Available online at : http://josi.ft.unand.ac.id/

Jurnal Optimasi Sistem Industri

| ISSN (Print) 2088-4842 | ISSN (Online) 2442-8795 |

Research Article

Physiological Signals as Predictors of Mental Workload: Evaluating Single Classifier and Ensemble Learning Models

Nailul Izzah [a,*], Auditya Purwandini Sutarto [a], Ade Hendi [b], Maslakhatul Ainiyah [c], Muhammad Nubli Bin Abdul Wahab [d]

[a] Department of Industrial Engineering, Universitas Qomaruddin, Gresik, Indonesia
[b] Department of Informatics Engineering, Universitas Qomaruddin, Gresik, Indonesia
[c] Department of English Education, Universitas Qomaruddin, Gresik, Indonesia
[d] Center of Modern Language and Human Science, Universiti Malaysia Pahang, Malaysia

[*] Corresponding Author: nailulizzah@uqgresik.ac.id

DOI: 10.25077/josi.v22.n2.p81-98.2023 **Submitted** : August 17, 2023 **Accepted** : October 16, 2023 **Published** : December 18, 2023

ABSTRACT

With a growing emphasis on cognitive processing in occupational tasks and the prevalence of wearable sensing devices, understanding and managing mental workload has broad implications for safety, efficiency, and well-being. This study aims to develop machine learning (ML) models for predicting mental workload using Heart Rate Variability (HRV) as a representation of the Autonomic Nervous System (ANS) physiological signals. A laboratory experiment, involving 34 participants, was conducted to collect datasets. All participants were measured during baseline, two cognitive tests, and recovery, which were further separated into binary classes (rest vs workload). A comprehensive evaluation was conducted on several ML algorithms, including both single (Support Vector Machine – SVM, and Naïve Bayes) and ensemble learning (Gradient Boost and AdaBoost) classifiers and incorporating selected features and validation approaches. The findings indicate that most HRV features differ significantly during periods of mental workload compared to rest phases. The SVM classifier with knowledge domain selection and leave-one-out cross-validation technique is the best model (68.385). These findings highlight the potential to predict mental workload through interpretable features and individualized approaches even with a relatively simple model. The study contributes not only to the creation of a new dataset for specific populations (such as Indonesia) but also to the potential implications for maintaining human cognitive capabilities. It represents a further step toward the development of a mental workload recognition system, with the potential to improve decision-making where cognitive readiness is limited and human error is increased.

Keywords: cognitive processing, mental workload, machine learning model, heart rate variability, autonomic nervous system

INTRODUCTION

With the modern economy and the growth of technology and knowledge-based professions, the mental workload is one of the most widely invoked concepts in ergonomics research and practice because of the greater emphasis on cognitive demands [1]. It has become integral to various sectors including technology, finance, law, healthcare, and various professional services [2]. This field increasingly requires critical thinking, problem-solving, data analysis, and other complex mental tasks. Mental workload, a specific facet of overall workload, delineates the delicate balance between task-imposed demands and an operator's ability to fulfil them [3]. Wickens' multiple resources theory further emphasizes the multifaceted nature of human information processing, illustrating how different resources can be exploited either simultaneously or sequentially. This theoretical construct assists system designers in predicting the compatibility or interference of simultaneous tasks.

Within the context of human factors and ergonomics, mental workload, and cognitive load are often interchangeable as both share a similar foundational concept regarding the amount of limited working memory for tasks [4], [5]. Cognitive load is a broader multidimensional construct that encompasses mental workload, mental effort, and performance, each with a unique identity and can be manipulated through task design and instructions [6]. While mental workload focuses on task demands, mental effort denotes the devoted resources, and performance reflects task execution and can be assessed through specific metrics either during the task or thereafter.

In an era surrounded by complex information and ubiquitous technology, the importance of understanding and managing mental workload continues to grow. The ability to measure and identify mental workload states has broad implications for individual and organizational success, including maximizing safety, efficiency, performance, and well-being [7]–[9]. The challenge lies in developing a mental workload monitoring system that is objective, real-time, and unobtrusive [10]. While self-report workload assessment such as NASA TLX has advantages primarily due to their ease and cost of administration, they are impractical in real-world applications, as they require explicit querying of use [7], [11]. Physiological signals, derived from the autonomous nervous system (ANS), offer a promising approach, given their objectivity, validity, non-invasive, and not interfering with the primary task [2], [9]. It has been also well-established as an indicator of mental workload fluctuations [2], [9]. Nonetheless, the utilization of physiological signals is not without its challenges. The approach is resource-intensive, requiring specialized, often costly, technology and specific expertise for data validation and interpretation. Complex data collection and analysis processes further complicate its implementation [9], [12]. Besides, despite being labelled as 'non-intrusive,' these methods may require users to wear sensors or equipment attached to their body, raising concerns about user comfort and practicality.

Interestingly, innovations in wearable technology offer a recent solution to some of these challenges. Devices such as smartwatches and chest straps are equipped with good validity sensors that can measure specific physiological signals [13], [14]. These devices are designed to integrate seamlessly into daily life, allowing for real-time, continuous data collection without interfering routine activities. This may enhance the practicality and workers' acceptability of physiological signal-based mental workload assessment methods. Moreover, recent progress in machine learning (ML) and Artificial Intelligence (AI) has significantly expanded the scope of human behaviour prediction systems [8], [15]. Importantly, ML and AI technologies offer the capability to address the complexity associated with collecting and processing physiological data for predicting mental workload [15], [16]. While models for stress detection are abundant (for review see [15], [17]), those focusing on cognitive or mental workload remain scarce. This distinction is critical as, despite sharing common physiological features, the underlying psychological mechanisms and activities for stress and mental workload are different [18], [19]. Stress often arises from emotional or environmental aspects unrelated to cognitive demands, whereas mental workload is specifically tied to task-specific requirements [18], [19]. Furthermore, the interpretation of what those features signify in the context of stress versus cognitive workload may be different.

Understanding Physiological Signals and HRV

Fluctuations in cognitive load are manifested through changes in the autonomous nervous system (ANS), or physiological signals. An increase in psycho-physiological load—such as performing a demanding task—leads to heightened activation of the sympathetic nervous system and inhibition of the parasympathetic system, a response known as the "fight-or-flight" reaction. This triggers the release of hormones, specifically epinephrine, and norepinephrine, leading to physiological alterations. These alterations were evidenced by changes in blood pressure, brain activity, skin conductance, respiration, and eye movement, accompanied by a reduction in heart rate variability (HRV) [2], [9], [20]. Conversely, the activation of the parasympathetic system and suppression of the sympathetic system initiates a process termed the "relax and digest" response, which induces the reverse physiological reactions to the fight-or-flight process [21], [22]. Research has demonstrated that HRV is not solely significant in the context of maintaining physical health but also in various aspects of well-being, including psychological health, cognitive function, and social interactions [23], [24].

As illustrated in Figure 1, HRV is calculated through beat-to-beat (RR interval) intervals in heart rate. It serves as a quantification of neurocardiac function and indicates bi-directional interactions between the heart and the brain, controlled by the ANS.

DOI: 10.25077/josi.v22.n2.p81-98.2023

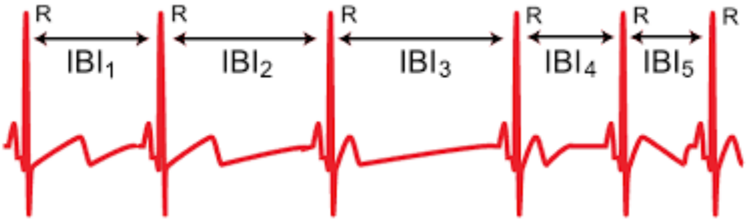


Figure 1. RR intervals, inter-beat intervals between all successive heartbeat

The analysis of HRV encompasses linear and non-linear domains. Table 1 provides a summary of HRV parameters in line with their respective analysis domains. These measures—such as RMSSD, pNN50, and HF—reflect vagal inputs to the heart, with parameters like LF indicating a mix of sympathetic and vagal parasympathetic activities., while SDNN represents cyclic components responsible for HRV. The ratio of low- and high-frequency power (LF/HF) is an estimator of the balance between the sympathetic and parasympathetic systems.

Related Work on HRV, Mental Workload, and Machine Learning

In a recent review, cardiac activity emerges as a primary physiological measure of mental workload (MWL) [2], [9]. Heart rate and some HRV parameters can quantify changes in HRV during different levels of mental workload. For example, HR increases with increasing task demands and could differentiate between rest and task periods in a simulated flight task [26]. NN intervals were seen to decrease during a high-demand multi-attribute task when compared to a low-demand task [27]. Fallahi and colleagues [28] found the lowest SDNN, RMSSD, and pNN50 when traffic control operators experienced high traffic density tasks, compared to baseline and low traffic density conditions. In the frequency domain, Veltman and Gaillard [29] stated that the MF band (0.07–0.14 Hz) is the most sensitive to changes in MWL while the effect in the LF band was observed by Splawn and Miller [30] at high task loads.

Table 1. Summary of some HRV parameters based on their respective analysis domains [25]

Analysis Domain	**Acronym**	**Unit**	**Description**
Time-domain	HR Max – HR Min	bpm	The average difference between the highest and lowest heart rates during each respiratory cycle
	SDNN	ms	The standard deviation of successive NN interval differences
	pNN50	%	Percentage of successive NN intervals that differ by more than 50 ms
	RMSSD	ms	Root mean square of successive NN interval differences
Frequency Domain	VLF power	ms^2	The absolute power of the very-low-frequency band (0.0033–0.04 H
	LF peak		The peak frequency of the low-frequency band (0.04–0.15 Hz)
	LF power	ms^2	The absolute power of the low-frequency band (0.04–0.15 Hz)
	LF power	nu	The relative power of the low-frequency band (0.04–0.15 Hz) in normal units
	LF power	%	The relative power of the low-frequency band (0.04–0.15 Hz)
	HF power	ms^2	The absolute power of the high-frequency band (0.15–0.4 Hz)
	LF/HF	%	
Non-Linear	SD1	ms	Poincaré plot standard deviation perpendicular to the line of identity
	SD2	ms	Poincaré plot standard deviation along the line of identity
	SD2/SD1	%	Ratio of SD1-to-SD2

Considering this association, researchers have been motivated to further utilize HRV as an index of cognitive processing through AI techniques [8], [31], [32]. In the machine learning domain, HRV-based models exhibit an accuracy rate that varies between 70-90% when combined with either other physiological signals or behaviors, and between 50-70% when using HRV exclusively [8], [32], [33]. Table 2 outlines prior studies related to cognitive load prediction based on HRV using various ML classifiers. The majority of protocols involved standardized cognitive tests such as the N-back test, Maastricht Acute Stress Test, and Psychomotor Vigilance Task, and only one study used a simulated task [34]. Moreover, most of the studies used multimodal signals including HRV, Galvanic Skin Response (GSR), electrooculography (EOG), and accelerometer. Prior studies also commonly employed full features or data-driven techniques to select HRV features. Feature selection methods itself can be categorized into filter-based methods, wrapper-based methods, and embedded methods [33], [35]. So far, however, very few studies utilized the knowledge domain to select HRV features which may increase their interpretability. There is a growing focus on the interpretability of models [36], [37], challenging the common belief that black box models are necessary for achieving high accuracy. In the machine learning context, the black box refers to the lack of transparency in the

Table 2. Summary of Related Work on Cognitive or Mental Workload Prediction

Lead Authors	**Subjects, Population, Task / Scenarios**	**Physiological Signals**	**Feature Extraction / Feature Engineering**	**Feature Selection & Final HRV Features**	**Validation**	**Class & Accuracy Scores**
Gjoreski [11]	- *N* = 23, Europeans (nationalities /countries not specified) - Task: CogLoad Test (variation of N-Back test)	ACC, GSR, TEMP, HRV	Scaling: min-max, session-specific standardization	Full data set vs Ranking method based on mutual information	LOOCV	Class: Cognitive load or not. ML & accuracy for session-specific standardization: full features, selected features -RF: 66.8%, 67.9% -kNN: 63.6%, 64.0% -NB: 58.5 %, 57.0% -LR: 64.0 % & 65.7% -AdaBoost: 65.6% & 67.3% -DT: 67.4% & 68.2% -XGB: 65.5 % & 66.4%
Pettersson [33]	- *N* = 23, Finland - Task: Maastricht Acute Stress Test	EOG, HRV	Not specified	- Sequential Forward Floating Search - Features: HR mean, HR std, RMSSD	8-fold CV	Class: baseline and task. ML & accuracy: HRV, EOG + HRV -SVM: 74.1 %, 85.9% -RF: 71.5%, 93.4% -XGB: 70.7%, 94.0%
Giannakakis [32]	- *N* = 24, Greece - Task: social exposure, stressful event recall, cognitive load, stressful videos	HRV	Without pairwise and Normalization using pairwise transformation	- mRMR - 11- HRV features: mean HR, LF, NN50, LFnorm, HRstd, pNN50, LF/HF, RMSSD, HFnorm, total power, HRV triangular index	10-fold CV	Class: stress (including cognitive) no stress. ML & accuracy: without pairwise, after pairwise -kNN: 66.7%, 73.8% -NB: 65.6%, 69.9% -SVM: 73.6%, 84.4% -RF: 75.1%, 70.0%
Posada-Quintero [31]	- *N* = 16, USA - Task: psychomotor vigilance task (PVT), n-back paradigm, and a visual search	HRV, EDA	Not specified	- Not specified - 4-HRV features: LF, LFnu, HF, HFnu	LOOCV	Class: baseline, vigilance, working memory, visual search. ML & accuracy -KNN: 66% -Linear SVM: 62% -LDA: 62%

Table 2 (cont.)

Lead Authors	Subjects, Population, Task / Scenarios	Physiological Signals	Feature Extraction / Feature Engineering	Feature Selection & Final HRV Features	Validation	Class & Accuracy Scores
Ross [34]	- N = 10, Canada - Task: Penetrating Trauma Simulation	HRV, GSR	Scaling: Normalized with baseline data	- LASSO - 18 HRV Features, not explicitly mentioned after selected	5-fold	Class: Cognitive load between novice and experts. ML & Accuracy: HRV, HRV + GSR - SVM: 72.8%, 79.8% - DT: 63.3%, 78.0% - RF: 72.4%, 66.7% - kNN: 53.3%, 83.9%

Notes: ACC=Accelerometer, TEMP = Skin temperature, EDA = Electrodermal Activity, EMG = Electromyography, RESP = Respiration, EOG = Electrooculogram, LASSO = Least Absolute Shrinkage and Selection Operators, CV = Cross-validation. LOOCV= Leave one out cross validation, PCA = Principal Component Analysis, RF = Random Forest, SVM = Support Vector machine, k-NN = K nearest neighbourhood, NB = Naïve Bayes, LR = Logistic Regression, XGB = Extreme Gradient Boosting, LDA = Linear Discriminant Analysis, DT = Decision Tree.

decision-making process and the internal processes used to achieve accurate prediction. Therefore, the importance of developing models that are interpretable and transparent has become a priority.

To our knowledge, there currently exists no publicly available HRV dataset specifically tailored to Asian populations, including regions such as Indonesia. Given that HRV is significantly influenced by ethnic characteristics, it is critical to develop and test various machine learning (ML) algorithms on population-specific data. This approach can potentially yield broader benefits across diverse sectors. Such a dataset is pivotal in advancing research, fostering innovation, and facilitating collaboration. By providing a common platform, it enables researchers, practitioners, and other stakeholders to work on shared goals[38].

Based on the above review, this research aims to develop machine learning (ML) models for predicting mental workload through HRV as a representation of physiological signals. Specific objectives include the evaluation of various ML algorithms, consisting of single classifiers and ensemble learning techniques, coupled with combination feature selections and validation strategies. The focus on an Indonesian population dataset for HRV-based prediction models establishes a novelty within the field. This research explores the use and promise of HRV-based models for predicting mental workload, making them relevant across various occupations and tasks that require significant cognitive effort. By highlighting the predictive value of physiological signals and investigating the interaction between machine learning and human behavior within a specific cultural context.

METHODS

Experimental Protocol

Participants

A total of 34 undergraduate students (age 19 – 24 years with the mean age of 21.9, standard deviation 1.38 years) took part in this study. The sample is a relatively balanced gender distribution (55.9% male). Among the participants, 26.5% were identified as active smokers, and all individuals were right-handed. Eligibility for the participants was determined based on the following conditions: 1) absence of neurological, heart, or psychiatric disorders; 2) not under chronic medical treatment; 3) no known allergies to adhesive substances or rubbing alcohol. Participation was voluntary. Written informed consents were obtained from all participants before the initiation of the experiment. The experimental protocol adhered to the ethical principles of the Declaration of Helsinki and received approval from the Local Research Ethics Committee.

Experimental Task

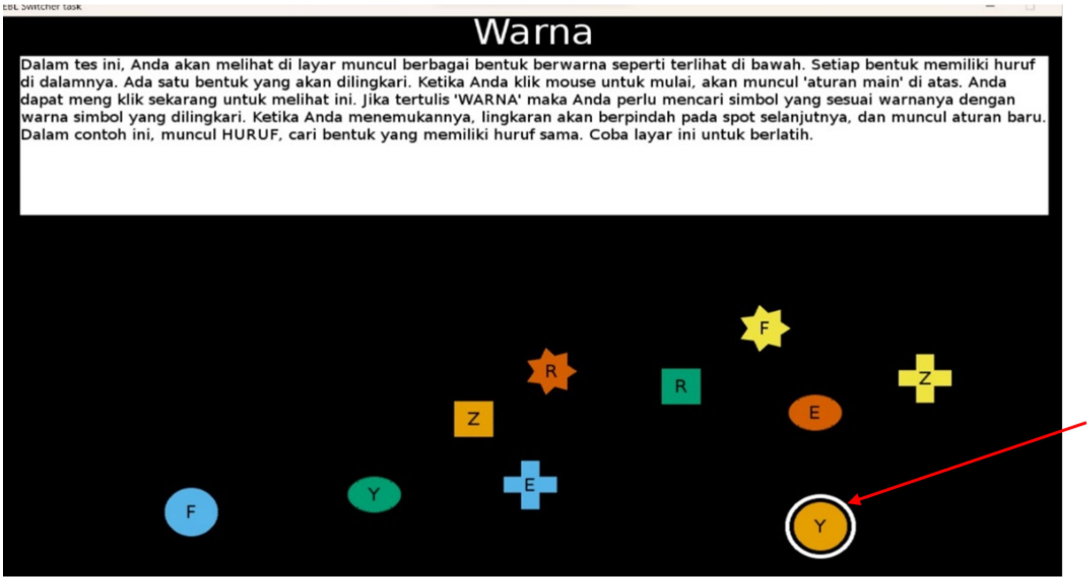
In this study, we assessed mental workload through standardized cognitive tests focusing on attention functions. Since attention and mental workload are closely intertwined, variations in attentional performance can provide insights into an individual's mental workload. When mental workload increases, it typically affects attentional capacities, making attention tests a reasonable proxy for assessing mental workload. Standardized tests facilitate a reliable way to evaluate mental workload which can be quantified and repeated under various conditions. Lastly, attention has several facets with several different types of tests that can detect subtle variances of mental workload, including selective, sustained, and divided attention [24].

The first cognitive task, the d2-Attention Test, delivered in a paper-and-pencil version. It is a neuropsychological test that assesses individuals' selective and sustained attention [39]. Participants were requested to cross out the letter "d" with two apostrophe marks among various distractors within 14 rows of 47 letters. They were given 20 seconds for each row to mark as many target symbols as possible and then move immediately to the next row. The second task, the Switcher Featuring task, is a computerized cognitive test that is part of the Psychology Experiment Building Language (PEBL) [40]. The objective of this task was to assess cognitive flexibility and divided attention by repeatedly switching between rule dimensions [40]. As shown in Figure 2, during the task, participants viewed a 14-inch laptop screen displaying ten distinct colored shapes. Each shape shared only one common dimension with another object, such as color, shape, or letter. Participants were prompted to select a matching object based on a shape, color, or letter displayed at the top of the screen after one object was circled. Subsequently, they were required to "switch" to a different feature, attempt to match the object based on that feature, and then return to the previous feature.

The task was divided into three sessions, each consisting of nine blocks or alternative configurations. Within each block, participants made ten responses. The task was structured as follows:

- Type 1: Condition Alternate Switch - In the first three blocks, participants switched between two of the three feature rules, with each block utilizing a different combination of pairs.
- Type 2: Condition Fixed Switch - In the subsequent three blocks, participants switched between the three feature rules in a consistent order, with the order changing for each block.
- Type 3: Condition Random Switch - In the final three blocks, participants switched between the three feature rules randomly, rendering the next rule unpredictable.

Before the main task, participants performed a brief practice round to familiarize themselves with the procedure. The main task lasted for approximately five minutes.



EBL Switcher task

Warna

Dalam tes ini, Anda akan melihat di layar muncul berbagai bentuk berwarna seperti terlihat di bawah. Setiap bentuk memiliki huruf di dalamnya. Ada satu bentuk yang akan dilingkari. Ketika Anda klik mouse untuk mulai, akan muncul 'aturan main' di atas. Anda dapat meng klik sekarang untuk melihat ini. Jika tertulis 'WARNA' maka Anda perlu mencari simbol yang sesuai warnanya dengan warna simbol yang dilingkari. Ketika Anda menemukannya, lingkaran akan berpindah pada spot selanjutnya, dan muncul aturan baru. Dalam contoh ini, muncul HURUF, cari bentuk yang memiliki huruf sama. Coba layar ini untuk berlatih.

Subjects should select an object with similar color ('warna') to the circled one

Figure 2. Switcher Featuring Task in PEBL Battery Software [40]
[The prompt delivered in Indonesia language (adjustable)]

DOI: 10.25077/josi.v22.n2.p81-98.2023

Experimental Procedure

After obtaining consent from each participant, the experimental procedure started with the placement of a Polar H-10 electrocardiogram (ECG) on the participant's chest. This equipment was vital for recording the interbeat interval. Prior to the data collection, the participants were instructed to abstain from the consumption of caffeine, smoking, and heavy meals for a minimum of two hours before the session. This instruction aligned with the methodological consideration for HRV research [41].

The experiment was divided into four stages, including baseline, two cognitive tasks, and recovery. For both the baseline and recovery measurements, participants were instructed to remain stationary in a sitting position for five minutes. The first cognitive task, known as the d2-Attention Test, was conducted for approximately five minutes. Subsequently, the Feature Switching Task was initiated, serving as the second cognitive task. Upon the completion of all stages, the sensors were removed, and participants were debriefed.

Machine Learning Model Development

Data Preprocessing

Data preprocessing is a must-do step before training a model. Its primary objective is to check the quality of the data and to find important information that can affect the performance of learning models [42]. Within this preprocessing stage, various aspects of the dataset are addressed, including handling missing values, scaling, and standardization. These procedures facilitate the preparation of the data, ensuring that it is ready for the learning process. The general architecture of machine learning model development for a mental workload prediction scheme is visually represented in Figure 3.

Feature Extraction

In our study, we utilized KUBIOS HRV Standard software (Version 3.5.0, Kubios, Finland) to generate 24 HRV features from RR interval data. The 24 HRV-based features were categorized as seven features within the time domain (including RR, mean HR, min HR, max HR, SDNN, RMSSD, pNN50), 14 within the frequency domain (including total power, total power log, VLF absolute power, VLF log, peak VLF, LF absolute power, LF log, LF nu, peak LF, HF, HF log, HF nu, peak HF, and LF/HF) and three non-linear features (SD2, SD1, SD2/SD1). The computation of each HRV feature was conducted within a 5-minute moving window, following the procedure: initially, an inter-beat interval (IBI) signal was extracted from the peaks of the ECG signal for each subject. Subsequently, each HRV feature was computed within a 5-minute moving window, employing a non-overlapping configuration. This 5-minute

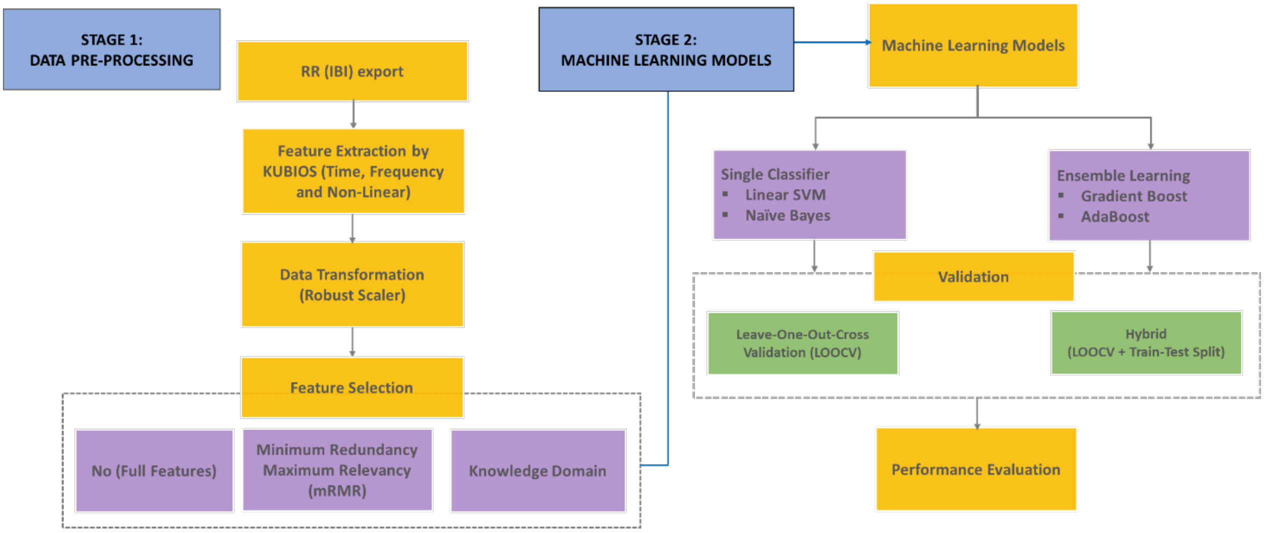


Figure 3. A general architecture of the mental workload prediction scheme employed in this study

recording duration is a minimum recommendation to obtain a reliable frequency-domain index [25], [41]. The final data consisted of 136 instances (4 conditions × 34 subjects).

Data Transformation

Given the majority of our HRV data exhibited skewness and contained outliers, we implemented data scaling to enhance the efficiency, performance, and interpretability of ML models [43]. machine learning models. Specifically, Robust Scaler was utilized to standardize the data. This method involves the removal of the median and scales the data according to the interquartile range (IQR), defined as the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile). Such an approach is for datasets that include a significant amount of noise and may have outliers due to physiological phenomena such as ectopic beats [25], [44].

Feature Selection

The performance of different classifiers depends on the features employed. The feature selection process is crucial to select the most important features, thereby improving classification outcomes and identifying a minimal feature set necessary to achieve predetermined classification accuracy [33], [35]. In this study, models were developed both with and without the application of feature selection methods. First, a set of 24 features, obtained through the use of Kubios software, was employed. Second, a knowledge domain was utilized to selectively identify appropriate features. This selection was conducted based on the recommendation of prior cognitive and HRV studies [20], [22], [41]. This approach was aimed to increase the interpretability of the models [25], [36] The selected features encompassed features from both the time and frequency domain. To facilitate comparison, the minimum redundancy maximum relevancy (mRMR) was additionally performed. This filter-based feature selection method, as proposed by [35], has been previously documented within the scientific literature, specifically in the context of selecting features utilizing HRV features [32], [45]. This method selects a subset of features by optimizing Mutual Information Quotient criterion using the highest correlation with the target variable but the lowest correlation among themselves. The selection of features is performed iteratively, employing a greedy search method based on optimizing an objective function, thus balancing both relevance and redundancy [35].

Machine Learning Classifiers

The development of models was conducted utilizing the following ML algorithms: Support Vector Machine (SVM), Naïve Bayes (NB), Gradient Boosting (GB), and AdaBoost. Recent reviews reveal that both single classifiers (e.g., SVM and Naïve Bayes) and ensemble learning models are the most prevalently employed techniques in HRV-based ML models[15], [17]. Generally, ensemble learning approaches have demonstrated superior predictive performance on supervised binary classification [46]. A brief description of each algorithm is as follows:

Support vector machine is a discriminative model, designed to find the optimal hyperplane to segregate data into different classes, especially in a high-dimensional space [43]. Naive Bayes is a family of probabilistic classifiers that applies Bayes' theorem, operating under strong independence assumptions between features. It assumes that the value of a specific feature is independent of the value of any other feature, depending on the class variable. Both Gradient boosting and AdaBoost algorithms are among the most prevalent ensemble decision trees-based learning techniques. In Gradient Boosting, trees are constructed sequentially, with each tree having the same weight and trying to correct the errors of its predecessor. Conversely, in AdaBoost, trees have weights. The method automatically adapts its parameters to the data according to the actual performance in the current iteration. Both the re-weighting of the data and the final aggregation weights are recalculated iteratively. Gradient Boosting serves as a generic algorithm that helps in finding the approximate solutions to the additive modeling problem whereas AdaBoost was the first designed boosting algorithm with a specific loss function. Gradient Boosting is considered more flexible than AdaBoost [43]. Figure 4 illustrates a visualization of each algorithm used in this study.

Validation Techniques

In this study, we implemented two different validation techniques: leave-one-out cross-validation (LOOCV) and a hybrid method, which is a combination between LOOCV and the conventional train-test split test. In the LOOCV

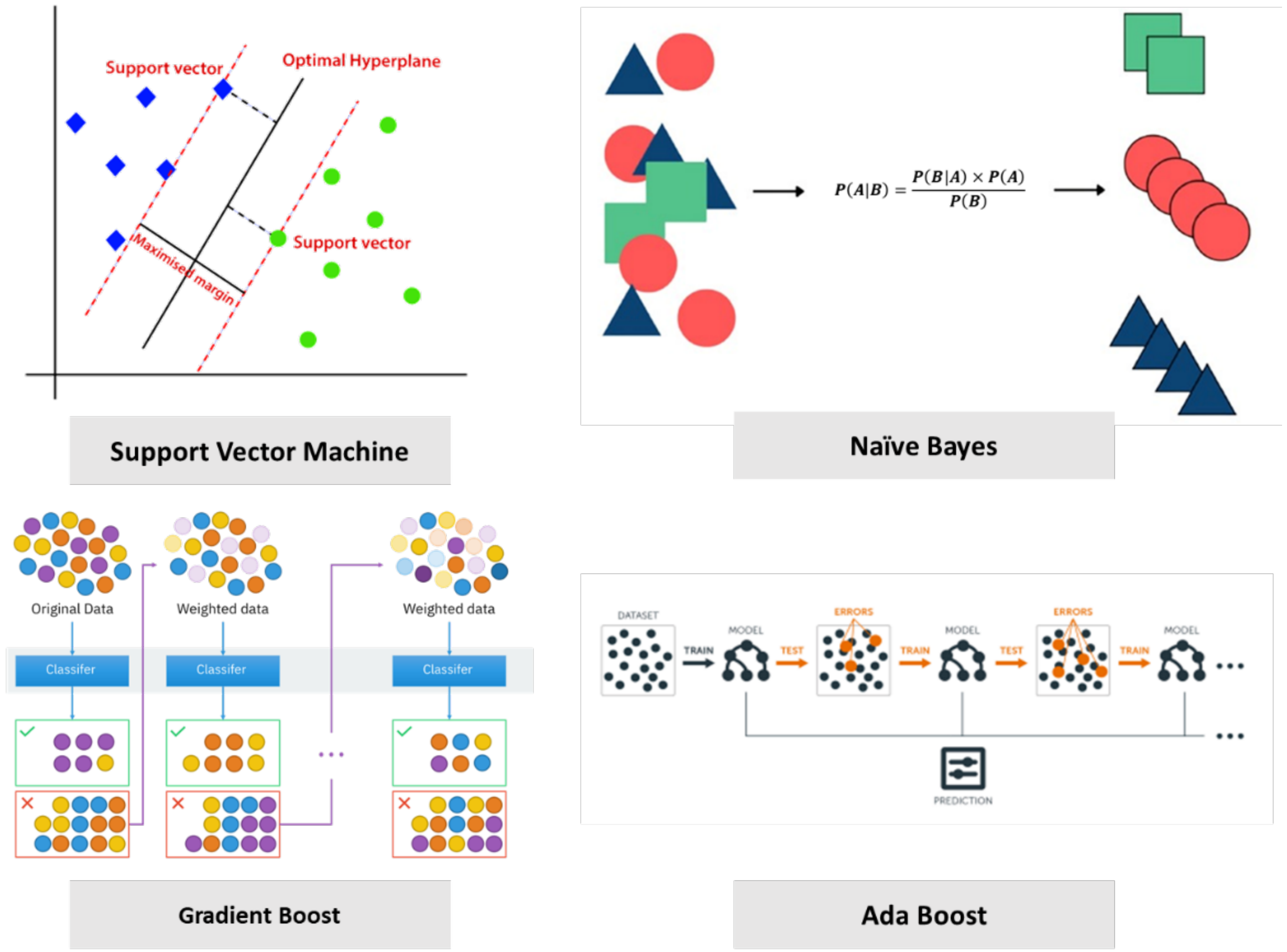


Figure 4. Illustration of each algorithm applied.
(Sources: SVM [47], Naïve Bayes [48], Gradient Boost and AdaBoost [49])

approach, each individual's HRV data is used once as the test set, while the remaining data points constitute the training set. This procedure is iteratively repeated for each data point, resulting in a number of separate learning experiments equivalent to the total data points. Such a procedure aligns well with the personalized approach, an essential issue when dealing with HRV data, given the unique characteristics of each individual's HRV. Consequently, LOOCV can potentially offer a more accurate validation mechanism. However, it should be noted that despite presenting low bias, LOOCV can have high variance because the training sets are so similar to each other. Besides, with a large number of observations, LOOCV can be computationally expensive and time-consuming as the model must be trained N times (where N represents the number of observations).

To address this issue, we utilized the hybrid technique that combines LOOCV and the train-test split test (in an 80: 20 ratio). This technique allows for more comprehensive model validation. Using this technique, LOOCV was applied solely to the training set (80% of the data) while the remaining 20% serves as unseen data for testing. This strategy considers individual variability through LOOCV and provides an unbiased performance evaluation using a hold-out test set. Moreover, this technique has more computational efficiency compared to the LOOCV [50].

Performance Evaluation

The evaluation of performance for each model was presented by its accuracy scores, an approach that quantifies how closely a model approximates the actual value. Accuracy is computed by the ratio number of correct predictions to the total prediction number [43]. For the classification tasks within this study, the scikit-learn library, a widely recognized tool in the field of machine learning, was employed[51].

RESULTS AND DISCUSSION

Results

Table 3 shows the descriptive statistics of each HRV index, both during rest and while participants engaged in cognitive tests. Since this study focused on building machine learning models with binary labels (rest and workload), we averaged the data derived from the d2-Attention and Switcher Featuring tests. The instances were balanced between 'rest' or 'no mental workload' and 'task' or 'mental workload'. When estimating the mental workload of a person, it is important to define the specific state of interest, referred to as the ground truth. In the context of this study, the ground truth is defined by the protocol implemented [33].

To evaluate whether the d2-Attention and Switcher Featuring tests could elicit physiological stress reactions, we performed Wilcoxon tests for all HRV parameters. The results were reported as z-values. As also displayed in Table 3, out of 24 features, 18 (75%) demonstrated significant differences between the two conditions (rest vs workload) ($p<0.05$). This finding indicates that the implemented protocol reflected changes in the majority of physiological signals.

Selected Features from Knowledge Domain and mRMR Methods

HRV features, selected using the knowledge domain, show trend median values consistent with expectations. For instance, when individuals were engaged in mental activities requiring cognitive processing, RMSDD, and HFnu tend to decrease, which indicated suppression in their vagal tone. Concurrently, LF tends to increase, implying a higher sympathetic activation [22]. In contrast, when employing the mRMR approach to train and test the model, using the top seven (a number equivalent to those selected from the knowledge domain), we obtained the following features: time domain (minimum heart rate), frequency-domain (VLF absolute power, VLF log, peak VLF, and ratio LF HF), and non-linear (SD2).

Table 3. Descriptive statistics and Results of Wilcoxon Test of HRV parameters

Variable	Condition	Mean	SD	Median	IQR	Z-score	Variable	Condition	Mean	SD	Median	IQR	Z-score
RR	R	740.56	91.44	720.50	80.00	-5.288***	LF	R	904.13	740.11	584.00	716.25	-4.668***
	W	702.69	93.32	708.00	110.75			W	571.28	465.38	418.50	558.50	
Mean HR	R	82.19	9.55	83.00	8.75	-5.152***	LF log	R	6.53	0.73	6.37	1.02	-4.568***
	W	86.91	11.88	85.00	13.50			W	7.78	14.08	6.05	1.23	
Min HR	R	70.57	8.74	71.00	10.75	-6.341***	LF nu	R	58.63	18.62	58.92	28.59	-0.611
	W	76.31	10.40	75.50	14.75			W	57.18	15.80	60.06	24.69	
Max HR	R	100.46	10.75	100.00	14.75	-0.948	Peak LF	R	0.08	0.03	0.08	0.06	-2.384*
	W	101.47	12.77	101.00	19.25			W	0.09	0.07	0.09	0.06	
SDNN	R	42.67	17.38	37.55	23.78	-5.515***	HF	R	816.01	1026.47	455.50	826.25	-3.853***
	W	34.00	13.42	30.70	18.93			W	513.21	594.76	349.00	537.25	
RMSSD	R	40.24	23.31	31.70	18.83	-3.703***	HF Log	R	6.63	4.40	6.17	1.61	-3.251**
	W	34.52	18.36	31.55	18.33			W	6.11	3.25	5.91	1.58	
pNN50	R	18.22	18.04	11.01	19.38	-2.357**	HF nu	R	41.28	18.61	41.01	28.89	-0.648
	W	14.98	16.13	10.87	19.15			W	47.07	38.53	40.25	24.71	
Total Power	R	1867.29	1592.45	1202.50	1629.00	-4.98***	Peak HF	R	0.26	0.08	0.28	0.16	-1.271
	W	1159.50	982.46	813.00	1140.75			W	0.27	0.08	0.29	0.17	
Total Power (log)	R	7.23	0.77	7.09	1.09	-5.142***	LF/HF	R	2.29	2.39	1.44	1.85	-1.35
	W	6.74	0.81	6.70	1.29			W	1.69	1.06	1.51	1.50	
VLF	R	145.07	132.51	97.00	102.75	-4.665***	SD1	R	28.50	16.51	22.45	13.35	-3.721***
	W	73.34	77.83	42.00	62.00			W	24.43	13.00	22.30	13.08	
VLF log	R	4.62	0.88	4.57	0.99	-5.066***	SD2	R	52.66	19.74	47.50	30.08	-2.429*
	W	3.84	0.97	3.74	1.23			W	41.00	15.02	36.40	23.75	
Peak VLF	R	0.03	0.01	0.04	0.01	-1.386	SD2/SD1	R	2.06	0.63	1.99	0.78	-5.774***
	W	0.04	0.00	0.04	0.01			W	2.62	6.20	1.81	0.73	

Note: R = Rest; W = Workload; p-value *Significant at p<0.0; **<0.01; ***<0.001

DOI: 10.25077/josi.v22.n2.p81-98.2023

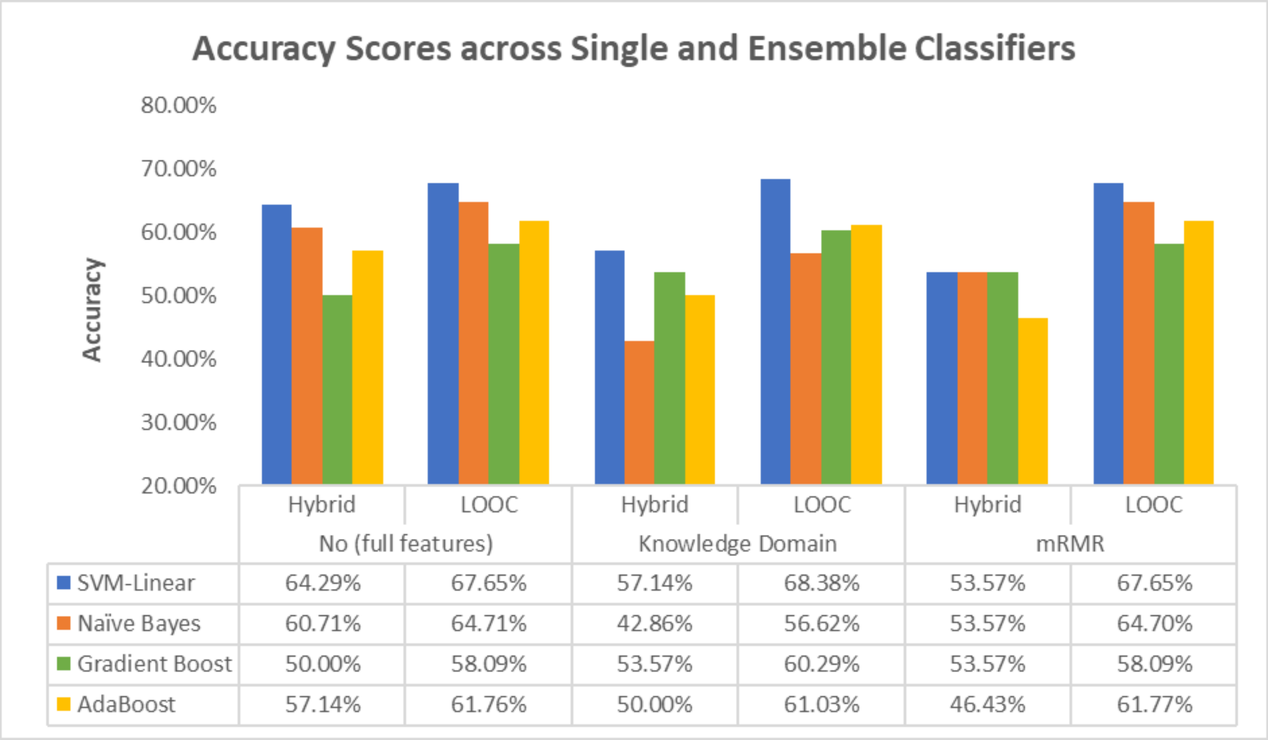
Table 4. Performance evaluation based on feature selection and validation techniques

Classifiers	**Feature Selection**			**Validation**	
	No (Full Features)	**Knowledge**	**mRMR**	**Hybrid**	**LOOC**
SVM	65.97%	62.76%	60.61%	58.33%	67.89%
NB	62.71%	49.74%	59.14%	52.38%	62.01%
GB	54.04%	56.93%	55.83%	52.38%	58.82%
AdaBoost	59.45%	55.52%	54.10%	51.19%	61.52%

Performance of Models

The accuracy scores for performance classification on the binary classes for all models are displayed in Table 4 and Figure 5. In general, the full features demonstrated superior performance when compared with fewer features obtained through either the knowledge domain or mRMR-based. However, the differences are relatively minimal and not consistent across different classifiers and validation techniques. With regard to the hybrid validation technique, the Gradient boost exhibits the weakest performance when employing the full set of features. In contrast, for the LOOC validation technique, the accuracy scores achieved using both the full feature set and a subset derived from the mRMR method are relatively similar (see Figure 5). In terms of validation techniques, all classifiers assessed by LOOC yielded higher accuracies compared to those validated using the hybrid method.

The optimal model was achieved utilizing the SVM (Support Vector Machine) approach, specifically employing knowledge-domain-based features and the LOOCV technique, resulting in an accuracy of 68.38%. The SVM appears to be the best classifier compared to Naïve Bayes and Ensemble classifiers. Contrary to expectations, the performance of ensemble learning models proved inferior to that of individual classifiers. The lowest accuracy score was presented by AdaBoost, using mRMR-based features and a hybrid technique (46.43%). his observation remains consistent, whether the full set of features or subsets derived from the knowledge domain or mRMR methods were employed, and is irrespective of whether the validation was conducted through hybrid or LOOC.



	Hybrid	LOOC	Hybrid	LOOC	Hybrid	LOOC
	No (full features)		Knowledge Domain		mRMR	
🟦 SVM-Linear	64.29%	67.65%	57.14%	68.38%	53.57%	67.65%
🟧 Naïve Bayes	60.71%	64.71%	42.86%	56.62%	53.57%	64.70%
🟩 Gradient Boost	50.00%	58.09%	53.57%	60.29%	53.57%	58.09%
🟨 AdaBoost	57.14%	61.76%	50.00%	61.03%	46.43%	61.77%

Figure 5. Comparison Accuracy Scores across Single and Ensemble Learning Classifiers

Discussion

A thorough evaluation was conducted on several ML algorithms, consisting of both single and ensemble learning classifiers while integrating selected features and some validation approaches. The findings suggest that the majority of HRV features were reflected differently during periods of mental workload compared to states of rest, including baseline and recovery phases. This observation is consistent with the results of prior studies [9], [20], [26]. In their review, Lohani, et al [20] suggest that cardiovascular measures (heart rate and HRV) may serve as robust indicators for the detection of near real-time cognitive changes in the real-world driving environment. Similarly, Mohanavelu et al [26] demonstrated that changes in HRV features such as SD2, SDNN, VLF, and total power are significant at all task load conditions during flight simulation involving 20 Indian fighter aircraft pilots. LFnu and HFnu were also able to distinguish the effect of low visibility and secondary cognitive task. Their studies contribute to our understanding of pilots' tasks and their cognitive demands during dynamic workload, as analyzed through HRV.

The linear SVM classifier demonstrated superior performance, achieving the highest accuracy among the evaluated algorithms, using the knowledge domain and leave-one-subject-out cross-validation approaches. In contrast with other models, including ensemble learnings, this outcome emphasizes that a simpler classifier can perform well in mental workload prediction. This observation confirms previous findings in cognitive state estimation [11], [33], [34], which were conducted using datasets from European, American, and Canadian samples, respectively. Interestingly, our finding did not support the widely recognized superiority of ensemble learnings over single classifiers [32], [46], which offer robustness, scalability, and ease of handling non-linearities. It seems that our HRV data is linearly separable, indicating a clear margin of separation between the two classes. In this context, SVM proves to be highly efficient and accurate. Nevertheless, since SVM also scales poorly for larger datasets. If future studies involve more subjects or more physiological signals or cognitive states, using SVM may be computationally expensive. In such scenarios. Algorithms like Gradient Boosting and AdaBoost may present a more reasonable starting point since they performed better without feature selection and are also known to scale more effectively for larger datasets [33].

It needs careful consideration that the accuracies achieved by all classifiers in this study although acceptable, range between 42.8 and 68.38%, which is relatively lower than those reported in prior HRV-based machine learning studies (e.g., [31]–[34]). This might be explained that our study exclusively utilized HRV signals, while more reliable and rigorous methods usually employ a fusion of multimodal signals. Such signals might include physiological measures (such as EEG, EDA, respiration, skin temperature, eye movement, and pupil diameter), behavioral manifestations (keystrokes and mouse dynamics, and sitting posture), facial expression, speech, and mobile phone use patterns] [11], [33], [52]. This approach, however, presents complex practical challenges including real-time multimodal data acquisition, data fusion, and data integration. Moreover, it raises important concerns regarding user privacy such as the implications of recording a person's computer keystrokes, video, and speech. Such methods may be impractical in actual business settings due to corporate computer security policies or global regulations workplace privacy laws [53]. Nevertheless, future studies need to explore the integration of multiple physiological measures to enhance the accuracy of cognitive load prediction. For instance, a combination of HRV and EDA or GSR has been shown to yield high accuracies without interfering with daily activities, as both signals can be recorded using a single device (e.g., Empatica E4) [11], [34]. Another plausible explanation contributing to the observed results is the study's focus on feature selection and validation issues, without an emphasis on hyperparameter tuning to optimize ML performance. Future studies should configure the hyperparameter values to produce the best model according to a predefined metric such as accuracy, while concurrently considering the balance between enhanced performance and computational costs.

Regarding the feature selection method, the best accuracy ranging from 50.0 to 67.65% was achieved with 24 features provided by the Kubios software. When comparing performance between HRV features selected by the knowledge domain and mRMR methods, relatively similar accuracies were observed. However, the knowledge domain offers advantages over mRMR in terms of its interpretability. The HRV features selected through the knowledge domain, including mean heart rate, SDNN, RMSSD, pNN50, HF, and LF, exhibited trends in the expected directions. These indicate correct markers of higher cognitive processing when individuals are engaged in mental workload [24]. In contrast, the mRMR method, which selected features like VLF absolute power, VLF log, peak VLF, ratio LF HF, and non-linear SD2, poses significant challenges in their interpretation. VLF is generally a representation of long-term regulation mechanisms, thermoregulation, and hormonal mechanisms [22], [25]; however, its peak is difficult to interpret. Moreover, the validity of VLF typically requires longer recording (exceeding 5 minutes [25], [41]. Furthermore, although non-linear SD2 provides information about the long-term variations in the NN interval fluctuations, as denoted by poincaré plot standard deviation along the line of identity, it does not represent any specific underlying physiological mechanism. Instead, it simply reflects the general complexity of the heart rate signal [25].

As predicted, a consistent trend of higher accuracy scores was observed with LOOCV across various models. Within LOOCV, 135 instances were used for training, and one is used for testing in each fold while in the hybrid technique, 108 instances were used for training. Since the majority of the data is used for training, the model can potentially learn more information, thereby achieving higher accuracy. Conversely, the hybrid technique reduced a number of training instances might lead to a less well-trained model. An additional factor to consider is the bias-variance trade-off. Since LOOCV is evaluated on only one instance at a time, the variance of the validation can be high, potentially leading to overly optimistic results. In contrast, the train-test split within the hybrid technique may provide a more balanced bias-variance trade-off, producing a more realistic estimation of the model's performance on previously unseen data. The difference in bias and variance might contribute to the lower accuracy scores compared to LOOCV. A further consideration is that LOOCV's repetitive fitting of the model (136 times) might induce overfitting, thus yielding a higher accuracy. In contrast, the hybrid technique involves fewer fittings of the model and might be less prone to overfitting, and might risk underfitting the data, resulting in lower accuracy [50]. Considering the advantages and disadvantages of both techniques, we recommend the use of LOOCV in HRV-based ML models, particularly when the dataset is relatively small as it can take into account individual differences. The hybrid technique should be used in scenarios involving large datasets or a computationally expensive model [50].

It is worth noting that this study does not specifically focus on differences in HRV across populations, rather, focusing on the ML models themselves. Exploring the extent to which demographic characteristics could influence HRV metrics would require an alternative methodology, such as a psychophysiological approach, which is beyond the scope of the current research. However, our unique dataset, representing HRV from a specific population (i.e., Indonesia), offers a perspective through which the effectiveness of ML models can be evaluated in various demographic settings. This highlights the potential benefits of tailoring ML algorithms to specific demographic groups, which is important yet often overlooked in the existing literature.

Limitations and Future Recommendations

This study has several limitations. One important limitation is the small sample size and homogenous sample characteristics (university students). Although the sample size of the current dataset is comparable to research on cognitive load [see Table 2], these findings should be confirmed in a larger study with more participants. This would allow for the generalizability of the conclusions. Further, the mental workload experiment was conducted in a controlled laboratory setting. This was to ensure the production of clean artifact-free datasets, thereby facilitating a fair comparison of different HRV measures and concluding the optimal physiological indicators of mental workload. Nevertheless, further studies need to replicate the experiment with more heterogenous samples, ideally in a real-work setting to enhance external validity.

While this current performance is considered acceptable, improvement of performance should be prioritized. This includes the exploration of several feature extraction strategies, such as employing segments of data (i.e., time windows) with various overlapping windows to enlarge a number of instances. While this study follows the recommended minimum recording duration of five minutes, other studies demonstrate that shorter time windows may produce good models [54]. Moreover, to enhance the generalization of mental workload model trained on a large population, the implementation of personalized models might be considered. This could involve a combination of samples from a large group, added with few individual-specific samples. In this context, calibration samples could function as the individual's "fingerprint," introducing unique attributes into the new model [53].

Implication

The results of this study highlight the connection between individuals' physiological characteristics, specifically HRV, and their experience of mental workload. This understanding may lead to the development of innovative strategies to adapt and support complex cognitive tasks, responding to real-time user engagement with various duties [2]. Monitoring mental workload in real-world environments could enhance human cognitive capabilities, particularly in decision-making scenarios where cognitive readiness is limited, and the likelihood of human error, due to factors like acute stress and other load factors, is elevated. This research represents a progressive step toward the future, exploring the utilization of physiological markers, derived from HRV, to distinguish between rest and mental workload.

Given the relatively uniform performance across all experimental combinations, the data recommend the use of a linear SVM classifier with selected features from the knowledge domain and LOOCV as a validation technique. This approach addresses two critical aspects: the creation of interpretable models within AI and the personalization of data that is person-dependent. For larger datasets, ensemble learning methods would be preferable.

Furthermore, this finding offers valuable insights into practical applications aimed at optimizing mental workload management. Such optimization can be achieved by the development of a wearable recognition system capable of accurately detecting increased mental workload in real-world situations and providing immediate feedback to the user. Finally, the dataset compiled during this study may foster interdisciplinary work, and encourage collaboration between researchers, practitioners, and other stakeholders in human factors and machine learning fields.

CONCLUSION

In the modern times, there has been a notable shift towards occupational roles that demand more complex cognitive processing, leading to the need for higher levels of mental workloads. This trend has been parallel with the rapid development of wearable sensing devices and advancements in artificial intelligence. Such developments lead to the growing interest in utilizing HRV as a promising approach for remotely and continuously monitoring workload. A substantial challenge in this domain, however, lies in the availability of relevant data for mental workload recognition, especially within specific populations, such as in Indonesia.

This current study aims to evaluate the performance of several HRV-based machine learning models: Support Vector Machine, Naïve Bayes, Gradient Boosting, and AdaBoost, employing a unique dataset gathered from experiments conducted within the Indonesian population. The research involves an analysis of each algorithm, applying HRV's full feature set and those selected from both knowledge domains and mRMR methods, and utilizing leave-one-out cross (LOOC) validation and hybrid validation techniques. The results reveal that the SVM classifier, coupled with knowledge domain selection and LOOC validation, is the best model. This finding emphasizes the potential of even simple machine learning models to predict mental workload through more interpretable features and LOOCV which can accommodate individual characteristics in HRV. The study provides insights into the development of a mental workload recognition system, potentially improving decision-making where cognitive readiness is constrained and the propensity of human error is elevated.

ACKNOWLEDGMENT

We express our gratitude to Rahma Indah Kusumawardani, Alimas Nurhati, Hadiatul Aula, and Lutfiyyah, students at Universitas Qomaruddin, for their valuable assistance in the experiment and data pre-processing phases.

CONFLICT OF INTEREST

The authors disclose no financial or personal conflicts of interest that could compromise the study's integrity.

FUNDING

This work was supported by the Directorate of Indonesian Higher Education under Research Grant for Junior Lectures [Grant number: 183/E5/PG.02.00.PL/2023].

References

[1] M. S. Young, K. A. Brookhuis, C. D. Wickens, and P. A. Hancock, "State of science: mental workload in ergonomics," Ergonomics, vol. 58, no. 1, pp. 1–17, Jan. 2015, doi: 10.1080/00140139.2014.956151.

[2] R. L. Charles and J. Nixon, "Measuring mental workload using physiological measures: A systematic review," Applied Ergonomics, vol. 74, no. September 2016, pp. 221–232, 2019, doi: 10.1016/j.apergo.2018.08.028.

[3] C. D. Wickens, "Multiple resources and mental workload," Human Factors, vol. 50, no. 3, pp. 449–455, 2008, doi: 10.1518/001872008X288394.

[4] G. Orru and L. Longo, "The Evolution of Cognitive Load Theory and the Measurement of Its Intrinsic, Extraneous and Germane Loads: A Review," in Human Mental Workload: Models and Applications, Springer International Publishing, 2019, pp. 23–48. doi: 10.1007/978-3-030-14273-5_3.

[5] P. Vanneste et al., "Towards measuring cognitive load through multimodal physiological data," Cognition, Technology and Work, vol. 23, no. 3, pp. 567–585, 2021, doi: 10.1007/s10111-020-00641-0.

[6] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven, "Cognitive Load Measurement as a Means to Advance Cognitive Load Theory," Educational Psychologist, vol. 38, no. 1, pp. 63–71, Jan. 2003, doi: 10.1207/S15326985EP3801_8.

[7] R. McKendrick, B. Feest, A. Harwood, and B. Falcone, "Theories and Methods for Labeling Cognitive Workload: Classification and Transfer Learning," Frontiers in Human Neuroscience, vol. 13, no. September, pp. 1–20, 2019, doi: 10.3389/fnhum.2019.00295.

[8] M. Gjoreski et al., "Cognitive Load Monitoring with Wearables-Lessons Learned from a Machine Learning Challenge," IEEE Access, vol. 9, pp. 103325–103336, 2021, doi: 10.1109/ACCESS.2021.3093216.

[9] D. Tao, H. Tan, H. Wang, X. Zhang, X. Qu, and T. Zhang, "A systematic review of physiological measures of mental workload," International Journal of Environmental Research and Public Health, vol. 16, no. 15, pp. 1–23, 2019, doi: 10.3390/ijerph16152716.

[10] K. F. A. Lee, W. S. Gan, and G. Christopoulos, "Biomarker-informed machine learning model of cognitive fatigue from a heart rate response perspective," Sensors, vol. 21, no. 11, pp. 1–16, 2021, doi: 10.3390/s21113843.

[11] M. Gjoreski et al., "Datasets for cognitive load inference using wearable sensors and psychological traits," Applied Sciences (Switzerland), vol. 10, no. 11, 2020, doi: 10.3390/app10113843.

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

[13] M. Schaffarczyk, B. Rogers, R. Reer, and T. Gronwald, "Validity of the Polar H10 Sensor for Heart Rate Variability Analysis during Resting State and Incremental Exercise in Recreational Men and Women," Sensors, vol. 22, no. 17, p. 6536, Aug. 2022, doi: 10.3390/s22176536.

[14] F. Schaule, J. O. Johanssen, B. Bruegge, and V. Loftness, "Employing Consumer Wearables to Detect Office Workers' Cognitive Load for Interruption Management," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, no. 1, pp. 1–20, 2018, doi: 10.1145/3191764.

[15] S. S. Panicker and P. Gayathri, "A survey of machine learning techniques in physiology based mental stress detection systems," Biocybernetics and Biomedical Engineering, vol. 39, no. 2, pp. 444–469, 2019, doi: 10.1016/j.bbe.2019.01.004.

[16] S. Ishaque, N. Khan, and S. Krishnan, "Trends in Heart-Rate Variability Signal Analysis," Frontiers in Digital Health, vol. 3, no. February, pp. 1–18, 2021, doi: 10.3389/fdgth.2021.639444.

[17] G. Vos, K. Trinh, Z. Sarnyai, and M. Rahimi Azghadi, "Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review," International Journal of Medical Informatics, vol. 173, no. February, p. 105026, 2023, doi: 10.1016/j.ijmedinf.2023.105026.

[18] A. W. K. Gaillard, “Comparing the concepts of mental load and stress,” Ergonomics, vol. 36, no. 9, pp. 991–1005, 1993, doi: 10.1080/00140139308967972.

[19] C. L. Bong, K. Fraser, and D. Oriot, “Cognitive Load and Stress in Simulation,” in Comprehensive Healthcare Simulation: Pediatrics, V. J. Grant and A. Cheng, Eds., in Comprehensive Healthcare Simulation. , Cham: Springer International Publishing, 2016, pp. 3–17. doi: 10.1007/978-3-319-24187-6_1.

[20] M. Lohani, B. R. Payne, and D. L. Strayer, “A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving,” Front. Hum. Neurosci., vol. 13, p. 57, Mar. 2019, doi: 10.3389/fnhum.2019.00057.

[21] C. Chen, C. Li, C.-W. Tsai, and X. Deng, “Evaluation of Mental Stress and Heart Rate Variability Derived from Wrist-Based Photoplethysmography,” in 2019 IEEE Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), IEEE, May 2019, pp. 65–68. doi: 10.1109/ECBIOS.2019.8807835.

[22] F. Shaffer and J. P. Ginsberg, “An Overview of Heart Rate Variability Metrics and Norms,” Frontiers in Public Health, vol. 5, no. 258, 2017, doi: 10.3389/fpubh.2017.00258.

[23] X. Arakaki et al., “The connection between heart rate variability (HRV), neurological health, and cognition: A literature review,” Front. Neurosci., vol. 17, p. 1055445, Mar. 2023, doi: 10.3389/fnins.2023.1055445.

[24] G. Forte, F. Favieri, and M. Casagrande, “Heart rate variability and cognitive function: A systematic review,” Frontiers in Neuroscience, vol. 13, no. JUL, pp. 1–11, 2019, doi: 10.3389/fnins.2019.00710.

[25] T. Pham, Z. J. Lau, S. H. A. Chen, and D. Makowski, “Heart rate variability in psychology: A review of HRV indices and an analysis tutorial,” Sensors, vol. 21, no. 12, pp. 1–20, 2021, doi: 10.3390/s21123998.

[26] K. Mohanavelu et al., “Cognitive Workload Analysis of Fighter Aircraft Pilots in Flight Simulator Environment,” Def. Sc. Jl., vol. 70, no. 2, pp. 131–139, Mar. 2020, doi: 10.14429/dsj.70.14539.

[27] S. H. Fairclough, L. Venables, and A. Tattersall, “The influence of task demand and learning on the psychophysiological response,” International Journal of Psychophysiology, vol. 56, no. 2, pp. 171–184, 2005, doi: 10.1016/j.ijpsycho.2004.11.003.

[28] M. Fallahi, M. Motamedzade, R. Heidarimoghadam, A. R. Soltanian, and S. Miyake, “Effects of mental workload on physiological and subjective responses during traffic density monitoring: A field study,” Applied Ergonomics, vol. 52, pp. 95–103, Jan. 2016, doi: 10.1016/j.apergo.2015.07.009.

[29] J. A. Veltman and A. W. K. Gaillard, “Physiological workload reactions to increasing levels of task difficulty,” Ergonomics, vol. 41, no. 5, pp. 656–669, 1998, doi: 10.1080/001401398186829.

[30] J. M. Splawn and M. E. Miller, “Prediction of perceived workload from task performance and heart rate measures,” Proceedings of the Human Factors and Ergonomics Society, no. April, pp. 778–782, 2013, doi: 10.1177/1541931213571170.

[31] H. F. Posada-Quintero and J. B. Bolkhovsky, “Machine learning models for the identification of cognitive tasks using autonomic reactions from heart rate variability and electrodermal activity,” Behavioral Sciences, vol. 9, no. 4, 2019, doi: 10.3390/bs9040045.

[32] G. Giannakakis, K. Marias, and M. Tsiknakis, “A stress recognition system using HRV parameters and machine learning techniques,” 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2019, pp. 269–272, 2019, doi: 10.1109/ACIIW.2019.8925142.

[33] K. Pettersson, J. Tervonen, J. Narvainen, P. Henttonen, I. Maattanen, and J. Mantyjarvi, “Selecting Feature Sets and Comparing Classification Methods for Cognitive State Estimation,” Proceedings - IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE 2020, pp. 683–690, 2020, doi: 10.1109/BIBE50027.2020.00115.

[34] K. Ross et al., “Toward Dynamically Adaptive Simulation: Multimodal Classification of User Expertise Using Wearable Devices,” Sensors, vol. 19, no. 19, p. 4270, Oct. 2019, doi: 10.3390/s19194270.

[35] M. Radovic, M. Ghalwash, N. Filipovic, and Z. Obradovic, “Minimum redundancy maximum relevance feature selection approach for temporal gene expression data,” BMC Bioinformatics, vol. 18, no. 1, p. 9, Dec. 2017, doi: 10.1186/s12859-016-1423-9.

[36] C. Rudin and J. Radin, “Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson from an Explainable AI Competition,” Harvard Data Science Review, vol. 1, no. 2, Nov. 2019, doi: 10.1162/99608f92.5a8a3a3d.

[37] Y. Mao et al., "How data scientists work together with domain experts in scientific collaborations: To find the right answer or to ask the right qestion?," Proceedings of the ACM on Human-Computer Interaction, vol. 3, no. 237, 2019, doi: 10.1145/3361118.

[38] B. Mahesh, E. Prassler, T. Hassan, and J. U. Garbas, "Requirements for a Reference Dataset for Multimodal Human Stress Detection," 2019 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2019, pp. 492–498, 2019, doi: 10.1109/PERCOMW.2019.8730884.

[39] R. Brickenkamp, "Test d2, Attentional Performance Test." Hogrefe, Göttingen, Germany, 1994.

[40] S. T. Mueller and B. J. Piper, "The Psychology Experiment Building Language (PEBL) and PEBL Test Battery," Journal of Neuroscience Methods, vol. 222, pp. 250–259, Jan. 2014, doi: 10.1016/j.jneumeth.2013.10.024.

[41] S. Laborde, E. Mosley, and J. F. Thayer, "Heart rate variability and cardiac vagal tone in psychophysiological research - Recommendations for experiment planning, data analysis, and data reporting," Frontiers in Psychology, vol. 8, no. 213, 2017, doi: 10.3389/fpsyg.2017.00213.

[42] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data Preprocessing for Supervised Leaning," vol. 1, no. 1, 2006.

[43] A. Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. Sebastopol, California: O'Reilly Media, Inc, 2019.

[44] S. Laborde, M. Raab, and N. P. Kinrade, "Is the ability to keep your mind sharp under pressure reflected in your heart? Evidence for the neurophysiological bases of decision reinvestment," Biological Psychology, vol. 100, no. 1, pp. 34–42, 2014, doi: 10.1016/j.biopsycho.2014.05.003.

[45] K. Dahal, B. Bogue-Jimenez, and A. Doblas, "Global Stress Detection Framework Combining a Reduced Set of HRV Features and Random Forest Model," Sensors, vol. 23, no. 11, p. 5220, May 2023, doi: 10.3390/s23115220.

[46] T. G. Dietterich, "Ensemble Methods in Machine Learning," in Lecture Notes in Computer Science, vol. 1857, no. 2, Berlin, Heidelberg, 2000, pp. 1–15. doi: 10.1007/3-540-45014-9_1.

[47] A. Saini, "Guide on Support Vector Machine (SVM) Algorithm." Accessed: Aug. 11, 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

[48] B. Alam, "Naive Bayes Classifier Python Tutorial 2023." Accessed: Aug. 11, 2023. [Online]. Available: https://hands-on.cloud/naive-bayes-classifier-python-tutorial/

[49] R. Kumar, "A Comparitive Study Between AdaBoost and Gradient Boost ML Algorithm." Accessed: Aug. 11, 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/adaboost-and-gradient-boost-comparitive-study-between-2-popular-ensemble-model-techniques/

[50] Y. Xu and R. Goodacre, "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning," J. Anal. Test., vol. 2, no. 3, pp. 249–262, Jul. 2018, doi: 10.1007/s41664-018-0068-2.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, and B. Thirion, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[52] W. Kraaij et al., "Personalized support for well-being at work: an overview of the SWELL project," User Modeling and User-Adapted Interaction, vol. 30, no. 3, pp. 413–446, 2020, doi: 10.1007/s11257-019-09238-3.

[53] K. Nkurikiyeyezu, A. Yokokubo, and G. Lopez, "Effect of person-specific biometrics in improving generic stress predictive models," Sensors and Materials, vol. 32, no. 2, pp. 703–722, 2020, doi: 10.18494/SAM.2020.2650.

[54] J. Tervonen, K. Pettersson, and J. Mäntyjärvi, "Ultra-short window length and feature importance analysis for cognitive load detection from wearable sensors," Electronics (Switzerland), vol. 10, no. 5, pp. 1–19, 2021, doi: 10.3390/electronics10050613.

AUTHORS BIOGRAPHY

Nailul Izzah is a lecturer in the Department of Industrial Engineering at Universitas Qomaruddin Gresik, Indonesia. She earned her Bachelor's degree in Mathematics Education from Universitas Islam Malang and a Master of Health from Universitas Airlangga, Surabaya. With over 15 years of experience teaching mathematics and statistics, her research interests lie in the areas of clustering and applied statistics.

Auditya Purwandini Sutarto serves as a lecturer in the Department of Industrial Engineering at Universitas Qomaruddin Gresik, Indonesia. She received her B.S. in Industrial Engineering from Institut Teknologi Bandung, followed by Master of Science in Mathematics from Universitas Gadjah Mada, and PhD in Technology Management from Universiti Malaysia Pahang, Malaysia. Her research contributions span across psychophysiology, ergonomics, and artificial intelligence, reflecting a deep commitment to interdisciplinary study.

Ade Hendi holds a B.S. and a Master Degree in computer science from Institut Sains dan Teknologi Palapa, Malang and Institut Teknologi Surabaya, respectively. Currently, he works at the Department of Informatics Engineering, Universitas Qomaruddin, Gresik, teaching subjects on system programming and mobile computing. His research interests cover broad topics on Internet of Thing and mobile computing.

Maslakhatul Ainiyah earned her Bachelor's Degree in Psychology from Universitas Darul Ulum, Jombang, and then hold a master degree in Psychology from Universitas 17 Agustus 1945, Surabaya. As a faculty member in the Department of English Education at Universitas Qomaruddin, Gresik, she specializes in teaching subjects related to psychology in education and counseling guidance. Her research focus on psychology in education.

Muhammad Nubli Bin Abdul Wahab has dedicated over 20 years to human development and technology. With a Bachelor's in Syariah, a Master's in Extension Education, and a Ph.D. in Management Information System, he is also a biofeedback practitioner with extensive training. Acknowledged nationally, his research focuses on the synergy of humans and technology, yielding accolades in competitions such as MTE, ITEX, PECIPTA, INPEX, and SIIF. His impactful contributions include software, gadgets, and biofeedback protocols that advance human potential. His broad expertise spans academic administration, teaching, supervision, consultancy, publications, and community services at University Malaysia Pahang.

DOI: 10.25077/josi.v22.n2.p81-98.2023