example from reading aloud to silent reading), with the advent of the digital revolution, where, how, and what we read is changing very rapidly. Chronic information overload, the ability to search and "surf" from one site to another as well as attention-seeking interface design all work against the practice of deeply reading a single, potentially complex piece of writing and drawing inferences and insights from it. While the ability to distill insights from a document takes mere milliseconds for a trained mind, the skill of deep reading takes years to develop. There is currently increasing concern that as a new generation grows up reading more and more on Mobile devices, deep reading skills may become underdeveloped [50]. While this concern seems well supported, there currently does not exist a reliable implicit method for measuring deep versus shallow reading habits across both Desktop and Mobile devices. Without being able to measure the potential decline of deep reading, we will remain uninformed about the true magnitude of the shift and the factors that might be accelerating it and we will not be able to create an informed strategy to preserve deep reading skills.

[1] https://ourworldindata.org/literacy

To address the problem of implicitly measuring deep reading across both Desktop and Mobile devices we first developed and tested a method for inducing both deep reading and shallow reading on both these platforms, then performed a series of user tests where we measured eye movement patterns in both conditions and finally we developed an algorithm that could predict deep or shallow reading on either platform using these patterns. Traditionally, deep versus shallow reading is assessed explicitly using reading comprehension tests. These tests consist of a carefully constructed set of questions that evaluate the reader's literal, inferential, and evaluative understanding of text [6]. Unfortunately, this evaluation method cannot scale as it is infeasible both to create such questions for every book or article or to expect readers to fill in a set of questions each time they consume a piece of writing. We instead need an implicit method of measuring reading modes.

Since cognition can not be measured directly, technologies need to use surrogate measures to sense and infer cognitive activities. The eyes have often been described as providing a "window into our mind" [47] and so eye movements have been used to detect reading [23] as well as differentiate between different reading goals [45]. And while eye fixation durations and saccade lengths have been shown to correlate with readers' comprehension levels [45, 46], robust models capable of classifying deep and shallow reading are yet to be made work on both Desktop and Mobile devices, which is where reading increasingly takes place [40]. With the prevalence of front-facing device cameras, continuous eye tracking becomes increasingly feasible.

We evaluated our approach in a lab-based experiment with 36 participants during which we recorded gaze and interaction data while reading on Desktop and a Mobile device. We extracted low and mid-level eye-movement features to train a machine-learning model to differentiate between deep and shallow reading. Our results show a strong correlation between eye-movement patterns and deep reading patterns and the resulting models allow us to detect deep reading with 0.82 AUC on Desktop and 0.73 AUC on a Mobile device. We present a detailed analysis of the differences in the reading patterns on Desktop vs. Mobile, which paves the way for continuous and unobtrusive tracking and quantification of reading activities.

2 RELATED WORK

The most validated methods of measuring reading comprehension are explict methods that use post hoc reading questions about the test to judge comprehension. The majority of work on implicit assessment of reading has focused on the interpretation of eye movements. Here we briefly introduce explicit methods, which we later use to validate our deep and shallow reading induction then cover the use of eye tracking for assessing reading comprehension for paper and Desktop reading. To the best of our knowledge, we are the first to present results for evaluating reading behaviour using eye tracking on the Mobile platform. Our work is mainly rooted in related research in eye movement tracking, comprehension assessment, and reading behaviour tracking.

2.1 Explicit Assessment of Reading Using Comprehension Tests

Reading comprehension can be loosely defined as the amount of information or meaning being extracted from the texts [20, 43]. Moreover, comprehension can be further categorized into different levels based on the amount of cognitive demands and interactions [31, 42]. Commonly, the underlying theory categorized comprehension into three levels – literal, inferential, and evaluative [5, 6, 11, 20].

- **Literal comprehension** requires readers to retrieve explicit information from the texts or recall what is stated in the text.

- **Inferential comprehension** requires readers to interpret the authors' meaning by connecting information that is implicit in the text.
- **Evaluative comprehension** requires readers to analyse information based on previous knowledge or experiences and thus relate what is being read to what is known.

Based on the level of comprehension, readers' reading behavior can be categorized into deep and shallow reading, where superficial comprehension level (literal) is achieved in shallow reading and deeper comprehension level (inferential and evaluative) is obtained through deep reading [39, 50]. Identifying deep and shallow reading is crucial in many research and applications in psychology, education, and user interface design [19, 39, 50]. However, differentiating such reading behaviors would require thorough comprehension tests (e.g., via comprehension questions), hence making it infeasible to be implemented and adopted easily. Moreover, individuals' reading behaviors are very dynamic [2, 21], which makes the analysis more challenging.

2.2 Implicit Assessment of Reading Comprehension using Eye Movements

The relationship between eye movements and human cognition has been extensively studied in the past. Many cognitive processes are infeasible to observe or measure directly but are highly correlated with eye movements [28, 35]. As a result, much research has used eye movements data to infer or analyse people's cognitive behavior, such as people's attention [1, 9], reading behavior [8, 26], and reading comprehension [3, 14–16, 30, 46].

Eye-tracking research has shown great potential in classifying reading behaviors and predicting readers' comprehension from their gaze data. For instance, eye movements are correlated with comprehension [23, 24, 34, 41], and more specifically, fixation duration and saccade length are known to be correlated with reading behavior [45] and comprehension level [45, 46]. The line of work that uses eye movements to classify comprehension was started by Underwood et al Underwood et al. [46]. They studied the relationship between eye fixations and reading comprehension, which identified that fixation duration could be an indicator of comprehension. However, the generalizability of their results was limited as they trained and tested on the same dataset. Copeland et al. [14–16], instead of focusing on comprehensions, studied predicting the accuracy of the answers. Makowski et al. [30] continued tackling the comprehension prediction problem. While succeeding in reader identities identification, their approach failed at predicting comprehension levels. Moreover, their texts were in German with approximately 158 words each, and there were only three comprehension questions. Hence, their design study may not be suitable to observe deep and shallow readings. Recently, Ahn et al. [3] studied predicting people's comprehension level using gaze patterns as well as other reading parameters such as reading difficulty. The comprehension was labeled in binary as high and low, based on the percentage of correctly answered questions. For overall comprehensions, they obtained an accuracy of 64% (11% better than the null accuracy) on testing against unseen passages (for the same participant), but failed to beat the null model when testing against unseen participants. Although account for comprehension levels, their study did not consider participants' reading modes and verify them via question analysis, also their results were limited in terms of generalizability.

Very little research has been done in utilizing these correlated features to help classify deep and shallow reading without the help of comprehension questions. Moreover, all of the studies above were only conducted on Desktops. And to the best of our knowledge, no works have evaluated similar experiments for Mobile devices.

2.3 Implicit Assessment of Reading Behavior using Eye Movements

Our goal is to classify deep reading, where material is being thoughtfully processed, versus shallow or skim reading where the reader either not comprehending the text or skipping through it or scanning for a singular fact. In truth, modes of reading are not often entirely separate, sometimes skim or shallow reading will occur during deep reading and conversely occasionally a skimming or shallow reader will take interest in an article and begin reading more deeply.

Skim reading is a type of shallow reading where readers are skipping through a text or perhaps just searching for a particular fact. Bieder et al Biedert et al. [8] showed that eye tracking could be effective for identifying skim reading of paper news articles. In this study, two groups of readers were asked to produce a small set of keywords to describe the news articles. One group was given a very short time to read the article (skim reading induction) whereas the other cohort was given unlimited time to read the article (no skim pressure - deep reading assumed). Using eye saccade metrics these researchers were able to differentiate between the two conditions with 86% accuracy. In another study, Kelton et al Kelton et al. [26] refined this study and looked at identifying skim versus deep reading on both a global (entire text) and local (specific region) scale. They followed a similar study design to Bieder et al Biedert et al. [8] and achieved an accuracy of 82.5% differentiating global skim vs. deep conditions and $72\% - 95\%$ accuracy for differentiating these modes in different local areas of the document. While these studies do show that time pressured readers did not stop (measured via saccade) to read the majority of the document, this only proved a differentiation between skim versus assumed deep reading because reader comprehension was never assessed.

3 DATA COLLECTION

For our data collection, we factored in three main limitations found in previous studies and introduced a new study design to systematically induce deep reading and shallow reading behaviour on both Desktop and Mobile devices. Our study design differed from previous studies as follows: Firstly, we used reading materials that were much longer ($\approx$ 1500 words) compared to previous studies[8, 26] where they used news article ($\approx 150 words$), since it is easier to induce and evaluate deep reading in them. Further, we gave more time to the reading due to our text length. Secondly, we used a set of carefully designed comprehension questions to systematically evaluate different types of comprehension and reading mode. Lastly, to ensure shallow reading behavior, we replaced "keywords summarizing" techniques commonly used in previous studies [8, 26] with "finding answers to preview-questions". This was done to counter the limitations of previous works [8, 26] where researchers induced shallow reading behaviour by asking participants to select three keywords that best describes the article. However, we believe that keywords of the articles can be easily obtained via inferring from the title and a few paragraphs, which does not explicitly validate shallow reading behaviour.

3.1 Tasks

We prepared four reading tasks in total – one deep reading task on Mobile, one deep reading task on Desktop, one shallow reading task on Mobile, and one shallow reading task on Desktop. The deep reading and shallow reading tasks were designed as follows:

Deep Reading: The deep reading task's purpose was to make participants thoroughly process and comprehend the reading material. Participants were, therefore, asked to take as much time as they needed to read an article. They were explicitly told that they would have to answer 20 in-depth multiple-choice questions about the text's content. Additionally, we alerted participants to the fact that—once the questions were shown—they could not revisit the article. Participants were encouraged to try and get as many questions correct as they could.

Shallow Reading: The shallow reading task was designed to induce participants' shallow reading behavior, i.e., obtaining literal information in the shortest amount of time. In this task, the participants were provided with three questions about the text beforehand and the goal was to find the answers to these three questions using as little time as they could. In the preview stage, only the question text was displayed but not the text. Once they started reading, participants were unable to jump back to the questions. They were meant to act as primers only. Additionally, we imposed a two-minute limit on the reading time. After reading, participants were asked to answer the three previously provided questions plus another 17 multiple-choice to test whatelse they took away from the test. Participants were informed that there were 20 questions in total but only the three previewed ones matter, and they were unable to refer back the article while answering the questions.

3.2 Reading Materials

We selected four articles from the easyCBM repository [4], which we randomly assigned across our two-by-two study design: deep versus shallow and Desktop versus Mobile reading. All articles were stories with approximately 1500 words in length and of 8th-grade reading level. Each article comes with 20 multiple-choice comprehension questions: seven questions tested literal, seven questions inferential, and six questions tested evaluative comprehension. Table 1 and Table 2 lists details for the articles used.

Table 1. Readability Statistics of Selected Article. FK stands for "Flesch Kincaid", GF stands for "Gunning Fog", SMOG stands for "Simple Measure of Gobbledygook", CL stands for "Coleman Liau", and AR stands for "Automated Readability".

Article	FK Reading Ease	FK Grade Level	GF Score	SMOG Index	CL Index	AR Index
1	76.4	6.3	8.6	6.7	9.5	6.2
2	76.8	7.5	10.2	6.3	9	8.5
3	68.5	7.4	10	7.5	11	7.4
4	77.8	5.6	7.9	5.9	9.5	5.2

Table 2. Text Statistics of Selected Article. Sent. stands for "Sentences"

Article	Sent.	Words	Complex Words	Complex Words %	Words per Sent.	Syllables per Word
1	131	1943	169	8.70%	14.83	1.36
2	79	1584	88	5.56%	20.05	1.30
3	99	1461	160	10.95%	14.76	1.46
4	138	1749	132	7.55%	12.67	1.37

The reading materials were displayed in alignment with previous readability studies [17, 32, 33, 37, 38] to best enhance readabilities (see Figure 1). On Desktop, we chose an even spaced sans serif font, Arial, and set the line width to be 600 px so that each line contained roughly 55 characters, which was shown by Dyson [17] to be most readable. In addition, we used 24 px font size and double spacing to improve both readabilities [38] and the accuracy of the eye tracker. All texts were left justified. While on Mobile, we kept the same font family, line space, and justification, the font size was made smaller to ensure the same number of characters were shown on each line. Texts navigation worked by vertically scrolling as this interaction is the de-facto standard in electronic interfaces now.

(a) Desktop Display Example (b) Mobile Display Example

Fig. 1. Examples of Desktop and Mobile Displays.

3.3 Experimental Setup

We recorded participants' eye movement data using the Tobii Pro X3-130 eye tracker[2] and the Tobii Pro Studio software. The eye tracker had a sampling rate of 120 Hz. The output data contains raw gaze coordinates (x, y), left and right pupil diameter, as well as the fixation and saccade information generated by the Pro Studio. The eye-tracker was mounted at the bottom of a 24-inch monitor for Desktop reading, and at the bottom of an iPhone 11 device for Mobile reading (see Figure 2). Both devices kept a log of participants' scrolling behavior, i.e., the x and y offset of the current viewing page.

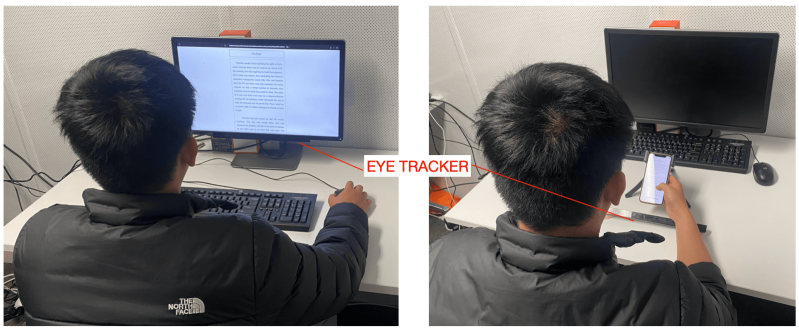


Fig. 2. Experiment Setup. Left: Desktop reading; Right: Mobile reading

3.4 Participants

We recruited 36 participants (17 women, 13 men), all of them were students and staffs from our university. The data from six participants was discarded due to insufficient reading, i.e., not being able to go through at least half of the article because they hit the two-minute time limit during the shallow reading task. The remaining 30 participants were all native or bilingual English speakers. Most participants had an education level of postgraduate degree (N=15), followed by bachelor degree (N=13), while the remaining were graduate certificate or diploma (N=1) and certificate III or IV (N=1).

[2]https://www.tobiipro.com/product-listing/tobii-pro-x3-120/

3.5 Procedure

Upon arrival, we informed participants about the purpose and the procedure of this study and asked them to sign a consent form. Next, we seated participants in a comfortable position and asked them to adjust the seat such that their heads were centered and approximately 60 – 65 cm away from the Desktop screen (and 50 cm from the Mobile phone screen). Before each reading task, a standard, built-in, 9-point eye-tracker calibration procedure was done. To start the study, participants were asked to fill in a pre-study survey regarding their demographics and past reading experiences. Then, they finished the Desktop reading task and the Mobile reading task in counterbalanced order. On each device, participants were asked to read two articles in counterbalanced order, one in deep reading mode and the other in shallow reading mode. The four articles were randomly assigned to each reading session. During each reading session, participants were instructed to use the mouse (Desktop) or swipe on the screen (Mobile) to scroll down. Further, while reading, participants were free to move their heads slightly as suggested by the robustness of the eye tracker. Lastly, participants were asked to finish a post-study survey about their experiences in this study. The study took around 45 – 60 minutes to finish and upon completion, participants were compensated with a 10 dollar coupon.

4 READING MODE DETECTION

In this section, we describe our step-by-step approach for detecting deep versus shallow reading. To begin with, we pre-processed the gaze-data by adjusting the raw gaze data with the scrolling behavior and converting it to fixations and saccades. Next, we extracted 50 low-level and 25 mid-level [44] gaze features from each sliding window of the processed data. Afterwards, based on the extracted features, we built and test various machine learning classifiers. Also, we tested the impact of different window and step sizes. Lastly, we selected the best-performed model and analyzed its performance in more detail.

4.1 Data Preprocessing

Although the Tobii Pro Studio software can pre-process the raw gaze-date to extract fixations, it does not consider the effect of scrolling events. For example, participants may stare at one point while scrolling. Therefore, we first incorporated the scrolling event into the raw gaze data, such that each gaze point is adjusted by the offset of the current page. Then, we generated fixations from the adjusted gaze data using the R software package *Saccades* [48] that uses the velocity-based algorithm for saccade detection proposed by Engbert and Kliegl [18]. Next, we identified saccades as transitions between consecutive fixations.

For most participants, deep reading took longer to complete than shallow reading. Hence, classification of the whole session would be trivial as the duration and fixation number themselves were already strong indicators. In order to evaluate the robustness of our classifiers, we used the sliding-window technique to partition our time-series gaze data into consecutive windows. The window size was fixed and all data points within this window were aggregated to generate features. We defined the start time difference between two consecutive windows as the step size. Note that if the step size equals window size, then there is no overlap between windows. Further, window size and step size may have a huge impact on classifications. For example, a smaller window size could give more fine-grained results but may not able to capture sufficient movement patterns. Previous studies [10, 27, 29] suggested different window sizes from 20 seconds to 60 seconds. Hence, in this study, we varied the window size and step size to measure their impact on performance.

4.2 Feature Extraction

To build the classifier based on preprocessed fixation and saccade data, we required a feature set that best captures the reading behavior. Usually, gaze features can be categorized into three levels – low-level, mid-level, and high-level [44]. Low-level gaze features are features that can be derived directly from the raw data, such as fixation duration, saccade length, and pupil diameter. Many similar studies have used mainly low-level features as their feature set [22, 25]. Mid-level gaze features were proposed by Srivastava et al. [44] to better capture various eye movement patterns. Mid-level gaze features consider several consecutive fixations and saccades together, and categorize their shape into patterns such as *compare*, *scan*, and *line reading*. High-level gaze features, on the other hand, are stimuli-specific features whose analysis normally requires the identification of an area of interest (AOI). For example, texts that cover a question can be treated as one AOI and the time spent in this AOI can be considered as one feature. In this study, we included 50 low-level features and 25 mid-level features as shown in Table 3 and Table 4. We did not consider high-level features as they were hard to extract and to be generalized, as AOI could vary from text to text.

Table 3. Selected Low-level Features

Category	Features	Measurement
Fixation Duration	number, mean, std, var, min, max	value
	rate, percentage, slope	value
	short, medium, long	count
	dispersion area (75%)	value
Saccade Length	mean, std, var	value
	follow, neighbor, opposite, opposite-neighbor	count
	right, left, up, down, up-right, down-right, up-left, down-left	short, medium, long count
Pupil Diameter	mean, std, var	left and right value

Table 4. Selected Mid-level Features. All features are measured by their count.

Category	Sub-category	Features
Shape Based	String	up, right, down, left
	Line	small, medium, long
	Comparison	left-right, right-left, up-down, down-up
	Scan	right-left, left-right, up-down, down-up, right-up, up-right, left-up, up-left, right-down, down-right, left-down, down-left
Distance Based	-	regression, else-where

For all fixations, saccades, and pupil diameters in each data chunk (i.e., a window), we calculated metrics, such as count, mean, standard deviation, variance, minimum value, and maximum value. Also, we derived the rate, percentage, and slope of all fixations. Fixation rate was defined as the ratio between the total fixation duration and the data duration (i.e., window size). Fixation percentage was the ratio between the count of fixations and the data duration. And fixation slope was the slope of the regression line fitted on all fixations. For fixation duration size, we set the thresholds as 200 ms and 400 ms, i.e., fixations within 200 ms were considered as short fixations, fixations between 200 ms and 400 ms

were medium fixations, and long fixations had a duration of at least 400 ms. For calculating fixation dispersion, we chose our dispersion area to be 75%, i.e., the area containing 75% fixations. The threshold of saccade size varies on devices: saccades were considered short if their length is less than 200 px on Desktop or less than 75 px on Mobile, long if length greater than 400 px (Desktop) or 175 px (Mobile), otherwise medium. The choice of such thresholds was based on the line width, short saccades went through less than 1/4 of the line, while long saccades took approximately half of the line. Moreover, saccades were categorized into eight directions. For instance, a saccade was in up-direction if it was pointed towards up-direction and the angle between it and the vertical line was at most 22.5 degrees. Two consecutive saccades were counted as one follow if they had the same direction. Similar reasoning could be applied to neighbor, opposite, and opposite-neighbor features.

Our mid-level features aligned with Srivastava et al. [44], except we chose a small line to be three right saccades (of any size) followed by one left long saccade, i.e., ArArArLl, where Ar stands for a right saccade of any length, and Ll stands for a left long saccade. Similarly, we defined a medium line to be four right followed by one long left, i.e., ArArArArLl, and a long line to be five right followed by one long left, i.e., ArArArArArLl.

4.3 Classification

Each window was labelled by its reading condition, i.e., deep versus shallow. We chose our baseline classifier as the majority vote as our data was imbalanced (80% deep versus 20% shallow). We selected Kernel Support Vector Machines (SVM), Logistic Regression, Random Forest, and XGBoost as our classifiers. It is known that hyper-parameter tuning (including model parameters, window size, etc.) should not be done on the same set that is used in the final evaluation, otherwise the results would be biased towards tuning. One commonly used approach to solve this problem is to divide the dataset into train, test, and validation sets, but due to the small size of our dataset, this would significantly reduce the available training data and hence the performances of our models. Therefore, we adopted an improved cross-validation approach called nested cross-validation [13]. In nested cross-validation, the dataset is firstly divided into k parts, such that in each round, one part is picked as the testing set while the rest are picked as the training set. Furthermore, before evaluating on the test set, a nested k'-cross-validation is performed within the training set for hyper-parameter tuning. Finally, the best hyper-parameter will be chosen to evaluate the test set. By doing this, the hyper-parameter tuning process does not have access to or is not exposed to the test set, hence the result is less biased. To evaluate the performance of each model, we used the nested cross-validation with $k = 30$ and $k' = 10$. So essentially a leave-one-participant-out validation was performed in the outer loop and a 10-cross validation was performed in the inner loop for hyper-parameter tuning. We reported accuracy, F1 score, and area under the curve (AUC) as our metrics.

During the hyper-parameter tuning, we varied the window size from 5 to 120 seconds with a step of 5 seconds. We also varied the step size from 10, 25, 50, to 100 percentage of the current window size. For each training set, the best window size and step were chosen to test the test set. Since our dataset is imbalanced, better accuracy can be obtained with the majority guesser when increasing the window size and decreasing the step size. Hence, we chose the best sizes based on AUC. Ties were broken by macro-F1 score first and then accuracy. We set the model hyper-parameters based on suggested default values and previous similar studies so that they were not tuned within the cross-validation. For Kernel SVM, we chose RBF kernel with $C = 10$ and gamma equals 0.01. For Logistic Regression, we set C to be 1 and penalty to be l_2. For Random Forest, we set the number of estimators to be 1000, the maximum feature parameter to be the square root of feature count, and the minimum number of samples required for internal node splitting as 2, for leaf splitting as 1. For XGBoost, we set the number of estimators to be 100, the maximum depth to be 3, gamma to be 0, subsample and subsample ratio of columns to be 1, regularization alpha to be 0, and learning rate to be 0.1.

In addition to mode classification, we also examined the possibility of identifying reading devices, i.e., which device the participant was reading at. This task would be trivial with saccade-related features, due to the huge size difference between Desktop and Mobile. Hence, we focused on using fixation and pupil features alone.

Lastly, we built models for cross-device reading mode classification, where we trained the models using data from one device, and tested the models using data from the other device.

5 RESULTS

In this section, we first check the validity of our study design, i.e., whether our task design successfully induced deep and shallow reading. Then, we present our results for the reading mode classification, where we demonstrate that our classification methods can effectively differentiate between deep reading and shallow reading. Furthermore, we analyze the impact of window size, step size, and mid-level gaze features on our classifiers. We also gained more insights by looking into important features for classification and misclassified instances. Lastly, we explored the role of gaze features in some other classification tasks, such as reading device classification, and cross-device mode classification.

5.1 Comprehension Score and Reading Speed

Since the deep and shallow reading behaviors were induced by our study design, we started by validating it using participants' comprehension results and reading behaviors. As shown below, during deep reading sessions, most participants read at a speed that prioritized comprehension and were thus able to achieve a high comprehension scores for all three comprehension levels (i.e., literal, inferential, and evaluative). While in shallow reading, participants read faster (i.e., performing skimming or scanning) but still obtained high-level literal comprehension. However, their inferential comprehension was significantly poorer in shallow reading. Therefore, we concluded that our study design had successfully induced the expected behavior.

Figure 3 shows the total number of correct answers across all participants. In both devices, participants had significantly better comprehension results in deep reading (Desktop $M = 14.89, SD = 1.72$; Mobile $M = 14.08, SD = 2.49$) than shallow reading (Desktop $M = 12.47, SD = 2.88$; Mobile $M = 11.11, SD = 2.23$). Paired t-test revealed statistically significant differences between the conditions for Desktop ($t(35) = 4.20, p < 0.001$) as well as Mobile ($t(35) = 5.18, p < 0.001$). Also, when reading shallowly, participants had significant better comprehension on Desktop ($M = 12.47, SD = 2.88$) than Mobile ($M = 11.11, SD = 2.23$) (paired t-test $t(35) = 2.37, p = 0.02$) . However, there is no such difference could be observed when reading deeply (paired t-test $t(35) = 1.80, p = 0.08$) .

Furthermore, we looked into the comprehension results for each question type, i.e., literal, inferential, and evaluative. The literal questions were further divided into two sets, the set of questions that were previewed to participants (Previewed Lit Q), and those that weren't. Figure 4 shows the comprehension results for each question type. When comparing deep versus shallow reading on each device, participants showed significantly better results for non-previewed literal questions (paired t-test: Desktop $t(35) = 3.01, p < 0.01$; Mobile $t(35) = 4.00, p < 0.01$) and inferential questions (paired t-test: Desktop $t(35) = 6.03, p < 0.001$; Mobile $t(35) = 3.20, p < 0.005$), but not for previewed literal questions (paired t-test $t(35) = 0.81, p > 0.1$) and evaluative questions (paired t-test $t(35) = 0.53, p > 0.1$) on Desktop. When comparing Desktop and Mobile reading, participants obtained better comprehension for evaluative questions in shallow reading (paired t-test $t(35) = 2.39, p < 0.05$) and inferential questions in deep reading (paired t-test $t(35) = 2.20, p < 0.05$), but not for others.

In addition, we investigated the reading speed of each participant, we found participants read at a much faster speed in shallow reading (paired t-test $t(35) = 9.29, p < 0.001$). As shown in Figure 5, most participants read at the speed

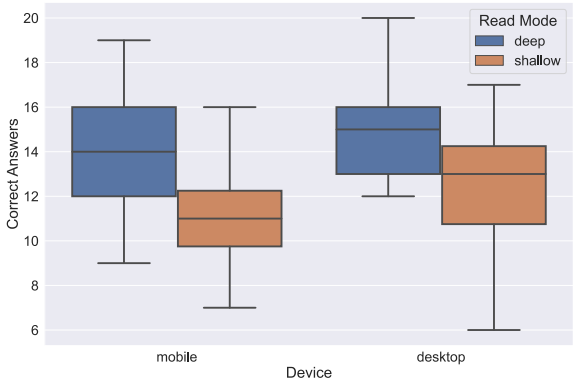


Fig. 3. Comprehension Results of All Participants

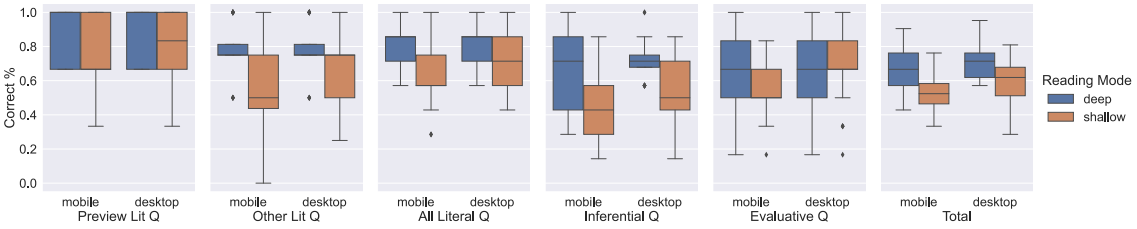


Fig. 4. Comprehension Results with Question Type Breakdown

normally considered as reading for comprehension (i.e., 300 words per minute [12]) (Desktop $M = 296.58, SD = 118.38$; Mobile $M = 306.37, SD = 131.74$) during the deep reading session. However, in shallow reading, most participants read at the speed normally considered as skimming (i.e., 450 words per minute [12]) and scanning (600 words per minute [12]) (Desktop $M = 602.93, SD = 221.38$; Mobile $M = 679.35, SD = 326.73$).

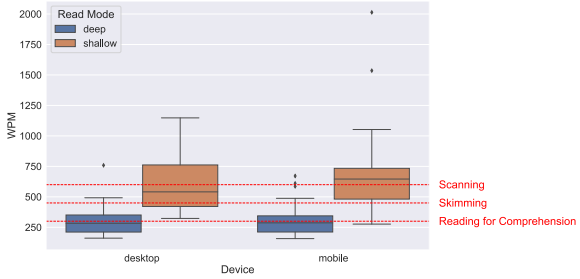
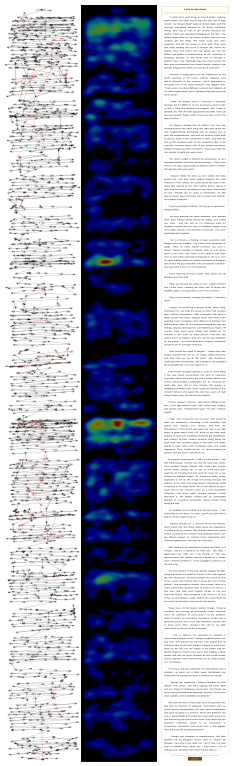


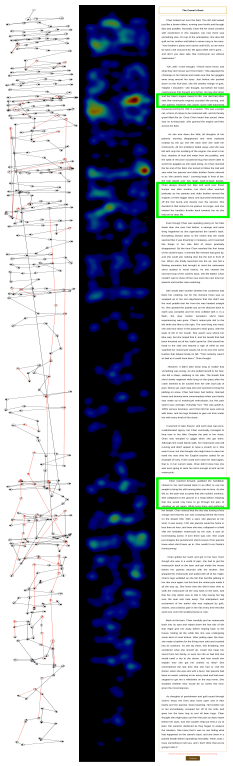
Fig. 5. Reading Speed of All Participants.

Two typical examples of deep and shallow readings were shown in Figure 6. We plotted the gaze map (i.e., fixations and saccades) and heatmap overlaid over the reading articles for each condition. As shown in Figure 6(a), when reading deeply, there were more fixations and more short saccades, hence the gaze plot looks relatively dense. On the other hand, when reading shallowly (Figure 6(b)), there were fewer fixations but more long saccades, also the regressions appeared

more frequently. It is worth noticing that as reflected by the shallow heatmap, participants paid much attention on the AOIs of previewed questions. While in deep readings, participants did have particularly focused regions but the overall distribution was more spread out. Another interesting finding is that for each paragraph in deep reading, participants usually put more effort into reading at the start of the paragraph than at the end of the paragraph.



(a) Deep Reading Example (p11, Desktop, deep)



(b) Shallow Reading Example (p11, Desktop, shallow)

Fig. 6. Examples of Deep and Shallow Readings. From left to right: gaze movement plot, gaze heatmap, and reading article. In shallow reading, the AOIs of previewed questions were marked with green boxes. Regressions were marked as red lines.

5.2 Reading Mode Classification

5.2.1 Classification Results. The performances of all models are summarized in Table 5, where we used bold text to mark the best classifier regarding one evaluation metric. Note that our baseline, a majority guesser, achieved a relatively high accuracy score due to the imbalanced data from larger number of windows from the longer deep reading sessions. This imbalance makes accuracy less reflective of performance than the AUC and F1 metrics.

All models outperformed our baseline approach, i.e., the majority guesser. Among all the models, XGBoost achieved the best performance in Desktop with an accuracy of 88.36%, a F1-score of 0.82, and an AUC of 0.82. In Mobile, Logistic Regression had the highest accuracy (86.18%) and F1-score (0.77). Meanwhile, the XGBoost had similar accuracy and F1-score, and it achieved the highest AUC score (0.73). The classifiers typically chose window sizes from 60 to 120 seconds, as well as a step size of 100% during the optimization. During training we also tried variants that balanced the

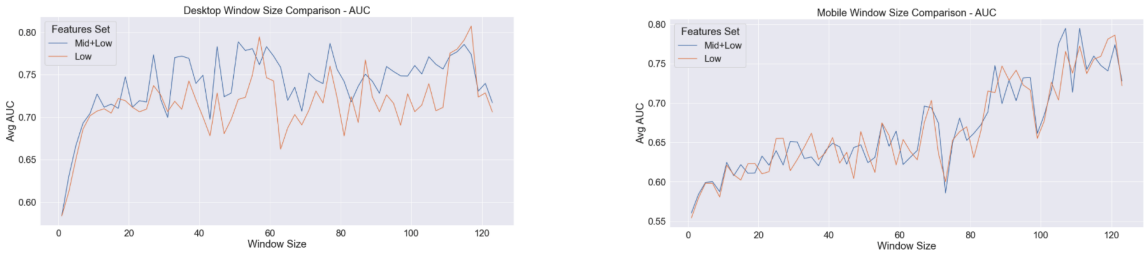
Table 5. Reading Mode Classification Results

Device	Model	Accuracy	Macro F1 Score	AUC
Desktop	Baseline (Majority Guessing)	0.8551	0.4610	0.5
	Linear SVM	0.8366	0.7501	0.7598
	RBF SVM	0.8824	0.8098	0.7811
	Logistic Regression	0.8797	0.8190	0.8042
	Random Forest	0.8623	0.7649	0.7392
	XGBoost	**0.8836**	**0.8239**	**0.8201**
Mobile	Baseline (Majority Guessing)	0.8081	0.4469	0.5
	Linear SVM	0.8320	0.7414	0.7222
	RBF SVM	0.8279	0.6922	0.6617
	Logistic Regression	**0.8618**	**0.7688**	0.7307
	Random Forest	0.8455	0.7416	0.7082
	XGBoost	0.8551	0.7558	**0.7311**

training data via down-sampling the majority class or up-sampling the minority class to account for the imbalance. However, both methods failed to improve our model in terms of our evaluation metrics.

5.2.2 Window Size and Step Size. To better understand the effect of step size and window size on the classifier, we examined them in detail for the best-performed model, XGBoost. We first varied the window size from 1 to 120 seconds fixing the step size to be equal to the windows size. Next we, took the best window size, and then varied the step size from 1 second to the best window size. For simplicity, we plotted only the accuracy and AUC scores, the F1 score graphs were very similar to the AUC graphs.

Figure 7 summarized the classification results of XGBoost model when varying the window size for window extraction. For Desktop, the model achieved a relatively good performance with any window size greater than 20 seconds. When we decreased the window size from 10 to 1, performance dropped significantly, indicating a small window size failed to capture any meaningful features. Meanwhile, window sizes had a much bigger influence on Mobile. Performance increased as we increased the window size from 1 to 120. Hence a bigger window size might be required for effective Mobile classification. Moreover, Desktop outperformed Mobile on window sizes below 80 seconds, and they performed roughly the same with larger window sizes.

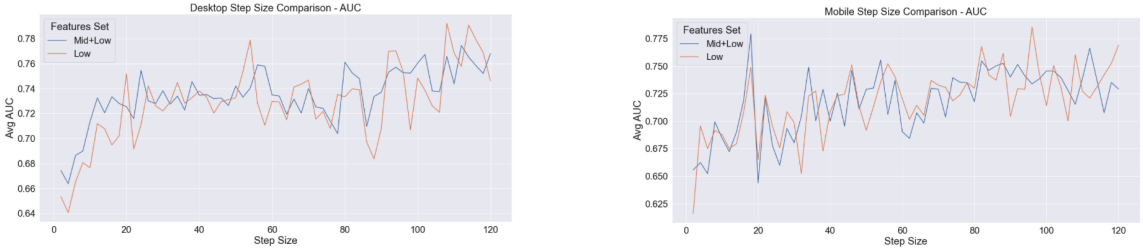


(a) Desktop Window Size Evaluation – AUC.

(b) Mobile Window Size Evaluation – AUC.

Fig. 7. Effect of Window Size on Classifier Performance

According to the result of our window size evaluation, we fixed the window size to be 120 seconds and varied the step size to measure its influence. Our results for step size comparison was summarized in Figure 8, where we still used XGBoost as our classifier. Similar trends were observed in both Desktop and Mobile. When increasing the step size from 1 second to 120 seconds, performance of both devices increased. Adding more data via overlapping windows does not help result in a better model.



(a) Desktop Step Size Evaluation – AUC.

(b) Mobile Step Size Evaluation – AUC.

Fig. 8. Effect of Step Size on Classifier Performance

5.2.3 Important Features. We further analyzed the important features outputted by XGBoost. Table 6 listed the top-ten features, which were chosen based on the proportion of samples they could split with highly correlated features being removed. When the window size is small (i.e., 5 seconds), fixation and pupil features were more important, showing that statistics about fixation duration and pupil diameter was crucial in differentiating deep and shallow reading. One possible explanation is that the window size was too small to pick up any meaningful movements, hence the model mainly used statistics features. As the window size increased, more and more saccade-related features were identified as important. When window size equals 60 seconds, more than half of the top ten features were saccade features.

Table 6. Top 10 Important Features in Mode Classification at Different Window Sizes. In pupil diameters, we used (L) for left eye and (R) for right eye. And we omitted "saccades" in saccade-related features, such as Up for Up saccades and Left for left saccades. Also, in fixation and saccade features, we use (S) for short, (M) for Medium, and (L) for Long.

Rank	5 Sec Window		15 Sec Window		60 Sec Window		120 Sec Window	
	Desktop	Mobile	Desktop	Mobile	Desktop	Mobile	Desktop	Mobile
1	Else-where	Fix slope	Else-where	Fix var	Else-where	Up-right (M)	Else-where	Up-right (M)
2	Pup mean (L)	Pup mean (L)	Pup mean (L)	Fix max	Up-left (M)	Up-left (M)	Up-left (M)	Fix slope
3	Pup mean (R)	Pup mean (R)	Sacc mean	Fix slope	Up-right (M)	Up (M)	Up (M)	Up (M)
4	Sacc mean	Sacc mean	Pup mean (R)	Pup mean (L)	Left (M)	Regression	Left (M)	Up-left (M)
5	Pup var (L)	Fix var	Left (M)	Pup mean (R)	Up (M)	Long Fix	Neighbor	Left (M)
6	Pup std (L)	Fix max	Fix var	Up-right (M)	Oppo-neigh	Fix min	Up-right (M)	Long Fix
7	Fix slope	Fix std	Fix max	Sacc mean	Fix max	Fix var	Oppo-neigh	Regression
8	Fix var	Fix min	Fix min	Fix std	Fix var	Fix slope	Fix var	Right (M)
9	Fix max	Fix mean	Fix std	Fix min	Med fix	Fix max	Fix std	Fix min
10	Fix rate	Pup std (R)	Pup var (L)	Else-where	Neighbor	Left (M)	Fix max	Fix var

Among them, features that are characteristic of "backtracking" were valued more, this includes medium-length saccades towards up, up-right, and up-left. Fixation duration features remained relatively important in large window

sizes, but pupil diameter features became less important. Moreover, the classification process relied heavily on distance-based mid-level gaze features such as else-where patterns and regression patterns, but not shape-related features. In addition, there are more important fixation duration features in Mobile reading than Desktop readings, meaning they are more important in Mobile classification. This might be because the screen size of Mobile is much smaller than Desktop, hence it is harder to identify and extract useful saccade patterns.

5.2.4 Example Instances. Besides the important features, we also looked into each instance (i.e., each window for each participant) and the classification results of XGBoost with a window size and a step size of 60 seconds. Figure 9 shows the typical correctly classified and misclassified instances. Figure 9(a) and Figure 9(b) gave the correctly classified deep and shallow reading window. Figure 9(c) showed one of our two misclassified deep windows, which occured when participant 14 skimmed the article after read it thoroughly. Figure 9(d) demonstrated one of the misclassified shallow windows, and as shown in the figure, participant 25 were actually reading at a very slow speed which appears to be more like deep reading itself.

(a) Correctly Classified Deep Instance (p34, Desktop, deep, window 1)

(b) Correctly Classified Shallow Instance (p34, Desktop, shallow, window 1)

(c) Misclassified Deep Instance (p14, Desktop, deep, window 10)

(d) Misclassified Shallow Instance (p25, Desktop, shallow, window 1)

Fig. 9. Classification Results Example Instances

5.3 Other Classification Tasks

5.3.1 Reading Device Classification. In device classification, we checked if we can predict the reading devices using various feature set. The window size and step size were fixed to 60 seconds, and we varied the feature set as fixation-related features only, fixation and pupil features, all low-level features, and low-level plus mid-level features. As shown in Table 7, this classification task was trivial when saccade-related features were allowed. The best classifier, Random Forest, achieved an accuracy of 100% with low-level features and low-level plus mid-level features. When using fixation features alone, we could still obtain 69.68% accuracy with Random Forest. Moreover, the classification accuracy could be boosted to 84.16% if pupil features were allowed and XGBoost was used for classification. The usefulness of pupil features could be reflected in Table 8, pupil features were the top two, 7th, and 8th important features.

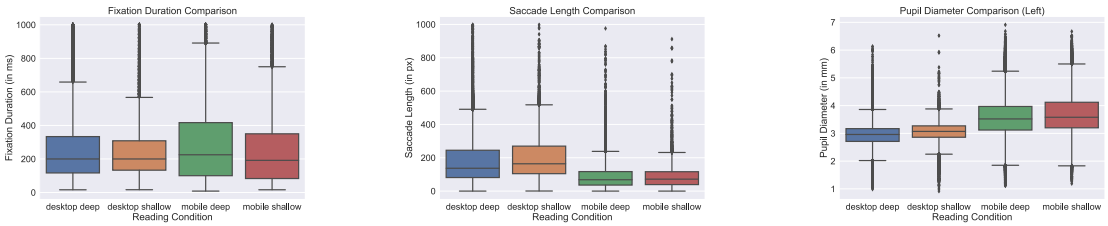
Table 7. Device Classification Accuracy.

Model	Fixation Only	Fixation + Pupil	Low Level	Low + Mid Level
Baseline (Majority Guessing)	0.5111	0.5111	0.5111	0.5111
RBF SVM	0.6610	0.6525	0.9949	0.9966
Logistic Regression	0.6678	0.6746	0.9915	0.9915
Random Forest	**0.6968**	0.8262	**1.0000**	**1.0000**
XGBoost	0.6934	**0.8416**	0.9949	0.9966

Table 8. Top 10 Important Features in Device Classification.

Rank	Fix Only	Fix + Pupil	Low level	Low + Mid level
1	Fix slope	Pup mean (L)	Right (M)	Right (M)
2	Med Fix	Pup mean (R)	Sacc var	Else-where
3	Fix std	Fix max	Sacc std	Sacc var
4	Fix max	Fix var	Left (M)	Left (M)
5	Fix var	Med fix	Sacc mean	Sacc std
6	Fix rate	Fix slope	Left (L)	Sacc mean
7	Fix mean	Pup std (R)	Right (L)	Left (L)
8	Fix min	Pup var (R)	Down-right (M)	Right (L)
9	Short fix	Fix min	Left (S)	Down-right (M)
10	Fix per	Fix mean	Opposite	Up-left (M)

5.3.2 Cross-device Mode Classification. In inter-device mode classification, we failed to obtain a model that significantly outperformed the baseline approach. However, this is expected and can be explained by the inherent different between devices. Figure 10 demonstrated the statistics summary of fixation duration, pupil diameter, and saccade length in all four conditions. For the same reading device, all features exhibited some differences between deep and shallow reading. For the same reading mode, there were a huge difference between Desktop and Mobile in all features. These difference explained why our device and mode classification succeed, and inter-device mode classification failed.



(a) Fixation Duration Comparison. (b) Saccade Length Comparison. (c) Pupil Diameter Comparison.

Fig. 10. Fixation, Saccades, and Pupil Data Summary

6 DISCUSSION

To advance reading research we need to be able to track reading behaviours in a scalable and non-intrusive manner. To develop implicit metrics about the quality of reading, we developed a method to reliably induce shallow and deep reading using three levels of comprehension. Since reading increasingly takes place electronically [40], we developed classifiers for both Desktop and Mobile devices.

6.1 Inducing and Validating Deep vs. Shallow Reading

Our evaluation showed that inferential comprehension is most prominently linked to deep reading activities regardless of reading device as it requires readers to empathise with the context of a story in order to derive insights not explicitly stated in the text. Inferential comprehension is essential for passage understanding as it requires readers to connect events in a narrative or understand a character's motive for a particular action [5]. The tasks we developed to induce deep and shallow reading were most effective in their discriminative power with regard to this crucial level of comprehension.

Readers' ability to answer non-previewed literal and inferential questions was further significantly impeded by shallow reading on both Desktop and Mobile. For previewed and evaluative questions we did, however, not see a difference between reading modes, which may imply that shallow readers have not just blindly skipped text but applied shallow reading as a selective form of text processing.

When reading on Mobile devices, our results show better comprehension for evaluative questions in shallow reading mode Especially high-level questions like "What kind of person is Kim?" could be answered well during shallow reading. This may indicate that readers have successfully adapted their reading ability to quickly determine how new information fit into their existing knowledge network and make use of prior experiences to quickly judge a passage or character.

Our method has been shown to effectively induce shallow and deep reading and thus can be reused by researchers to further the research on these two reading modes. The reading materials can be found in the Appendix along with the comprehension questions and their groupings into three levels. We further open-sourced our reading interface, which is listed in our Git repository[3].

6.2 Linking Deep Reading to Eye Gaze Pattern

By combining participants' comprehension results and their reading speed with eye movement data we were able to determine eye gaze patterns that are characteristic of deep reading activities. In general, participants read slowly at a speed for comprehension in deep reading, and relatively fast when doing shallow reading. By looking at the gaze

[3]link anonymized

plot and heatmap of each participant, we found that participants tend to read line by line carefully in deep reading, i.e., there were much more short saccades and less long saccades, and the overall reading patterns looked more like scanning slowly from left to right, line by line. While in shallow reading, we found that participants jump quickly between paragraphs and only spent relatively more time on texts related to previewed questions. Their reading patterns were more like scanning in a zig-zag manner with longer saccades and less fixations. In addition, these features played an important role in the later classification. We found that, classifiers saw a task as deep reading if it had less upwards and left saccades (i.e., backtracking-like saccades), more and longer fixations, and smaller pupil diameter. The bigger the window size, the more important saccade-related features were. Desktop classifiers utilized saccade features more, as it were easier to pick up on larger screen size. Moreover, distance-based mid-level features were also considered by the classifiers as they identified regression patterns.

6.3 Reading Mode Classification

Our results demonstrate the existence of a strong relationship between eye movements and reading mode. This allowed us to build classifiers to distinguish deep from shallow reading. One of the central contributions of our work is to provide reliable classifiers for reading on Desktop as well as on Mobile devices.

Our classifiers achieved high performance in recognizing deep versus shallow reading in general settings. In Desktop classification, among all four classifiers, the XGBoost achieved an accuracy of 88.36%, a F1 score of 0.82 and an AUC of 0.82, which outperformed our baseline measurement, majority guesser (accuracy = 85.51%, F1 score = 0.46 and AUC = 0.50), by the most. Note our baseline approach achieved a high accuracy due to imbalanced dataset. In Mobile, Logistic Regression performed the best with accuracy = 86.18%, F1 score = 0.77 and AUC = 0.73. XGBoost achieved a similar performance with accuracy = 85.51%, F1 score = 0.76 and AUC = 0.73. The majority guesser had accuracy = 80.81%, F1 score = 0.48, and AUC = 0.50.

Misclassifications may be due to the large variance among individuals' reading behaviors, which has been widely reported about [36], and the natural mix of deep and shallow reading in daily reading activities. This was verified by looking at each individual misclassified instance. During deep reading session, some participants (e.g., participant 14 and 32) still performed a "revision" after the reading the entire article once, and the "revision" was more like skimming and scanning than deeply reading. Similarly, when reading shallowly, some participants (e.g., participant 22, 25, 36, etc.) were not able to perform shallow reading due to their deep engagement with the presented stories, hence they still read slowly and deeply. Moreover, we the performance difference between Desktop and Mobile may be exacerbated by the smaller screen size and closer distance to the eye tracker. Hence the abilities of both our eye tracking hardware (i.e., eye tracker's accuracy in terms of correctly identifying fixations and saccades) and software (i.e., correctly picking up useful movements by feature extraction) were limited. To further improve performance on Mobile, more fine-grained eye-tracking technique and Mobile-tailored feature set were required. Moreover, we experimented with various window size and step size and found that a size of 120 seconds is the most suitable one for both. This aligns with previous studies [10, 27, 29] indicating that a large window size is required for such high-level activity recognition. A shorter window size may not be sufficient to capture important movements. The claim on shorter window was confirmed by important features outputted by XGBoost – when we increased the window size from 5 to 60 seconds, fixation and pupil features were out-weighted by saccades and mid-level features, indicating the presence of movement-related features and their importance to our model. Our results on step size indicated that adding more data via sliding window overlapping did not work well. To obtain a more robust and accurate, more data needs to be collected and analyzed. In addition, we found that statistics about pupil diameter size played an important role in our model, which was not

often considered in previous works. Also, we found that less saccade features were considered as crucial in Mobile. This aligns with our suspicion that our hardware and software ability was limited on the Mobile device.

Additionally, our results on device classification showed that we could tell the reading device effectively (accuracy 69.68%) by using fixation features along. It was mainly because fixation duration was shorter in Desktop reading. One possible explanation for this is the text size. In order to keep the same amount of characters on each line, the text size on Mobile was smaller than Desktop. Previous studies revealed that the smaller the font size, the longer the fixation duration [7, 37]. Hence it was expected that fixation duration could help identify deep and shallow reading. With the help of pupil features, the performance of our device classifier could be boosted to 84.16%. The pupil features were so useful because the pupil diameter was larger and had larger variance in Mobile. This might be because of the text size as well as screen settings such as brightness. Further study was required to examine it. The task became trivial when considering saccade features, due to the huge difference in screen size between Desktop and Mobile.

6.4 Limitations

First of all, as an in-lab study participants read under controlled environmental and device conditions. Additionally, we imposed several soft constraints on participants for accurate data recording (e.g., do not move head too much). Hence, a lot of varying environmental influences (e.g., distractions, differences in lighting, etc.) encountered during everyday reading activities are not reflected in our data. The stationary set-up of using a Tobii eye tracker under controlled condition was a necessary first step to validate the successful induction of deep and shallow reading as well as link deep reading activities to eye gaze patterns. The resulting classifiers build the basis for building more robust and Mobile tracking solutions, which we also plan to expand to other sensor types (e.g., eye tracking through the front-facing rgb camera).

Moreover, in our study design we limited the shallow reading time to two minutes, which may have caused changes in reading behavior due to the explicit time pressure imposed. The high scores in literal comprehension, however, indicate that participants were able to successfully undertake the reading tasks.

Regarding our feature selection, we suspect that a better set could be obtained if we tailored our features specifically for reading task, such as developing more appropriate mid-level and high-level features, which we left for future work. Moreover, since a large window size was required for accurate classification, our current approach is more appropriate for post-hoc than real-time classification, which we deem feasible considering our goal of tracking reading behaviour changes long-term.

6.5 Future Directions

In future work, we will explore more accurate and less constrained methods for classifying reading behaviour in-the-wild. By improving our classifiers through richer mid-level features along with collecting larger datasets, we intend to build more robust classifiers that run with alternative sensors, such as an rgb camera or Apple's true-depth camera. This would allow us to take the tracking of deep reading episodes into the field by deploying our models on consumer devices.

By providing a method to induce deep and shallow reading activities along with the link between eye gaze patterns and deep reading, our work paves the way towards implicitly tracking reading behaviour over time. Follow-up studies that we can now undertake include investigating the effect of reading content, layout, or style on reading behaviour in natural settings. For example, are users conditioned to various reading behaviours based on the website they visit (i.e.

social media vs news feeds vs blogs)? Similarly, with robust classification we can study different settings where reading behaviours may be influenced by contextual factors, such as while at work, home, or commuting.

7 CONCLUSIONS

Understanding reading behaviour on Desktop and Mobile devices through eye-movement is important to foster and understand deep reading in natural settings. Our study is the first to perform a detailed investigation into reading modes for both Desktop and Mobile devices. This was enabled through a study design we introduced that could successfully induce people's deep and shallow reading behavior as evidenced through comprehension results and reading speed. Moreover, we showed that these two reading modes could be effectively identified with classifiers trained on features derived from eye-movement data. Finally, we discussed key characteristics of deep and shallow readings through the analysis of the most important features for our classifiers. Our work paves the way to study long-term changes in reading behaviour through implicit metrics in natural settings.

ACKNOWLEDGMENTS

We acknowledge ...

REFERENCES

[1] Yomna Abdelrahman, Anam Ahmad Khan, Joshua Newn, Eduardo Velloso, Sherine Ashraf Safwat, James Bailey, Andreas Bulling, Frank Vetere, and Albrecht Schmidt. 2019. Classifying attention types with thermal imaging and eye tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–27.

[2] Peter Afflerbach. 2015. *Handbook of individual differences in reading: Reader, text, and context.* Routledge.

[3] Seoyoung Ahn, Conor Kelton, Aruna Balasubramanian, and Greg Zelinsky. 2020. Towards predicting reading comprehension from gaze behavior. In *ACM Symposium on Eye Tracking Research and Applications*. 1–5.

[4] J Alonzo, G Tindal, K Ulmer, and A Glasgow. 2006. easyCBM online progress monitoring assessment system. *Eugene, OR: Center for Educational Assessment Accountability* (2006).

[5] Mary DeKonty Applegate, Kathleen Benson Quinn, and Anthony J Applegate. 2002. Levels of thinking required by comprehension questions in informal reading inventories. *The Reading Teacher* 56, 2 (2002), 174–180.

[6] Deni Basaraba, Paul Yovanoff, Julie Alonzo, and Gerald Tindal. 2013. Examining the structure of reading comprehension: do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing* 26, 3 (2013), 349–379.

[7] D Beymer, D Russell, and P Orton. 2008. An eye tracking study of how font size and type influence online reading. People and computers XXII: culture, creativity, interaction: proceedings of HCI 2008. In *the 22nd British HCI Group annual conference*, Vol. 2.

[8] Ralf Biedert, Jörn Hees, Andreas Dengel, and Georg Buscher. 2012. A robust realtime reading-skimming classifier. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. 123–130.

[9] Mark R Blair, Marcus R Watson, R Calen Walshe, and Fillip Maj. 2009. Extremely selective attention: eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35, 5 (2009), 1196.

[10] Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Tröster. 2010. Eye movement analysis for activity recognition using electrooculography. *IEEE transactions on pattern analysis and machine intelligence* 33, 4 (2010), 741–753.

[11] Douglas Carnine, Jerry Silbert, Edward J Kameenui, and Sara G Tarver. 1997. Direct instruction reading. (1997).

[12] Ronald P Carver. 1990. *Reading rate: A review of research and theory.* Academic Press.

[13] Gavin C Cawley and Nicola LC Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research* 11 (2010), 2079–2107.

[14] Leana Copeland and Tom Gedeon. 2013. Measuring reading comprehension using eye movements. In *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 791–796.

[15] Leana Copeland, Tom Gedeon, and B Sumudu U Mendis. 2014. Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error. *Artif. Intell. Res.* 3, 3 (2014), 35–48.

[16] Leana Copeland, Tom Gedeon, and Sumudu Mendis. 2014. Fuzzy output error as the performance function for training artificial neural networks to predict reading comprehension from eye gaze. In *International Conference on Neural Information Processing*. Springer, 586–593.

[17] Mary C Dyson. 2004. How physical text layout affects reading from screen. *Behaviour & information technology* 23, 6 (2004), 377–393.

[18] Ralf Engbert and Reinhold Kliegl. 2003. Microsaccades uncover the orientation of covert attention. *Vision research* 43, 9 (2003), 1035–1045.

[19] Maureen P Hall, Aminda O'Hare, Nicholas Santavicca, and Libby Falk Jones. 2015. The power of deep reading and mindful literacy: An innovative approach in contemporary education. *Innovación educativa (México, DF)* 15, 67 (2015), 49–60.

[20] Harold L Herber. 1978. *Teaching reading in content areas.* Prentice Hall.

[21] Edmund Burke Huey. 1908. The psychology and pedagogy of reading: With a review of the history of reading and writing and of methods, texts, and hygiene in reading. (1908).

[22] Shoya Ishimaru, Kensuke Hoshika, Kai Kunze, Koichi Kise, and Andreas Dengel. 2017. Towards reading trackers in the wild: detecting reading activities by EOG glasses and deep neural networks. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 704–711.

[23] Halszka Jarodzka and Saskia Brand-Gruwel. 2017. Tracking the reading eye: Towards a model of real-world reading.

[24] Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review* 87, 4 (1980), 329.

[25] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. 2013. A review on video-based human activity recognition. *Computers* 2, 2 (2013), 88–131.

[26] Conor Kelton, Zijun Wei, Seoyoung Ahn, Aruna Balasubramanian, Samir R Das, Dimitris Samaras, and Gregory Zelinsky. 2019. Reading detection in real-time. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. 1–5.

[27] Peter Kiefer, Ioannis Giannopoulos, and Martin Raubal. 2013. Using eye movements to recognize activities on cartographic maps. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 488–491.

[28] Reinhold Kliegl, Antje Nuthmann, and Ralf Engbert. 2006. Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of experimental psychology: General* 135, 1 (2006), 12.

[29] Kai Kunze, Yuzuko Utsumi, Yuki Shiga, Koichi Kise, and Andreas Bulling. 2013. I know what you are reading: recognition of document types using mobile eye tracking. In *Proceedings of the 2013 international symposium on wearable computers*. 113–116.

[30] Silvia Makowski, Lena A Jäger, Ahmed Abdelwahab, Niels Landwehr, and Tobias Scheffer. 2018. A discriminative model for identifying readers and assessing text comprehension from eye movements. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 209–225.

[31] Sandra McCormick. 1992. Disabled readers' erroneous responses to inferential comprehension questions: Description and analysis. *Reading Research Quarterly* (1992), 55–77.

[32] Aliaksei Miniukovich, Antonella De Angeli, Simone Sulpizio, and Paola Venuti. 2017. Design guidelines for web readability. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 285–296.

[33] Aliaksei Miniukovich, Michele Scaltritti, Simone Sulpizio, and Antonella De Angeli. 2019. Guideline-based evaluation of web readability. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[34] Gary E Raney, Spencer J Campbell, and Joanna C Bovee. 2014. Using eye movements to evaluate the cognitive processes involved in text comprehension. *Journal of visualized experiments: JoVE* 83 (2014).

[35] Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124, 3 (1998), 372.

[36] Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr. 2012. Psychology of reading. (2012).

[37] Luz Rello and Mari-Carmen Marcos. 2012. An eye tracking study on text customization for user performance and preference. In *2012 Eighth Latin American Web Congress*. IEEE, 64–70.

[38] Luz Rello, Martin Pielot, and Mari-Carmen Marcos. 2016. Make it big! The effect of font size and line spacing on online readability. In *Proceedings of the 2016 CHI conference on Human Factors in Computing Systems*. 3637–3648.

[39] Judith C Roberts and Keith A Roberts. 2008. Deep reading, cost/benefit, and the construction of meaning: Enhancing reading comprehension and deep learning in sociology courses. *Teaching Sociology* 36, 2 (2008), 125–140.

[40] Ellen Rose. 2011. The phenomenology of on-screen reading: University students' lived experience of digitised text. *British Journal of Educational Technology* 42, 3 (2011), 515–526.

[41] Ladislao Salmerón, Johannes Naumann, Victoria García, and Inmaculada Fajardo. 2017. Scanning and deep processing of information in hypertext: an eye tracking and cued retrospective think-aloud study. *Journal of Computer Assisted Learning* 33, 3 (2017), 222–233.

[42] VE Snider. 1988. The role of prior knowledge in reading comprehension: A test with LD adolescents. *Direct Instruction News* 611 (1988).

[43] Catherine Snow. 2002. *Reading for understanding: Toward an R&D program in reading comprehension.* Rand Corporation.

[44] Namrata Srivastava, Joshua Newn, and Eduardo Velloso. 2018. Combining low and mid-level gaze features for desktop activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–27.

[45] Alexander Strukelj and Diederick C Niehorster. 2018. One page of text: Eye movements during regular and thorough reading, skimming, and spell checking. *Journal of Eye Movement Research* 11, 1 (2018).

[46] Geoffrey Underwood, Alison Hubbard, and Howard Wilkinson. 1990. Eye fixations predict reading comprehension: The relationships between reading skill, reading speed, and visual inspection. *Language and speech* 33, 1 (1990), 69–81.

[47] Boris M. Velichkovsky and John Paulin Hansen. 1996. *New Technological Windows into Mind: There is More in Eyes and Brains for Human-Computer Interaction.* Association for Computing Machinery, New York, NY, USA, 496–503. https://doi.org/10.1145/238386.238619

[48] Titus von der Malsburg. 2015. Saccades: An R package for detecting fixations in raw eye tracking data.

[49] Maryanne Wolf. 2018. *Reader, come home: The reading brain in a digital world.* Harper New York, NY.

[50] Maryanne Wolf, Mirit Barzillai, and John Dunne. 2009. The importance of deep reading. *Challenging the whole child: reflections on best practices in learning, teaching, and leadership* 130 (2009), 21.