



Robust PPG-Based Mental Workload Assessment System Using Wearable Devices

Win-Ken Beh , Student Member, IEEE, Yi-Hsuan Wu, and An-Yeu Wu , Fellow, IEEE

Abstract—Heart rate variability (HRV) has been used in assessing mental workload (MW) level. Compared with ECG, photoplethysmogram (PPG) provides convenient in assessing MW with wearable devices, which is more suitable for daily usage. However, PPG collected by smartwatches are prone to suffer from artifacts. Those signal corruptions cause invalid Inter-beat Intervals (IBI), making it challenging to evaluate the HRV feature. Hence, the PPG-based MW assessment system is difficult to obtain a sustainable and reliable assessment of MW. In this paper, we propose a pre- and post-processing technique, called outlier removal and uncertainty estimation, respectively, to reduce the negative influences of invalid IBIs. The proposed method helps to acquire accurate HRV features and evaluate the reliability of incoming IBIs, rejecting possibly misclassified data. We verified our approach in two open datasets, which are CLAS and MAUS. Experiment results show proposed method achieved higher accuracy (66.7% v.s. 74.2%) and lower variance (11.3% v.s. 10.8%) among users, which has comparable performance to an ECG-based MW system.

Index Terms—Photoplethysmogram (PPG), mental workload assessment, Signal Quality Index (SQI), outlier removal.

I. INTRODUCTION

MENTAL Workload (MW) refers to the portion of operator information processing capacity or resources that are required to meet system demands [1]. A high mental workload means more information processing capacity or resources in performing a task. The MW assessment helps understand human operators in terms of processing capability or subjective psychological experiences [2]. Hence, MW assessment is an essential consideration for avoiding task error or working under overload conditions. As illustrated in Fig. 1, it possesses numerous applications, ranging from safety to smart technology, including driver awareness [3]–[5], mental health monitoring [6], and Brain-Computer Interfacing (BCI) [7].

Manuscript received 31 August 2021; revised 22 November 2021; accepted 21 December 2021. Date of publication 28 December 2021; date of current version 5 May 2023. This work was supported in part by the Ministry of Science and Technology of Taiwan under Grants MOST-109-2622-8-002-012-TA and MOST-110-2221-E-002-184-MY3, and in part by PixArt Imaging Inc., Hsinchu, Taiwan under Grant Pix-108053. (Win-Ken Beh and Yi-Hsuan Wu contributed equally to this work.) (Corresponding author: An-Yeu Wu.)

The authors are with the Department of Electrical Engineering, Graduate Institute of Electronics Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: kane@access.ee.ntu.edu.tw; rickwu@access.ee.ntu.edu.tw; andywu@ntu.edu.tw).

Digital Object Identifier 10.1109/JBHI.2021.3138639

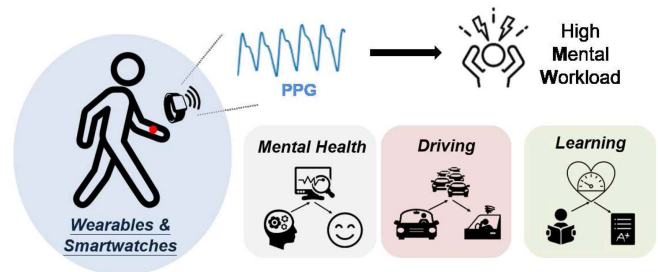


Fig. 1. Application scenario of the PPG-based MW system.

Cardiovascular parameters are affected by the mental state of a person. Studies have suggested that Heart Rate Variability (HRV) is a helpful index for assessing mental workload [8]–[10]. HRV acquired from ECG is widely used and has a long history for the assessment of mental workload. It can measure the time and frequency domain information from heart activity by the Inter-beat Intervals (IBI), indicating sympathetic and parasympathetic nervous activity for mental workload assessment.

Although we can use ECG to achieve the assessment of mental workload, however, we prefer photoplethysmogram (PPG) over ECG for long-term monitoring scenarios. It is because the measurement of ECG needs multiple contact points across the heart to form an electric loop. The way to measure ECG is unacceptable to smartwatch users when they need to put their finger on smartwatches for a long time. With regards to this, some PPG-based MW assessment systems have been introduced [11]–[13]. The benefit of using PPG-based MW assessment is the simple measurement that does not interfere with users' activity. On the contrary, the PPG peaks are less discernible than that of ECG, making it harder to detect the peak from PPG than ECG. Moreover, PPG is highly likely to be contaminated by motion noise during prolonged monitoring. Compared with ECG, valid beat interval retrieved from PPG is only around 50%, while ECG can result up to 99% of valid beat intervals [14]. The incorrect beats interval will propagate the error to the extracted HRV features, making the PPG-based system less robust.

Conventionally, researchers focus on finding valuable features or models to perform accurate MW assessments. However, there is less attention to the problem of misclassifying MW levels by contaminated PPG signals. In this paper, we developed pre- and post-processing techniques to reduce the negative impact of contaminated PPG, then present a robust PPG-based MW assessment system using smartwatches, as illustrated in Fig. 2. It consists of three parts:

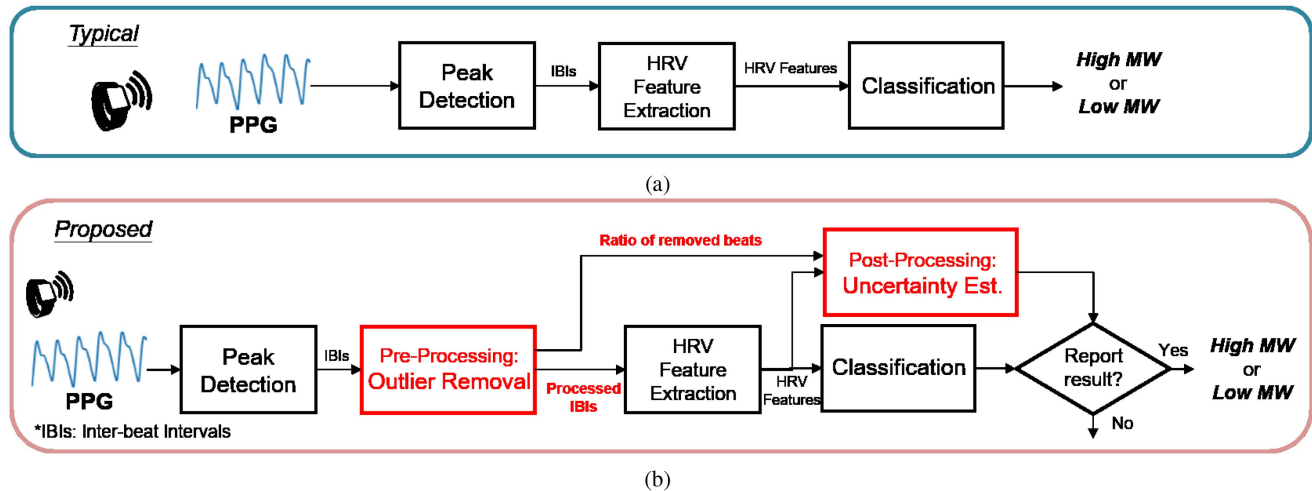


Fig. 2. Processing flow of the (a) conventional MW system and (b) the proposed MW system.

- 1) Typical MW assessment system
- 2) Pre-processing: Outlier Removing Mechanism
- 3) Post-processing: Uncertainty Estimation

The pre-process technique, called outlier removal, aims to detect and remove invalid PPG beat intervals to get an accurate HRV feature. Conventionally, outlier removal relies on numerical techniques [15], [16]. Different from conventional methods, we build an outlier removal mechanism by training a classifier. We have used signal quality indices (SQIs) as features, and the invalid beat interval label by synchronized ECG.

After removing several invalid beat intervals, it could affect the detection reliability. Hence, we propose a post-process technique, called the uncertainty estimation model, to estimate the reliability of the incoming processed signal. The proposed post-process technique will determine the uncertainty level of the signal. A higher uncertainty level means the processed signal has a higher probability of reporting random outcomes, which also means unreliable. Once the detection result is not reliable enough, we will not report the detection result. With the combination of pre- and post-process techniques of PPG signals, we can achieve similar performance to ECG-based MW detection.

In summary, our contributions are listed below:

- 1) We propose a pre-processing technique for removing IBIs outlier. Different to conventional numerical methods, the proposed outlier removal is composed of machine learning classifier and Signal Quality Indices (SQIs) as features, which can accurately remove invalid beats. As a result, the calculated HRV feature has only 0.05% absolute relative error, while conventional methods have 1.8% of error.
- 2) We created a ECG-assisted labeling method to retrieve the outlier label for outlier removal system. This method allows us automatically generate a sequence of outlier label cross-checking the IBIs between ECG and PPG, which reduce a lot of effort in manual labeling.
- 3) We propose a post-processing technique, called uncertainty estimation, for rejecting probably misclassified data. To the best of our knowledge, we are the first to investigate the misclassified probability of the processed

IBIs. The experiment result shows that the estimated score from proposed technique is highly correlated to the wrong classification ratio. By rejecting unreliable data, we further improves the accuracy of MW detection and lowering the standard deviation among peoples.

The remaining part of this paper is organized as follows: We will first introduce typical MW assessment system in Section II. After that, we explain the proposed pre-processing and post-processing techniques in Section III and Section IV respectively. Next, the overall system performance will be illustrated in Section V. Finally, we conclude our work in Section VI.

II. REVIEW: TYPICAL PPG-BASED MW ASSESSMENT SYSTEM

In this section, we will introduce the typical processing flow of PPG-based MW assessment system. The processing flow of typical PPG-based MW assessment system is illustrated in Fig. 2(a). First, the system will perform a peak detection to get inter-beat-intervals (IBIs). Next, the IBIs sequences are then used to extract HRV feature. Lastly, these features will be use to classify MW levels. We will describe these processing block in following subsection.

A. Peak Detection

PPG is susceptible to noise such as movement, breathing, and wristband tightness, which can cause baseline wandering in signals. As illustrated in Fig. 3(a), a severe baseline wander will lead to a wrong/inaccurate peak position. Hence, before detecting the peak, we have applied Empirical Mode Decomposition [17] for baseline wander removal. We chose EMD because the peak locations can be reserved after employing it to raw PPG. Next, we used the algorithm from [18] for the peak detection, shown in Fig. 3(b). We can get IBIs from either detect PPG peak or valley. In our work, we chose to detect PPG valley instead of peak because PPG valley has higher contrast for the detection. The IBIs from valley position are considered to be more reliable than peaks for HRV analysis [19].

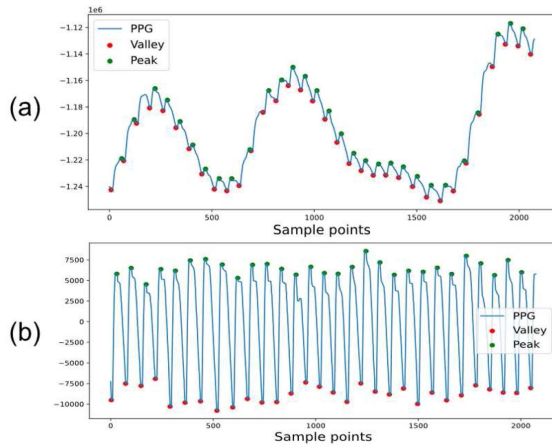


Fig. 3. Peak detection on: (a) raw signal and (b) processed signal by EMD.

B. Feature Extraction of HRV

After getting the IBIs, we are going to extract HRV feature from IBIs. HRV is correlated with MW changes and tends to diminish when a higher MW is generated [20]. It is calculated by analyzing the time series of the beat-to-beat intervals from ECG or PPG. There are various methods for extracting HRV, which can be divided into three analysis methods of time-domain, frequency-domain, and nonlinear-domain.

- 1) *Time Domain*: The features indicate the total variability of heartbeats using statistical methods [21], including the standard deviation of IBIs (SDNN), the square root of the mean squared differences between adjacent IBIs (RMSSD), the standard deviation of differences between adjacent IBIs (SDSD), the count or percentage of successive beats lengths that differed more than 50 ms (NN50, pNN50), the IBIs triangular index and the triangular interpolation of IBI histogram (TriIndex, TINN).
- 2) *Frequency Domain*: A spectrum estimation was calculated for the IBI series. We estimated the power spectrum by fast Fourier transform Welch's periodogram techniques. The spectrum was then divided into very low frequency (VLF, 0-0.04 Hz), low frequency (LF, 0.04-0.15 Hz), and high frequency (HF, 0.15-0.4 Hz). Since the signal length for calculating HRV features was 2 minutes, the VLF would be calculated with an incorrect value and therefore not taken. The total power (TF) and LF/HF ratio were also calculated. The LF and HF were also represented as the normalized units (nLF, nHF) to extract the sympathovagal component of the HRV better.
- 3) *Non-linear Domain*: We use the time-delay embedding methods to capture the nonlinear properties of the PPI time series. The nonlinear methods are introduced for estimating the complexity of the time series and constructing the relation with mental states. The nonlinear measurements include Poincaré plot [22] and Correlation dimension [23].

The extracted HRV features were listed in Table I. They are extracted from each session using 2 minutes windows with 30 seconds overlap, as suggested in [24].

TABLE I
TABLE OF EXTRACTED FEATURES

Feature Type	Features	p value
<i>Time-Domain</i>	SDNN	0.0271*
	NN50	0.0300*
	pNN50	0.0060*
	RMSSD	0.6114
	SDSD	0.9311
	TINN	0.2059
	TriIndex	2.89 E-4*
<i>Frequency-Domain</i>	Total Freq. (TF)	0.0514
	High Freq. (HF)	0.0500*
	normalized High Freq. (nHF)	0.0015*
	Low Freq. (LF)	0.0079*
	normalized Low Freq. (nLF)	0.6054
	LF/HF	0.0015*
<i>Nonlinear-Domain</i>	Poincaré plot (SD1)	0.6186
	Poincaré plot (SD2)	0.5988
	Poincaré plot (SD1/SD2)	0.6119
	Correlation Dimension (CD)	2.90 E-15*

*: p-value < 0.05 (significant difference).

TABLE II
SVM PARAMETERS

Parameter	Setups
Kernel	Linear
Regularization C	(10e-3 ~ 300)
Class	2
Class weight	(0: 0.67, 1: 0.33)

C. Classification

The extracted HRV features would then be used for MW classification. Considering the physiological response was inherently participant independent, the participant-based feature standardization was applied [25]. The detection of MW was a two-class classification problem using the difficulty of the N-back tasks as the ground truth. Since higher N would stimulate a higher MW state, we considered the 2-, 3-back tasks as the “high” MW states and the 0-back task as the “low” MW state. Support vector machine (SVM) [26] classifiers with a linear kernel were used for MW classification. The linear SVM adjustable regularization term C was fine-tuned by grid-search in the range [10e-3, 300]. The parameters of SVM are listed in Table II. The single modality classifier was trained separately for each channel (ECG, fingertip PPG, and wrist-worn PPG).

III. PRE-PROCESSING: OUTLIER REMOVAL MECHANISM

The IBI outlier is defined as an abnormal beat interval (artifact). Typical sources of outliers include additional or missed beats that are caused by erroneous peak detection under motion. Generally, detecting outlier is basically rely the value of IBI. Some extreme IBI values will be regarded as outlier. In our work, we compared the IBIs that retrieved from PPG to ECG, as illustrated in Fig. 4. The long beat and the short beat can be regarded as outliers. These IBI outliers frequently occur when using wearable sensors such as PPG, which can contribute to the decline in the performance of HRV-based applications. In Fig. 5, we showed the HRV feature value calculated from two IBI sequences retrieved from PPG and ECG. The blue line represents the feature extracted from ECG's IBIs (RRI),

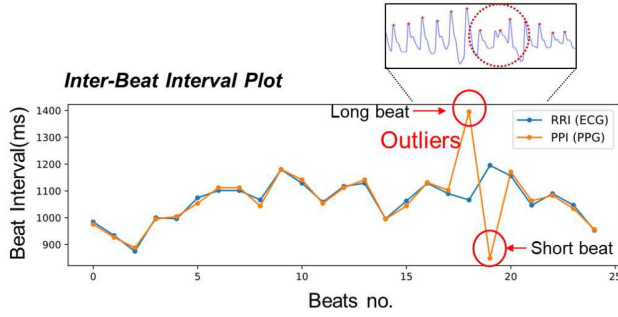


Fig. 4. Comparison of RRI (blue), PPI with outliers (orange).

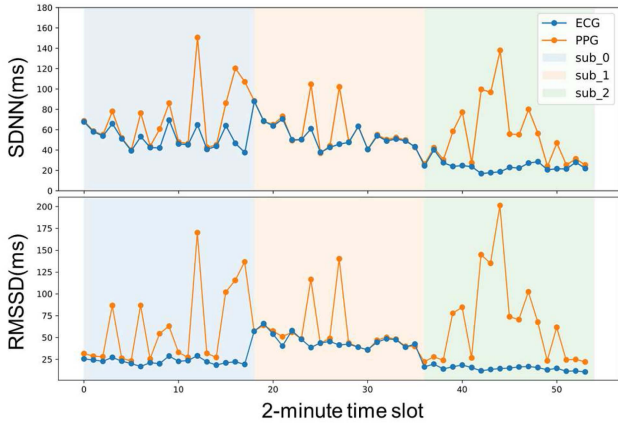


Fig. 5. Estimated Time-domain HRV features (SDNN and RMSSD) from ECG and PPG signals.(Each data point is extracted from a 2-minutes length signal.).

while the orange line represents PPG's IBIs (PPI), which are accompanied with outliers. From Fig. 5, we observe that the HRV feature value calculated from PPG has much difference to the feature value calculated from ECG. It is because outliers often occur in PPG signals, those outliers will have a dramatic impact on HRV feature value. It infers that the removal of outliers is critical for making the detection on wearable devices robust. The traditional methods detect outliers by examining the value changes in the IBI sequence. However, these numerical methods are not suitable for all conditions. Threshold values and parameters also need to be adjusted over time. In addition, it is difficult to determine whether the detected outliers are genuine outliers or just longer or shorter beat intervals. A way to catch the outliers is from the waveform domain since erroneous beat intervals are caused by incorrect peak detection of corrupted waveforms.

In our work, we detect PPG valleys to obtain the IBI sequence, we called it valley-to-valley interval(VVI). The PPG waveform between valleys is a PPG pulse. Therefore, it motivates us to detect outliers from PPG pulses. As shown in Fig. 6, if the waveform of a PPG pulse between the detected valleys is complete, the corresponding beat interval is considered to be a correct IBI. On the contrary, if the waveform is incomplete and distorted, the corresponding beat interval is considered an IBI outlier.

Hence, the proposed outlier removing mechanism will calculate PPG pulses characteristic or signal quality indices (SQIs).

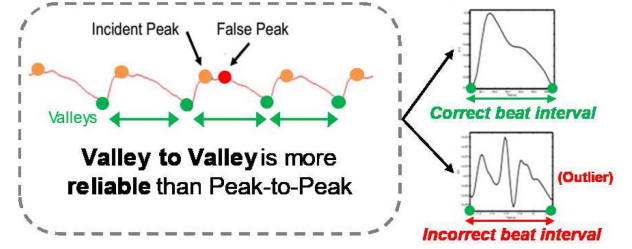


Fig. 6. Valley-to-Valley intervals (VVI) in PPG and its relationship with outliers.

Then, those SQIs of pulses will be sent to a trained detection model to determine whether it is an outlier. The detection model is trained by the SQIs and outlier labels, or the outlier labels are obtained from comparing the IBIs to synchronized ECG signals. Finally, we removed those outliers based on model's prediction. The process flow of proposed outlier removing mechanism is illustrated in Fig. 7.

A. Signal Quality Indices (SQI) of PPG Pulse

To determine whether the PPG pulse is corrupted enough to induce the outlier in the IBI sequence, we have to evaluate the waveform of the PPG pulses. There have been several metrics used for assessing the PPG signal quality. For example, in [27] the author used a series of rules to determine ECG/PPG signal quality. Several metrics, such as HR, IBIs, and template matching correlation, are checked. Since we aim to detect outliers rather than determine signal quality in our work, some statistical metrics from [28] are good enough to use. Those metrics are Perfusion, Skewness, Kurtosis, and Entropy, which can also be used to describe the PPG waveform feature. They are defined as follows:

$$Perfusion = [(y_{\max} - y_{\min}) / |\bar{x}|] \times 100, \quad (1)$$

where \bar{x} is the statistical mean of the x signal (raw PPG signal), and y is the filtered PPG signal.

$$Skewness = \frac{1}{N} \sum_{i=1}^N \left[x_i - \frac{\hat{\mu}_x}{\sigma} \right]^3, \quad (2)$$

where $\hat{\mu}_x$ and σ are the empirical estimate of the mean and standard deviation of x_i , respectively, and N is the number of samples in the PPG signal.

$$Kurtosis = \frac{1}{N} \sum_{i=1}^N \left[x_i - \frac{\hat{\mu}_x}{\sigma} \right]^4, \quad (3)$$

$$Entropy = - \sum_{n=1}^N x[n]^2 \log_e(x[n]^2). \quad (4)$$

The features above statistically describe the waveform of PPG for determining if the waveform is corrupted. In addition to statistical methods, some template-based features [29] has been used to assess the quality of the PPG signal. These methods build a PPG pulse template for each subject and evaluate the quality by comparing it with the template waveform. Dynamic

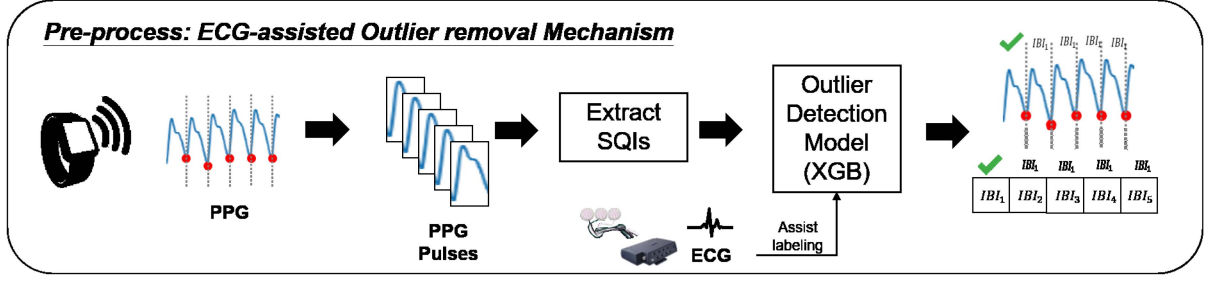


Fig. 7. Process flow of outlier removal mechanism: It includes a SQLs extractor and outlier detection model (XGB).

TABLE III
FEATURE LIST TABLE OF PPG WAVEFORM

Feature Type	Features	
Waveform Domain	Statistical	Perfusion
		Entropy
		Kurtosis
		Skewness
	Template	DTW, DDTW (derivative PPG) Pearson's r, Pearson's r of derivative PPG
Temporal Domain		d_IBI (difference)
		r_IBI (ratio)
		m_IBI (median)

time warping (DTW) and Pearson's correlation coefficient are applied to compare the template waveform similarity.

The other method to compare the similarity is using Pearson's correlation coefficient, which can be defined as follows:

$$\text{Pearson's } r = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^N (x_i - \mu_x)^2} \sqrt{\sum_{i=1}^N (y_i - \mu_y)^2}}, \quad (5)$$

where x is the PPG signal to be compared, and y is the template PPG waveform.

In addition to those features that assess the PPG signal quality, we also include the information of consecutive IBIs, which is inspired by the idea of the traditional filter – based outlier detection method. We consider the difference between adjacent IBIs, the ratio between IBIs, and the difference between the median of nearby IBIs that is motivated by the Kubios filter, which can be defined as follow:

$$d_IBI(i) = IBI(i) - IBI(i-1), \quad (6)$$

$$r_IBI(i) = IBI(i)/IBI(i-1), \quad (7)$$

$$m_IBI(i) = IBI(i) - \text{median} \left[IBI \left(i - \frac{w}{2}, \dots, i + \frac{w}{2} \right) \right], \quad (8)$$

where w is the window size for computing the median. The total extracted features for assessing are listed in Table III.

B. ECG-Assisted Outlier Labeling

Next, before training a machine learning model to determine outliers, we have to collect labels for training, knowing what

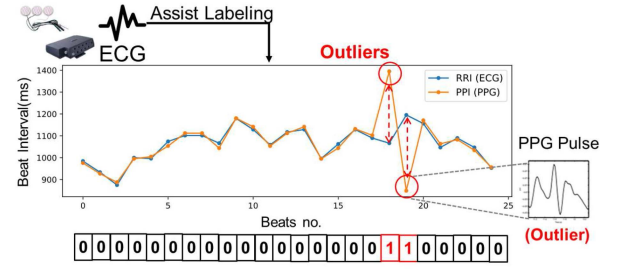


Fig. 8. ECG-assisted labeling to identify the outliers.

type of PPG pulses corresponds to the outliers. That is, each PPG pulse must have a corresponding label indicating whether it is an outlier or not. Conventionally, many human resources are required to obtain these labels from the enormous number of PPG pulses, which is time-consuming and a waste of labor. In addition to these drawbacks, the standards of different annotators are inconsistent, and the distorted waveforms perceived by humans do not necessarily lead to outliers in the IBI. Therefore, we proposed the ECG-assisted method to label the outliers. We can obtain two IBI sequences from ECG (RRI) and PPG (VVI) by the simultaneously collected ECG signal. Since ECG is much more stable than PPG, the points with large differences in RRI and VVI can be considered as outliers, as shown in Fig. 8.

Before we used ECG as a reference, we used the method from [27] to systematically check the signal quality of ECG. The method from [27] is employed a series of rules to determine the quality of ECG signals, which include: “HR between 40-180 bpm”, “All RR intervals are $\leq 3s$ ” and “Max RR interval/Min RR interval $< 2.2s$ ”. The ECG signal segments are recognized as good quality if the salient feature (R peaks) can be detected.

Next, we extracted RRI from ECG and aligned it with the VVI sequence. We perform shifting and comparing the Root Mean Squared Error (RMSE) of the two sequences, and the shifted amount corresponds to the minimum RMSE. After the alignment, the second stage is comparing the RRI and VVI values. If the difference is greater than a predefined threshold, the beat interval in VVI is labeled as an outlier. If we encounter the case of an additional or missing peak, the length of the RRI and VVI sequence will be different. In this case, it is hard to compare these two sequences to get outlier labels. Hence, we will repeat beat interval alignment in the first stage after every detected outlier. The two stages are repeated after all beat intervals are compared. With the assistant of ECG signal,

we can automatically obtain the label of outliers, which can precisely label outliers and save much time. After we have the label of outliers, we can train our outlier detection model from the information of PPG pulses and corresponding outlier labels.

C. Outlier Detection Model - Extreme Gradient Boosting (XGBoost)

After obtaining the PPG features and corresponding outlier labels, we can train the classification model to detect outliers. Among all the machine learning algorithms, gradient boosting tree – based model [30] has been shown in many applications in different domains. XGBoost [31] is an efficient and scalable gradient boosting machine, which has won lots of machine competitions in recent years [32]. It is an ensemble model consisting of sets of *classification and regression tree (CART)*. While XGB is used for supervised learning problems, and we use training data x_i to predict a target variable y_i , the model can be described in the form:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F. \quad (9)$$

K is the total number of trees, f_k for the k^{th} tree is a function in the functional space F , and F is the set of all possible CARTs. In training, each of new – trained CART will try to complement the so – far residual. Objective function optimized at $(t+1)^{\text{th}}$ CART is described:

$$obj = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^t \Omega(f_i), \quad (10)$$

where $l()$ denotes the training loss function, y_i the is ground truth and $\hat{y}_i^{(t)}$ is the prediction value at step t . $\Omega()$ given by:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (11)$$

is the regularization term, where T is the number of leaves and w_j is the score on the j^{th} leaf. When Eq. 10 is optimized, Taylor's expansion is used to use gradient descent for different loss functions. Furthermore, feature selection is no need when we use the XGBoost approach. During the training period of XGBoost, good features would be chosen as a node in trees, which means features not used are abandoned.

In this paper, we use the scikit – learn API for XGBoost classification. The inputs of XGBoost have eleven features, and the outputs are prediction results for outlier detection. The performance will be shown in the next section.

D. Validation of the Pre-Processing Method

We validate the pre-processing technique in this subsection. The purpose of outlier removal is to reduce the feature error and gap between PPG and ECG. Hence, we evaluate the effectiveness of different outlier removal methods by comparing the extracted HRV feature value between ECG and PPG. The absolute relative error (ARE) [33] is used to calculate the feature error. We calculate the error of each HRV feature for different outlier

TABLE IV
COMPARISON OF FEATURE ERROR BETWEEN DIFFERENT OUTLIER REMOVAL METHODS

Method	Do nothing	Quotient Filter [15]	Kubious Filter [16]	Proposed	Outlier Label
Mean ARE (%)	144.92	28.13	24.95	23.17	23.15
Compare to Outlier Label	+122%	+4.98%	+1.8%	+0.05%	-

removal methods to assess the effectiveness. ARE is defined as:

$$ARE(i, j) = \left| \frac{f_{i,j}^{ECG} - f_{i,j}^{PPG}}{f_{i,j}^{ECG}} \right| \times 100(\%), \quad (12)$$

and $meanARE$ is defined as:

$$meanARE = \frac{\sum_{i=1}^N \sum_{j=1}^M ARE(i, j)}{N \times M}. \quad (13)$$

$f_{i,j}^{ECG}$ denotes subject j 's i^{th} HRV feature extracted from ECG, $f_{i,j}^{PPG}$ denotes subject j 's i^{th} HRV feature extracted from PPG, N is the HRV feature number and M is the subject number.

We compare the HRV features error in five cases. First, we extract HRV features from the VVI without removing outlier (Do nothing) and compute the error. Then we applied the quotient filter [15] and Kubios filter [16] to detect and remove the outliers of VVI and calculated the error of extracted HRV features, respectively. The fourth case was applying the proposed outlier removal method on VVI and calculating the extracted HRV features error. Lastly, we removed the outlier based on the outlier labels identified from the ECG and calculated the error of extracted outlier, which was regarded as the minimum error since all outliers were removed (Outlier label). The results is listed in Table IV.

We can observe that the feature error is significant without removing the outlier points. After applying the filter-based approaches (quotient and Kubios) to remove outliers, the error of the HRV feature values can be significantly reduced. However, there is still a gap between that and the best-case scenario. The proposed methods can achieve the slightest feature error with a 0.05% difference from the best case.

IV. POST-PROCESSING: UNCERTAINTY ESTIMATION

In the previous section, the outlier removal method is introduced to remove the irregular beat intervals. However, the removed VVIs imply data loss of those segments and have two levels of impact. First, the extracted HRV features will have some errors due to data loss. Second, there will be uncertainty in the classification caused by the error of input features. Proposed post-processing is designed to quantify the uncertainty brought by those removed VVIs. A high uncertainty score will infer a higher probability of causing the wrong detection result. This estimation process is illustrated in Fig. 9: We will first estimate the HRV feature error causing by missing IBIs. Next, according to the estimated feature error, we will analyze the influence on classification and estimate the probability of processed signal wrong classifying.

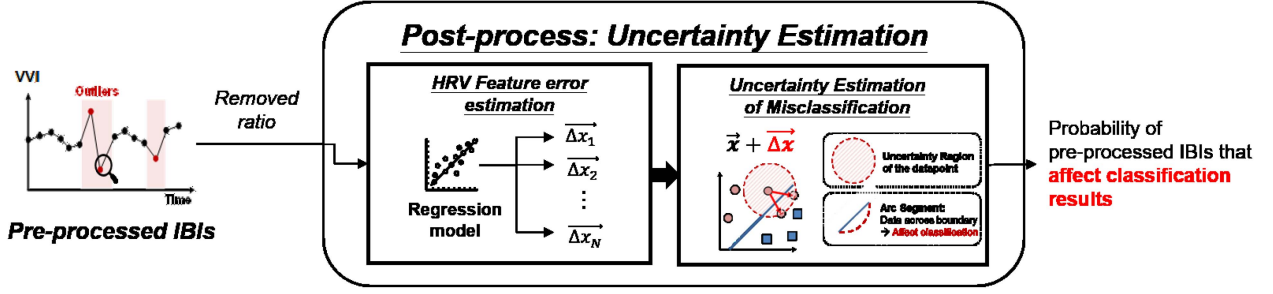


Fig. 9. Flowchart of post-process mechanism. It includes HRV feature error estimation and uncertainty estimation of misclassification.

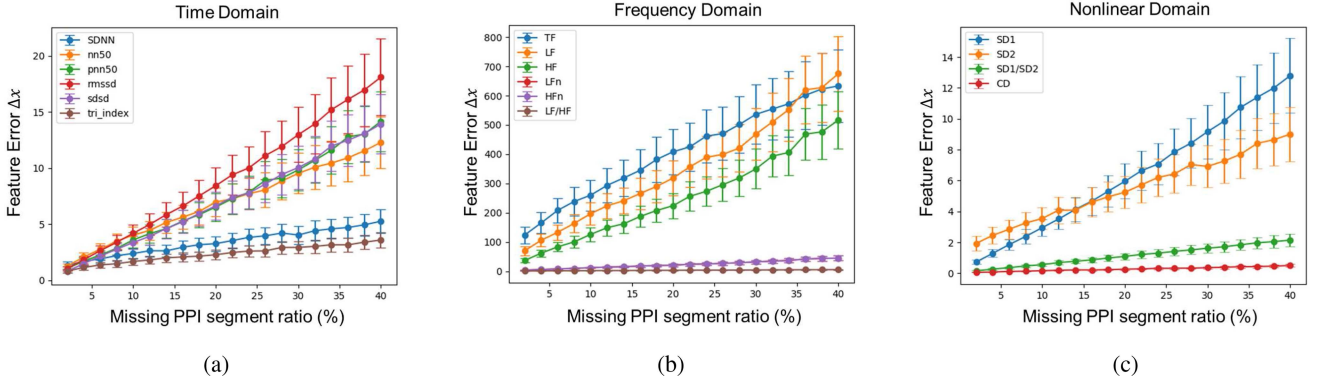


Fig. 10. Simulation result of feature error with different ratios of missing IBIs. (a) Time-domain, (b) Frequency-domain, and (c) Nonlinear-domain.

A. Error Estimation of HRV Features

To estimate the feature error, we perform an experiment that randomly removes certain proportions of the IBI sequence and compares the features extracted with those from the complete IBI. According to the simulation result in Fig. 10, we can observe the trend of the feature error of each HRV feature to missing VVI ratio. The growth of each feature error increases linearly as the portion of missing segments increase, but with different slopes. Our observation is consistent with the research in [34]. Hence, once we got the missing ratio of processed IBIs, we estimate this error by the regression model. The error of each feature can be estimated for a given missing ratio of IBI.

B. Influence of Feature Error on SVM Classification

The linear SVM [26] is used for mental workload classification in this paper. The mathematical properties of the linear SVM allow us to explore the influence of feature errors on classification. In the testing phase, the binary classification function of linear SVM is defined as:

$$y = f(\mathbf{x}) = \text{sign}(\vec{\mathbf{w}} \cdot \vec{\mathbf{x}} - b), \quad (14)$$

where $\vec{\mathbf{w}}$ ($\vec{\mathbf{w}} \in \mathbb{R}^n$) denotes the weight of each feature, which is also the normal vector to the hyperplane, $\vec{\mathbf{x}}$ ($\vec{\mathbf{x}} \in \mathbb{R}^n$) denotes the input vector, $\frac{b}{\|\vec{\mathbf{w}}\|}$ determines the offset of the hyperplane, and y is either 1 or -1, each indicating the class where $\vec{\mathbf{x}}$ belongs to.

If the input feature vector comes with a known error $\Delta \vec{\mathbf{x}}$, the input vector becomes $\vec{\mathbf{x}}' = \vec{\mathbf{x}} + \Delta \vec{\mathbf{x}}$. The uncertainty of features makes data points shift [35]. The error of input feature

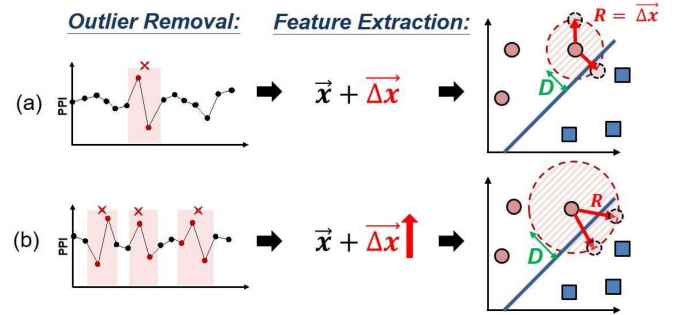


Fig. 11. Influence of feature error on classification: (a) smaller feature error and (b) bigger feature error.

will affect the classification if:

$$\text{sign}(\vec{\mathbf{w}} \cdot (\vec{\mathbf{x}} + \Delta \vec{\mathbf{x}}) - b) \neq \text{sign}(\vec{\mathbf{w}} \cdot \vec{\mathbf{x}} - b), \quad (15)$$

which means that the data point across the hyperplane and the classification result will be wrong.

Therefore, if we can estimate the feature error, we can evaluate whether the error will affect the classification. Since the feature error is a scalar, instead of a vector, the uncertainty of the datapoint can be approximated as a circular region with a radius of $\|\Delta \vec{\mathbf{x}}\|$.

As the outlier removal system removes more outliers, the error of the extracted features becomes larger. The larger feature error increases the uncertainty of the datapoint. We can expect that the radius of uncertainty region in SVM increases as well, which will raise the risk of misclassification, as shown in Fig. 11.

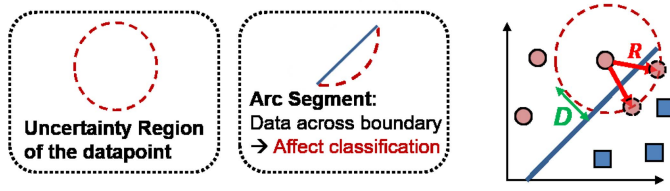


Fig. 12. Uncertainty estimation model of linear SVM.

The larger radius of the uncertainty region will lead to a higher risk of misclassification. It makes the feature error a factor that can be used to evaluate the risk of misclassification. In addition, another factor to consider is the distance from the data point to the decision boundary, denoted by D in Fig. 11. The closer the data point is to the boundary, the greater the risk of misclassification. Therefore, to estimate the probability of misclassification in linear SVM, we will consider two factors, the feature error can be estimated using regression, and the distance from the boundary can be obtained from the trained parameters in SVM.

C. Uncertainty Estimation of Wrong Classification

This section aims to quantify the probability of misclassification by a simple model with the two factors that affect the SVM classification. The first is the uncertainty of the input feature, which is the feature error that can be estimated from the ratio of removed beats. It is defined as:

$$R = \|\vec{\Delta x}\|, \quad (16)$$

where $\|\vec{\Delta x}\|$ is the feature error obtained by the regression model. The other is the distance from the data point to the decision boundary, which can be regarded as the confidence level of the input data, defined as:

$$D = \frac{|\vec{w} \cdot \vec{x} + b|}{\|\vec{w}\|}, \quad (17)$$

where \vec{w} is the weight of the linear SVM, \vec{x} is the input vector, HRV features, and the b is the interception of the linear SVM.

The misclassification occurs when the datapoint crosses the decision boundary under the influence of feature error and falls in the arc segment area, as illustrated as solid arc line in Fig. 12. Thus, the probability of misclassification is modeled as the ratio of the arc segment to the circular segment. Those feature points close to the decision boundary, which is the case of small D , are more likely to cross the decision boundary, causing misclassification. Moreover, for those cases with significant feature error R , misclassification is more likely to happen. The calculation of uncertainty level can be formulated as below:

$$UncertaintyLevel = \frac{1}{\pi} \cos^{-1} \frac{D}{R}. \quad (18)$$

D. Validation of Post-Processing Method

To validate whether the model can correctly estimate the probability of misclassification, we experiment to test the correlation between the model's result and the actual result. The PPG data pool contains data with different rates of VVI removed

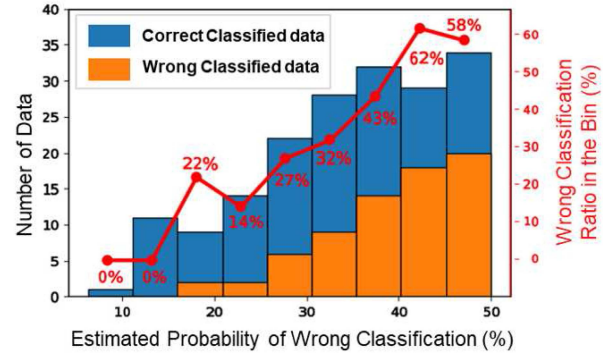


Fig. 13. Validation of the uncertainty estimation model.

depending on the degree of contamination of the movement. The first step is to divide the data pool into bins with varying misclassification probabilities, i.e., the first bin is for data with 0% misclassification probability, and the last bin is for data with 50% misclassification probability.

The second step is to perform the mental workload classification on each bin's data separately. The purpose is to check the error rate in each bin. Ideally, the bin with a higher estimated misclassification probability will have a higher error rate. The results are shown in Fig. 13. The x-axis is the estimated probability of the proposed estimation model (uncertainty level), and the y-axis on the left side represents the data number in each bin. The blue bars represent correctly classified data, and the orange bar represents misclassified data. The red number and the y-axis on the right side represent the error rate in each bin. The error rate and the estimated uncertainty level have a Pearson's r correlation coefficient of 0.96, indicating that the uncertainty estimation model was validated.

V. EVALUATION ON OPEN DATABASE

The experiment result of overall system is presented in this section. First, we compare the MW assessment system with and without the pre- and post- processing blocks. Next, we concatenate these two techniques into typical MW assessment processing flow, which was illustrated in Fig. 2(b), to show the improvement of overall accuracy.

We have used CLAS [36] and MAUS [37] dataset for our experiment. CLAS dataset provides ECG, GSR, and ear-lobe PPG. The MW level was elicited through a series of interactive tasks, such as Math and Logic problem. On the other hand, MAUS dataset provides ECG, fingertip-PPG, wrist-PPG, and GSR signals of 22 subjects. The MW level is induced by performing the N-back task. We choose to use these datasets because it contains partially-contaminated PPG, which has a different outlier ratio in VVI over subjects. The PPG signals from these two datasets have room for improvement with some processing techniques. Hence, they fit into our scenario, and we can show how much improvement by the proposed technique. Furthermore, these datasets contain synchronized ECG signals that can be used to find the limits of improvement that the processing technology can make. With regards to this, we chose these two datasets for the evaluation of the proposed technique.

TABLE V
COMPARISON RESULT OF THE FRAMEWORKS WITH AND WITHOUT THE PRE- AND POST- PROCESSING BLOCKS

		Modality	Accuracy(%)	F1-score(%)	Rejection Rate(%)	Sample Size
CLAS [37]	MW System	PPG	64.7 \pm 10.0	66.9 \pm 8.0	-	155 (31 Subjects)
	MW System + Outlier Removal	PPG	70.7 \pm 6.1	73.2 \pm 5.2	-	
	MW System + Outlier Removal + Uncertainty Estimation	PPG	78.3 \pm 4.6	79.9 \pm 3.8	31.6	
		ECG	81.3 \pm 1.6	82.1 \pm 1.4	-	
MAUS [38]	MW System	PPG	66.7 \pm 11.3	67.5 \pm 11.9	-	342 (19 Subjects)
	MW System + Outlier Removal	PPG	69.4 \pm 11.0	69.9 \pm 10.7	-	
	MW System + Outlier Removal + Uncertainty Estimation	PPG	74.2 \pm 10.8	74.5 \pm 10.4	28.4	
		ECG	75.3 \pm 8.9	75.4 \pm 8.9	-	

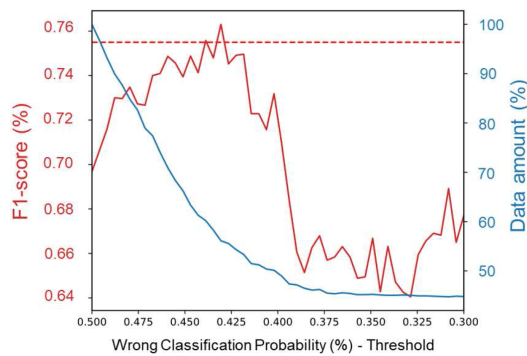


Fig. 14. Threshold determining of the uncertainty estimation.

Before the comparison, we determine the threshold of the uncertainty model for rejecting the high probability of misclassification data, we test the system under different thresholds. The threshold ranges from 0.5 to 0.3. A threshold of 0.5 means rejecting data with a probability of misclassification greater or equal to 0.5, which is the maximum value of the uncertainty estimation model. As the threshold moves from 0.5 to 0.3, indicating we are rejecting more data. The threshold of 0.3 means we reject data with a misclassification probability greater or equal to 0.3. As a result, we are rejecting the data of probability range from 0.3 to 0.5. The result is shown in Fig. 14.

The x-axis is the threshold of wrong classification probability from the uncertainty estimation model, the y-axis on the left side is the average F1-score of testing, and the y-axis on the right side is the ratio of remaining data (for prediction instead of rejection). The red dash line represents the f1-score of the ECG. We lower the threshold from 0.5 to 0.3, meaning we reject the data from the higher to a lower probability of misclassification. The remaining data decreases as the threshold decreases. As data with a high probability of misclassification are rejected, the F1-score of MW prediction increases. The F1-score of PPG approaches the ECG as the threshold decreases. After a threshold of 0.4, there is a dramatic drop in the F1-score, suggesting that some correctly classified data are rejected. Therefore, we select a threshold of 0.458 considering the performance and remaining data amount. The experiment setup is listed in Table VI.

TABLE VI
EXPERIMENT SETUP

Processing Block	Parameter	Setup
Pre-processing: Outlier Removal	Classifier	XGBoost
	Feature Number	11
	Class	2 (non-outlier/outlier)
Post-processing: Uncertainty Estimation	Threshold	0.458
Mental Workload Detection	Classifier	Linear SVM
	Feature Number	17
	Class	2 (Low/High MW)
	Validation	Leave-one-subject-out

Next, we compare the mental workload detection system with and without the pre- and post- processing blocks. We treated the performance of the ECG-based system as the performance bound of our system. Hence, we also listed the performance of the ECG-based system in Table VI to see how much the proposed technique can get to the performance bound. We have examined the accuracy of the proposed system in CLAS and MAUS datasets. The result is listed in Table V. From the experiment result, we can see that each block makes a performance improvement. The outlier removal mechanism make smaller feature error, hence making MW assessment more accurate. After removing outlier, some uncertainty is generated, proposed post-processing block: uncertainty estimation can be used to evaluate the probability of misclassification. For those misclassify probability higher than 0.458 will be discarded. In this case, no MW assessment result will be reported. The rejection rate indicates the percentage of data that not reporting MW level. It is common to reject PPG if the signal quality is not qualified [27]–[29], [38]. Signal quality assessment methods from [27]–[29], [38] are used to discard poor quality signals for acquiring accurate vital measurement. However, there is still no research investigating how signal quality affects classification-like measurement, such as MW assessment. To the best of our knowledge, we are the first to bridge the gap of this problem. We developed a quantitative method, uncertainty estimation, to quantify the probability of noisy PPG signal affecting classification results.

TABLE VII
COMPARISON RESULT BETWEEN PROPOSED SYSTEM AND RELATED WORK

Articles	Devices	Features	Methods	Task	Classes	Accuracy	Participants
Cinaz et al. (2011) [6]	ECG Chest Belt	ECG (HRV) GSR	LDA, kNN, SVM	Office-work	3	71%	7
Dibyanshu et al. (2019) [11]	Samsung Gear S2	PPG (HRV) Breath Pattern	Decision Tree	Addition Task	2	78%	16
Ekiz et al. (2019) [12]	Samsung Gear S2	PPG (HRV)	LSTM	Daily Life	2	70%	17
Schaule et al. (2018) [13]	Mircrosoft Band 2	PPG (HRV) GSR	SVM, Random Forest	N-back	2	66%	10
Our Work (CLAS)	Shimmer3	PPG (HRV)	Linear SVM	Math problem	2	78%	31
Our Work (MAUS)	PixArt PPG watch	PPG (HRV)	Linear SVM	N-back	2	74%	19

By using the proposed pre- and post- processing techniques, the PPG-based MW assessment system can achieve higher accuracy and lower variance among users. For CLAS dataset, we improve the typical MW system accuracy with 13.6% and lowering the variance with 5.4%. For the MAUS dataset, we improve the typical MW system accuracy with 7.5% and lowering the variance with 0.5%. The reason that we improve the accuracy so much is we rejecting around 30% of data. When we are rejecting probably misclassified data, remain data has larger portion to be correctly classified, and the reported result is more reliable. Lastly, after removing outlier from IBIs and rejecting around 30% high probability misclassify data, proposed PPG-based MW system achieve comparable performance to ECG-based MW system.

Lastly, as illustrated in Table VII, we made a table to compare the proposed system to related work. All of these works use different devices, features, and methods to assess MW under different tasks. It is hard to compare between these works under different settings, methods, and datasets. However, the proposed system that only uses a single modality PPG can achieve relatively high accuracy.

VI. CONCLUSION

This paper presents a robust mental workload detection system based on PPG signal. Two enhancements to the conventional processing flow are proposed, the first part identifies and removes outliers in VVI by the ML-based method, which minimizes the HRV feature errors and improves the accuracy and F1-score by 3%. Next, processed data with a high probability of misclassification are rejected to reduce the false alarm, which further improves the accuracy and F1-scores by 5%. The proposed system are validated on two open datasets with similar conclusion, which achieves comparable performance to ECG-based MW system.

REFERENCES

- [1] F. T. Eggemeier, G. F. Wilson, A. F. Kramer, and D. L. Damos, "Workload assessment in multi-task environments," in *Multiple-Task Performance*. Boca Raton, FL, USA: CRC Press, 1991, pp. 207–216.
- [2] R. J. Lysaght, S. G. Hill, A. Dick, B. D. Plamondon, and P. M. Linton, "Operator workload: Comprehensive review and evaluation of operator workload methodologies," U.S. Army Res. Inst. for the Behav. and Social Sci., Tech. Rep. 851, 1989.
- [3] J. Paxion, E. Galy, and C. Berthelon, "Mental workload and driving," *Front. Psychol.*, vol. 5, pp. 1–11, Dec. 2014.
- [4] M. Zokaei, M. J. Jafari, R. Khosrowabadi, A. Nahvi, S. Khodakarim, and M. Pouyakian, "Tracing the physiological response and behavioral performance of drivers at different levels of mental workload using driving simulators," *J. Saf. Res.*, vol. 72, pp. 213–223, 2020.
- [5] M.-J. Jafari, F. Zaeri, A. H. Jafari, A. T. P. Najafabadi, S. Al-Qaisi, and N. Hassanzadeh-Rangi, "Assessment and monitoring of mental workload in subway train operations using physiological, subjective, and performance measures," *Hum. Factors Ergonom. Manuf. Serv. Industries*, vol. 30, no. 3, pp. 165–175, 2020.
- [6] B. Cinaz, B. Amrich, R. L. Marca, and G. Tröster, "Monitoring of mental workload levels during an everyday life office-work scenario," *Pers. Ubiquitous Comput.*, vol. 17, no. 2, pp. 229–239, 2013.
- [7] L. Fridman, B. Reimer, B. Mehler, and W. T. Freeman, "Cognitive load estimation in the wild," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–9.
- [8] P. A. Hancock and N. Meshkati, *Human Mental Workload*. Amsterdam, The Netherlands: North-Holland, 1988.
- [9] S. Massaro and L. Pecchia, "Heart rate variability (HRV) analysis: A methodology for organizational neuroscience," *Organizational Res. Methods*, vol. 22, no. 1, pp. 354–393, 2019.
- [10] H. Qu, X. Gao, and L. Pang, "Classification of mental workload based on multiple features of ECG signals," *Informat. Med. Unlocked*, vol. 24, 2021, Art. no. 100575.
- [11] D. Jaiswal, A. Chowdhury, D. Chatterjee, and R. Gavas, "Unobtrusive smart-watch based approach for assessing mental workload," in *Proc. IEEE Region 10th Symp.*, 2019, pp. 304–309.
- [12] D. Ekiz, Y. S. Can, and C. Ersoy, "Long short-term network based unobtrusive perceived workload monitoring with consumer grade smart-watches in the wild," *IEEE Trans. Affect. Comput.*, p. 1, 2019, doi: 10.1109/TAFC.2021.3110211.
- [13] F. Schaule, J. O. Johanssen, B. Bruegge, and V. Loftness, "Employing consumer wearables to detect office workers' cognitive load for interruption management," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–20, 2018.
- [14] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, "Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study," *Sensors*, vol. 19, no. 8, p. 1849, 2019.
- [15] J. Piskorski and P. Guzik, "Filtering poincare plots," *Comput. Methods Sci. Technol.*, vol. 11, no. 1, pp. 39–48, 2005.
- [16] M. P. Tarvainen, J.-P. Niskanen, J. A. Lipponen, P. O. Ranta-Aho, and P. A. Karjalainen, "Kubios HRV-heart rate variability analysis software," *Comput. Methods Programs Biomed.*, vol. 113, no. 1, pp. 210–220, 2014.
- [17] N. E. Huang et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London. Ser. A, Math., Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, 1998.
- [18] W.-K. Beh, Y.-H. Wu, and A.-Y. (Andy) Wu, "MAUS: A dataset for mental workload assessment on n-back task using wearable sensor," *IEEE Dataport*, May 10, 2021, doi: 10.21227/q4td-yd35.
- [19] X. Chen, T. Chen, F. Luo, and J. Li, "Comparison of valley-to-valley and peak-to-peak intervals from photoplethysmographic signals to obtain heart rate variability in the sitting position," in *Proc. 6th Int. Conf. Biomed. Eng. Informat.*, 2013, pp. 214–218.

- [20] R. Castaldo, L. Montesinos, T. S. Wan, A. Serban, S. Massaro, and L. Pecchia, "Heart rate variability analysis and performance during a repeated mental workload task," in *Embec & NBC 2017*. Berlin, Germany: Springer, 2017, pp. 69–72.
- [21] A. J. Camm *et al.*, "Heart rate variability: Standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology," 1996.
- [22] C. K. Karmakar, A. H. Khandoker, J. Gubbi, and M. Palaniswami, "Complex correlation measure: A novel descriptor for poincaré plot," *Biomed. Eng. Online*, vol. 8, no. 1, pp. 1–12, 2009.
- [23] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," in *The Theory of Chaotic Attractors*. Berlin, Germany: Springer, 2004, pp. 170–189.
- [24] R. Castaldo, L. Montesinos, P. Melillo, C. James, and L. Pecchia, "Ultra-short term HRV features as surrogates of short term HRV: A case study on mental stress detection in real life," *BMC Med. Informat. Decis. Mak.*, vol. 19, no. 1, pp. 1–13, 2019.
- [25] I. Mijić, M. vSarlija, and D. Petrinović, "MMOD-COG: A database for multimodal cognitive load classification," in *Proc. 11th Int. Symp. Image Signal Process. Anal.*, 2019, pp. 15–20.
- [26] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
- [27] C. Orphanidou, T. Bonnici, P. Charlton, D. Clifton, D. Vallance, and L. Tarassenko, "Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 3, pp. 832–838, May 2015.
- [28] M. Elgendi, "Optimal signal quality index for photoplethysmogram signals," *Bioengineering*, vol. 3, no. 4, p. 21, 2016.
- [29] Q. Li and G. D. Clifford, "Dynamic time warping and machine learning for signal quality assessment of pulsatile signals," *Physiol. Meas.*, vol. 33, no. 9, pp. 1491–1501, 2012.
- [30] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [31] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.
- [32] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau, "The Higgs Boson machine learning challenge," in *Proc. Neural Inf. Process. Syst. Workshop High-Energy Phys. Mach. Learn.*, 2015, pp. 19–55.
- [33] D. Morelli, A. Rossi, M. Cairo, and D. A. Clifton, "Analysis of the impact of interpolation methods of missing RR-intervals caused by motion artifacts on HRV features estimations," *Sensors*, vol. 19, no. 14, p. 3163, 2019.
- [34] H. J. Baek and J. Shin, "Effect of missing inter-beat interval data on heart rate variability analysis using wrist-worn wearables," *J. Med. Syst.*, vol. 41, no. 10, pp. 1–9, 2017.
- [35] J. Bi and T. Zhang, "Support vector classification with input data uncertainty," *Adv. Neural Inf. Process. Syst.*, vol. 17, no. 1, pp. 161–168, 2005.
- [36] V. Markova, T. Ganchev, and K. Kalinkov, "CLAS: A database for cognitive load, affect and stress recognition," in *Proc. Int. Conf. Biomed. Innov. Appl.*, 2019, pp. 1–4.
- [37] W.-K. Beh, Y.-H. Wu, and A.-Y. A. Wu, "MAUS: A dataset for mental workload assessment on N-back task using wearable sensor," 2021. [Online]. Available: <https://dx.doi.org/10.21227/q4td-yd35>
- [38] Y.-C. Yang, W.-K. Beh, Y.-C. Lo, A.-Y. A. Wu, and S.-J. Lu, "ECG-aided PPG signal quality assessment (SQA) system for heart rate estimation," in *Proc. IEEE Workshop Signal Process. Syst.*, 2020, pp. 1–6.