HCIS | Human-centric Computing and Information Sciences

RESEARCH (Research Manuscript) Open Access

Human-centric Computing and Information Sciences (2023) 13:27
DOI: https://doi.org/10.22967/HCIS.2023.13.027
Received: May 2, 2022; Accepted: July 20, 2022; Published: Jun 30, 2023

Cognitive Load Recognition Based on T-Test and SHAP from Wristband Sensors

Jeong-Kyun Kim, Kangbok Lee, and Sang Gi Hong*

Abstract

The understanding of cognitive load plays a key role in increasing the potential of human-centric systems. Recently, cognitive load detection has attracted the attention of researchers. Competitions are being held and relevant data are being released in this regard. Managing cognitive load through wearable devices in daily life contributes, amongst others, to industrial safety. Wearable bands require high prediction results with less data because of their limited battery and processing power. Therefore, by detecting important features that characterize cognitive load, we aimed to obtain a high load detection classification accuracy using few features. In total, we detected 179 features such as heart rate variabilities, descriptive statistical, and frequency features. Important features were detected using the independent t-test and SHapley additive exPlanation (SHAP). Furthermore, an accuracy of 70.3% was obtained with only ten important features using the LightGBM classifier. Heart rate variability and galvanic skin response were used as the important features. Additionally, the discrete wavelet transform was used as a more important frequency-domain feature than the discrete cosine transform. The proposed cognitive load detection method achieved higher accuracy with fewer features using a lighter classifier than those reported by existing CogLoad data studies.

Keywords

Cognitive Load, SHAP, XAI, Wearable Sensors, Mental Fatigue, Stress

1. Introduction

Cognitive load refers to the level of mental effort needed when a user performs a task that requires learning or decision-making. Moderate cognitive load increases productivity; however, cognitive overload may lead to accidents in daily life [1, 2]. Predicting and quantifying cognitive load via human-centric systems that aim to adapt automatically to a user's cognitive state can be useful in various applications for human-computer interaction [3]. Cognitive load can be used as a major marker to avoid stress, fatigue, frustration (generally caused by cognitive overload), and boredom (occurring at low cognitive load levels) [4].

Currently, questionnaires such as the National Aeronautics and Space Administration Test Load Index (NASA-TLX) [5], subjective rating scale [6], and subjective mental effort rating scale [7] are used to measure cognitive load. A questionnaire is a subjective evaluation method based on user feedback. A measurement

***Corresponding Author:** Sang Gi Hong (sghong@etri.re.kr)
Intelligent Convergence Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Korea

method that employs a questionnaire is easy to evaluate and inexpensive. However, the results obtained using such methods are difficult to quantify because of the high possibility that each item is biased by the subjective viewpoint of the participants, including the measurer's task, behavior, and work memory. Measurement by questionnaire requires the participant to be capable of responding to each item [8].

To consider the shortcomings of the subjective evaluation of cognitive load, previous studies have investigated various biosignals, such as electroencephalogram (EEG), heart rate (HR), respiration rate, skin conductance, and blinking [3, 4]. When undergoing physical or mental loading, the sympathetic nervous component of the autonomic nervous system is stimulated. When the sympathetic nervous system is stimulated, the cardiovascular system and organs become more responsive. This causes the HR, blood pressure, and blood flow to increase [9].

Cognitive load recognition using smart devices can be added as a function of smartwatches, which have recently become common to be used in industrial, defense, and disaster response settings to reduce the risks caused by cognitive overload. Cognitive load research using smart devices has been conducted to detect handcrafted time and frequency features from the HR, R-R interval (RRI), galvanic skin response (GSR), and temperature, as well as to infer cognitive load using machine learning algorithms. By using smartwatch biosignals, during the CogLoad@UbiComp2020 competition, cognitive load was detected via various signal processing, feature selection, and machine learning techniques [10]. The CogLoad dataset thus obtained has been used for subsequent research on cognitive load [3, 4].

Recently, studies [10, 11] using deep learning have been conducted; however, the accuracy has not been impressive. Therefore, higher-level models should be developed. Particularly, it is difficult to commercialize smart-watch-based cognitive load evaluation using high-resolution signals because of the amount of data transmission needed. Therefore, a method that uses metadata to detect cognitive features is required.

Feature selection has been used to detect the cognitive load, reduce the model size, and obtain high accuracy. Recently, studies using explainable artificial intelligence (XAI) have attracted attention for the interpretation of the results of machine learning techniques, and the XAI algorithm has been used to evaluate the importance of features and their effects [12, 13].

In summary, this study makes multi-fold contributions:

- We use CogLoad data for validation to propose state-of-the-art approaches with higher accuracy than those reported in previous studies [3, 4].
- We propose an algorithm that acquires a variety of features from smart devices and effectively detects important features and combines statistical analysis and XAI.
- Using XAI algorithms, we analyze the importance and contribution to cognitive load of the biological signals acquired from smart bands.
- Using significant features obtained by the proposed method, we obtain high accuracy with fewer features, helping to manage cognitive loads in everyday life even in circumstances where the processing power of wearable devices is limited.

The rest of this paper is organized as follows: Section 2 provides an overview of the related studies on biosignal-based cognitive load and approaches using CogLoad data. Section 3 describes the CogLoad data, the proposed approaches, and the evaluation method in detail. Section 4 presents the results obtained using the proposed method. Section 5 discusses biosignal features for cognitive load. Section 6 presents the conclusions, limitations, and directions for future studies.

2. Related Work

Cognitive load, stress, and fatigue are a collective set of text features because cognitive load leads to stress and fatigue [14]. The study of biosignals related to cognitive load employs a similar approach to that of studying stress, fatigue, and emotion [15]. This involves acquiring biosignals from external stimuli and analyzing them to detect the resulting biosignal changes. Recently, studies have been conducted to classify the state of external stimuli, using artificial intelligence [16].

The heartbeat interval and GSR are the biosignals mainly used in the detection of stress and cognitive load by using wearable devices. The RRI, which is the interval between heartbeats [17], has a steadily changing pattern even in stable situations [4]. In general, the more stable the state, the larger and more complex the changes between the heartbeats. Conversely, when under stress or during physical activity, the RRI shows a regular and constant pattern. When the cognitive load increases, the sympathetic nervous system causes sweating, and consequently, the GSR increases [18]. Heart rate variability (HRV) is yet another important index for stress analysis, and stress is evaluated by detecting the HRV features in the time-frequency domain [19]. The mean of RRI (mRR), the standard deviation of RRI (SDRR), and the root mean square (RMS) of successive differences (RMSSD) are used as time features. The signal power in the very low frequency (VLF), low frequency (LF), high frequency (HF) bands, and the LF/HF ratio are used as frequency features. Hao [9] collected photoplethysmogram (PPG) signals to monitor mental load and detected HRV from PPG to obtain linear features (time domain and frequency domain) and nonlinear features—Poincaré plot, scatter plot, and sample entropy (SampEn) analysis, while Pearson correlation and t-test results showed that there were useful features such as HR, mRR, RMSSD, HF, etc.

Anusha et al. [20] used an electrodermal activity (EDA), electrocardiogram (ECG), and skin temperature (ST) to monitor work stress. A maximum work stress recognition accuracy of 97.13% was calculated using a combination of EDA and ST with data segments of 60s duration. Pettersson et al. [21] used electro-oculogram (EOG) and ECG signals to detect cognitive stress during the Maastricht Acute Stress Test. The results show that the best performance is achieved when the features of both EOG and ECG signals are employed and detected using support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGBoost) for classification in two states (relaxed and stressed); SVM achieved the best results. Fan et al. [22] investigated the effects of mental workload on EEGs and ECGs. Three levels of mental workload were obtained by monitoring tasks with different levels of difficulty. Among the ECG features, mRR, RMSSD, HF_norm, and SampEn decreased significantly as the task difficulty increased, and LF_norm and LF/HF showed significant increases. Principal component analysis was used to reduce the dimensions of the features, and the results were used as inputs to the SVM, achieving an accuracy of 80%.

Table 1. Existing studies on biosignal-based cognitive load recognition

Study	Signals (Hz)	Window lengths (sec)	Model	Accuracy
Anusha et al. [20]	ECG(64), GSR(1000), ST(1000),	30–300	LDA, kNN, QDA, SVM	0.97
Pettersson et al. [21]	ECG(1000), EOG(1000),	45	SVM, RF, XGBoost	0.97
Fan et al. [22]	EEG(512), ECG(1000)	360	SVM	0.80
Borisov et al. [4]	ST(1), GSR(1), HR(1), HRV(1)	30	Ensemble model	0.66
Tervonen et al. [3]	ST(1), GSR(1), HR(1), HRV(1)	30	XGBoost	0.67

As shown in Table 1, studies [20–22] report high accuracy in cognitive load detection but did not use wearable devices and have high sample rates. A competition for cognitive load detection using wearable devices was held at UbiComp 2020 to propose and evaluate wearable device-based cognitive load algorithms. Gjoreski et al. [10] reported the cognitive load recognition results of combining feature detection, feature selection, and classifiers of various methods for 13 teams participating in the CogLoad competition. Using the CogLoad dataset, RRI, GSR, ST, and HR were detected by the wearable band, and 825 instances obtained from 23 participants were provided. Features including handcrafted feature extraction, statistical (such as mean, variance, kurtosis, median, and sum), and frequency (such as the power spectral density ratio of HRV) features were detected. The numbers of detected features varied between 4 and 129. Some participants selected important features using maximal information coefficients, a sequential backward floating search, and Gini impurity. SVM, RF, XGBoost, and light gradient boosting machine (LightGBM) were used as the classifiers [23]. The model using the ensemble yielded particularly high classification results. Borisov et al. [4] achieved the highest competitive performance. In the proposed

method, descriptive statistical features of minimum, maximum, mean, standard deviation, sum, and skew were obtained for the GSR, HR, RRI, and ST signals, and the cognitive load was detected using eight LightGBM ensemble models. The proposed model's accuracy was 66%.

Tervonen et al. [3] presented a comparative accuracy analysis of ultrashort (<30 seconds) window lengths in cognitive load sensing using wearable devices and CogLoad data. In total, 157 features such as the statistical features of each signal, one and two derivatives, and HRV features, were obtained and used as inputs to the XGBoost classifier. The 25-second window showed the highest accuracy (67.6%), similar to previous studies using the same dataset. Overall, the model accuracy tended to decrease as the window length decreased with the lowest performance (60.0%) observed with the 5-second window.

3. Materials and Methods

The proposed method detected 179 features such as HRV, descriptive statistical, and frequency features for HR, RRI, GSR, and ST detected with the wristband; 19 time-frequency features were detected as HRV features, while 10 statistical and 30 frequency features were obtained from the four signals. Important features were detected to ensure that the cognitive load was recognized at a high rate with fewer features in the wristband signal. Important-feature detection algorithms were selected using XAI and statistical-based feature selection algorithms. Independent t-test, Boruta, and SHapley additive explanation (SHAP) were combined for use as the feature selection method. Furthermore, RF, XGBoost, and LightGBM, which are tree ensemble models, were used as the backbone models for SHAP. Cognitive load was detected using SVM, RF, XGBoost, LightGBM, and ensemble (LightGBM), with the important features as the input (Fig. 1).

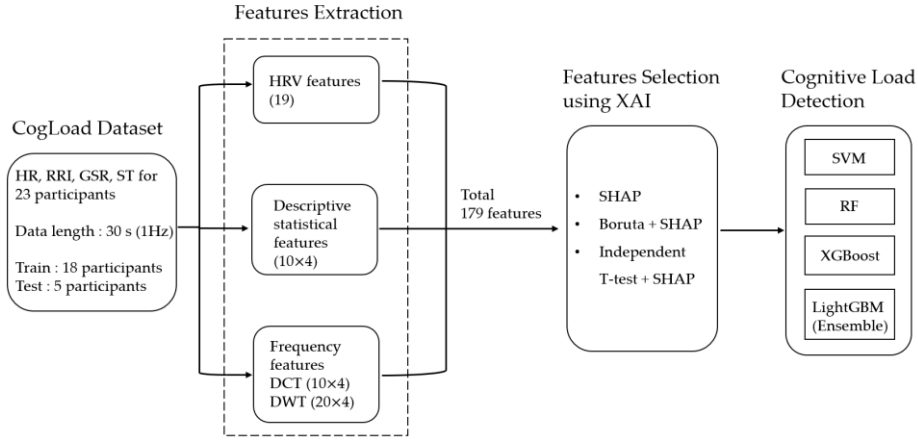


Fig. 1. Cognitive load detection flowchart.

3.1 CogLoad Data

CogLoad data [10] is a dataset obtained from the RRI, HR, GSR, and ST values of 23 participants. N-back tasks (n=2, 3) were performed as cognitive capacity tests, and biosignals were acquired during cognitive load estimation tests in addition to six elementary cognitive tasks (ECTs) [24]. The six ECTs were the Gestalt completion, hidden pattern, Finding A's, number comparison, pursuit, and scattered X tests. Variously challenging (easy, medium, and difficult) cognitive load estimation tests were conducted. After each task, the cognitive load was assessed using the NASA Task Load Index (NASA-TLX) [5]. The participant was asked to rest for 3 minutes after completing the questionnaire and the biosignals were obtained using a commercial wristband worn on the non-dominant arm during each task. A total of 825 instances were recorded, the rest periods were labeled as "no load," and the task periods were labeled as

"cognitive load." The dataset was divided for training and testing; 632 instances from 18 participants were used for the training. Each instance consisted of four 30-second signals, and the signal was provided at the sampling rate of 1 Hz.

3.2 Feature Extraction

To detect the cognitive load, HRV, descriptive statistics, and frequency features were detected from the CogLoad data. For the HRV features, 19 time-frequency features were detected in the RRI data [14, 25]. The HRV parameters are listed in Table 2. Descriptive statistical and frequency features were detected for all four signals. The 10 descriptive statistical parameters [12] were max, min, SD, absolute sum of values (AbSum), RMS, kurtosis, skewness, gradient from maximum value to minimum value (MMgr), difference between maximum value and minimum value (DMM), and maximum for the difference between two successive values (Mdif). The frequency features of the signals were selected from the coefficients of the discrete cosine transform (DCT) [26] and discrete wavelet transform (DWT) [27].

Table 2. HRV features

	Abbreviation	Definition
Time features	mHR	Mean of heart rate
	SDHR	Standard deviation of heart rate
	mRR	Mean of RRI
	SDRR	Standard deviation of RRI
	CVRR	Coefficient of variance of RRI
	SDSD	Standard deviation of differences between adjacent RRI
	RMSSD	Root mean square of the successive differences
	RR50	Number of pairs of adjacent RRI differing by more than 50 ms in the entire recording
	pRR50	Percentage of RR50
Frequency features	aVLF	Average of very low frequencies
	aLF	Average of low frequencies
	aHF	Average of high frequencies
	pVLF	Percentage of very low frequencies
	pLF	Percentage of low frequencies
	pHF	Percentage of high frequencies
	nLF	aLF/(aLF+aHF)
	nHF	aHF/(aLF+aHF)
	LF/HF ratio	aLF/aHF
	dLFHF	abs(pLF-pHF)

DCT is commonly used for data compression because of its greater capacity to concentrate the signal energy into a few transform coefficients. Contrary to the Fourier transform, DCT uses only real numbers. Ahmed et al. [26] defined the DCT as follows:

$$F[j] = w[j] \sum_{k=1}^{n} x[k] \cos\left\{\frac{\pi}{2n}(2k-1)(j-1)\right\}, \quad (1)$$

$$w[j] = \begin{cases} \frac{1}{\sqrt{n}} & j = 1 \\ \frac{\sqrt{2}}{\sqrt{n}} & j = 2, 3, \ldots, n \end{cases}, \quad (2)$$

where $x[k]$ is the biosignal, and $F[j]$ is the DCT coefficient. After performing DCT, ten low-frequency coefficients were selected because most of the biosignal information tends to be concentrated in a few low-frequency components of the DCT.

In 1982, geophysicist Jean Morlet first introduced wavelet transform as a family of functions constructed from translations and dilatations of a single function known as the “mother wavelet” [27]. Wavelet transform has been widely used in the field of biomedical signals. DWT decomposition consists of low- and high-pass filters (LPF and HPF) that divide the signal information by half. The LPF and HPF coefficients are the approximate and detailed components, respectively. The approximation signal was decomposed into a hierarchical set of approximate and detailed components [28]. Daubechies six (db6) were used as mother wavelets for DWT composition:

$$A[n] = \sum_k x[k] \times g[2n - k], \tag{3}$$

$$D[n] = \sum_k x[k] \times h[2n - k], \tag{4}$$

where $g[n]$ is the half-band LPF, $h[n]$ is the half-band HPF, $A[n]$ is the output LPF (the approximation components) and $D[n]$ is the output HPF (the detail components). Moreover, $x[n]$ denotes the original biosignal (Fig. 2).

Considering the DWT features, min, mean, max, variance, and AC variance were obtained for LL, HL, LH, and HH. A total of 179 features were detected by obtaining 10 DCT coefficients and 20 DWT features for the four signals, including 19 HRV and 40 descriptive statistics features.

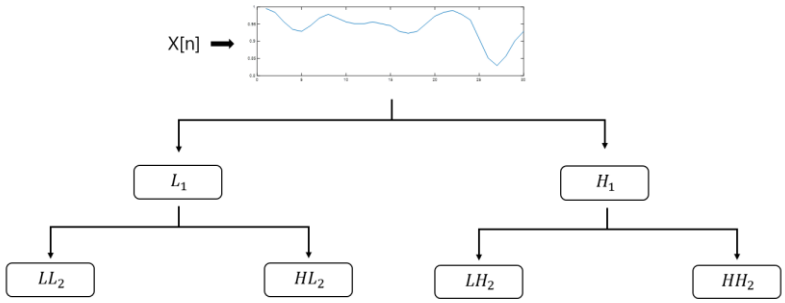


Fig. 2. DWT decomposition.

3.3 Feature Selection

Algorithms for detecting important features include statistical or machine learning-based methods. Considering the statistical method, the validity of each feature difference in the two target groups can be verified by performing an independent t-test on the two target groups and removing unnecessary features [29]. Machine learning methods are classified into model-specific and agnostic methods based on dependency. A model-specific method using a tree-based model includes a method that uses the gain value within the tree and a permutation method based on a variable subset [30]. Permutation randomly changes the value of a feature and measures the effect of the change on the prediction. Among the model-specific methods, the Boruta algorithm is specially designed for feature selection [31]. When the model does not have explanatory power, the prediction results are interpreted post-hoc. Most interpretable techniques in machine and deep learning are post-hoc. Moreover, SHAP is a model-agnostic and post-hoc algorithm that has attracted attention [12]. For detecting fatigue, feature selection was performed using an independent t-test, Boruta, and SHAP.

The applied Boruta algorithm was designed as a wrapper for RF classification [31]. The Boruta algorithm steps include the following:

1. A new column (shadow features) is created by copying the original sample features.
2. The shadow features are randomly mixed.
3. RF is performed on the shadow features.
4. Max Z-score among shadow attributes (MSZA) is calculated as the largest of the Z values obtained by RF.

5. The original feature also extracts the Z value using RF.
6. If the Z value is larger than MSZA, the feature is considered important.
7. Repeat steps 1 to 6 to obtain statistically significant progress.

The Boruta algorithm evaluates the relevance of variables in the information system by comparing the importance measure provided by the RF to that obtained for artificially added random attributes.

SHAP is a method of interpreting results from machine learning models [12]. In addition, SHAP is used for interpreting the contribution of features in a model. A feature is selected by assessing its contribution and its positive and negative influence on the model are considered. SHAP was proposed by Lundberg et al. [32] and was based on the concept of the Shapely value in game theory.

The main advantages of the SHAP method are the local explanation and consistency of the global model structure. The SHAP value is a numerical expression of the contribution of each feature to the total outcome. The contribution of each feature can be expressed as the degree of change in the total outcome when the contribution of that feature is excluded. Equation (5) represents the SHAP value, where $\emptyset_i$ is the SHAP value for the data, n is the feature, S represents all the features (excluding the i-th feature), $v(S)$ is the contribution to the result (excluding the i-th feature), and $v(S \cup \{i\})$ is the contribution of all features, including the i-th feature. The degree of contribution of the i-th feature is the value obtained by subtracting the sum of the contributions, excluding the i-th feature from the total contribution:

$$\emptyset_i(v) = \sum_{S \epsilon N\{i\}} \frac{|S|!(n-|S|-1)!}{n!} \big(v(S \cup \{i\}) - v(S)\big). \quad (5)$$

3.4 Predictive Method for Cognitive Load Detection

Cognitive load was detected using various classification models, such as SVM, RF, XGBoost, LightGBM, and ensemble models, using features obtained from biosignals as inputs. SVM is a binary classifier that determines the optimal separation hyperplane that maximizes the margin between two classes [33]. Before the advent of deep learning and ensemble classifiers, it was used in various fields as the best-performing model. RF is an ensemble classifier with a bagging method that combines multiple trees to obtain the results of each classification model using multiple votes or weighted averages [34]. Because RF learns using data and features from the training data, it lowers the risk of overfitting. XGBoost is a model for boosting decision trees that improves the speed of gradient boosting machines. Boosting models generate robust classifiers by iteratively updating the parameters of previous classifiers to increase their accuracy and reduce the slope of the loss function [34]. LightGBM considers the short-comings of GBMs' one-sampling and exclusive feature bundling techniques. It also produces better performance and processes faster than XGBoost in various fields [35].

The SVM parameters were kernel = linear, gamma = 1.0, and C = 5.0. The RF parameters were number of trees = 50, max_depth = 30, and number of features = square root of the number of proposed features. The XGBoost parameters were booster = gbtree, objective = binary:logistic, eta = 0.018, max_depth = 15, gamma = 0.009, subsample = 0.98, and colsample_bytree = 0.86. The LightGBM parameters were objective = binary, num_leaves = 544, max_bin = 22, colsample_bytree = 0.320, learning_rate = 0.220, and lambda = 76.

3.5 Evaluation

For evaluation, 632 instances from 18 participants were used as the training data, and 193 instances from five participants were used as the test data. In the previous studies based on CogLoad data, either k-fold cross-validation, or leave-n-subjects-out was used as the evaluation strategy. We used the same evaluation method as in [4]; the highest performance during the challenge was achieved using this method. In this method, five-fold cross-validation was used for learning, and the accuracy and ROC-AUC were obtained using test data.

4. Results

Fig. 3 shows the results of detecting the cognitive load using SVM, RF, XGBoost, and LightGBM, while increasing the number of features following the decreasing order of *p*-values for 179 features based on the t-test results. For SVM the accuracy decreased at a certain threshold as the number of features increased. RF showed the highest accuracy when the number of features was 82; nonetheless, there was a large deviation based on feature selection. As the number of features increased to 23, the accuracy of XGBoost and LightGBM increased linearly, and LightGBM exhibited higher accuracy than XGBoost.

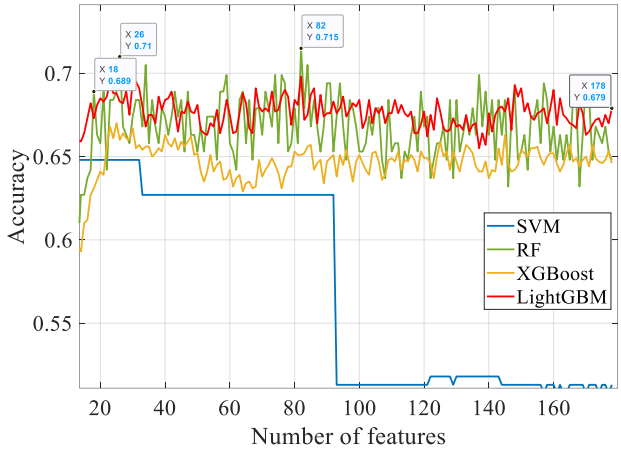


Fig. 3. T-test-based cognitive load detection results.

Fig. 4 shows the classification results obtained using SVM, RF, XGBoost, and LightGBM following the order of the highly important features, using SHAP. The SHAP method is model-agnostic and is used to obtain feature importance based on RF, XGBoost, and LightGBM. The performance of the RF and XGBoost-based SHAP were not significantly impacted by the number of features. The SHAP-based RF classification results exhibited a large deviation similar to that in the t-test result. The SHAP of the LightGBM produced highly accurate stable results, even with a small number of features. Particularly, when using only 100 features with small *p*-values of the independent t-test (Ttest-SHAP) the accuracy of the SHAP results deviated little when more features were added. Using Boruta-SHAP, the XGBoost-based method could not detect the important features because with few features, the method's accuracy was low, and LightGBM-based methods obtained results similar to those of the simple SHAP.

When used as an ensemble classifier as proposed by Borisov et al. [4], the results of classification by adding important features obtained from SHAP, Ttest-SHAP, and Boruta-SHAP (in the order of importance) are shown in Fig. 5. The ensemble model has high cognitive load detection accuracy with few features; nevertheless, it shows that the accuracy decreases as more features are added.

Table 3 compares the Borisov and Tervonen methods with ours when using the proposed 179 feature. A higher ROC-AUC is obtained when using the features of the proposed method, compared to previous study results. Considering the various feature importance detection methods shown in Figs. 4 and 5, the importance detection methods showing the highest performance are LightGBM-based SHAP and Ttest-SHAP. Table 4 shows the results of the single and ensemble LightGBM based on the increase in important features obtained by SHAP and Ttest-SHAP. The important features used in Table 4 are presented in Table 5. The single LightGBM obtains high accuracy, and the ensemble LightGBM obtains a high ROC-AUC. SHAP was most accurate when ten features were used as the input. However, Ttest-SHAP's accuracy and ROC-AUC were stable when the important features were increased from 10 to 20.

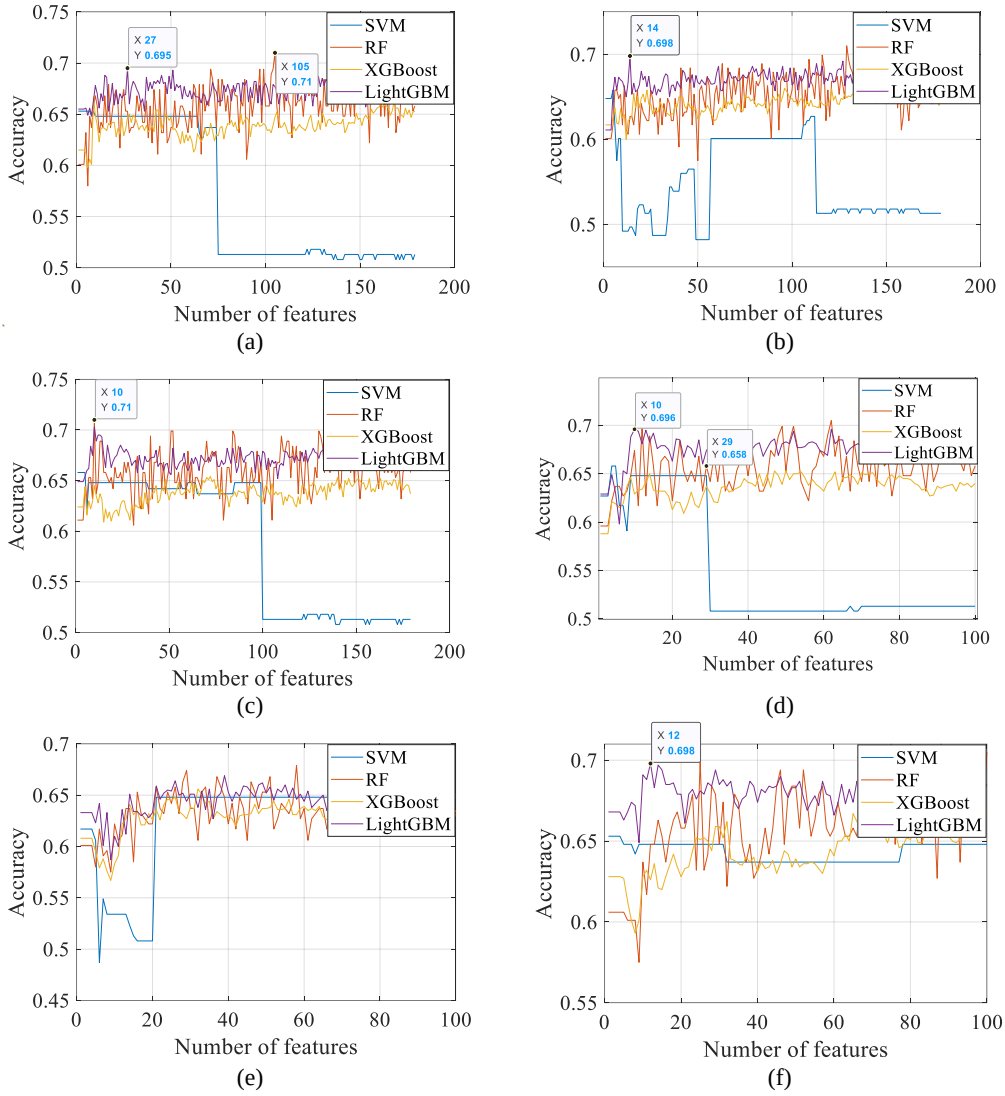


Fig. 4. Cognitive load detection result based on SHAP, Boruta, and t-test combinations: (a) SHAP with RF, (b) SHAP with XGBoost, (c) SHAP with LightGBM, (d) Ttest-SHAP with LightGBM, (e) Boruta-SHAP with XGBoost, and (f) Boruta-SHAP with LightGBM.

Table 3. Cognitive load recognition results using the proposed 179 features

Study	Feature	Model	Evaluation strategy	Evaluation	
				Accuracy	ROC-AUC
Tervonen et al. [3]	HRV, Descriptive statistical, Frequency	XGBoost	Leave-two-subjects-out	0.67	-
Borisov et al. [4]	Descriptive statistical	RF	5-fold cross-validation	0.64	0.67
		Ensemble model		0.66	0.69
Proposed	HRV, Descriptive statistical, Frequency	SVM	5-fold cross-validation	0.52	0.52
		RF		0.67	0.67
		XGBoost		0.64	0.64
		LightGBM		0.66	0.66
		Ensemble model		0.66	0.71

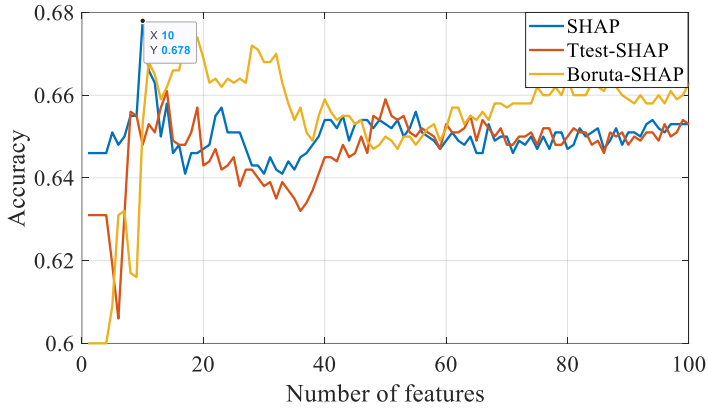


Fig. 5. Cognitive load detection results based on ensemble LightGBM.

Table 4. Cognitive load recognition results using the feature importance method

Method	Model	Evaluation	Number of important features						
			4	5	6	7	8	9	10
SHAP	LightGBM	Accuracy	0.649	0.654	0.660	0.662	0.684	0.675	0.703
		ROC-AUC	0.649	0.655	0.661	0.663	0.684	0.675	0.703
	Ensemble	Accuracy	0.646	0.651	0.648	0.650	0.655	0.655	0.678
		ROC-AUC	0.679	0.680	0.679	0.686	0.693	0.693	0.721
Ttest-SHAP	LightGBM	Accuracy	0.649	0.630	0.598	0.653	0.650	0.688	0.696
		ROC-AUC	0.649	0.631	0.598	0.653	0.650	0.688	0.696
	Ensemble	Accuracy	0.631	0.619	0.606	0.631	0.656	0.655	0.648
		ROC-AUC	0.666	0.660	0.648	0.669	0.700	0.708	0.707

Table 5. Important features of cognitive load obtained with SHAP

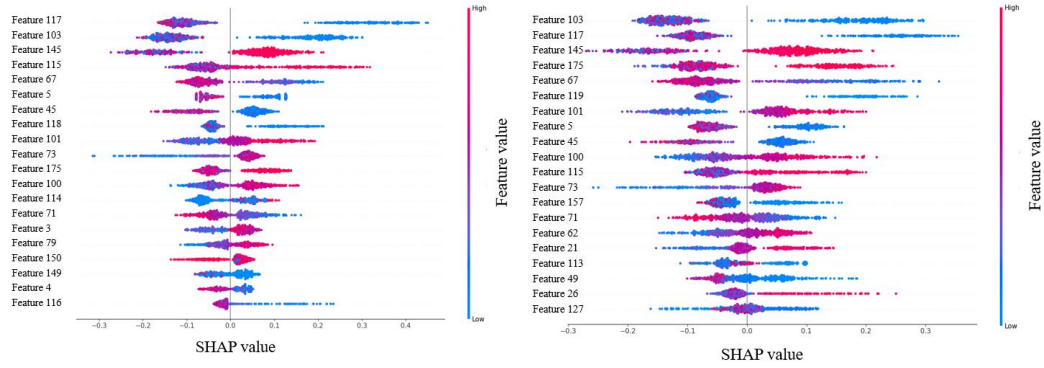
Method	Important features											
	1	2	3	4	5	6	7	8	9	10	11	12
SHAP	117	103	145	115	67	5	45	118	101	73	175	100
Ttest-SHAP	103	117	145	175	67	119	101	5	45	100	115	73
Boruta-SHP	5	103	62	114	115	100	176	67	153	82	91	6

5. Discussion

A wristband-based cognitive load recognition algorithm is proposed to manage cognitive load in daily life using biosignals acquired using smartwatches and bands. Smart wearable devices must function with minimal data due to processor performance limitations. Therefore, it is difficult to process signals obtained at high sampling rates and it is necessary to increase cognitive load recognition accuracy by detecting important features from biosignals obtained at low sampling rates. To detect important features related to cognitive load, the independent t-test, Boruta, and SHAP were combined. SHAP shows the best performance in interpreting handcrafted feature-based machine learning. However, when the number of features increases compared to the amount of data, it leads to errors in calculating the features' importance and effect. Consequently, when important features are applied to the classifier, the accuracy further deviates according to the number of features. To solve this problem, SHAP was applied after removing features without significant differences using the t-test. Accordingly, the accuracy variation, as determined by the number of features, decreased. The CogLoad dataset was used to verify the proposed

method. This dataset has recently been used by researchers to validate cognitive load recognition algorithms using commercial smart bands. The proposed method obtained higher accuracy with CogLoad than that achieved in existing studies [3, 4, 10].

The importance of biosignal features in recognizing cognitive load was shown when Tervonen et al. [3] demonstrated the high importance of HRV-related features after analyzing 157 features such as descriptive statistics of the biosignals, HRV, and skin conduction response features using Gini impurity of XGBoost. Borisov et al. [4] analyzed the importance of the descriptive statistical features of biosignals using SHAP, and the results showed HR's skew, min, median, and HRV's std as being highly important. Hao et al. [9] analyzed the linear and nonlinear features of HRV and showed the importance of linear features such as HR, NN, RMSSD, and HF. The feature importance and contribution were analyzed using SHAP and Ttest-SHAP for the proposed HRV, technical statistics, and frequency features. These results are shown in Table 5 and Fig. 6.



(a) (b)

Fig. 6. (a) SHAP plot with LightGBM and (b) Ttest-SHAP plot with LightGBM.

In Fig. 6, an analysis of the SHAP plot shows the high and low feature values respectively indicated in red and blue. The right side of the X-axis represents a positive contribution, indicating that the cognitive load contributes toward the right of X. The top-five features of SHAP were Feature 117 (RRI, DWT, HH2, max), 103 (RRI, DWT, LL2, variance), 145 (GSR, DWT, HL2, min), 115 (RRI, DWT, HH2, min), and 67 (RRI, DCT, coefficient 8). The top five features of the Ttest-SHAP were Feature 103 (RRI, DWT, LL2, variance), Feature 117 (RRI, DWT, HH2, max), 145 (GSR, DWT, HL2, min), 175 (ST, DWT, HH2, min), and 67 (RRI, DCT, coefficient 8). In the two methods for analyzing feature importance, features 103, 117, and 145 showed high contributions. Feature 103 is characteristic of the low-frequency region of the heartbeat, indicating that the change in the RRI is small during cognitive load. Feature 117 has the maximum value of the high-frequency component of RRI. When the maximum value is large, HRV is large; likewise, when the maximum value is small, HRV is small. Therefore, when Feature 117 is reduced, the cognitive load makes a positive contribution, indicating that the HRV is small. Feature 145 shows the minimum value of the low-frequency value of the GSR change; furthermore, as the value increases, the risk of cognitive load increases. As shown elsewhere [3, 4, 9], this study confirms that HRV can be used as an important cognitive load recognition feature and that DWT features are more effective for detecting cognitive load features than HRV time-frequency and descriptive statistics features.

6. Conclusion

This study proposed an algorithm for detecting cognitive load using biosignals tracked by wristbands. Algorithms that achieve high cognitive load recognition accuracy with minimal processes are required because a wristband has limited processing power and battery capacity. Therefore, it is necessary to detect

important cognitive load recognition features. In the CogLoad@UbiComp2020 competition, various feature detection and machine learning techniques were proposed to detect the cognitive load. We proposed 179 features in total and important-feature detection methods while using fewer features to obtain higher accuracy than previous CogLoad-based studies. As a method for detecting important features, SHAP has been attracting scholarly attention. Although SHAP is a good feature importance detection algorithm, when there are several features, the accuracy of importance detection is poor. By using an independent t-test to reduce the number of features, SHAP achieves a more stable cognitive load detection performance.

High accuracy and ROC-AUCs were obtained using the proposed features. Even with only 10 important features obtained using Ttest-SHAP, the cognitive load detection performance was higher than that of previous studies. If the important features are carefully selected, good performance can be achieved using a single LightGBM rather than an ensemble. Additionally, SHAP was used to verify the contribution of features to cognitive load detection models. For analyzing the CogLoad data using Ttest-SHAP, HR change and GSR were used as the important features during a cognitive load, and the DWT feature was used as being more important than the DCT. Notably, the current study was limited to CogLoad data and did not include a clinical interpretation of the important features detected. Therefore, in the future, we intend to apply the proposed algorithm to various types of open data including clinical interpretations.

Author's Contributions

Conceptualization, JKK, KL, SGH. Methodology, JKK, KL, SGH. Software, JKK. Validation, JKK, SG Hong. Formal analysis, JKK. Investigation, JKK. Resources, KL, SGH. Data curation, JKK. Writing of the original draft, JKK. Writing of the review and editing, JKK. Visualization, JKK. Supervision, KL, SGH. Project administration, KL, SGH. Funding acquisition, KL, SGH.

Funding

This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00751). This research was supported by Korea Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea governments (KCG, MOIS, NFA) (No. RS-2022-001549812, Development of technology to respond to marine fires and chemical accidents using wearable devices).

Competing Interests

The authors declare that they have no competing interests.

References

[1] J. C. Castro-Alonso, B. B. de Koning, L. Fiorella, and F. Paas, "Five strategies for optimizing instructional materials: Instructor-and learner-managed cognitive load," *Educational Psychology Review*, vol. 33, no. 4, pp. 1379-1407, 2021.

[2] A. R. Conway and R. W. Engle, "Working memory and retrieval: a resource-dependent inhibition model," *Journal of Experimental Psychology: General*, vol. 123, no. 4, pp. 354-373, 1994.

[3] J. Tervonen, K. Pettersson, and J. Mantyjarvi, "Ultra-short window length and feature importance analysis for cognitive load detection from wearable sensors," *Electronics*, vol. 10, no. 5, article no. 613, 2021. https://doi.org/10.3390/electronics10050613

[4] V. Borisov, E. Kasneci, and G. Kasneci, "Robust cognitive load detection from wrist-band sensors," *Computers in Human Behavior Reports*, vol. 4, article no. 100116, 2021. https://doi.org/10.1016/j.chbr.2021.100116

[5] N. von Janczewski, J. Kraus, A. Engeln, and M. Baumann, "A subjective one-item measure based on NASA-TLX to assess cognitive workload in driver-vehicle interaction," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 86, pp. 210-225, 2022.

[6] Z. Skora, K. Ciupinska, S. H. Del Pin, M. Overgaard, and M. Wierzchon, "Investigating the validity of the perceptual awareness scale: the effect of task-related difficulty on subjective rating," *Consciousness and Cognition*, vol. 95, article no. 103197, 2021. https://doi.org/10.1016/j.concog.2021.103197

[7] L. Longo and G. Orru, "Evaluating instructional designs with mental workload assessments in university classrooms," *Behaviour & Information Technology*, vol. 41, no. 6, pp. 1199-1229, 2022.

[8] G. N. Dimitrakopoulos, I. Kakkos, Z. Dai, J. Lim, J. J. deSouza, A. Bezerianos, and Y. Sun, "Task-independent mental workload classification based upon common multiband EEG cortical connectivity," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1940-1949, 2017.

[9] T. Hao, X. Zheng, H. Wang, K. Xu, and S. Chen, "Linear and nonlinear analyses of heart rate variability signals under mental load," *Biomedical Signal Processing and Control*, vol. 77, article no. 103758, 2022. https://doi.org/10.1016/j.bspc.2022.103758

[10] M. Gjoreski, B. Mahesh, T. Kolenik, J. Uwe-Garbas, D. Seuss, H. Gjoreski, M. Lustrek, M. Gams, and V. Pejovic, "Cognitive load monitoring with wearables: lessons learned from a machine learning challenge," *IEEE Access*, vol. 9, pp. 103325-103336, 2021.

[11] A. Salfinger, "Deep learning for cognitive load monitoring: A comparative evaluation," in *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, Virtual Event, 2020, pp. 462-467. https://doi.org/10.1145/3410530.3414433

[12] J. K. Kim, M. N. Bae, K. Lee, J. C. Kim, and S. G. Hong, " Explainable artificial intelligence and wearable sensor-based gait analysis to identify patients with osteopenia and sarcopenia in daily life," *Biosensors*, vol. 12, no. 3, article no. 167, 2022. https://doi.org/10.3390/bios12030167

[13] Y. Y. Jo, Y. Cho, S. Y. Lee, J. M. Kwon, K. H. Kim, K. H. Jeon, S. Cho, J. Park, and B. H. Oh, "Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram," *International Journal of Cardiology*, vol. 328, pp. 104-110, 2021.

[14] A. W. K. Gaillard, "Stress, workload, and fatigue as three biobehavioral states: a general overview," in *Stress, Workload, and Fatigue*. Boca Raton, FL: CRC Press, 2000, pp. 623-639.

[15] O. Parlak, "Portable and wearable real-time stress monitoring: a critical review," *Sensors and Actuators Reports*, vol. 3, article no. 100036, 2021. https://doi.org/10.1016/j.snr.2021.100036

[16] S. Gedam and S. Paul, "A review on mental stress detection using wearable sensors and machine learning techniques," *IEEE Access*, vol. 9, pp. 84045-84066, 2021.

[17] S. Lee and D. Park, "A real-time abnormal beat detection method using a template cluster for the ECG diagnosis of IoT devices," *Human-centric Computing and Information Sciences*, vol. 11, article no. 4, 2021. https://doi.org/10.22967/HCIS.2021.11.004

[18] Y. Shi, N. Ruiz, R. Taib, E. Choi, and F. Chen, "Galvanic skin response (GSR) as an index of cognitive load," in *Proceedings of the Conference on Human Factors in Computing Systems (CHI) Extended Abstracts*, San Jose, CA, 2007, pp. 2651-2656.

[19] J. Kim, K. B. Lee, S. Lee, H. Yang, and S. G. Hong, "A novel stress measurement system with handhold electrodes in massage chairs," in *Proceedings of 2016 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju, South Korea, 2016, pp. 859-863.

[20] A. S. Anusha, J. Jose, S. P. Preejith, J. Jayaraj, and S. Mohanasankar, "Physiological signal based work stress detection using unobtrusive sensors," *Biomedical Physics & Engineering Express*, vol. 4, no. 6, article no. 065001, 2018. https://doi.org/10.1088/2057-1976/aadbd4

[21] K. Pettersson, J. Tervonen, J. Narvainen, P. Henttonen, I. Maattanen, and J. Mantyjarvi, "Selecting feature sets and comparing classification methods for cognitive state estimation," in *Proceedings of 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, Cincinnati, OH, 2020, pp. 683-690.

[22] X. Fan, C. Zhao, X. Zhang, H. Luo, and W. Zhang, "Assessment of mental workload based on multi-physiological signals," *Technology and Health Care*, vol. 28, no. S1, pp. 67-80, 2020. https://doi.org/10.3233/THC-209008

[23] F. Liu, L. Zhao, X. Cheng, Q. Dai, X. Shi, and J. Qiao, "Fine-grained action recognition by motion saliency and mid-level patches," *Applied Sciences*, vol. 10, no. 8, article no. 2811, 2020. https://doi.org/10.3390/app10082811

[24] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psycho-physiological measures for assessing cognitive load," in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, Copenhagen, Denmark, 2010, pp. 301-310.

[25] D. Wehler, H. F. Jelinek, A. Gronau, N. Wessel, J. F. Kraemer, R. Krones, and T. Penzel, "Reliability of heart-rate-variability features derived from ultra-short ECG recordings and their validity in the assessment of cardiac autonomic neuropathy," *Biomedical Signal Processing and Control*, vol. 68, article no. 102651, 2021. https://doi.org/10.1016/j.bspc.2021.102651

[26] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. 23, no. 1, pp. 90-93, 1974.

[27] J. Morlet, G. Arens, E. Fourgeau, and D. Glard, "Wave propagation and sampling theory—Part I: Complex signal and scattering in multilayered media," *Geophysics*, vol. 47, no. 2, pp. 203-221, 1982.

[28] T. M. Li, H. C. Chao, and J. Zhang, "Emotion classification based on brain wave: a survey," *Human-centric Computing and Information Sciences*, vol. 9, article no. 42, 2019. https://doi.org/10.1186/s13673-019-0201-x

[29] T. Heeren and R. D'Agostino, "Robustness of the two independent samples t-test when applied to ordinal scaled data," *Statistics in Medicine*, vol. 6, no. 1, pp. 79-90, 1987.

[30] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artificial Intelligence Review*, vol. 55, pp. 3503-3568, 2022.

[31] S. M. Kochav, Y. Raita, M. A. Fifer, H. Takayama, J. Ginns, M. S. Maurer, M. P. Reilly, K. Hasegawa, and Y. J. Shimada, "Predicting the development of adverse cardiac events in patients with hypertrophic cardiomyopathy using machine learning," *International Journal of Cardiology*, vol. 327, pp. 117-124, 2021.

[32] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, et al., "Explainable AI for trees: From local explanations to global understanding," 2019 [Online]. Available: https://arxiv.org/abs/1905.04610.

[33] R. Shailendra, A. Jayapalan, S. Velayutham, A. Baladhandapani, A. Srivastava, S. Kumar Gupta, and M. Kumar, "An IoT and machine learning based intelligent system for the classification of therapeutic plants," *Neural Processing Letters*, vol. 54, pp. 4465-4493, 2022.

[34] D. Zhang, L. Qian, B. Mao, C. Huang, B. Huang, and Y. Si, "A data-driven design for fault detection of wind turbines using random forests and XGboost," *IEEE Access*, vol. 6, pp. 21020-21031, 2018.

[35] A. Shehadeh, O. Alshboul, R. E. Al Mamlook, and O. Hamedat, "Machine learning models for predicting the residual value of heavy construction equipment: an evaluation of modified decision tree, LightGBM, and XGBoost regression," *Automation in Construction*, vol. 129, article no. 103827, 2021. https://doi.org/10.1016/j.autcon.2021.103827