IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 31, 2021, accepted February 5, 2021, date of publication February 11, 2021, date of current version February 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3058664

Intelligent Biofeedback Augmented Content Comprehension (TellBack)

HAYTHAM HIJAZI[1], RICARDO COUCEIRO[1], JOÃO CASTELHANO[2], PAULO DE CARVALHO[1], MIGUEL CASTELO-BRANCO[2], AND HENRIQUE MADEIRA[1]

[1]Centre of Informatics and Systems, University of Coimbra (CISUC), 3000-214 Coimbra, Portugal
[2]Coimbra Institute for Biomedical Imaging and Translational Research (CIBIT), Institute of Nuclear Sciences Applied to Health (ICNAS), University of Coimbra, 3000-214 Coimbra, Portugal

Corresponding author: Haytham Hijazi (haytham@dei.uc.pt)

This work was partially funded by the BASE project under Grant POCI - 01-0145 - FEDER- 031581, in part by the Centro de Informtica e Sistemas da Universidade de Coimbra (CISUC), and also in part by Coimbra Institute for Biomedical Imaging and Translational Research (CIBIT), Institute of Nuclear Sciences Applied to Health (ICNAS), University of Coimbra under Grant PTDC/PSI-GER/30852/2017 | CONNECT-BCI.

ABSTRACT Assessing comprehension difficulties requires the ability to assess cognitive load. Changes in cognitive load induced by comprehension difficulties could be detected with an adequate time resolution using different biofeedback measures (e.g., changes in the pupil diameter). However, identifying the Spatio-temporal sources of content comprehension difficulties (i.e., when, and where exactly the difficulty occurs in content regions) with a fine granularity is a big challenge that has not been explicitly addressed in the state-of-the-art. This paper proposes and evaluates an innovative approach named In**tell**igent Biofeed**back**Augmented Content Comprehension (TellBack) to explicitly address this challenge. The goal is to autonomously identify regions of digital content that cause user's comprehension difficulty, opening the possibility to provide real-time comprehension support to users. TellBack is based on assessing the cognitive load associated with content comprehension through non-intrusive cheap biofeedback devices that acquire measures such as pupil response or Heart Rate Variability (HRV). To identify when exactly the difficulty in comprehension occurs, physiological manifestations of the Autonomic Nervous System (ANS) such as the pupil diameter variability and the modulation of HRV are exploited, whereas the fine spatial resolution (i.e., the region of content where the user is looking at) is provided by eye-tracking. The evaluation results of this approach show an accuracy of 83.00% $\pm$ 0.75 in classifying regions of content as difficult or not difficult using Support Vector Machine (SVM), and precision, recall, and micro F1-score of 0.89, 0.79, and 0.83, respectively. Results obtained with 4 other classifiers, namely Random Forest, k-nearest neighbor, Decision Tree, and Gaussian Naive Bayes, showed a slightly lower precision. TellBack outperforms the state-of-the-art in precision & recall by 23% and 17% respectively.

INDEX TERMS Biomedical measurement, cognitive load, content comprehension, eye-tracking, heart rate variability, machine learning.

I. INTRODUCTION

Imagine a software technology that augments your ability to comprehend complex concepts and ideas. You are reading a sentence in a technical document and the software installed in your tablet, laptop, or smartphone automatically detects specific passages or words in a paragraph that make the entire paragraph cumbersome for you and promptly displays an explanation, shows an example, or provides you with a definition that will make you understand the whole idea. We call this approach In**Tell**igent Biofeed**Back** Augmented Content Comprehension (TellBack), and the present paper proposes the idea and evaluates the accuracy of TellBack in the identification of specific content regions (e.g., line, expression, etc.) that are considered difficult to understand by the user. The accurate identification of content locations that cause user's comprehension difficulties is a crucial step to show the feasibility of the proposed approach.

TellBack uses cognitive load as a key element to identify specific parts of digital content that cause comprehension difficulty. The goal is to predict that the user is needing support in understanding the content element/passage that is causing comprehension difficulties while reading digital content.

The associate editor coordinating the review of this manuscript and approving it for publication was Giuseppe Desolda.

Digitized reading resources are immensely growing. Due to the popularity of handheld portable devices, such as e-readers and tablets, reading is becoming more on-screen than on printed material. However, various studies, such as [1], showed that on-screen reading is associated with short attention and poor comprehension. Comprehension of digital resources deserves more research attention, especially with the ever-growing reliance on digital resources comprehension in learning, training, and industry. For instance, in the software industry, code review is one of the essential tasks in software engineering to ensure the quality of the delivered service. Code review requires full code comprehension by the reviewer. The code and its description are generally digitized resources that require the reviewer to immensely read and comprehend.

Digital resources comprehension requires the ability of users or learners to integrate and evaluate information across different types of representations. The integration and evaluation are often mentally demanding tasks, which impose extra cognitive load on learners.

Comprehension, in general, is a higher cognitive process that imposes a cognitive load and involves different cognitive states, which we hypothesize in TellBack that these states can be accurately captured by non-intrusive biofeedback devices through physiological manifestations of the Autonomous Nervous Systems (ANS). In the context of this paper, biofeedback is defined as the process of providing task-relevant feedback to the user based on her/his cognitive state (captured through physiological ANS reactions) using commercially available wearables such as bracelets, watches, and rings.

Our main contributions in this work can be summarized by the following points:

- The paper introduces a new technique to detect the cognitive load and thereby the comprehension difficulties at elemental parts of the content (e.g., lines, expressions, etc.) and not a global assessment of cognitive load as investigated in the state-of-the-art.
- The use of eye-tracking is not to assess the cognitive load as it is generally used in the state-of-the-art (which has the disadvantage of a slow response time) but to provide an accurate spatial resolution where the user is looking at when the HRV and pupillometry biomarkers indicate peaks in the cognitive load.
- The evaluation of the improvements in precision and recall achieved with TellBack, when compared with a very recent related study in [12], shows 23% better precision and 17% better recall.
- This work addresses the limitations of a single modality using data fusion of non-intrusive biosensors and real-time biomarkers from HRV and pupillometry. Unlike other recent studies [2], [3], [12] that exploit the multimodality of biosensors, we used simple features (in terms of computation), but discriminant enough to detect the increase of the user's cognitive load/mental effort in real-time.
- The use of machine learning and AI techniques to optimize the prediction of "when and where" the user is encountering difficulty on screen.

The final goal of Tellback is to augment the user's ability to comprehend complex concepts and ideas through a new type of intelligent interface that seems to guess when and where the user is making an unusual effort to grasp the meaning and to provide the user with accurate contextual help. This would simplify the life of readers, learners, trainees, and many others, and would potentially accelerate the learning curve of intellectual skills. Providing such functionality involves building an intelligent predictive model that can take human biomarkers as input, process those biomarkers, and provide a Spatio-temporal indicator of comprehension difficulty as an output.

Biofeedback devices have been extensively used in research to study the cognitive process involved in various activities such as software development (e.g., see a recent survey on measuring the cognitive load of software developers [9]). Recent work in the context of programmer's errors in software development [4], [5], [6] showed that the integration of HRV, pupillometry, and eye-tracking allows the identification of the code lines that correspond to an increase of cognitive load of individual programmers. Although the goal of the research published in [4], [5], [6] is the annotation of source code for the prediction of programmers' errors, it shows that it is possible to associate programmers' mental load in real-time to specific lines of code or lexical tokens [6]. This has motivated us to explore the same concept in the much broader context of general content comprehension and intelligent user interfaces, as proposed in TellBack.

The evaluation presented in this paper is intended to provide a first answer to the following question: **How accurate and precise are biomarkers extracted from HRV, pupillometry, and eye-tracking in detecting, in real-time, where exactly the content comprehension difficulties occur on screen?**

Detecting cognitive load in real-time using multimodal sources of biosensors has recently emerged as one of the promising approaches in adaptive and cognition-aware e-learning [2], [3], [12], [13]. Those very recent studies were mainly focusing on providing feedback on learners' engagement and cognitive load changes. However, the idea of TellBack is to localize the content (either in learning context or other real-life contexts) to be able to provide contextual help promptly.

The structure of this paper is as follows: the second section discusses the background concepts and the related work. The third section introduces the proposed approach and the methods that were applied. The fourth section addresses the dataset and the experimental protocol that was followed. Results and discussions are explored in the fifth section. Whereas the sixth section discusses the limitations and threats to the validity of this approach. Finally, the conclusion and the future directions are presented in the seventh section of this paper.

II. BACKGROUND AND RELATED WORK

This section provides background information related to the biometric measures that were exploited in this work and provides an overview of the related work. Section II-A briefly introduces HRV and its use in cognitive load detection and assessment. Section II-B concisely presents pupillography and eye-tracking and their use in cognitive load assessment. Section II-C discusses the related work.

A. HEART RATE VARIABILITY (HRV)

Changes in the cognitive load influence the way the ANS regulates the cardiovascular system and causes detectable variations of the heart rate known as HRV (Heart Rate Variability). HRV is modulated by the two components of the ANS (i.e., the sympathetic and the parasympathetic nervous systems) and it is based on calculating the changes of periods between consecutive cardiac cycles. HRV can be assessed in both the time and the frequency domain. In the frequency domain, for example, the Low Frequency (LF) variability of the heart rate is associated with the blood pressure control (i.e., sympathetic), whereas High Frequency (HF) of the heart rate variability is associated with respiratory sinus arrhythmia (i.e., parasympathetic).

The idea of using HRV to detect changes in the cognitive load is not new [21], [29]. However, as indicated in [22], the within-subject measurements of HRV are still uncertain because each subject exhibits distinct HRV rhythms. Thus, non-linear methods of HRV analysis would be prominent in unveiling the complexity of the HRV rhythms. HRV is also shown to be sensitive to many factors such as gender, circadian rhythm, age, prior activities of the subjects, and breathing conditions. Although these factors may change from individual to individual, this is not a problem for the use of HRV in TellBack. We use LF/HF ratio means and spikes to detect peaks in the cognitive load of specific individuals, and thus between-subject measurements are not relevant for TellBack.

B. PUPILLOGRAPHY AND EYE-TRACKING

Eye pupil response has been recognized as an indicator of cognitive and attentional efforts. Various research attempts such as [23], [24] established the evident relationship between pupil activity and attentional cognitive efforts. Beatty described in [23] that when a person recalls something from memory or attempts to parse sentences, the pupil dilates slightly and returns to its normal size after the task is done. This reaction was called task-evoked pupillary response (TEPR) [25]. The spectral analysis of the pupil diameter (PD) is considered a good index for both the mental efforts and fatigue state.

Eye-tracking devices have also been used in research to study the eye gaze during distinct mental tasks such as code comprehension [4], [5], [6]. In principle, eye-tracking is the process of tracking the eye movement and determining where the user is looking at or the absolute point of gaze (POG), which refers to the point in the visual scene at which the user's gaze is focused on. Some technologies like in [7] and [8] used eye gaze to pinpoint difficulties in content. For example, the iDict technology which is described in [8] employs the eye gaze features to assist nonnative English users who encounter difficulties in interpreting English words. The assistive application accordingly translates that word based on the eye movements. In TellBack the eye-tracker is used to localize the regional/elemental parts of content that might have caused the comprehension difficulties.

C. RELATED WORK

Table 1 below includes a related work comparison. However, very few research attempts have been conducted to address the problem of prediction of specific passages in texts that could be difficult to comprehend by users. An early attempt was made by Sibert and Jacob [7]. These authors developed a gaze movements-based assistant that delivers visual and auditory prompting controls to help recognition and pronunciation of words for remedial reading instruction whenever a disabled reader encounters a difficulty. Later, Hyrskykari introduced iDict [8] which analyzes eye gaze to predict difficulties in English words for nonnative speakers and accordingly translate or pronounce difficult words. Those related works [7], [8] lack the use of the information sources multimodality (i.e., including different biomarkers). Moreover, they lack the use of machine learning techniques to predict the difficulty. Due to the non-linearity of cognitive load measurements, the use of statistical analysis is not enough. In TellBack, we utilize different AI techniques such as feature-level fusion, ensemble decisions, and various classification models.

Table 1 (below) presents state-of-the-art works that used multimodal biosensors in assessing cognitive load. Some of these works were mentioned in a mapping study [9] that provided a thematic analysis of measuring the cognitive load of software developers while performing mental tasks. The study included 33 articles from 11 search engines.

The authors in [9] showed that 55% of the studies used EEG in monitoring the cognitive load, 36% used a combination of sensors, 6% used eye-tracking, and 3% used fMRI. The recent works started to investigate the multimodality of the input sources in different scopes. Table 1 (above) summarizes those studies and other studies to show the biomarkers that were used to assess the cognitive load, the methodology, the results achieved, and the limitations.

Since the idea of TellBack is new, it was not easy to perform a meaningful comparison between TellBack and the previous studies. Nevertheless, Table 1 briefly discusses the limitations of the previous works in a possible utilization scenario of identification of user's content comprehension difficulties. As noticed, most of the studies that used multimodal input sensors address software development, code comprehension, and e-learning scenarios. There are very few studies that addressed other scopes and other types of content with the limitation of a single modality to assess the cognitive load.

TABLE 1. Related studies comparison.

Studies	Biomarkers	Scope & artifacts	Methodology	Results	Limitations
Mills (2020) [12]	Electrodermal Activity (EDA), eye-tracking	English texts	This study aimed at detecting mind-wandering status while reading digital resources using EDA and eye-tracking.	The chosen model had a weighted precision of 72.2% and a weighted recall of 67.4%.	This study did not assess the real-time cognitive load but detected the moments when the reader wanders. Nonetheless, this work is considered the most related to the TellBack in terms of using lightweight multimodal biosensors in reading digital content. Thus, a performance comparison was performed.
Sharma (2020) [13]	eye-tracking, EEG, wristband, facial expression	Textual Questions, learning	This work attempted to estimate the learners' efforts through physiological signals analysis. It used k-means to cluster students' behavior, and then traditional classifiers.	Overall precision, recall, and F1-score are 0.89, 0.84, and 0.90, respectively	This study was limited to assess the mental efforts of learners when answering questions and did not explore other stimuli and mental tasks.
Fritz (2014) [14]	EEG, EDA Eye-tracker	Software development	The authors classified code comprehension difficulty by assessing psychophysiological measures of programmers.	Classifiers could predict task difficulty for a new programmer with 64.99% precision and 64.58% recall and a new task with 84.38% precision and 69.79% recall.	This study did not use the eye-tracker to localize the difficult regions or elemental parts of code. It used an eye-tracker to assess the mental workload which could give an accurate estimate of cognitive load but might not be a timely manner.
Nourbakhsh (2014) [15]	Eyeblink, EDA	Arithmetic tasks	In this study, eyeblinks, and EDA were used to discriminate workload levels using different arithmetic operations.	Achieved an accuracy of 75% in binary classification and of 50% in four-level classification.	The scope of artifacts was limited to arithmetic tasks which hardly motivates to extrapolate to other types of stimuli.
Muller (2016) [10]	Heart Rate, Skin conductance	Software development	The authors focused on the online prediction of code quality concerns using biometric measures.	The biometric classifier could correctly (38.6%) code elements with quality concerns on method level and (40%) on class level. The precision was not high with 13.0%, and 22.0%, respectively.	They segmented the code elements based on an assumption that the developer is affected by the code lines he just selected or edited. This method could be inaccurate because selecting or editing code might not answer whether the developer is attempting to understand that line or just doing a random selection.
Lee (2018) [17]	EEG, Eye-tracking	Software development	Authors used EEG and eye-tracking to predict the programmer's expertise and task difficulty in software development.	Predicted task difficulty and programmer level of expertise with 64.9% and 97.7% precision and 68.6% and 96.4% recall, respectively.	This study relied on EEG which could not be practical in a software development environment due to its intrusiveness. Furthermore, the precision and the recall of the task difficulty are considered relatively low.
TellBack	HRV, Pupillography Eye-tracking	Open scope of content	In our study, we aimed at detecting the moment and the regional parts of digital content that caused the difficulty in understanding using simple real-time features.	Tellback achieved an accuracy of 83.00% ± 0.75 in predicting the difficulty, and 89%, 79%, and 83% precision, recall, and F1-score, respectively.	TellBack could outperform the recent related study in [12] with 23% better precision and 17% better recall. However, limitations are discussed further in section VI.

For example, [10] used EEG to assess the cognitive load of users while interpreting different representations of data visualizations (i.e., data distributions). Another example, pupil dilation was used in [11] as a metric to assess linguistics representations comprehension.

In summary, the studies that have introduced the idea of detecting the Spatio-temporal sources of comprehension difficulty in digital content lack the fusion from multimodal biosensor as well as the generality of content being addressed in comprehension difficulty. The relevant study in [12] used the multimodal sources of information to detect the mind wandering moments of readers and not their comprehension difficulties. The studies that used the multimodality model of sensors such as [13], [14], [15], [17] were in general supporting software development and narrow specific domains. Even in other domains, there was no such ability to localize the content that caused the comprehension difficulty at a very fine granularity level.

Another evident aspect is that the use of advanced AI techniques is not commonly used in such proposals. For example, and back to the mapping study in [9], 40% of the studies that assessed the cognitive load of software developers do not use any sort of machine learning techniques, whereas the rest are limited by using only classification techniques for different sets of biosensing features. Most used classification techniques are (according to [9]) Support Vector Machine (SVM) 15%, Naive Bayes 15%; Multi-algorithms for classification 9%; K-means 3%; Decision Tree 3%; Logistic Regression 3%; Neural Network 3%; Random Forest 3%; Linear Regression 3%, and Relevance Vector 3%.

To measure the cognitive load associated with mental efforts (e.g., content comprehension), we must distinguish between the objectivity (subjective vs objective) methods and the causality (direct vs indirect) methods. In our evaluation experiments, the subjective perception of the cognitive load was assessed using self-reporting through the NASA-TLX [18] questionnaire, whereas objective measurement was obtained using peripheral physiological responses driven by the autonomic nervous system (ANS) captured by biofeedback low-intrusive sensors.

In TellBack we propose using HRV and eye-tracking (including pupillography) since there are multiple low- intrusive solutions to implement these sensors. A total of 83% of the explored studies used eye-tracker and eye-related features (e.g., blink), 66% used skin conductance or EDA measures, and 50% included EEG. The comparison with TellBack was not easy because the ideas presented were not similar. However, we used [12] as a good example of a related recent study to compare TellBack with. The precision and Recall were the valid measures to perform the comparison which is shown in Table1.

III. METHODOLOGY

As described earlier, TellBack is a new idea to automatically identify a user's difficulties in comprehending specific passages/elements during the reading of contents on tablets, laptops, or smartphones. Our vision for TellBack is a future intelligent and attentive user-interface [16] for enhancing comprehension using, for example, pop-ups containing relevant information on the topic extracted from the web could. People could indeed stop reading when encountering a comprehension difficulty and could google the meaning of the obscure elements in the content. However, in practice, this rarely happens, due to the rush of the moment or due to the interaction limitations of small devices such as smartphones, which is the de facto computational platform in both the developed and developing world. Therefore, the key message is often lost because of these unsolved difficulties, which hinders proper decision making and leads to all sorts of mistakes, drastically disturbing the learning process and having a huge negative impact on society.

Figure 1 below shows a schematic representation of TellBack main components. This representation can be divided into three distinct phases: bio-signals acquisition, signal processing, machine learning models, and content annotation based on comprehension difficulty classification.

Recent results appeared in [6] in the context of assessing cognitive state (mainly high cognitive load and distraction states) of software programmers show that it is possible to associate such cognitive information, in real-time, to specific lines of code or lexical tokens, to identify such code lines as being more susceptible of having software bugs. In the work reported herein, we will expand such preliminary results obtained in the specific context of prevention of software bugs, to a much broader context of general content comprehension. The approach followed in TellBack can be summarized in the following components as shown in Figure 1:

1) Real-time ANS biomarkers measurements (HRV, Pupillography).
2) Features extraction and selection to assess cognitive load spikes and patterns.
3) Machine learning techniques to improve detection of high mental effort/cognitive load from the extracted features.
4) Eye-tracker to localize content regions that are associated with high mental effort.

ANS signals are captured from subjects while performing content comprehension on screen. The signals used to assess the cognitive load are sampled with timestamps allowing the identification of the moments when the user is performing the comprehension task. The signals obtained are fed into the feature extraction and selection module. In this module, the domain knowledge is used to choose the most relevant biomarkers out of the signals. Using those discriminant features extracted from the collected pupillography and HRV synchronized with eye-tracker (that will be illustrated in the next section), elemental digital contents associated with such features are classified into two main binary classes, i.e., ''Difficult'' and ''Not Difficult''. Those classes are considered the comprehension difficulty state as shown in Figure 1.

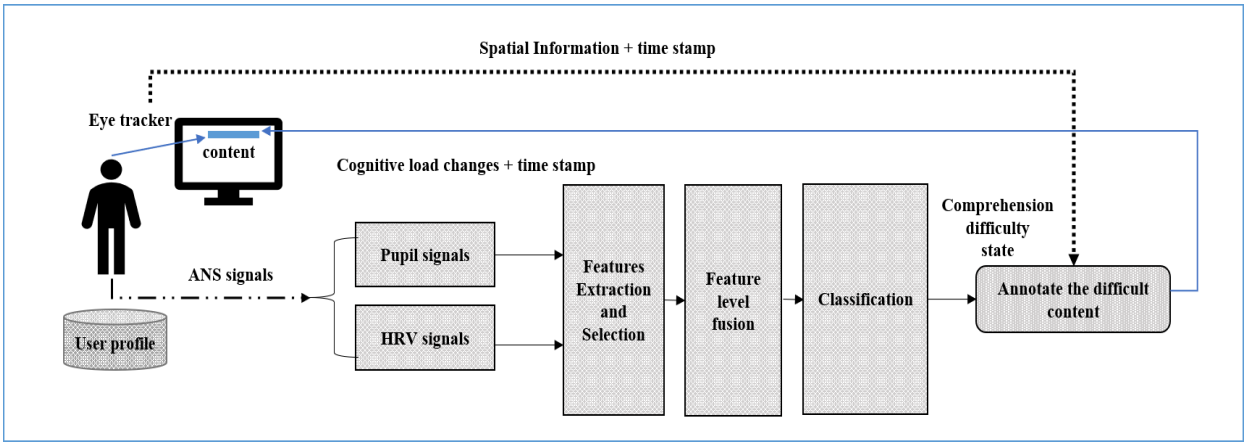

FIGURE 1. A schematic representation of TellBack.

The eye-tracker is employed as the information source to provide gaze geometrical information (i.e., spatial information) related to elements classified with comprehension ''difficulty''.

In the classification module, a set of classifiers are suggested to obtain the best results in terms of accuracy, precision, and recall. To ensure the interpretability of the decision, we included the use of the Decision Tree classifier among other classifiers, regardless of its known low accuracy in some problems.

It is well known that isolated ANS manifestations are non-specific, i.e., multiple stimulus and physiological processes might induce similar ANS modulation of a specific ANS physiological manifestation such as HRV or pupil diameter variability. To reduce this non-specificity, we propose to combine distinct ANS manifestations using a data fusion approach to exploit specific patterns of activation.

In this sense, the evaluation presented in this paper represents a less rich scenario that only uses HRV, pupillometry, and eye-tracking, which suggests that our first results provide a conservative view on the accuracy, precision, and recall of TellBack.

IV. DATASET, PROTOCOL AND METHOD

This section describes the dataset that was used to prove the concept of TellBack and discusses the experimental protocol and the methods used to analyze the resulting data.

A. DATASET

The dataset that was used for this study is an already existing publicly available dataset, which was originated from the BASE project mentioned in [4], [5], [6]. The dataset was developed in one of the project's studies to monitor physiologic reactions and mental effort that are associated with code comprehension in different complexities using non-intrusive biosensors. The biosensors that were used in the experiment are the ECG and Eye-tracking with Pupillography. The controlled experiment involved 30 subjects experienced in Java. The subjects experimented in the same room with the same conditions while being equipped with mentioned wearable non-intrusive sensors. The experiment protocol of the BASE study included the following steps:

1) **Baseline activity** to let the subjects look at an empty grey screen with a black cross in the center for 30 seconds to detach participants from any activity.
2) **Reference Activity** that comprises a screen with a Natural Language text for 60 seconds.
3) **Baseline activity** to let the subjects look at an empty grey screen with a black cross in the center for 30 seconds to detach participants from any activity.
4) **Code comprehension** comprises three code programs with different complexities (c1, c2, c3).
5) **Baseline activity** to let subjects look at an empty grey screen with a black cross in the center for 30 seconds to detach participants from any activity.
6) **Survey** using NASA-TLX to assess the subjects' mental efforts while comprehending the code.
7) **Control questions** to check if the subject understood the codes correctly.

This experiment yielded 90 datasets corresponding to the 30 volunteers with 3 tasks (i.e., c1, c2, c3). In this experiment, the changes in HRV and Pupil Diameter (PD) biomarkers were analyzed during code comprehension tasks to detect moments that correspond to the high mental effort of the programmer. Usually, those moments correspond to HRV and pupil spikes (i.e., peaks). These moments were mapped to corresponding spatial coordinates of code locations using eye-tracking information.

B. PROTOCOL AND METHOD

The main goal of the BASE project was to assess the cognitive load of programmers and annotate the code lines that are more prone to have software bugs. The annotation of those potentially problematic code lines would alert the programmers online to revise those code lines. However, in our study, TellBack uses the same technique (i.e., using biofeedback

wearables to assess cognitive load), but with a different purpose and scope. The main goal of TellBack is to predict content comprehension difficulty based on assessing the user cognitive load in real-time and to use the eye-tracker synchronously to localize the content elemental parts that caused the comprehension difficulty.

Since the protocol used to produce the dataset available from the BASE study was not designed for the identification of user's comprehension difficulties while reading the code (the goal in the BASE project was programmer's errors and bugs), we established a specific protocol (adding a new layer to the initial protocol) that utilizes the dataset already available and enhances such dataset with additional data to evaluate the accuracy of our new approach. More specifically, the additional data consists of the identification of the code regions that are considered easy/difficult to understand by the subjects. Our specific protocol assumes the following, which takes the form of a data-gathering interview:

1) The same subjects of the BASE study were invited to a remote data gathering interview.
2) The selection of the subjects was narrowed down to those who have most likely the same experience in Java programming compared with the experiment time (i.e., they have not gained additional experience in Java).
3) The selected subjects were asked to fill a form indicating their current position, personal evaluation of the gained experience in programming languages since the time of the BASE study, and the current programming languages that they have been practicing for the time.
4) The same code snippets (c1, c2, c3) that were used in the BASE study were presented to the subjects. These snippets represent codes of different complexity levels ranging from $Vg = 3$ for c1 to $Vg = 14$ for c3 based on McCabe metric.
5) The main task was asking the subjects to read and comprehend the code snippets again with the same amount of time given in the original study.
6) An additional layer of the original protocol was added by asking the subjects to manually label the code lines with three colors: red for most difficult, orange for difficult, and green for easy. Highlighting the code lines was based on the subjectively perceived difficulty level.
7) The same control questions were presented to the subjects to check their code comprehension.
8) The data analysis was initially performed by extracting the pupillography and HRV signals that are associated with the time window of each annotated region in point 6 based on the gaze information.
9) By repeating those steps for each selected subject, a new dataset was developed, which comprises the pupillography and HRV features that correspond to different code lines and regions labeled by the subjects according to the difficulty.

For simplicity in the analysis phase, we considered orange and red colors indicated in point 6 above as ''Difficult'', whereas green was considered as ''Not Difficult''.

We hypothesize that if the subject in the new data-gathering interview highlighted a certain segment as ''Difficult'', then it would have been difficult at the time of the experiment for him or her to comprehend that segment of code, and thus this would enable us to compare between the code segments highlighted ''difficult'' and the HRV and pupilometer signals acquired at the original study time. Likewise, labeled code segments with ''Not Difficult'' are also compared with the corresponding HRV and pupillography signals. This would lead us to prove whether HRV and pupillography features could be used to identify with adequate time resolution different content pieces where users are encountering comprehension difficulties. The new layer of protocol mentioned in point 6 above would enable us to observe to which extent the eye-tracker was able to identify the regions/lines of code that correspond to high mental effort.

The data-gathering interview took place through Zoom due to the Covid-19 outbreak. A total of 11 subjects were interviewed, as shown in Table 2, and 30 different samples were obtained. Only one of the subjects was excluded because there was no corresponding sufficient information from the original study. Table 2 shows the subjects' information profile. The average time of interviews was 60 minutes. At first, subjects were oriented to the task that is required from them through an explanatory video. After that, they were provided with the codes (i.e., c1, c2, and c3) to read and comprehend them separately. A stopwatch timer was used in each step. The subjects were constantly asked to highlight the code lines whenever they feel that these code lines needed more mental efforts to grasp.

TABLE 2. Subjects information profile.

Participant ID	*Current position*	*Rated experience*	*Programming languages of experience*	*Available diagrams*
ID4	Java programmer	High	Java	Code1
ID9	Software Engineer	Moderate	Robby, AngularJS, HTML5, C#	Code1, 2
ID11	Nokia, Software engineering	High	Python	Code2
ID14	Master's researcher	Moderate	Python, Shell script, JavaScript	Code1
ID17	Software engineering	High	.NET, C#	Code1, 2
ID20	Master student - developer	High	Java for Android, Java	Code1
ID21	Consultant	Moderate	Java, Python	Code1, 3
ID23	Medical Apps programmer	Moderate	Java and AngularJS	Code1, 3
ID29	Assistant Professor	Moderate	Python, Java	Code1, 2
ID30	Backend developer - python	Nothing	Python	Code2

The subjects were asked about the current programming languages that they have been working on since the time of the original study. We were conservative about those who acquired excessive experience in Java (i.e., the language of the code snippets) because they may not experience the same difficulty that they encountered at the time of the original experiment.

After gathering the needed information, the analysis of the results was carried out as follows and as illustrated in pseudo-code 1:

1) Extracting the time instants of pupillography and HRV singles for each labeled code line or region by the subject.
2) Calculating the following features (number of HRV LF/HF ratio spikes which represents the ratio between the low and the high frequency of the HRV signal, the number of pupillography LF/HF spikes, mean of LF/HF spikes values of pupillography, mean of LF/HF spikes values of HRV).
3) Labeling those features with a binary difficulty level (Difficult, not Difficult) as annotated by the subjects in the data-gathering session.
4) Performing unpaired t-test to examine if there is a significant statistical difference between the two difficulty states according to the given features in point 2 (assuming a normal distribution).
5) Applying Grid Search to find the optimum hyperparameters.
6) Using leave-one-out cross-validation to evaluate the resulted model.
7) Using K-Fold cross-validation to compare with point 6.
8) Identifying the signals patterns (i.e., HRV and pupillography) that are associated with the difficulty in understanding using different classifiers (SVM both the linear and the radial basis function, Random Forest, KNN, Naïve Bayes, Decision Tree).
9) Running the model 50 times and calculating the accuracy mean and the accuracy standard deviation.
10) Calculating the precision, recall, micro F1-score, macro F1-score.

The following Pseudo-Code represents steps 1-3 in detail.

Pseudo-Code 1 Mapping Extracted Features with Regional Parts of Content on Screen

Input: local_maxima(signal): signal ϵ {HRV, pupil}, S,
S = (s1, s2,..,sn)
S. time: instant time of S vector
EyeGazeTime = Time_instant(signal) signal ϵ {eye tracker}, eye gaze time instants – time instants of subjects looking at screen.
Output: eye_gaze_coordinates associated with S

1: For each s ϵ S do
2: T(s)$\leftarrow$ S. time
3: For each s ϵ S do
4: EGT(T(s)) $\leftarrow$ EyeGazeTime
5: Find samples n where
6: EGT(T(s))<= T(s)+δ AND EGT(T(s)) >= T(s)- δ);

δ is the sample range where T(s) spans.

7: For each sample range n ϵ EGT(T(s)) do
8: Find eye_gaze_coordinates(n) where subjects were looking at that time instant (when spikes occur).
9: Return (eye_gaze_coordinates(n))

V. RESULTS AND DISCUSSION

As previous studies have shown that pupillography and HRV are timely manner indicators for the cognitive load (and are non-intrusive), TellBack started with those measures to assess the cognitive load associated with the understanding difficulty. However, those measures are insufficient to annotate the elemental parts of content that caused the comprehension difficulty. Therefore, eye-tracking was introduced to tell us where the user is looking whenever the pupil or HRV signal spikes above the threshold, indicating an abrupt increase in the user's cognitive load.

After performing the analysis, the preliminary results showed that most of the code regions that were difficult for the subjects to understand were also mentally demanding at the time of the interview (shown by self-annotations). Figure 2 (above) shows an example of one of the subjects. In this Figure, we can see that the mentally demanding code regions (Figure 2-D) correspond in general to high gaze density (Figure 2-B and C). Similarly, the pupil and HRV spikes tended to increase above a certain threshold (Figure 2-E) when the task is considered difficult or most difficult. The threshold was determined based on data observations.

Likewise, the pupil and HRV spikes were defined based on empiric data observation and analysis for different subjects. Using gaze and saccades to infer the user's cognitive load requires the integration in time, which means that the precision in the time domain is poor. On the contrary, pupil, and HRV features provided an accurate assessment of cognitive load in the time domain (moments when the spikes occur).

As observed in Figure 2, the code lines annotated as difficult or most difficult (orange and red color respectively) by the subject in our data-gathering interviews correspond to a high number of HRV and pupillography spikes (orange appears in this example). That means the higher number of HRV and pupil spikes as seen in (Figure 2-E) within a specific time window, the more potential evidence of high cognitive load induced on the subject by understanding efforts. Note that to identify the point in the code that corresponds to an HRV or pupil spike (i.e., the screen area where the user was looking at when the spike occurred) we needed to find the gaze region in part B of Figure 2 that is vertically aligned with the spike, and then to go to the right and find the code lines (part C and D of the figure) that represent the eye-tracking information and the self-labeling of code

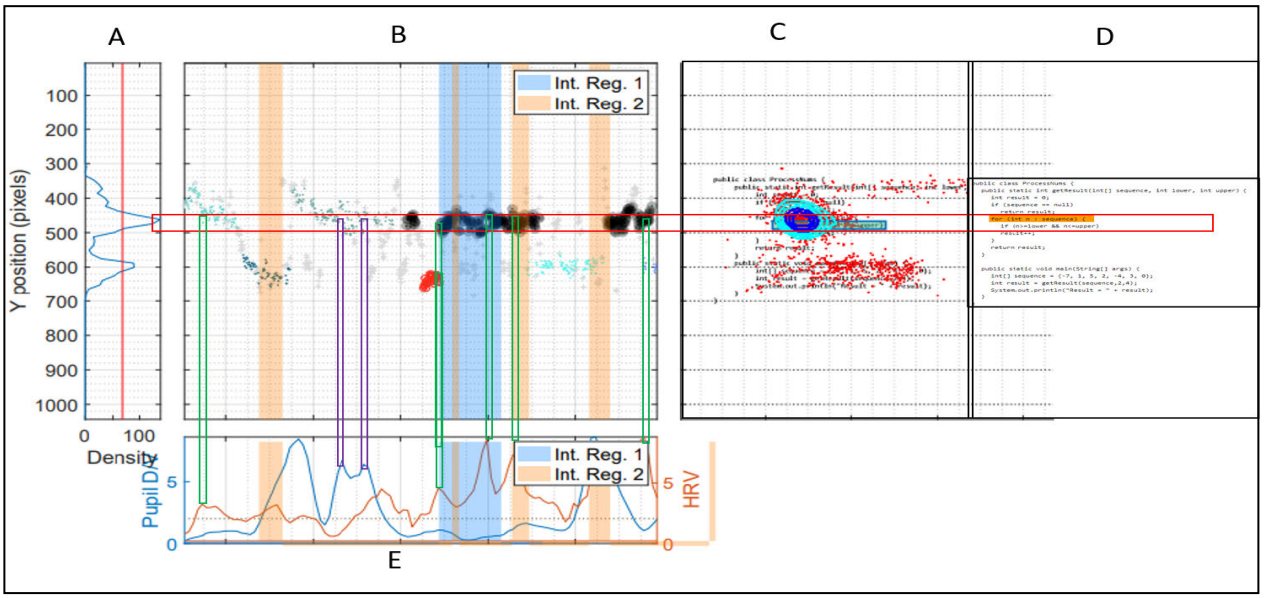

FIGURE 2. Pupil and HRV signals of the subject no 9 in comparison with his self-labeling: A) Eye gaze density corresponding to the y-axis of the code. B) Clusters over time and gaze velocity; b) the pupillography and HRV extract LH features extracted. C) The red dots that represent the eye gaze geometrical distribution superimposed on the code. D) The additional layer of the original protocol shows labeled code lines by subjects. E) The corresponding pupillography and HRV signals and spikes.

regions, respectively. Before applying the empirical analysis, we could see from Figure 2 (C and D) that there is a clear mapping between what eye-tracker tells us about the mental effort in specific lines/regions of code and the labeled code lines/regions by subjects as appeared in part D of the same Figure. To evaluate the biomarkers that correspond to each region, the following features were analyzed: the number of HRV LF/HF spikes, the number of pupillography LF/HF spikes, the mean of LF/HF ratio spikes values of pupillography, mean of LF/HF ratio spikes values of HRV based on the time instants window of the subjects' gaze that is corresponding to the annotated code regions.

An unpaired t-test with a confidence interval of 95% was performed to examine if there is a statistically significant difference between the number of pupillography LF/HF ratio spikes in the two difficulty groups (i.e., difficult, not difficult). The unpaired t-test with $p - value = 0.0095$ shows that the means are statistically different. In the ''difficult'' group ($M = 9.5$, $SD = 6.5$), whereas in the ''not difficult'' group ($M = 3.055556$, $SD = 6.5$). Likewise, the unpaired t-test shows that the means of the number of HRV LF/HF ratio spikes with $p - value = 0.0002975$ are statistically different in the two difficulty states., The same test shows also that there is a statistically significant difference between the means of pupillography LF/HF ratio spikes values in the two difficulty groups with ($M = 5.10200$, $SD = 2.88$) in the ''difficult'' group, and with ($M = 4.07444$, $SD = 2.86$) in the ''not difficult'' group with p-value $= 0.3733$. Therefore, the alternative hypothesis is true which assumes the statistically significant differences. The features correlation shown in Figure 3 above shows clear discrimination between the ''Difficulty'' and ''No difficulty'' behavior. The correlation scores between the biomarkers such as HRV and Pupil spikes number and the difficulty perceived behavior of subjects shown by their self-annotations of the code regions indicate a strong correlation as shown in Figure 3.

Likewise, the box plot in Figure 4 shows selected features distributions against the difficulty groups of regions of codes as perceived by subjects. Figure 4a shows that the higher the number of pupil spikes within a code region, the higher the difficulty perceived by the subject. Similarly, Figure 4b indicates that the increasing number of HRV spikes is proportional to the difficulty state in comprehending regions of code. From this Figure, we can see the differences in the means, medians, and minimum values of the two different groups. The relatively high spread of HRV data from the center can be explained in terms of the high sensitivity of the

HRV signal to different cognitive states such as anxiety, stress, and fatigue. Those biomarkers (i.e., features) were fed into 5 different classifiers namely, Support Vector Machine (SVM) with both a linear kernel and a Radial Basis Function kernel, Random Forest, K-Nearest Neighbor (KNN), Decision Tree, and Gaussian Naíve Bayes. Due to the limited dataset, and to avoid any sort of data overfitting, this work used the leave-one-out cross-validation (LOOCV) method [19]. The LOOCV uses one sample of the training data for validation and uses the rest of the data for training until it covers all the data records. However, for making further comparisons with other methods of cross-validation, K-fold cross-validation [26] was also examined.

All classifiers were run 50 times to calculate the mean and the standard deviation of the accuracies as shown

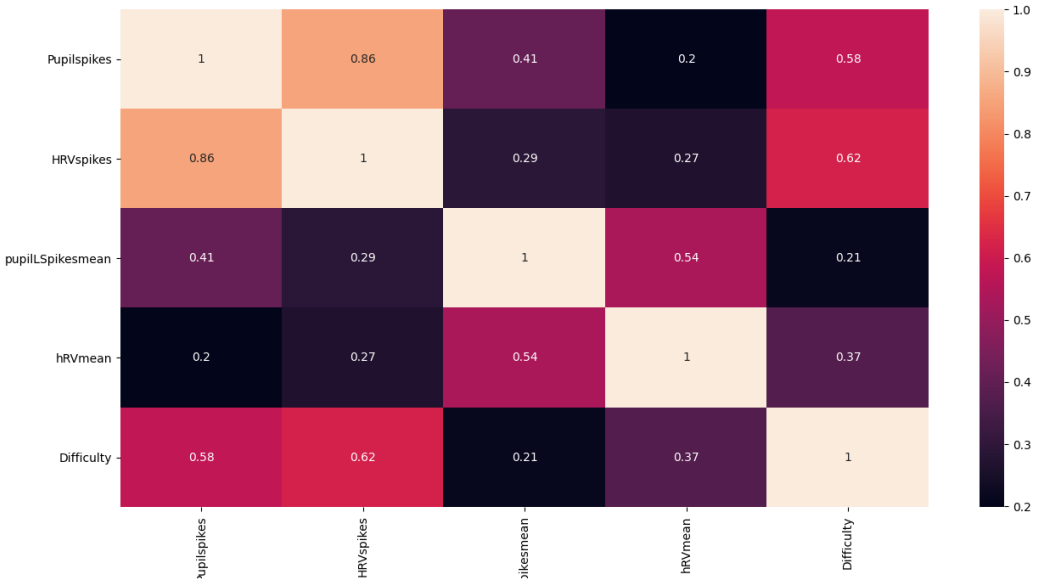

FIGURE 3. A features correlation matrix with the difficulty perceived behavior.

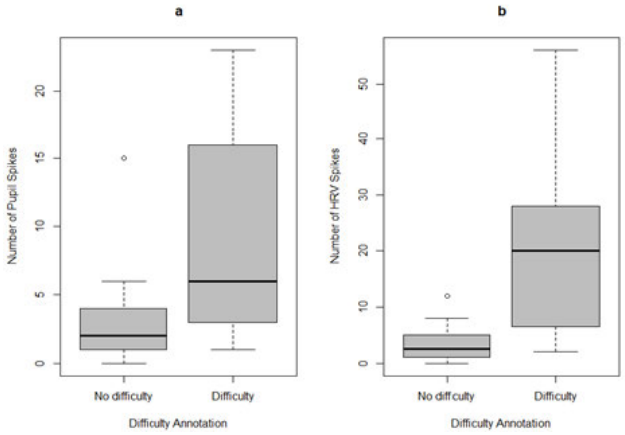

FIGURE 4. Selected pupil and HRV features distribution.

in Table 3 along with the precision, recall, F1-score (Macro and Micro).

As shown in the table above, we can notice that including the HRV features only could achieve the best results in terms of accuracy, precision, recall, and F1-score with SVM using the RBF.

HRV alone could explain the perceived difficulty behavior in understanding at the code line level as annotated by the subjects. Likewise, we can see the SVM with the RBF kernel performed the best when we fused the pupil and the HRV data, whereas the worst results were achieved by using pupillography measurement only. When we applied the 5-fold cross-validation, the performance of the HRV is still giving good results. However, the fused features of both HRV and Pupillography using k-nearest neighbors have given the best performance results in these settings.

Classification parameters were achieved through hyperparameter Grid Search fine-tuning (Table 4). Despite the best results achieved using HRV only, the feature fusion is considered a more robust approach to avoid the non-specificity of the HRV signals. As can be observed from the results, the model is performing well in terms of accuracy, precision, and recall. This gives us an overall evaluation of our approach that it is valid and well-performing although with a limited number of samples. In our approach evaluation, we can show that pupillography and HRV features set including only the number of HRV LF/HF ratio spikes, the number of pupillography LF/HF ratio spikes, the mean of LF/HF ratio spikes values of pupillography, the mean of LF/HF ratio spikes values of HRV could provide us with an assessment of the difficulty experienced by subjects in understanding code at the code line level and on time.

The reported results have been verified three times from three different machines running the code. It is worth mentioning that Grid Search was applied to turn the hyperparameters of each classifier to achieve the best accuracy as shown in Table 4 below.

Those values were used in each model presented in Table 3 along with the different cross-validation techniques to build the initial model of TellBack. As we can observe from the results reported herein, that TellBack shows a very good performance in predicting the difficulty encountered in comprehension using non-intrusive biosensors coupled with the use of machine learning techniques. The results reported herein can be used to build the initial model of TellBack intelligent tool that responds to the variations in the biomarkers to predict the comprehension difficulty promptly.

VI. LIMITATIONS AND THREATS TO VALIDITY

Although this first evaluation of the Tellback approach shows very promising results, there are still limitations that should

TABLE 3. Tellback performance evaluation.

Measurement	Classifier	Accuracy	Precision	Recall	F1-score (macro)	F1-score (micro)
Leave-One-Out Cross-validation						
HRV	SVM – RBF	83.00% ±0.75	0.89	0.79	0.80	0.83
	SVM-Linear	80.00% ±0.81	0.82	0.76	0.77	0.80
	Random Forest	77.00% ±0.85	0.84	0.76	0.78	0.80
	K-nearest neighbor k=9	82.00% ±0.76	0.88	0.78	0.80	0.82
	Decision Tree	80.50% ±1.8	0.82	0.76	0.77	0.80
	Naïve Bayes	76.00% ±0.89	0.76	0.73	0.74	0.76
Pupillography	SVM – RBF	71.10% ±0.90	0.73	0.62	0.73	0.76
	SVM-Linear	75.00% ±0.87	0.76	0.67	0.68	0.75
	Random Forest	73.00% ±0.88	0.73	0.75	0.70	0.75
	K-nearest neighbor k=9	70.00% ±0.92	0.74	0.64	0.63	0.7
	Decision Tree	67.00% ±0.94	0.70	0.71	0.69	0.7
	Naïve Bayes	73.00% ±0.88	0.73	0.69	0.7	0.73
HRV & Pupillography	SVM – RBF	80.00%±3.2	0.82	0.76	0.77	0.80
	SVM-Linear	69.16%±8.9	0.65	0.64	0.64	0.66
	Random Forest	77.00% ±0.85	0.72	0.71	0.71	0.73
	K-nearest neighbor k=9	78.33%±5.9	0.89	0.79	0.80	0.83
	Decision Tree	67.13%±10	0.65	0.65	0.65	0.66
	Naïve Bayes	76.00% ±0.85	0.77	0.74	0.74	0.76
K-Fold Cross-validation, Splits =5						
HRV	SVM – RBF	83.33% ±6.6	0.82	0.76	0.77	0.80
	SVM-Linear	73.33% ±0.81	0.77	0.73	0.71	0.73
	Random Forest	76.6% ±6.66	0.73	0.69	0.70	0.73
	K-nearest neighbor k=9	80.00% ±3.33	0.89	0.79	0.80	0.83
	Decision Tree	73.50% ±3.23	0.72	0.70	0.71	0.71
	Naïve Bayes	76.66% ±8.16	0.76	0.73	0.74	0.76
Pupillography	SVM – RBF	76.66% ±8.1	0.70	0.65	0.65	0.70
	SVM-Linear	70.00% ±2.24	0.79	0.72	0.73	0.76
	Random Forest	76.6% ±.66	0.65	0.63	0.64	0.66
	K-nearest neighbor k=9	73.33% ±8.35	0.77	0.68	0.68	0.73
	Decision Tree	60.0% ±2.4	0.54	0.54	0.55	0.56
	Naïve Bayes	70.66% ±9.16	0.69	0.66	0.67	0.70
HRV & Pupillography	SVM – RBF	80.00%±6.67	0.82	0.76	0.77	0.80
	SVM-Linear	63.16%±11.9	0.62	0.62	0.62	0.63
	Random Forest	73.00% ±1.24	0.72	0.70	0.71	0.71
	K-nearest neighbor k=9	83.33% ±10.6	0.89	0.79	0.80	0.83
	Decision Tree	83.11% ±12.13	0.89	0.79	0.80	0.83
	Naïve Bayes	76.44% ±11.23	0.68	0.68	0.68	0.70

be considered and discussed as the main threats to validity of the present study.

First, the dataset used was acquired in a controlled experimental environment and has the classic limitations regarding the made-up setup. The content used for the different comprehension tasks (software source code) might not be perfectly realistic or representative of real-world software, as the three programs used to represent tasks with different complexities are all of them relatively small programs when compared to real software.

TABLE 4. Tuned parameters using grid search.

Classifier	Hyperparameter
SVM	{'C': 1, 'gamma': 0.001, 'kernel': 'rbf'}
Random Forest	{'criterion': 'gini', 'max_depth': 4, 'max_features': 'auto', 'n_estimators': 200}
KNN	{'n_neighbors': 9, 'weights': 'uniform'}
Decision Tree	max_leaf_nodes=2, random_state=42

Naturally, a controlled experiment cannot use very large code samples as the duration of the task would be prohibitively long. In any case, the most complex code used in the experiment has a McCabe cyclomatic complexity of 14, which is already quite complex and truly representative of complex software units.

A second aspect to be considered is the fact that we had to specify the second protocol to define the second round of sessions with the participants. The goal was to collect the necessary additional information to use the available dataset for the evaluation of content (i.e., code) comprehension. These two steps (i.e., the initial experiments that produced the dataset and the second round of sessions with the participants) were separated by around 12 months. In any case, the second protocol was designed to consider the possible evolutions of the participants in terms of improved skills in Java programming. In practice, the possible deviation introduced by these two steps experimental approach is in the direction of leading to conservative evaluation results. The reason is that more mature participants (i.e., the same participant with additional 12 months of general programming experience) will tend to mark as difficult to understand code snippets that for sure were difficult for him in the first round. In other words, if TellBack classifies code lines as being difficult to understand, for sure the participant considers that code as difficult. A third limitation concerns the number of participants in the study. The big impact of needing a second step to ask participants to identify code lines that they consider difficult to understand had the side effect that only 11 participants (with 30 samples) from the original group of 30 participants were recruited so far for the second round of sessions. We are aware of this limitation and we are still contacting additional participants to increase the number of participants a bit more. In any case, the standard deviation observed is relatively contained, which is a good sign.

VII. CONCLUSION AND FUTURE DIRECTIONS

Biofeedback signals have shown a very good potential to assess the cognitive load in different mentally demanding tasks including content comprehension. In this paper, we proposed the TellBack approach that uses biofeedback measures such as pupillography and HRV to assess changes in cognitive load in a timely manner and uses such abrupt changes in the cognitive load to identify user's difficulties in understanding content. Spatial resolution was achieved using the eye-tracker to pinpoint exactly where the user was looking at when the peaks of cognitive load are detected. The combination of eye-tracker and cognitive load assessment enabled adequate Spatio-temporal resolution to identify the elemental part of the content that hindered the comprehension process.

This experimental evaluation of the accuracy of TellBack in a code understanding scenario showed quite promising results, reaching an accuracy of $83.00\% \pm 0.75$ in the identification of code lines that have been classified by experiment participants as difficult to comprehend. These quite positive results show that our TellBack concept is valid and proved to be accurate enough to consider future developments in the form of a prototype.

As a future direction, we believe that fusing the multimodal information sources applied with additional complementary information sources such as the Electrodermal Activity (EDA) could enhance even more the precision and recall of the approach. Furthermore, there are additional discriminant features that might be extracted from the information source applied in this study that potentially can discriminate between distinct levels of comprehension difficulty when the user is trying to understand the content. For instance, eye gaze features (e.g., saccades, regression, blinks, and fixation duration/count) might be employed in combination with our first proven measures for assessing the cognitive load. On the other hand, we believe that the integration of context information such as the profile of the user (e.g., user's background knowledge, expertise, and reading pattern) might further enhance the approach and minimize false positives. We assume that after further stages of development, we would achieve a technology that provides popup supportive messages, translations, or even retrieve relevant web content whenever it assesses the risk that the person is encountering difficulty at some point (temporal and spatial). This would make the difficult content closer and more organized to the user in real-time. Therefore, capturing users' preferences and fields of interest is necessary to improve the difficulty prediction of TellBack. For instance, TellBack would predict that the most preferred topics to users would be the least expected to exhibit comprehension difficulties and vice versa. Thus, at that stage, an automatic tool is required to guess the topic that is presented in front of the user. This tool would be an intelligent text mining technique such as the proposed in [27]. The work in [27] is a two-stage framework for topic extraction from the scientific literature. Alternatively, TellBack would use efficient supervised clustering techniques for general text classification such as the work presented in [28].

We argue that there is a wide spectrum of usages to this technology that can be exploited in many realms such as analyzing the comprehension behavior of people with neurological disorders like Autism or Alzheimer's Disease. Moreover, this technology can be used to create a scoring system for employees or trainees about their reading and understanding of distinctive degrees of certain task documents or critical reports.

ACKNOWLEDGMENT

The authors would like to thank the volunteers that participated in the experiments.

REFERENCES

[1] N. Carr, *The Shallows: How the Internet is Changing the Way We Think, Read and Remember*. London, U.K.: Atlantic Books Ltd, 2010.

[2] N. Herbig, T. Düwel, M. Helali, L. Eckhart, P. Schuck, S. Choudhury, and A. Krüger, "Investigating multi-modal measures for cognitive load detection in e-learning," in *Proc. 28th ACM Conf. User Modeling, Adaptation Pers.*, Jul. 2020, pp. 88–97.

[3] W. L. Romine, N. L. Schroeder, J. Graft, F. Yang, R. Sadeghi, M. Zabihimayvan, D. Kadariya, and T. Banerjee, "Using machine learning to train a wearable device for measuring students' cognitive load during problem-solving activities based on electrodermal activity, body temperature, and heart rate: Development of a cognitive load tracker for both personal and classroom use," *Sensors*, vol. 20, no. 17, p. 4833, Aug. 2020.

[4] R. Couceiro, G. Duarte, J. Durães, J. Castelhano, C. Duarte, C. Teixeira, M. C. Branco, P. Carvalho, and H. Madeira, "Pupillography as indicator of programmers' mental effort and cognitive overload," in *Proc. 49th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN)*, Jun. 2019, pp. 638–644.

[5] R. Couceiro, G. Duarte, J. Durães, J. Castelhano, C. Duarte, C. Teixeira, M. C. Branco, P. Carvalho, and H. Madeira, "Biofeedback augmented software engineering: Monitoring of programmers' mental effort," in *Proc. IEEE/ACM 41st Int. Conf. Softw. Eng., New Ideas Emerg. Results (ICSE-NIER)*, May 2019, pp. 37–40.

[6] R. Couceiro, P. Carvalho, M. C. Branco, H. Madeira, R. Barbosa, J. Durães, G. Duarte, J. Castelhano, C. Duarte, C. Teixeira, N. Laranjeiro, and J. Medeiros, "Spotting problematic code lines using nonintrusive programmers' biofeedback," in *Proc. IEEE 30th Int. Symp. Softw. Rel. Eng. (ISSRE)*, Berlin, Germany, Oct. 2019, pp. 93–103.

[7] L. E. Sibert and R. J. K. Jacob, "Evaluation of eye gaze interaction," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2000, pp. 281–288.

[8] A. Hyrskykari, *Eyes in Attentive Interfaces: Experiences From Creating iDict, a Gaze-Aware Reading Aid*. Tampere, Finland: Tampere Univ. Press, 2006.

[9] L. Gonçales, K. Farias, B. da Silva, and J. Fessler, "Measuring the cognitive load of software developers: A systematic mapping study," in *Proc. IEEE/ACM 27th Int. Conf. Program Comprehension*, May 2019, pp. 42–52, doi: 10.1109/ICPC.2019.00018.

[10] S. C. Müller and T. Fritz, "Using (bio)metrics to predict code quality online," in *Proc. 38th IEEE/ACM Int. Conf. Softw. Eng.*, May 2016, pp. 452–463.

[11] V. Demberg and A. Sayeed, "The frequency of rapid pupil dilations as a measure of linguistic processing difficulty," *PLoS ONE*, vol. 11, no. 1, Jan. 2016, Art. no. e0146194.

[12] C. Mills, J. Gregg, R. Bixler, and S. K. D'Mello, "Eye-mind reader: An intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering," *Hum.-Comput. Interact.*, vol. 35, pp. 1–27, 2020.

[13] K. Sharma, Z. Papamitsiou, J. K. Olsen, and M. Giannakos, "Predicting learners' effortful behaviour in adaptive assessment using multimodal data," in *Proc. 10th Int. Conf. Learn. Anal. Knowl.*, Mar. 2020, pp. 480–489.

[14] T. Fritz, A. Begel, S. C. Müller, S. Yigit-Elliott, and M. Züger, "Using psycho-physiological measures to assess task difficulty in software development," in *Proc. 36th Int. Conf. Softw. Eng.*, May 2014, pp. 402–413.

[15] N. Nourbakhsh, Y. Wang, and F. Chen, "GSR and blink features for cognitive load classification," in *Proc. IFIP Conf. Hum.-Comput. Interact.* Berlin, Germany: Springer, Sep. 2013, pp. 159–166.

[16] R. Toala, F. Gonçalves, D. Durães, and P. Novais, "Adaptive and intelligent mentoring to increase user attentiveness in learning activities," in *Proc. Ibero-Amer. Conf. Artif. Intell.* Cham, Switzerland: Springer, Nov. 2018, pp. 145–155.

[17] S. Lee, D. Hooshyar, H. Ji, K. Nam, and H. Lim, "Mining biometric data to predict programmer expertise and task difficulty," *Cluster Comput.*, vol. 21, no. 1, pp. 1097–1107, Mar. 2018.

[18] S. G. Hart, "NASA-task load index (NASA-TLX); 20 years later," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*. Los Angeles, CA, USA: Sage Publications, Oct. 2006, vol. 50, no. 9, pp. 904–908.

[19] T.-T. Wong, "Performance evaluation of classification algorithms by *k*-fold and leave-one-out cross validation," *Pattern Recognit.*, vol. 48, no. 9, pp. 2839–2846, Sep. 2015.

[20] H. F. Posada-Quintero, T. Dimitrov, A. Moutran, S. Park, and K. H. Chon, "Analysis of reproducibility of noninvasive measures of sympathetic autonomic control based on electrodermal activity and heart rate variability," *IEEE Access*, vol. 7, pp. 22523–22531, 2019.

[21] M. P. Tarvainen, P. O. Ranta-Aho, and P. A. Karjalainen, "An advanced detrending method with application to HRV analysis," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 2, pp. 172–175, Feb. 2002.

[22] S. Mukherjee, R. Yadav, I. Yung, D. P. Zajdel, and B. S. Oken, "Sensitivity to mental effort and test–retest reliability of heart rate variability measures in healthy seniors," *Clin. Neurophysiol.*, vol. 122, no. 10, pp. 2059–2066, 2011.

[23] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources.," *Psychol. Bull.*, vol. 91, no. 2, pp. 276–292, 1982.

[24] A. T. Duchowski, K. Krejtz, I. Krejtz, C. Biele, A. Niedzielska, P. Kiefer, M. Raubal, and I. Giannopoulos, "The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2018, pp. 1–13.

[25] J. Reilly, A. Kelly, S. H. Kim, S. Jett, and B. Zuckerman, "The human task-evoked pupillary response function is linear: Implications for baseline response scaling in pupillometry," *Behav. Res. Methods*, vol. 51, no. 2, pp. 865–878, Apr. 2019.

[26] K. T. Chui, D. C. L. Fung, M. D. Lytras, and T. M. Lam, "Predicting at-risk university students in a virtual learning environment via a machine learning algorithm," *Comput. Hum. Behav.*, vol. 107, Jun. 2020, Art. no. 105584.

[27] A. Onan, "Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering," *IEEE Access*, vol. 7, pp. 145614–145633, 2019.

[28] A. Onan, "Hybrid supervised clustering based ensemble scheme for text classification," *Kybernetes*, vol. 46, no. 2, pp. 330–348, Feb. 2017.

[29] G. F. Walter and S. W. Porges, "Heart rate and respiratory responses as a function of task difficulty: The use of discriminant analysis in the selection of psychologically sensitive physiological responses," *Psychophysiology*, vol. 13, no. 6, pp. 563–571, Nov. 1976.

HAYTHAM HIJAZI received the B.Eng. degree in computer systems engineering with an excellent grade, in 2012, and the M.Sc. degree in information technology from the University of Stuttgart, Germany.

From 2012 to 2019, he worked with Palestine Ahliya University, Bethlehem, as a Lecturer, a researcher, a Data Centre Director, and a Quality Assurance Manager. He received the DAAD scholarship. Since 2019, he has been a Ph.D. Research Fellow with the Center for Informatics and Systems, University of Coimbra (CISUC), Portugal. He participated in many national and international projects, such as NETKITE, WeDetect, and BASE. His research interests include artificial intelligence, machine learning, biomedical signal processing, and health informatics. His current thesis work is focused on biofeedback intelligent systems for augmenting content comprehension.

RICARDO COUCEIRO received the M.Sc. degree in biomedical engineering the Ph.D. degree in information science and technology from the University of Coimbra, in 2008 and 2015, respectively. Since 2006, he has been a Research Member of the Center for Informatics and Systems, University of Coimbra (CISUC). Since 2006, he participated in several national and international projects, such as MyHeart, Heartcycle, and Welcome. He is the Project Manager of the H2020 project LINK. His publications include several book chapters and over 40 papers in refereed national and international journals and conferences. His research interests include biomedical signal processing, pattern recognition, and modeling applied to the analysis of the cardiovascular systems.

JOÃO CASTELHANO received the Ph.D. degree in health sciences (biomedical sciences). He has a Clinical Investigator and GCP certifications to conduct clinical trials with human participants. He is currently a Biomedical Engineer in the field of medical imaging and radiation. He is also a Junior Researcher in medical engineering with the University of Coimbra. He conducts research and gives technical support to many projects in medical imaging and radiation applied to Neurosciences. His research interests include experimental design, acquisition, and analysis of multimodal imaging data (EEG, fMRI, MRS, PET, TMS, ECoG, and fNIRS).

PAULO DE CARVALHO is currently an Associate Professor with Habilitation with the Faculty of Science and Technology, University of Coimbra. His main research interests include intelligent algorithms for personal health solutions for chronic cardiovascular and respiratory diseases and clinical decision support systems. His publications include several book chapters and over 200 papers in refereed international journals and conferences. He has been involved in several organizations and program committees of international conference in the health informatics domain. He has served as a project and a work package coordinator, as well as a team leader in several national, such as FCT, P2020, and EU projects, such as FP7, H2020, EIT-Health. He raised over 5.5 million € in competitive funding from national, EU programs and industry, such as Philips, AlticeLabs. He is the Director of the Working Group on Health Informatics of the IFMBE. He serves as an Associate Editor for the IEEE JOURNAL ON BIOMEDICAL AND HEALTH INFORMATICS.

MIGUEL CASTELO-BRANCO received the Ph.D. degree from the Max-Planck Institute for Brain Research, Frankfurt, Germany. He is currently the Director of the Coimbra Institute for Biomedical Imaging and Translational Research (CIBIT). He has been the Director of IBILI, a leading Vision Research Institute in Portugal. He is the Scientific Coordinator of the National Functional Brain Imaging Scientific initiative. He was also the Director of ICNAS, Medical Imaging Infrastructure, University of Coimbra. Under his leadership, IBILI was classified as an Excellent Research Unit by international evaluation panels. He is currently an Associate Professor with the University of Coimbra. He has held a Professorship in Psychology, University of Maastricht, The Netherlands, in 2000. He was a Postdoctoral Fellow with the Max-Planck-Institute for Brain Research, Germany, where he had also performed his Ph.D. work, from 1994 to 1999. His achievements are well reflected in publications in top General Journals, such as *Nature* and *PNAS*, Top Clinical Translational research journals, such as the *Journal of Clinical Investigation*, *Brain*, *Human Molecular Genetics*, as well as others in the field of vision research, such as *Investigative Ophthalmology and Visual Sciences*, *Journal of Vision*, *Vision Research Archives of Ophthalmology*, *Ophthalmology*, Human Neurophysiology and Neuroscience, such as the *Journal of Neuroscience*, the *Journal of Neurophysiology*, *Human Brain Mapping*, *Neuroimage*, *Cerebral Cortex*, *Neuron*, and others.

HENRIQUE MADEIRA is currently a Full Professor with the University of Coimbra, where he has been involved in research on dependable computing, since 1989. He has coordinated or participated in tens of projects funded by the Portuguese Government and by the European Union. His main research interests include experimental evaluation of dependable computing systems, including security evaluation and benchmarking, fault injection techniques, error detection mechanisms, and transactional systems dependability, subjects on which he has authored or coauthored more than 180 papers in refereed conferences and journals (Google Scholar profile). From 1999 to 2002, he was the Vice-Chair of the IFIP Working Group 10.4 Special Interest Group (SIG) on Dependability Benchmarking from the establishment of the SIG. He was a Program Co-Chair of the International Performance and Dependability Symposium track of the IEEE/IFIP International Conference on Dependable Systems and Networks, DSN-PDS2004. He has organized several workshops and scientific events. He was a Conference Coordinator of the IEEE/IFIP DSN, the major conference of the dependability area, in 2008. He has also been asked to be a referee for many international conferences and journals. He has served on program committees of the major conferences of the dependability and database areas.