# Real-World Classification of Student Stress and Fatigue Using Wearable PPG Recordings

Justin Laiti*, *Member, IEEE*, Yu Liu†, Pádraic J. Dunne*, Elaine Byrne*, Tingting Zhu†
*Royal College of Surgeons in Ireland, Dublin, Ireland
†Institute of Biomedical Engineering, University of Oxford, Oxford, England

***Abstract*—Wearable-based affective computing offers a promising solution for monitoring and managing stress and fatigue in adolescent student populations. This can help to support student well-being without increasing reliance on mobile phones by providing on-device insights into stress and fatigue levels. To do so, the processing and classification methods must be lightweight enough to be executed in real-time on wearables. This study outlines the findings of the implementation of a student-informed wearable and mobile app, Wellby. Students tested this for one month while completing routine photoplethysmography (PPG) recordings and check-ins on their perceived levels of stress and fatigue. This study proposes a lightweight processing pipeline, intended for wearable-based deployment, while examining its classification performance on real-world student PPG data from Wellby. The pipeline performs denoising, fixed noise elimination, and peak detection to calculate time-domain heart rate variability (HRV) metrics. It was first evaluated on public datasets, the Wearable Stress and Affect Detection (WESAD) dataset and the AKTIVES dataset, achieving an area under the receiver operating characteristic curve (AUC-ROC) of up to 91.58% for stress classification on WESAD and 76.61% on AKTIVES. In the Wellby dataset, the adapted processing pipeline achieved an AUC-ROC of 77.02% for stress classification and 71.58% for fatigue classification using only time-domain HRV features. Furthermore, the inclusion of a signal quality metric and baseline well-being questionnaires improved the AUC-ROC for stress classification to 91.60% in the best performing model. These findings demonstrate the potential for wearables to implement real-time affective computing, providing timely feedback to students in real-world settings based on PPG and contextual data. The code used in this study is available on GitHub [https://github.com/j-laiti/PPG-affect-classification].**

***Index Terms*—Wearables, student stress and fatigue detection, PPG, edge computing, machine learning**

## I. Introduction

THE increased adoption of wearable devices, such as wristbands, glasses, rings and headbands, has expanded the availability of sensors for continuous monitoring of physiological data [1]. These devices capture health-related data through measurement techniques such as photoplethysmography (PPG), accelerometry (ACC), and electroencephalography (EEG) for the collection of heart activity, movement, and brain activity, respectively [2]. Physiological signals from wearables can provide insights into psychological processes, making these devices valuable tools for monitoring affective states such as stress and fatigue in daily life [3]–[6].

The detection of acute psychological stress using data from wearable devices has been gaining attention within affective computing research [7]. Psychological stress manifests itself through physical reactions in the body that can be detected from brain activity, eye movements, respiration rate, skin conductivity, and heart activity [5]. Among these, skin conductivity and heart activity are the most commonly used for wearable-based stress detection [8]. For example, heart rate variability (HRV) is a widely used signal for stress detection [8], emotion recognition [3], [9], and fatigue detection [10]. PPG is a technique that uses non-invasive optical sensors to detect changes in blood flow through light absorption which is commonly used in wearables for HRV-based monitoring due to ease of integration and suitability for continuous monitoring [11].

With the increasing use of wearables for affect detection, educational settings have emerged as a key research area in affective computing, as both students and teachers face increasing stress levels [5], [12]. Biofeedback tools, which provide real-time insight based on physiological data, show promising results for stress management in both educators and students [13], [14]. Academic stress is commonly experienced by adolescents cross-culturally and is associated with poor sleep quality [15], nutrition disorders [16], and low academic performance [13]. Importantly, the physiological stress response undergoes a period of development in adolescence that makes it unique to that of adults, with heightened and more prolonged stress-induced hormonal responses that can increase susceptibility to stressors [17]. The detection of the unique stress and fatigue responses in educational settings presents an opportunity to address student-specific well-being needs that may have long-term health implications [18]. Although adolescents are increasingly seeking mobile apps and wearables to support their well-being needs, few student-focused tools support engagement and change in health behavior [19], [20]. PPG-based affective computing tools offer promising support for students but remain underexplored in educational settings. To understand PPG-based stress and fatigue detection for students, more research is needed in real-world environments, as most existing studies have been conducted in controlled laboratory settings [5], [21], [22].

Machine Learning (ML) techniques are widely used for affect state classification due to the efficient handling of complex physiological patterns compared to traditional threshold-based methods [23]. Therefore, wearable devices for affect detection in educational settings must integrate ML models,

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.
Manuscript received March 28, 2025; revised 2025.

complicating the implementation of lightweight stress and fatigue detection. During the school day, students often have limited access to their smartphones, so wearable devices must function on their own. However, the small form factor of wearables restricts their storage and processing capabilities [24]. Real-time stress and fatigue classification requires several steps, including continuous data collection, processing, and communication with peripheral devices. Thus, wearable-based processing methods must be optimized for edge computing, or perform these steps locally on the wearable before sending the output to a cloud server, thereby balancing accuracy with power and computational constraints [25].

Despite the promise of PPG-based stress detection, signal processing in real-world conditions remains a challenge due to motion artifacts and inconsistent signal quality [26], [27]. Many consumer wearables generate stress, readiness, or drowsiness scores. However, the algorithms used to derive these metrics from everyday wearable data often lack transparency [28] and consistency between devices [29]. This presents a need for more robust signal processing and validation methods to encourage the implementation of accurate and context-specific insights for the people using these devices.

These challenges can be summarized as follows:

- The limited number of real-world studies that have explored PPG-based wearable detection of fatigue or stress in educational settings hinders evidence-based solutions.
- Wearables must execute complex pipelines despite hardware limitations in power and computation.
- Existing stress and fatigue scores lack transparency in how they are generated.

To address these gaps, this study presents a novel, lightweight processing pipeline for PPG signals collected from students during the final month of the school year. The processing pipeline is adapted to be computationally efficient and suitable for PPG-based affect detection on wearable devices by implementing simplified filtering, noise elimination, peak detection, and affect classification using time-domain HRV features. Signal quality and contextual information, including baseline well-being questionnaires, are also included in the feature set for this pipeline.

For PPG data collection, we use a custom-built, wrist-worn PPG sensor which is a low-cost purpose-specific device designed to support edge computing. This enables local signal processing, reduced latency, and enhanced security and privacy to provide on-device insights [30], [31]. We evaluate our proposed pipeline for stress and fatigue detection based on both our student dataset and open-source datasets for stress detection [32], [33] to assess generalizability.

This paper contributes the following:

- A real-world implementation and evaluation of PPG-based wearable stress and fatigue detection optimized for wearable use in educational settings.
- An adapted lightweight processing pipeline for HRV extraction and ML stress and fatigue classification compared to baseline methods.
- An explainability analysis of HRV features and contextual factors that influence the output of the classification models.

To our knowledge, no prior work has provided an open-source PPG-processing pipeline validated with an adolescent cohort and optimized for deployment on resource-constrained wearable devices. Our approach addresses this gap by focusing on data from adolescents in real-life scenarios and presenting a lightweight pipeline for affect classification based on time-domain HRV, baseline well-being, and signal quality metrics.

The remainder of this paper is organized as follows. Section II outlines related research on affect detection using wearables in educational settings and for PPG-based stress and fatigue detection. Section III details the methods used for data collection, PPG signal processing, and performance evaluations on both public and Wellby datasets. Section IV presents the results of the PPG processing, the correlation of HRV features with affect labels, the performance of ML models across each dataset, and the impact of the PPG features on the model performance. Section V discusses and interprets these findings, while Section VI concludes the paper with key takeaways, limitations, and recommendations for future work.

## II. Related Work

This research investigates a lightweight PPG-based stress and fatigue classification pipeline applied to public, lab-based datasets and data collected in this study from students in their daily life. A recent systematic review of wearable technologies in education highlighted that most studies focus on university students, mainly use EEG devices, and are conducted in controlled laboratory settings [34]. Among the few studies that incorporate HRV-based measurements, one explored stress levels during exams, comparing across student levels (e.g., undergraduate vs. graduate) or exam types [35]. Another study implemented PPG-based HRV biofeedback sessions using slow breathing with live HRV metrics feedback to reduce social stress and anxiety in primary school students [36]. Other techniques, such as eye tracking [37] and facial recognition [38], have been used to assess student attention and fatigue levels during class times. A study in technical education settings used PPG-based stress detection during lessons and reported over 86% accuracy [39]. Despite the growth of research in this area, there remains a gap in the application of PPG-based wearables to monitor stress and fatigue in students' daily lives, particularly outside of structured classroom activities and exams.

In addition to a lack of real-world studies, a review of PPG-based wearable stress detection emphasized the need for greater attention to data processing methodologies [21]. While edge-based signal processing was highlighted for its speed and direct access to raw data, most pipelines still rely on computationally intensive methods suitable only for offline or cloud-based analysis. Neural network architectures, including artificial neural networks (ANNs), convolutional neural networks (CNNs), Long Short-Term Memory networks (LSTMs), and transformer architectures are widely applied for feature extraction and affect detection of PPG signals. One study demonstrated higher performance using a simple ANN compared to traditional machine learning methods for binary stress classification based on the WESAD dataset, with accuracies of 95.21% and 93.20%, respectively [40]. Despite promising research in neural networks, other traditional methods of feature processing without deep learning have resulted in an accuracy of 96.50% on the same WESAD dataset using PPG alone [41]. Prior research implemented ensemble base peak detection, adaptive noise filtering, and classification using machine learning with the best performance from Linear Discriminant Analysis [41]. Furthermore, other research has investigated hybrid feature approaches, including one study which combined features generated from a CNN with traditional ECG and EDA feature extraction from WESAD, resulting in an accuracy of 90.05% [42]. Alternative approaches using spectrograms have also been explored for PPG feature extraction [43], [44]. While there are various approaches for classifying affective state based on PPG data, the processing pipelines used in these studies are computationally intensive and do not address on-device feasibility. Additionally, while the WESAD dataset is widely used to study affect detection [45], it is limited to an adult cohort and a controlled setting [32]. Therefore, there remains a need to evaluate lightweight pipelines in comparison to other mainstream methods to address the feasibility of implementation for PPG-based stress and fatigue edge detection in various contexts.

## III. Data Acquisition & Methods

### A. Participants

Senior-level students (n = 43) from three secondary schools in Ireland participated in this project (Table I). Schools A and B are mainstream secondary schools (co-ed and all-girls, respectively, catering for students aged 12-18 years), while school C is a Youthreach Centre which is an alternative school that supports early-school leavers from mainstream schools. Students in Schools A and B were eligible if they were in their fifth year (age range 16-18 years), while all students in School C (Youthreach) were eligible (age range 16-19 years). The project consisted of two stages. In stage one, groups of students from each of these schools participated in a needs assessment and co-design process to tailor a wearable device and mobile app, which they named ‘Wellby’, to their well-being needs [46]. Stage two is outlined in this paper, where students tested Wellby for four weeks. Table I outlines the demographics of the participants and indicates the number of students who participated in the co-design previously. The legal guardians of the students completed consent forms and students completed assent forms to participate in this study. Students were informed that their participation was voluntary, that they could leave at any time during the research, and that their participation or non-participation would not affect any other aspect of their schoolwork or performance.

### B. Study Protocol

In this study, the students had access to the Wellby mobile and mobile app for four weeks in May 2024 (Fig. 1). PPG recordings collected throughout the month were linked to in-app mood, stress, and fatigue check-ins. Student feedback on the usability of Wellby, including app engagement metrics and post-study interviews, will be reported in a forthcoming paper. At the beginning of the month, students completed three questionnaires which have been validated for adolescents, including the Perceived Stress Scale (PSS) [47], Pittsburgh Sleep Quality Index (PSQI) [48], and the Engagement, Perseverance, Optimism, Connectedness, and Happiness (EPOCH) Measure of Adolescent Well-Being [49].

The wearable device, assembled by RCSI researchers, collected PPG recordings from the wrist. This device, enclosed in a 3D printed nylon casing, included the XIAO NRF52840 microcontroller and the MAX30101 PPG sensor which was chosen for its form factor and the ability to use red, infrared, and green LEDs for PPG. The MAX30101 has undergone previous validation for wrist-based HRV measurements [50], [51]. The recordings were initiated in the mobile app after the Bluetooth connection was secured to the wearable. Students received reminders three times a week to complete check-ins on the app by starting a one-minute PPG recording while sitting.

After the recording, the students completed a brief check-in, including scales for level of relaxation and alertness to indicate stress and fatigue level and other contextual information for each recording. The raw PPG recordings, extracted HRV features, and check-in data were uploaded to the cloud and stored in a Google Firestore account accessed by the researchers. Students were able to see the computed metrics after the recording, including heart rate, signal quality index (SQI), and two time-domain HRV metrics, including root mean squared of standard differences (RMSSD) and standard deviation of n-to-n peaks (SDNN). The signal quality was calculated based on the standard deviation and mean of both the overall signal and peak-to-peak intervals, which were expressed to students as either low, good or excellent. A summary of previous recording metrics and educational information was available to students in the app. The data collection process is further outlined in Fig. 1.

### C. PPG Processing Pipeline

We developed a lightweight PPG processing pipeline designed to accommodate the computational requirements of the NRF52840 microcontroller used in the Wellby wearable device building on previous work [41], [52]–[54]. Common PPG processing steps in these works include running a bandpass filter, moving average, noise elimination, peak detection, and HRV feature extraction. These features serve as inputs for training and evaluating ML classification models. An open-source pipeline outlined by Heo et al. [41] uses the Wearable Stress and Affect Detection (WESAD) dataset [32] and incorporates noise filtering, adaptive noise elimination, ensemble peak detection, and HRV feature extraction from time, frequency and non-linear domains. In contrast, our proposed pipeline uses a simplified noise elimination method, a single threshold-based peak detection algorithm, and focuses exclusively on time-domain HRV features (Fig. 2). Table II outlines the time-domain features extracted from the PPG data.

This adapted pipeline was validated on the open-source WESAD [32] and AKTIVES datasets [33], and the Wellby data collected in this study. The raw PPG data in the WESAD

TABLE I
DEMOGRAPHICS OF PARTICIPANTS BY SCHOOL

| Category | School A (n=16) | School B (n=11) | School C (n=16) | Total (n=43) |
|---|---|---|---|---|
| **Age** | | | | |
| 16 | 5 | 2 | 5 | 12 |
| 17 | 11 | 7 | 7 | 25 |
| 18 | 0 | 2 | 1 | 3 |
| 19 | 0 | 0 | 3 | 3 |
| **Ethnicity** | | | | |
| White Irish | 12 | 5 | 13 | 30 |
| White - Other European | 4 | 2 | 1 | 7 |
| White African | 0 | 1 | 0 | 1 |
| Black or Black Irish | 0 | 1 | 1 | 2 |
| Asian | 0 | 2 | 1 | 3 |
| **Gender** | | | | |
| Female | 8 | 10 | 9 | 27 |
| Male | 8 | 0 | 5 | 13 |
| Non-binary | 0 | 0 | 1 | 1 |
| Prefer not to say | 0 | 1 | 1 | 2 |
| **Previous Research Engagement** | | | | |
| Prior co-design participants | 15 | 10 | 15 | 40 |
| New participants | 1 | 1 | 1 | 3 |

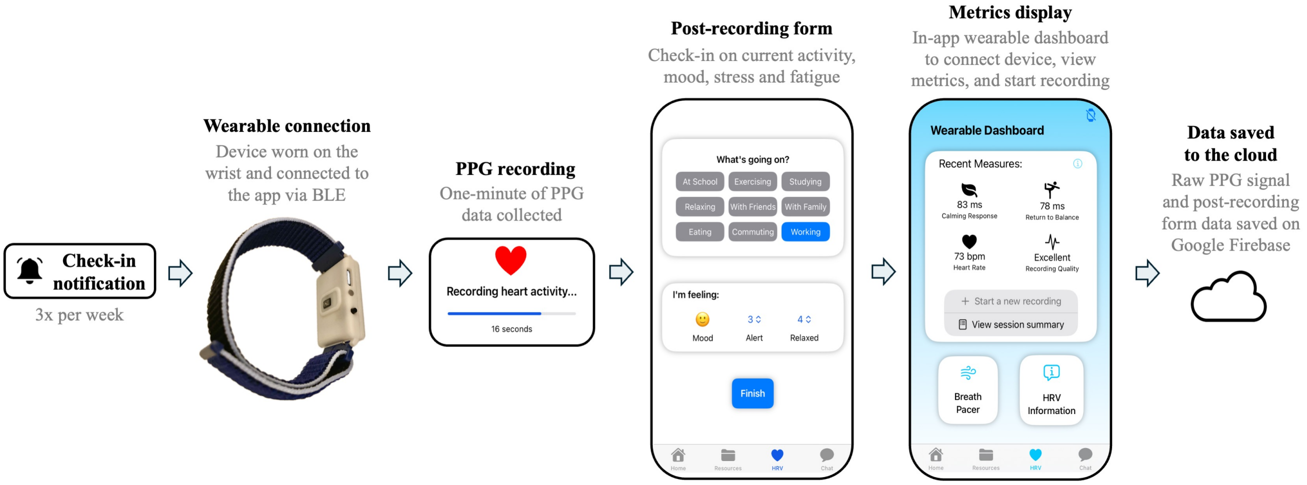

Fig. 1. Data collection process overview: (a) Students received check-in reminders three times per week. (b) To begin a session, students wore the wearable device and paired it with the mobile app. (c) During each session, students completed a one-minute PPG recording via Bluetooth Low Energy (BLE). (d) Following the recording, students completed post-recording self-reports rating their alertness and relaxation levels from 1-5 to assess stress and fatigue. (e) The app displayed metrics from the latest recording (RMSSD, SDNN, HR, and SQI). (f) The raw PPG signals, extracted features, and self-report data were then uploaded to the cloud for storage and analysis

dataset were segmented into 2-minute components to enable comparison with similar work [41], [55], while the AKTIVES dataset was separated by 30-second intervals based on the short duration of the labels [33]. 1-minute recordings were retained for the Wellby dataset. These windows have been shown to be sufficient for HRV time-domain calculations [56] and align with the computational limitations of wearable devices. Furthermore, windows of 30-60 seconds have demonstrated high performance for stress detection, which may be attributed to a correlation with the onset time of the physiological stress response [57].

The signals were standardized, and a 2nd order Butterworth bandpass filter of 0.5-10 Hz was applied based on previous research in PPG signal processing [58] and PPG-based stress detection [41]. The lower cutoff can help to reduce signal drift, respiratory artifacts, and low-frequency motion artifacts while the upper cutoff can eliminate high-frequency noise [59]. Other studies recommend alternative ranges, such as 0.2-10 Hz [60] and 0.5-8 Hz [59], which were tested for the WESAD dataset and show a less than 1% and 5% average change, respectively, compared to 0.5-10 Hz in the time-domain HRV metrics used in this study.

Following the bandpass filter, a moving average filter with a window of 5 data points was applied to smooth the signal.

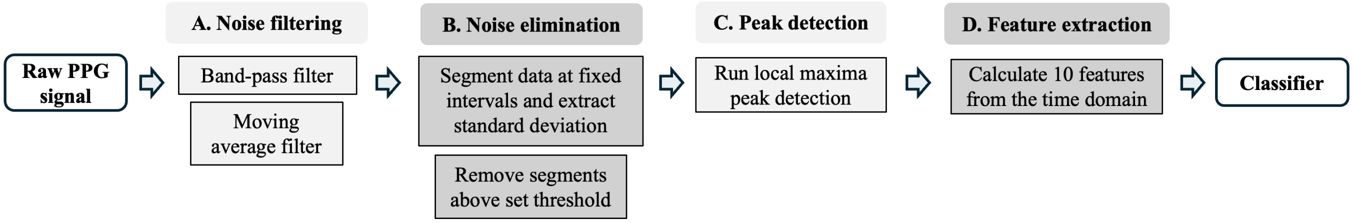
Raw PPG signal
A. Noise filtering
Band-pass filter
Moving average filter
B. Noise elimination
Segment data at fixed intervals and extract standard deviation
Remove segments above set threshold
C. Peak detection
Run local maxima peak detection
D. Feature extraction
Calculate 10 features from the time domain
Classifier

Fig. 2. PPG processing pipeline presented including noise filtering, noise elimination, peak detection and feature extraction.

The signal was then segmented into 3-second components, and the standard deviation was calculated for each segment. To eliminate high-noise components, the highest 15% of segments were removed [61] and the same moving average filter was reapplied to smooth the final signal. Peak detection was implemented using a threshold-based algorithm that identifies local maxima above the global average of the signal [62].

TABLE II
TIME-DOMAIN HRV METRICS AND THEIR CALCULATION FORMULAS

| Metric | Calculation |
|---|---|
| Mean HR | Mean HR $= \frac{60}{NN}$ |
| Std HR | Std HR = Standard deviation of HR values |
| RMSSD | RMSSD $= \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(NN_{i+1} - NN_i)^2}$ |
| SDNN | SDNN = Standard deviation of all $NN$ |
| SDSD | SDSD = Standard deviation of $NN_{i+1} - NN_i$ |
| Mean NN | $\overline{NN} = \frac{1}{N}\sum_{i=1}^{N} NN_i$ |
| Mean SD | Mean SD = Mean of $NN_{i+1} - NN_i$ |
| Median NN | Median NN = Median of $NN_i$ |
| pNN20 | pNN20 $= \frac{\text{Number of } \lvert NN_{i+1}-NN_i \rvert > 20ms}{N-1} \times 100\%$ |
| pNN50 | pNN50 $= \frac{\text{Number of } \lvert NN_{i+1}-NN_i \rvert > 50ms}{N-1} \times 100\%$ |

### D. PPG Datasets

The Wellby dataset collected in this study includes data from 43 students who were encouraged to record PPG data three times a week for four weeks using the Wellby wearable and mobile app. In total, 100 one-minute PPG recordings (50 Hz) were uploaded via the mobile app. However, due to motion artifacts, initial hardware issues, and several missing self-reported stress/fatigue ratings, only 38 recordings were retained for analysis. These recordings represent data from 19 different students and were examined to confirm that distinct PPG waveforms were present in each.

The public WESAD dataset consists of 15 participants who wore the wrist-based Empatica E4 device while undergoing controlled baseline, amusement, and stress-inducing conditions. The dataset provides PPG, skin conductance, temperature, and accelerometer data. The combined activities resulted in approximately 36 minutes of PPG data per WESAD participant [32]. This paper focuses only on the PPG signal from baseline and stress states, aligning with a prior evaluation of the WESAD dataset [41].

The public AKTIVES dataset includes 25 child participants who wore the wrist-based Empatica E4 device while completing therapeutic games in a controlled setting. Participants completed two computer-based games lasting approximately 7 minutes each [33]. The dataset provides PPG, skin conductance, temperature, and accelerometer data, although only the PPG signal was used in this study. Participant data were included in the analysis in this paper if PPG data was available with at least one stress label during the gaming activities.

Although both of the public datasets outlined have been used for stress detection, they have several limitations compared to the Wellby dataset collected in this study. The WESAD and AKTIVES data were collected in a controlled laboratory environment during predefined stress-inducing activities and neither include fatigue labels. In addition, WESAD only includes adult participants. In contrast, the Wellby dataset offers real-world data from secondary school students. The characteristics of each dataset are summarized in Table III.

TABLE III
DATASETS USED FOR PPG-BASED AFFECT CLASSIFICATION

| Property | Wellby | WESAD | AKTIVES |
|---|---|---|---|
| Participants | 19 | 15 | 15 |
| Age (mean ± SD) | 16.9 ± 0.8 | 27.5 ± 2.4 | 10.2 ± 1.3 |
| Sampling rate | 50 Hz | 64 Hz | 64 Hz |
| Total recordings used | 38 (Stress: 15/38, Fatigue 24/38) | 30 (15 stress, 15 baseline) | 16 |
| Duration per recording | 1 minute | Stress: $\sim$10 min, Baseline: $\sim$20 min | $\sim$7 minutes |
| Context | Everyday life | Controlled lab setting | Controlled lab setting |
| Label used as ground truth | Self-reported stress/fatigue | Activity-based stress/baseline | Expert-labeled stress |

### E. ML Classification of Stress and Fatigue

The features extracted from the PPG recordings were used to train various ML models to determine stress and fatigue using binary classification algorithms. The subjective self-assessment of stress and fatigue completed by the students after each recording was used as the ground truth for the Wellby dataset. Consistent with previous work [55], the 1–5 scale for self-reported calmness and alertness was binarized, with scores of 1–3 classified as ”stress/fatigue” and scores of 4–5 as ’no stress/fatigue’ to achieve binary classification.

A range of traditional ML algorithms were compared, including Random Forest (RF) [63], AdaBoost (AB) [64], K-Nearest Neighbors (kNN) [65], Linear Discriminant Analysis

(LDA) [66], Support Vector Machine (SVM) [67], and Gradient Boosting (GB) [68].

These models were selected based on their previous application in affect detection studies [5], [41], [69] and relatively low computational cost compared to deep learning models. Additionally, a systematic review found that SVM, RF, and kNN were most often used for stress detection, with consistent high performance in tree-based models and the highest performance reported using LDA [57]. Similarly, SVM, kNN and tree-based models are frequently used for fatigue detection from wearable data [4].

### *F. Model Training and Evaluation*

Stratified group k-fold validation and leave one group out (LOGO) evaluation methods were used to ensure the balance of classes and separate participants in the data splits. To address class imbalance, models were trained using weighted loss functions that assigned higher penalties to the minority class (stress/fatigue). The training and evaluation architectures for public datasets followed previous studies to enable direct comparisons [33], [41]. For the Wellby dataset, hyperparameter tuning was performed using grid search to maximize the average precision. Model performance was assessed using average precision, balanced accuracy, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) to assess model robustness.

For each model evaluation, we included baseline comparisons from deep learning approaches including one CNN-based model [70] and one hybrid model [42] incorporating both time-domain features and CNN features. Efficiency metrics were compared across the methods and the datasets in this study. The model implementations and efficiency measurements can be found in GitHub [https://github.com/j-laiti/PPG-affect-classification].

In addition, models were trained on different feature sets included in the Wellby data to compare the impact of the combination of contextual information with the HRV features extracted from the processing pipeline. Feature importance was quantified using Shapley values for the SVM model to interpret their impact on the classification outcomes.

## IV. Results

This section presents the evaluation of stress and fatigue detection using a lightweight processing pipeline for PPG signal classification across three datasets. We first examine the distribution of stress and fatigue labels by student demographics in the Wellby dataset, then demonstrate our processing pipeline on sample PPG signals from WESAD, AKTIVES, and Wellby datasets. We analyze correlations between HRV features and affective labels, evaluate ML classification performance, assess demographic bias and computational efficiency, and conclude with feature importance analysis using Shapley values. The complete code is available on GitHub [https://github.com/j-laiti/PPG-affect-classification].

### *A. Stress and Fatigue Label Distribution*

To contextualize the ML classification evaluation of the Wellby dataset, the distribution of stress and fatigue labels was first examined across the demographics of the students, including school, gender, age, and ethnicity (Table IV). Only the data associated with the final selected high-quality PPG signals is summarized with the number of recordings listed next to each category. In particular, the rate of recordings with the stress label was highest in School 2 (the all-girls secondary school), while the rate of fatigue-labeled recordings was highest at School 3 (the co-ed Youthreach Centre). Similarly, the percent of stress-labeled recordings was highest for females, while the percent of fatigue-labeled recordings was highest for males. Although these patterns suggest some variation of stress and fatigue across demographic groups, no statistical tests were performed due to the limited sample sizes in several subgroups. Therefore, these distributions are presented descriptively to support the interpretation of the subsequent classification results.

TABLE IV
Distribution of Stress and Fatigue Labels Across Student Demographics in the Wellby Dataset

| Category | Stress (%) | Fatigue (%) |
|---|---|---|
| **School** | | |
| School 1 (n=12) | 33.30 | 66.70 |
| School 2 (n=8) | 62.50 | 37.50 |
| School 3 (n=18) | 33.30 | 83.30 |
| **Gender** | | |
| Female (n=16) | 62.50 | 62.50 |
| Male (n=18) | 22.20 | 72.20 |
| Prefer not to say (n=4) | 25.00 | 75.00 |
| **Age** | | |
| 16 (n=9) | 22.20 | 66.70 |
| 17 (n=20) | 35.00 | 80.00 |
| 18 (n=5) | 80.00 | 20.00 |
| 19 (n=4) | 50.00 | 75.00 |
| **Ethnicity** | | |
| Asian (n=1) | 0.00 | 100.00 |
| Black or Black Irish (n=5) | 80.00 | 20.00 |
| White - Other European (n=2) | 0.00 | 100.00 |
| White African (n=3) | 33.30 | 66.70 |
| White Irish (n=27) | 37.00 | 74.10 |

### *B. PPG Signal Processing*

The processing pipeline presented in this paper was applied to PPG data from each dataset. The pipeline included (1) noise reduction to account for signal drift and apply smoothing, (2) noise elimination to remove noise artifacts, and (3) peak detection to identify pulse peaks used in HRV feature extraction. Fig. 3 illustrates the application of this pipeline on raw PPG signals from the included datasets, demonstrating reliable peak detection and enabling robust HRV feature calculation.

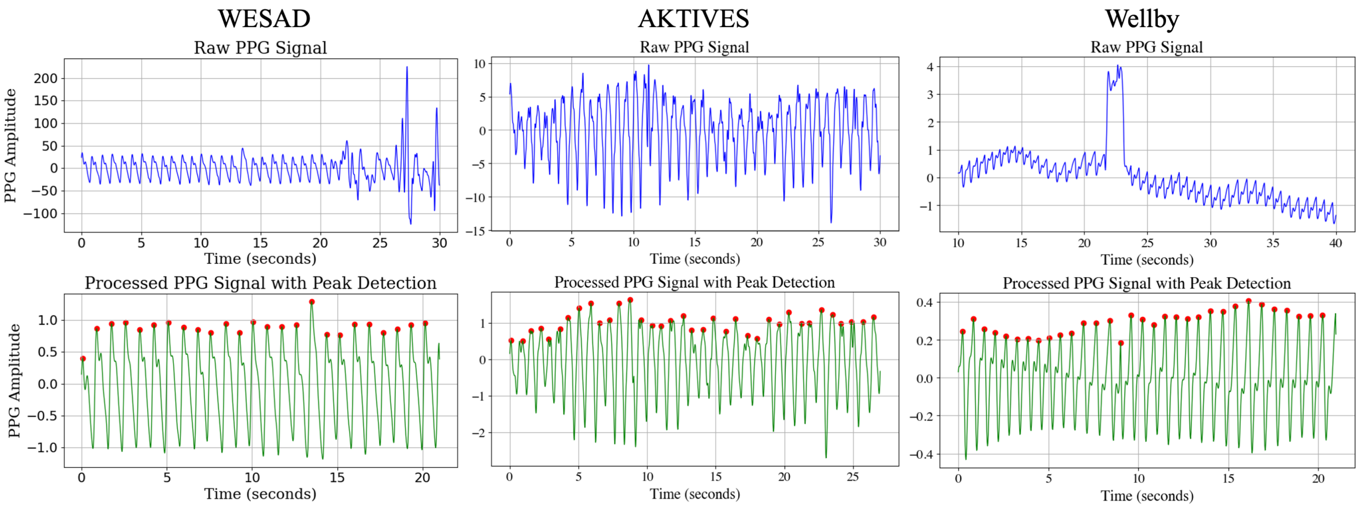

Fig. 3. PPG processing pipeline applied to 30 second segments of data for each dataset.

## C. Correlation Between HRV Features and Labels

To explore the relationship between individual HRV features and self-reported stress and fatigue labels, point-biserial correlation coefficients were calculated for the features in each dataset. As shown in Fig. 4, most HRV features exhibited weak correlations with fatigue labels in the Wellby dataset and stress labels in the AKTIVES dataset. Moderate correlations were observed between select HRV features and stress labels in the Wellby and WESAD dataset, with generally lower correlations in the Wellby results. Notably, most of the features in the Wellby dataset showed negative correlations with stress labels, while the features in the WESAD dataset showed a general positive correlation with stress labels. These HRV features are typically expected to decrease with increased stress [56]. The correlation analysis suggests that individual HRV features alone are insufficient for robust stress or fatigue detection, reinforcing the need for multivariate ML approaches to capture more complex, non-linear physiological patterns.

## D. Classification Performance on the Public Datasets

The extracted features from the WESAD and AKTIVES datasets were used to train six classification algorithms (RF, AB, kNN, LDA, SVM, GB) for stress detection based on PPG signals compared to baseline methods. We used Leave-One-Group-Out (LOGO) cross-validation for WESAD and 10 repeated random 70% training and 30% testing splits for AKTIVES to enable direct performance comparison with previous research [32], [71].

On the WESAD dataset, our proposed pipeline achieved it's highest classification with SVM (AUC-ROC: 91.58%, Accuracy: 93.44%, F1-Score: 86.99%) (Table V). The hybrid baseline showed competitive performance with GB achieving the highest overall (AUC-ROC: 91.63%, Accuracy: 93.61%, F1-Score: 93.61%), while the CNN baseline achieved slightly lower performance metrics than the proposed pipeline and hybrid baseline.

Based on the AKTIVES dataset, the proposed pipeline with GB achieved the highest performance across all evaluation metrics (AUC-ROC: 76.61%, ACC: 80.67%, F1-Score: 78.93%) compared to the hybrid baseline with GB (AUC-ROC: 71.34%, ACC: 78.67%, F1-Score: 77.72%) (Table VI). The CNN baseline showed substantially degraded performance (AUC-ROC: 44.26%, ACC: 48.73%, F1-Score: 46.14%), indicating challenges of applying deep learning architectures to small datasets with high inter-subject variability.

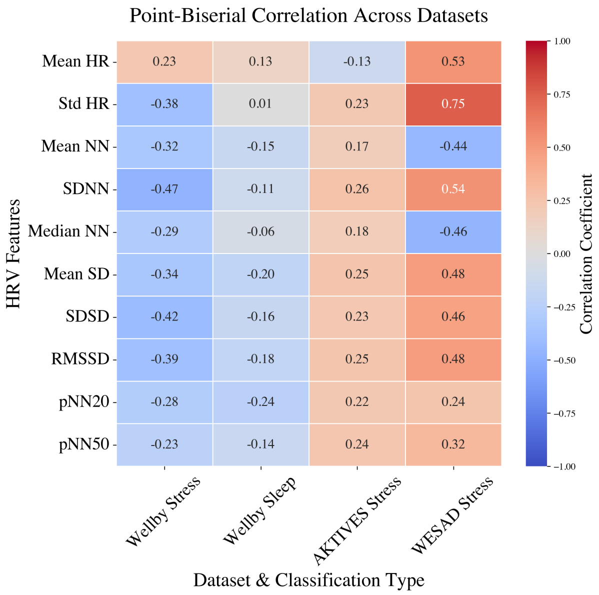

Fig. 4. Point-biserial correlation coefficients between HRV features and stress/fatigue binary labels in the Wellby, AKTIVES, and WESAD datasets.

## E. Classification Performance on the Wellby Dataset

Following high classification performance on the public datasets, the same signal processing pipeline and classification methods were applied to the raw PPG recordings in the Wellby dataset. Since this dataset is imbalanced, with a low number of positive class samples for stress detection and negative class samples for fatigue detection, average precision (AP) was

TABLE V
Performance of stress classification algorithms on the WESAD dataset. Results are presented as $\mu \pm \sigma$ (95% CI).

| Method | Algorithm | AUC-ROC | F1 Score | Accuracy |
|---|---|---|---|---|
| Proposed Pipeline | RF | 89.26 ± 14.49 (81.23, 97.29) | 83.84 ± 21.48 (71.94, 95.74) | 91.37 ± 10.69 (85.45, 97.29) |
| | AB | 86.34 ± 15.03 (78.02, 94.67) | 80.42 ± 20.76 (68.93, 91.92) | 88.12 ± 14.54 (80.07, 96.17) |
| | kNN | 88.21 ± 16.04 (79.32, 97.09) | 81.81 ± 24.58 (68.20, 95.42) | 91.38 ± 10.43 (85.61, 97.16) |
| | LDA | 89.99 ± 11.14 (83.82, 96.16) | 85.28 ± 16.03 (76.40, 94.16) | 92.53 ± 6.96 (88.68, 96.39) |
| | SVM | **91.58 ± 12.23 (84.80, 98.35)** | **86.99 ± 17.85 (77.10, 96.87)** | **93.44 ± 7.82 (89.12, 97.77)** |
| | GB | 88.74 ± 12.78 (81.66, 95.82) | 82.92 ± 18.67 (72.58, 93.26) | 90.50 ± 9.49 (85.25, 95.75) |
| Hybrid Baseline (CNN + TD) | RF | 90.06 ± 12.32 (83.24, 96.89) | 86.20 ± 17.81 (76.34, 96.07) | 92.89 ± 9.00 (87.91, 97.87) |
| | AB | 90.64 ± 11.63 (84.20, 97.09) | 86.30 ± 16.79 (77.00, 95.60) | 92.41 ± 9.04 (87.40, 97.41) |
| | kNN | 81.53 ± 13.27 (74.18, 88.88) | 73.93 ± 20.37 (62.65, 85.21) | 88.14 ± 8.64 (83.35, 92.93) |
| | LDA | 86.51 ± 15.57 (77.88, 95.13) | 78.78 ± 26.78 (63.95, 93.61) | 89.79 ± 10.47 (83.99, 95.58) |
| | SVM | 87.58 ± 10.43 (81.81, 93.36) | 82.95 ± 15.63 (74.29, 91.60) | 91.89 ± 6.32 (88.39, 95.39) |
| | GB | **91.63 ± 11.38 (85.33, 97.94)** | **88.28 ± 15.90 (79.48, 97.09)** | **93.61 ± 8.86 (88.71, 98.52)** |
| CNN Baseline | CNN | **87.46 ± 7.99 (79.47, 95.45)** | **81.56 ± 11.91 (69.65, 93.47)** | **91.12 ± 5.22 (85.90, 96.34)** |

TABLE VI
Performance of stress classification algorithms on the AKTIVES dataset. Results are presented as $\mu \pm \sigma$ (95% CI).

| Method | Algorithm | AUC-ROC | Accuracy | F1 Score |
|---|---|---|---|---|
| Proposed Pipeline | RF | **76.61 ± 4.83 (73.16, 80.07)** | **80.67 ± 2.38 (78.96, 82.37)** | **78.93 ± 2.09 (77.43, 80.42)** |
| | AB | 69.04 ± 3.78 (66.34, 71.75) | 77.83 ± 3.69 (75.19, 80.47) | 75.74 ± 3.16 (73.48, 78.00) |
| | kNN | 74.61 ± 4.38 (71.47, 77.74) | 79.83 ± 4.04 (76.94, 82.72) | 78.66 ± 4.59 (75.38, 81.95) |
| | LDA | 69.32 ± 3.50 (66.81, 71.83) | 78.67 ± 2.46 (76.91, 80.43) | 75.29 ± 3.01 (73.14, 77.44) |
| | SVM | 70.67 ± 5.94 (66.42, 74.91) | 72.33 ± 4.39 (69.19, 75.47) | 71.61 ± 2.97 (69.48, 73.74) |
| | GB | 73.86 ± 6.15 (69.46, 78.26) | 79.17 ± 4.10 (76.23, 82.10) | 77.41 ± 3.65 (74.79, 80.02) |
| Hybrid Baseline (CNN + TD) | RF | 62.79 ± 7.39 (57.50, 68.07) | 73.50 ± 5.47 (69.59, 77.41) | 70.96 ± 5.31 (67.17, 74.76) |
| | AB | 58.81 ± 6.21 (54.37, 63.26) | 73.33 ± 3.33 (70.95, 75.72) | 68.56 ± 2.98 (66.43, 70.69) |
| | kNN | **71.34 ± 6.47 (66.72, 75.97)** | **78.67 ± 4.96 (75.12, 82.21)** | **77.72 ± 4.65 (74.39, 81.05)** |
| | LDA | 61.59 ± 5.23 (57.84, 65.33) | 71.83 ± 4.93 (68.30, 75.36) | 71.44 ± 4.15 (68.47, 74.41) |
| | SVM | 59.30 ± 14.13 (49.20, 69.41) | 74.50 ± 4.65 (71.17, 77.83) | 72.59 ± 4.66 (69.26, 75.92) |
| | GB | 63.60 ± 6.78 (58.75, 68.45) | 75.67 ± 3.35 (73.27, 78.06) | 73.10 ± 3.51 (70.59, 75.61) |
| CNN Baseline | CNN | **44.26 ± 13.14 (34.86, 53.66)** | **48.73 ± 22.19 (32.86, 64.60)** | **46.14 ± 24.71 (28.46, 63.82)** |

used as the primary performance measure. AP summarizes the precision-recall curve and is commonly used in imbalanced classification tasks [72], allowing us to evaluate the model's ability to correctly identify positive stress and fatigue cases. AUC-ROC was also included to allow for direct comparison with the public dataset results. Grid search was used for hyperparameter tuning, as it provides an exhaustive method of evaluating parameters that is reproducible and feasible on our smaller datasets, optimizing models based on AP. The mean and standard deviation of the Wellby performance metrics are reported for consistency with baseline dataset comparisons and previous research [41].

The classification results were compared between two different feature sets: (1) time-domain (TD) features only and (2) TD features, SQI, and baseline questionnaires (BQs). The baseline DL methods were also run for comparison with the proposed pipeline. TD features are widely used for stress and fatigue classification and are computationally efficient compared to frequency-domain or non-linear HRV features. The SQI helps account for the increased variability in noise present in everyday recordings, while BQs provide context about individual differences that may influence physiological responses. Three-fold stratified group k-fold cross-validation was used to maintain class balance across folds while ensuring that no data from the same participant appeared in both the training and test sets.

Stress classification performance improved with the inclusion of SQI and baseline questionnaires, leading to better average precision across models. For TD features only, kNN achieved the highest AP score of 80.99%. SVM achieved the highest AP scores of 91.15% when SQI and BQs were added (Table VII). AUC-ROC results followed a similar trend, kNN performing best for TD features alone (77.02%), while SVM achieved the highest performance of 91.60% with the additional features (Table VII). The hybrid baseline evaluation resulted in lower overall AP and AUC-ROC, with SWM achieving an AP of 81.13% and AUC-ROC of 83.00%. Similar to the AKTIVES training set, the CNN baseline evaluation indicates substantial overfitting for stress classification on the Wellby dataset.

TABLE VII
STRESS CLASSIFICATION PERFORMANCE ($\mu \pm \sigma$) ON THE WELLBY DATASET ACROSS DIFFERENT APPROACHES USING AVERAGE PRECISION (AP) AND AUC-ROC.

| Method | Algorithm | AP | AUC-ROC |
|---|---|---|---|
| TD Features Only | RF | 63.57 ± 4.97 | 65.67 ± 11.30 |
| | AB | 58.60 ± 11.62 | 63.76 ± 12.77 |
| | kNN | **80.99 ± 5.40** | **77.02 ± 7.71** |
| | LDA | 67.41 ± 7.24 | 68.65 ± 9.34 |
| | SVM | 77.04 ± 2.06 | 71.56 ± 3.55 |
| | GB | 52.22 ± 5.86 | 49.34 ± 4.96 |
| TD + SQI + BQs | RF | 75.80 ± 12.27 | 75.60 ± 15.77 |
| | AB | 62.15 ± 7.92 | 69.58 ± 5.34 |
| | kNN | 76.45 ± 10.42 | 73.88 ± 11.18 |
| | LDA | 84.59 ± 6.15 | 86.71 ± 4.41 |
| | SVM | **91.15 ± 3.38** | **91.60 ± 3.48** |
| | GB | 65.48 ± 15.89 | 64.65 ± 12.92 |
| Hybrid Baseline (CNN + TD) | RF | 60.47 ± 5.46 | 61.97 ± 2.78 |
| | AB | 61.84 ± 9.84 | 59.49 ± 5.08 |
| | kNN | 68.82 ± 8.85 | 69.84 ± 14.33 |
| | LDA | 70.82 ± 15.59 | 74.87 ± 13.61 |
| | SVM | **81.13 ± 8.32** | **83.00 ± 9.46** |
| | GB | 70.66 ± 18.44 | 74.21 ± 11.27 |
| CNN Baseline | CNN | **44.25 ± 31.32** | **33.27 ± 23.76** |

Fatigue classification followed the same evaluation process as stress classification (Tables VIII). The highest AP (85.84%) was achieved by SVM when trained on TD features, baseline questionnaires, and SQI combined. GB performed best for TD features alone with an AP of 83.19% and an AUC-ROC of 71.58%. AP values were consistently higher than AUC-ROC, suggesting that precision is higher than the overall separability between fatigue and non-fatigue classes. The hybrid baseline performance was generally lower than the proposed pipeline performance, with a maximum AP of 77.73% and AUC-ROC of 56.80%. Conversely, the CNN baseline performance was comparable to the proposed pipeline with an AP of 82.58% and AUC-ROC of 73.36%, suggesting the potential influence of class imbalance in a small dataset on this classification method, since there is a positive class imbalance for fatigue data in contrast to the negative class imbalance for stress detection.

Leave-One-Group-Out (LOGO) cross-validation was also performed for both stress and fatigue classification. Due to the small sample size in the Wellby dataset, there was a limited number of samples per participant. Most of the participants had only 1 to 3 recordings and a few had data covering both class labels (non-stress/stress or non-fatigue/fatigue). As a result, LOGO performance varied widely, particularly for participants with only one class label (non-stress/stress or non-fatigue/fatigue), rendering metrics that depend on both true positives and true negatives less reliable. Balanced accuracy was used as the primary evaluation metric for LOGO classification, given class imbalance and participant variability. TD features, baseline questionnaires, and SQI were included in both stress and fatigue classification evaluations that showed highly variable results (Table IX). The highest balanced accuracy was 81.14% for stress classification (SD = 32.43) and 75.00% for fatigue classification (SD = 37.17) (Table IX). The large standard deviation suggests that the LOGO cross-validation results lacked stability for this dataset.

TABLE VIII
FATIGUE CLASSIFICATION PERFORMANCE ($\mu \pm \sigma$) ON THE WELLBY DATASET ACROSS DIFFERENT APPROACHES USING AVERAGE PRECISION (AP) AND AUC-ROC.

| Method | Algorithm | AP | AUC-ROC |
|---|---|---|---|
| TD Features Only | RF | 74.79 ± 12.37 | 52.79 ± 14.08 |
| | AB | 73.37 ± 11.24 | 59.26 ± 16.60 |
| | kNN | 68.17 ± 12.47 | 42.97 ± 20.68 |
| | LDA | 74.00 ± 7.34 | 49.42 ± 12.06 |
| | SVM | 75.32 ± 7.87 | 54.51 ± 8.69 |
| | GB | **83.19 ± 12.12** | **71.58 ± 12.09** |
| TD + SQI + BQs | RF | 77.48 ± 9.70 | 61.08 ± 14.74 |
| | AB | 81.57 ± 9.29 | 66.95 ± 17.59 |
| | kNN | 68.83 ± 12.53 | 50.33 ± 19.42 |
| | LDA | 78.80 ± 10.38 | 58.61 ± 16.00 |
| | SVM | **85.84 ± 11.20** | **68.58 ± 23.07** |
| | GB | 75.56 ± 7.67 | 52.65 ± 12.69 |
| Hybrid Baseline (CNN + TD) | RF | 73.18 ± 10.99 | 46.99 ± 19.11 |
| | AB | 77.48 ± 9.37 | 62.15 ± 12.16 |
| | kNN | 69.28 ± 3.94 | 37.43 ± 13.05 |
| | LDA | 69.62 ± 11.15 | 44.97 ± 21.31 |
| | SVM | 72.91 ± 8.54 | 42.28 ± 20.21 |
| | GB | **77.73 ± 18.46** | **56.80 ± 36.50** |
| CNN Baseline | CNN | **82.58 ± 16.28** | **73.36 ± 18.95** |

TABLE IX
BALANCED ACCURACY ($\mu \pm \sigma$) FROM LOGO CROSS-VALIDATION FOR STRESS AND FATIGUE CLASSIFICATION ON THE WELLBY DATASET.

| Algorithm | Stress | Fatigue |
|---|---|---|
| RF | 60.09 ± 39.05 | **75.00 ± 37.17** |
| AB | 71.05 ± 37.40 | 67.11 ± 39.78 |
| kNN | 65.35 ± 39.02 | 63.16 ± 42.43 |
| LDA | **81.14 ± 32.43** | 66.05 ± 43.61 |
| SVM | 80.70 ± 36.78 | 67.89 ± 38.23 |
| GB | 71.05 ± 37.40 | 68.42 ± 37.91 |

TABLE X
COMPUTATIONAL EFFICIENCY COMPARISON OF THE PROPOSED PPG PROCESSING PIPELINE AGAINST BASELINE METHODS ACROSS DATASETS.

| Method | WESAD | | | AKTIVES | | | Wellby | | |
|---|---|---|---|---|---|---|---|---|---|
| | Train Time (s) | Inference Time (s) | Model Size (KB) | Train Time (s) | Inference Time (s) | Model Size (KB) | Train Time (s) | Inference Time (s) | Model Size (KB) |
| Proposed Pipeline | 3.906 | 0.003 | 12.612 | 0.319 | 0.001 | 10.350 | 0.099 | 0.001 | 3.344 |
| Hybrid Baseline | 21.616 | 0.007 | 117.561 | 3.215 | 0.003 | 44.031 | 2.074 | 0.003 | 11.620 |
| CNN Baseline | 6184.827 | 0.046 | 6791.880 | 303.347 | 0.016 | 6791.900 | 42.021 | 0.011 | 6791.820 |

### F. Model Efficiency

Beyond evaluating model performance on the Wellby dataset, we also assessed computational efficiency to explore the feasibility of deploying these models on wearable devices. The NRF52840 includes a 64 MHz ARM Cortex-M4 processor with 256 KB of RAM and 1 MB of flash memory. To validate the suitability of our approach with these hardware constraints, we compared the proposed processing pipeline against an existing PPG-based stress detection pipeline proposed in Heo et al. [41] using a 2-minute segment of WESAD data. The adapted pipeline achieved a 96.9% reduction in processing time and a 55.6% reduction in memory usage for the extracted time-domain features compared to the time, frequency, and non-linear domains extracted in Heo et al. [41].

The training time, inference time, and model sizes for stress classification were evaluated for each method and dataset (Table X). All experiments were conducted on a laptop with an Apple M1 Pro chip (8-core CPU, 16GB unified memory), using Python 3.13.1 and scikit-learn 1.6.1. CPU was used instead of GPU to simulate conditions similar to wearable deployment, where GPU processing is not feasible. SVM was chosen as the reference model due to its high-performance metrics across the datasets. The training time represents the combined time for feature extraction and model training. The 1D CNN in the hybrid baseline for feature extraction and the dilated CNN used in the CNN baseline were trained for 30 epochs to ensure consistent comparison across methods. The inference time represents the average time per sample for feature extraction and classification. Model size refers to the memory footprint of the trained model. The proposed pipeline achieved the shortest training times, fastest inference times, and smallest model sizes across all datasets. For the WESAD dataset, training times were 3.906, 5.846, and 221.157 seconds for the proposed pipeline, hybrid baseline, and CNN baseline, respectively. Similarly, model sizes for the Wellby dataset were 3.344, 11.620, and 6791.820 KB, respectively. This analysis highlights the computational efficiency of the proposed pipeline for practical application in wearable or resource-constrained environments.

### G. Demographic Bias

To examine potential demographic bias due to imbalanced data, we compared the classification accuracy of the best-performing model (SVM with all features) across demographic subgroups (Table XI). To ensure sufficient sample sizes in each subgroup, only male and female genders were reported, and ethnicity was grouped into non-white and white categories. The stress classifier was 3.2% more accurate for male participants, whereas the fatigue classifier was 7.7% more accurate for female participants. Ethnicity and school showed larger differences in classification accuracy compared to gender. The stress and fatigue classification accuracy for white participants was 4.2% and 18.8% greater than non-white participants, respectively. School-based analysis revealed that School 2 had 10.8% greater stress detection accuracy while School 1 had 19.5-27.8% greater fatigue prediction accuracy compared to other schools.

TABLE XI
CLASSIFICATION ACCURACY OF THE BEST-PERFORMING MODEL (SVM WITH ALL FEATURES) BY DEMOGRAPHIC GROUPS.

| Category | Sample Size (n) | Stress Accuracy | Fatigue Accuracy |
|---|---|---|---|
| **Gender** | | | |
| Female | 16 | 68.8% | 68.8% |
| Male | 18 | 72.2% | 61.1% |
| **Ethnicity** | | | |
| Non-white | 6 | 66.7% | 50.0% |
| White | 32 | 71.9% | 68.8% |
| **School** | | | |
| School 1 | 12 | 66.7% | 77.8% |
| School 2 | 8 | 87.5% | 50.0% |
| School 3 | 18 | 66.7% | 58.3% |

### H. Feature Importance Analysis

Shapley values were used to quantify feature importance in stress and fatigue classification from the Wellby dataset. These values were calculated for SVM models trained using three-fold stratified group k-fold cross-validation. SVM was selected due to its strong performance for both stress and fatigue classification and the use of a linear kernel, which facilitates interpretation of feature impact via Shapley values. Feature importance was assessed for models trained on TD features, SQI, and baseline questionnaire data. The Shapley value heatmaps represent the feature importance across each recording session in the Wellby dataset.

In stress classification, the features with the greatest impact on model output included mean HR, median NN, baseline

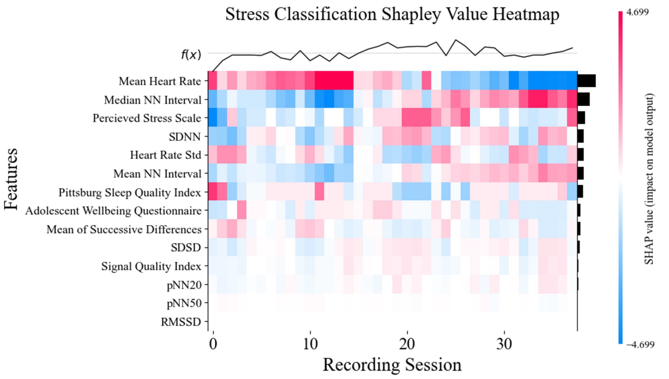

Fig. 5. Shapley value heatmap for stress classification showing the contribution of each feature across individual data samples.

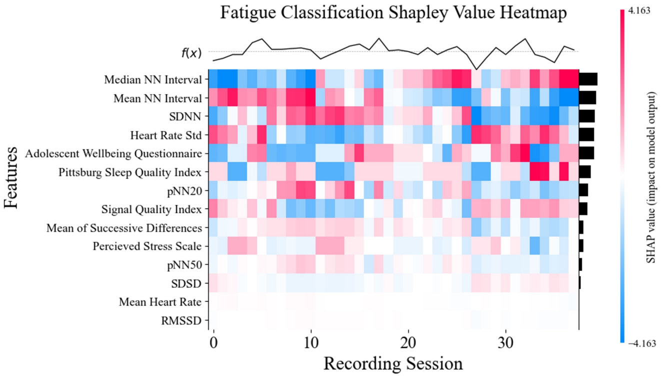

Fig. 6. Shapley value heatmap for fatigue classification showing the contribution of each feature across individual data samples.

Perceived Stress Scale, SDNN, and standard deviation of HR. The TD metrics (mean HR, median NN, SDNN, and standard deviation of HR) are widely used HRV metrics associated with stress [73], [74]. In particular, the Perceived Stress Scale, measured at the beginning of the month, had a significant influence on the model, suggesting that students with higher baseline stress were more likely to report moments of stress throughout the month. The heatmap of the feature importance for stress classification is shown in Fig. 5.

In fatigue classification, the feature importance ranking differed slightly from those observed in stress classification. The features with the greatest impact included the median NN interval, mean NN interval, SDNN, standard deviation of HR, and the baseline well-being questionnaire. These HRV metrics have previously been associated with the detection of sleep-related fatigue and drowsiness [75]. Interestingly, median NN interval, SDNN, and standard deviation of HR were among the five most impactful features for stress and fatigue classification. The EPOCH Adolescent Well-being Questionnaire also had a high impact on the model output, suggesting that students with lower baseline well-being scores were more likely to report fatigue throughout the month. The heatmap of the importance of the features for fatigue classification is shown in Fig. 6. These results highlight the value of combining TD HRV features and contextual variables to improve the stress and fatigue classifications of the model.

## V. Discussion

The results demonstrate the application of a PPG processing pipeline that was adapted for future on-device classification of stress and fatigue based on real-world data from adolescents. The lightweight processing pipeline makes affect detection more accessible for student populations by eliminating reliance on cloud computing.

The point-biserial correlation between individual PPG features and stress or fatigue labels were moderate to weak across the included datasets, reflecting the need for ML models to detect complex physiological correlations with affective state. This aligns with previous research which indicates that affective states are related to multiple interacting physiological and contextual factors [76]. The weaker correlations in the Wellby dataset compared to the WESAD dataset may also suggest the complexity of physiological responses in real-world settings compared to controlled lab settings with pre-defined stressors, consistent with previous literature on HRV in naturalistic settings [77]. Additionally, the negative correlations of Wellby HRV features generally followed expected trends of HRV features decreasing in response to physiological stressors [56]. The positive correlations in the WESAD and AKTIVES dataset may indicate that these controlled experiments are detecting increases in cognitive load [78] or attention [79] during the experimental tasks and are not generalizable for real-world stress detection.

The presented processing pipeline was computationally efficient, with substantially lower training time and model size compared to the DL baselines, while still including necessary noise reduction and ML classification algorithms. The pipeline relies solely on TD HRV features, avoiding computationally intensive spectral analysis or DL feature extraction. It maintained a high classification performance compared to DL baseline methods for the public WESAD and AKTIVES datasets, collected in a controlled environment, and the Wellby dataset, collected during the daily lives of the students in this study. The performance of stress classification on the Wellby dataset was lower than that of the WESAD dataset, but comparable to the performance on the AKTIVES dataset. This may reflect the increased complexities of detecting stress in a youth cohort [17], increased motion artifacts [26], and less controlled stressors in real-world environments [77].

The high performance of SVM across the WESAD and Wellby datasets may demonstrate SVM's advantages including robustness to small sample sizes and effective margin maximization in high-dimensional feature spaces [80]. The strong performance of GB, particularly in fatigue detection, aligns with previous research on the strength of tree-based models in stress and fatigue classification [4], [57]. This may also be due to the ability of these models to capture complex non-linear patterns and feature interactions without distributional assumptions about the input data. This is particularly relevant given that non-linear relationships are suggested to exist between PPG time-domain metrics and stress-induced sympathetic nervous system activity [81], which tree-based models can effectively capture.

The stress and fatigue classification using the time domain features extracted from Wellby achieved results comparable to other real-world classification methods [4], [22], [82]. The inclusion of a SQI improved the classification performance across most models. This may be because the SQI allowed the model to adjust the contribution of HRV features based on signal quality, helping to mitigate the influence of noise on the classification. When baseline questionnaires on student stress, sleep, and overall well-being were also included in the feature set, the classification of stress yielded the highest results. The inclusion of contextual data may explain why this model performed better than those in a similar study [55]. This underscores the benefit of complementing HRV metrics with contextual data to enhance classification performance, consistent with similar research on stress classification [83].

The performance of the baseline methods generally decreased on smaller datasets relative to the performance of the proposed pipeline. The size of the AKTIVES and Wellby datasets may be insufficient to train more complex CNN architectures and lead to overfitting and poor generalization. These findings highlight an advantage of the proposed pipeline for real-world wearable deployment, where training data and computational resources are often limited. There remains a significant need for research into lightweight deep learning models that are both computationally efficient and able to recognize complex physiological patterns.

The Shapley values indicated that HRV metrics, such as SDNN and median NN interval, correlated with both stress and drowsiness detection, which is consistent with the literature [10], [56]. Shapley value analysis also revealed that the baseline well-being questionnaires had a strong impact on model output. This further emphasizes the relevance of combining physiological signals with contextual data to better capture affective states in real-world environments.

The study sample size limited the generalizability of outcomes associated with the Wellby dataset. In particular, the LOGO cross-validation yielded highly unstable performance metrics due to the small number of recordings per participant, the imbalance between stress and non-stress labels, and the range of label balance within individual participants. The high variability indicates potential overfitting to individual participant characteristics rather than learning generalizable HRV patterns associated with stress and fatigue across students. Future work should validate these methods in larger, more diverse student cohorts to test model performance on unseen data using LOGO validation. Given the small sample size and the focus of the study on lightweight, on-device classification methods, deep learning models were not evaluated in this context. Future studies with larger student datasets could explore deep learning approaches and personalized models for affect detection.

This study highlighted varied model performance based on participant demographics in the Wellby dataset. This performance variation may suggest physiological differences in response to affective states, cultural differences in self-reporting patterns, or the relatively small sample size of particular subgroups. Since resting HRV varies between demographic groups including ethnicity [84], age [85], and gender [85], the underrepresentation of certain demographic subgroups can lead to model bias and limit generalizability to broader populations. The imbalanced distribution of fatigue and stress labels across demographic groups may further influence model generalizability, as evidenced by the substantial performance differences observed across subgroups. The observed performance differences across schools may be particularly relevant given that School 3 represents a Youthreach Centre serving early school-leavers, who may experience different stressors compared to traditional secondary school students [46].

Beyond sample size and demographic representation, another limitation is that this study did not account for individual differences in student health or physiology. Health status is a determinant of HRV patterns including levels of inflammation [86], presence of cardiovascular disease [87], metabolic conditions [88], and mental health disorders [89]. Future affect detection studies should incorporate health status data to account for the potential influence of these conditions on individual baseline HRV and stress responses.

Future research should investigate lightweight PPG processing pipelines across diverse student populations with particular attention to demographic-specific model performance and personalized approaches. Future studies could also implement measures to enhance the reliability of self-reported measures since many affect detection studies rely on these measures as ground truth outside of laboratory settings [57]. To account for this, self-reported measures can be strengthened by including multiple biosignals and implementing baseline calibration techniques [90]. Exploration of affect detection in various educational settings could foster models that are both generalizable and lightweight. To implement appropriate affect detection for students, engineers should collaborate with stakeholders such as students, teachers, psychologists, and behavioral scientists to understand how to provide effective biofeedback insights for improving student well-being. Further collaboration between industry and academia could also help increase the transparency and validity of affect detection on commercial devices [29].

Taken together, this study addresses important gaps in the literature, including the limited exploration of PPG-based affect detection in real-world settings [21], [22], [91] and the lack of lightweight classification algorithms that are deployable on wearable devices [92]. The co-designed Wellby wearable and mobile app represents a student-centered approach to affective computing, providing a customized intervention aligned with student preferences and practical needs. In future studies, the Wellby device could be used to test real-time, on-device deployment of affect detection models in educational contexts. By demonstrating the feasibility of lightweight processing on real-world student data, this work contributes to the broader goal of making physiological monitoring more accessible in everyday settings, such as classrooms or rural environments, where cloud connectivity may be limited.

## VI. Conclusion

This study contributes to the growing interest in the application of wearables for emotion recognition in everyday settings by demonstrating the effectiveness of a lightweight machine learning approach for stress and fatigue detection [3]. The results of this study demonstrate the ability of a computationally efficient processing pipeline to achieve reliable stress and fatigue classification from PPG recordings captured from secondary school students during daily life. This pipeline demonstrated strong classification performance on the public WESAD and AKTIVES dataset, as well as the Wellby dataset despite the small sample size. It also demonstrates the benefit of including signal quality metrics and contextual information alongside HRV features to improve affect detection in everyday environments. Further research should explore larger samples and more personalized machine learning approaches for implementing affective computing pipelines on wearable devices, including usability and ethical considerations to ensure that they appropriately address student well-being needs. This can help bridge the gap between affective computing and the application of wearables in real-world environments such as educational settings.

## Acknowledgments

The authors would like to acknowledge the support of the teachers of each of the participating secondary schools for their assistance in organizing the study and sharing information with the students. The participation and feedback of students throughout the process has been an invaluable part of this research. The authors also gratefully acknowledge the RCSI Postgraduate Secondment Award, which enabled the collaboration between RCSI and the University of Oxford.

## References

[1] S. Park and S. Jayaraman, *Wearables: Fundamentals, advancements, and a roadmap for the future*. Elsevier, 2021, pp. 3–27.

[2] J. Dunn, R. Runge, and M. Snyder, "Wearables and the medical revolution," *Personalized medicine*, vol. 15, no. 5, pp. 429–448, 2018.

[3] S. Saganowski, B. Perz, A. G. Polak, and P. Kazienko, "Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1876–1897, 2023.

[4] N. R. Adão Martins, S. Annaheim, C. M. Spengler, and R. M. Rossi, "Fatigue monitoring through wearables: A state-of-the-art review," *Frontiers in physiology*, vol. 12, p. 790292, 2021.

[5] S. Gedam and S. Paul, "A review on mental stress detection using wearable sensors and machine learning techniques," *IEEE Access*, vol. 9, pp. 84 045–84 066, 2021.

[6] M. Spapé, K. Mäkelä, and T. Ruotsalo, "Nemo: A database for emotion analysis using functional near-infrared spectroscopy," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 1166–1177, 2024.

[7] M. L. González Ramírez, J. P. García Vázquez, M. D. Rodríguez, L. A. Padilla-López, G. M. Galindo-Aldana, and D. Cuevas-González, "Wearables for stress management: A scoping review," in *Healthcare*, vol. 11. MDPI, 2023, Conference Proceedings, p. 2369.

[8] e. a. Kim, Hye-Geum, "Stress and heart rate variability: a meta-analysis and review of the literature." *Psychiatry investigation*, vol. 15.3, p. 235, 2018.

[9] F. Li, G. Liu, Z. Zou, Y. Yan, X. Huang, X. Liu, and Z. Liu, "A classification framework for depressive episode using r-r intervals from smartwatch," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 1387–1399, 2024.

[10] J. Vicente, P. Laguna, A. Bartra, and R. Bailón, "Drowsiness detection using heart rate variability," *Medical & biological engineering & computing*, vol. 54, pp. 927–937, 2016.

[11] D. T. Weiler, S. O. Villajuan, L. Edkins, S. Cleary, and J. J. Saleem, "Wearable heart rate monitor technology accuracy in research: a comparative study between ppg and ecg technology," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 61. SAGE Publications Sage CA: Los Angeles, CA, 2024, Conference Proceedings, pp. 1292–1296.

[12] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, and N. B. Hussin, "Affective computing in education: A systematic review and future research," *Computers & Education*, vol. 142, p. 103649, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360131519302027

[13] R. T. Bradley, R. McCraty, M. Atkinson, L. Arguelles, R. A. Rees, and D. Tomasino, "Reducing test anxiety and improving test performance in america's schools," *Boulder Creek: Institute of Heart-Math*, 2007.

[14] K. Horgan, S. Howard, and F. Gardiner-Hyland, "Pre-service teachers and stress during microteaching: An experimental investigation of the effectiveness of relaxation training with biofeedback on psychological and physiological indices of stress," *Applied Psychophysiology and Biofeedback*, vol. 43, no. 3, pp. 217–225, 2018. [Online]. Available: https://doi.org/10.1007/s10484-018-9401-9

[15] R. D. Romeo, "The teenage brain: The stress response and the adolescent brain," *Current Directions in Psychological Science*, vol. 22, no. 2, pp. 140–145, 2013, doi: 10.1177/0963721413475445. [Online]. Available: https://doi.org/10.1177/0963721413475445

[16] N. Ramadhani and T. Mahmudiono, "Academic stress is associated with emotional eating behavior among adolescent," *Media Gizi Indonesia*, vol. 16, no. 1, pp. 38–47, 2021.

[17] M. C. Pascoe, H. S. E., , and A. G. Parker, "The impact of stress on students in secondary school and higher education," *International Journal of Adolescence and Youth*, vol. 25, no. 1, pp. 104–112, 2020, doi: 10.1080/02673843.2019.1596823. [Online]. Available: https://doi.org/10.1080/02673843.2019.1596823

[18] S. M. Sawyer, R. A. Afifi, L. H. Bearinger, S.-J. Blakemore, B. Dick, A. C. Ezeh, and G. C. Patton, "Adolescence: a foundation for future health," *The lancet*, vol. 379, no. 9826, pp. 1630–1640, 2012.

[19] V. A. Goodyear, K. M. Armour, and H. Wood, "Young people learning about health: the role of apps and wearable devices," *Learning, Media and Technology*, vol. 44, no. 2, pp. 193–210, 2019. [Online]. Available: https://doi.org/10.1080/17439884.2019.1539011

[20] R. Raeside, N. Halim, S. Partridge, and J. Redfern, "713 a review of mobile apps for prevention of chronic disease in adolescents across a range of behaviours," *Heart, Lung and Circulation*, vol. 29, 2020.

[21] M. Namvari, J. Lipoth, S. Knight, A. A. Jamali, M. Hedayati, R. J. Spiteri, and S. Syed-Abdul, "Photoplethysmography enabled wearable devices and stress detection: a scoping review," *Journal of Personalized Medicine*, vol. 12, no. 11, p. 1792, 2022.

[22] M. Bolpagni, S. Pardini, M. Dianti, and S. Gabrielli, "Personalized stress detection using biosignals from wearables: A scoping review," *Sensors*, vol. 24, no. 10, p. 3221, 2024.

[23] A. Alslaity and R. Orji, "Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions," *Behaviour & Information Technology*, vol. 43, no. 1, pp. 139–164, 2024. [Online]. Available: https://doi.org/10.1080/0144929X.2022.2156387

[24] J. Williamson, Q. Liu, F. Lu, W. Mohrman, K. Li, R. Dick, and L. Shang, "Data sensing and analysis: Challenges for wearables," in *The 20th Asia and South Pacific Design Automation Conference*. IEEE, 2015, Conference Proceedings, pp. 136–141.

[25] S. Seneviratne, Y. Hu, T. Nguyen, G. Lan, S. Khalifa, K. Thilakarathna, M. Hassan, and A. Seneviratne, "A survey of wearable devices and challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2573–2620, 2017.

[26] K. B. Kim and H. J. Baek, "Photoplethysmography in wearable devices: A comprehensive review of technological advances, current challenges, and future directions," 2023.

[27] S. M. A. Iqbal, I. Mahgoub, E. Du, M. A. Leavitt, and W. Asghar, "Advances in healthcare wearable devices," *npj Flexible Electronics*, vol. 5, no. 1, p. 9, 2021. [Online]. Available: https://doi.org/10.1038/s41528-021-00107-x

[28] Y. Abdelaal, M. Aupetit, A. Baggag, and D. Al-Thani, "Exploring the applications of explainability in wearable data analytics: Systematic literature review," *J Med Internet Res*, vol. 26, p. e53863, 2024. [Online]. Available: https://www.jmir.org/2024/1/e53863https://doi.org/10.2196/53863http://www.ncbi.nlm.nih.gov/pubmed/39718820

[29] C. Doherty, M. Baldwin, R. Lambe, D. Burke, and M. Altini, "Readiness, recovery, and strain: an evaluation of composite health scores in consumer wearables," *Translational Exercise Biomedicine*, 2025. [Online]. Available: https://doi.org/10.1515/teb-2025-0001

[30] E. Covi, E. Donati, X. Liang, D. Kappel, H. Heidari, M. Payvand, and W. Wang, "Adaptive extreme edge computing for wearable devices," *Frontiers in Neuroscience*, vol. 15, p. 611300, 2021.

[31] G. Premsankar, M. Di Francesco, and T. Taleb, “Edge computing for the internet of things: A case study,” *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1275–1284, 2018.

[32] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, “Introducing wesad, a multimodal dataset for wearable stress and affect detection,” in *Proceedings of the 20th ACM international conference on multimodal interaction*, 2018, Conference Proceedings, pp. 400–408.

[33] B. Coşkun, S. Ay, D. Erol Barkana, H. Bostanci, Uzun, A. B. Oktay, B. Tuncel, and D. Tarakci, “A physiological signal database of children with different special needs for stress recognition,” *Scientific Data*, vol. 10, no. 1, p. 382, 2023. [Online]. Available: https://doi.org/10.1038/s41597-023-02272-2

[34] M. A. Hernández-Mustieles, Y. E. Lima-Carmona, M. A. Pacheco-Ramírez, A. A. Mendoza-Armenta, J. E. Romero-Gómez, C. F. Cruz-Gómez, D. C. Rodríguez-Alvarado, A. Arceo, J. G. Cruz-Garza, M. A. Ramírez-Moreno, and J. d. J. Lozoya-Santos, “Wearable biosensor technology in education: A systematic review,” 2024.

[35] S. Hammoud, R. Karam, R. Mourad, I. Saad, and M. Kurdi, “Stress and heart rate variability during university final examination among lebanese students,” 2019.

[36] A. Aranberri-Ruiz, A. Aritzeta, A. Olarza, G. Soroa, and R. Mindeguia, “Reducing anxiety and social stress in primary education: A breath-focused heart rate variability biofeedback intervention,” 2022.

[37] R. Gopinathan.N and E. Sherly, “Visual attention score and fatigue level measure of students through eye analysis–machine learning approach,” in *2022 IEEE 19th India Council International Conference (INDICON)*, 2022, Conference Proceedings, pp. 1–5.

[38] L. Juntong, “Application of artificial neural network in student’s behavior analysis in english class,” in *2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI)*, 2022, Conference Proceedings, pp. 01–03.

[39] S. Mandal, B. Ghosh, and R. Naskar, “A photoplethysmography (ppg) sensor based stress level monitoring system for undergraduate students of technical education,” in *2023 IEEE 20th India Council International Conference (INDICON)*, 2023, Conference Proceedings, pp. 197–202.

[40] P. Bobade and M. Vani, “Stress detection with machine learning and deep learning using multimodal physiological data,” in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2020, Conference Proceedings, pp. 51–57.

[41] S. Heo, S. Kwon, and J. Lee, “Stress detection with single ppg sensor by orchestrating multiple denoising and peak-detecting methods,” *IEEE Access*, vol. 9, pp. 47 777–47 785, 2021.

[42] L. Zhao, X. Niu, L. Wang, J. Niu, X. Zhu, and Z. Dai, “Stress detection via multimodal multitemporal-scale fusion: A hybrid of deep learning and handcrafted feature approach,” *IEEE Sensors Journal*, vol. 23, no. 22, pp. 27 817–27 827, 2023.

[43] M. Chen, Q. Zhu, M. Wu, and Q. Wang, “Modulation model of the photoplethysmography signal for vital sign extraction,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 969–977, 2020.

[44] Pankaj, A. Kumar, M. Kumar, and R. Komaragiri, “Optimized deep neural network models for blood pressure classification using fourier analysis-based time–frequency spectrogram of photoplethysmography signal,” *Biomedical Engineering Letters*, vol. 13, no. 4, pp. 739–750, 2023. [Online]. Available: https://doi.org/10.1007/s13534-023-00296-6

[45] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, and W. Zhang, “A systematic review on affective computing: Emotion models, databases, and recent advances,” *Information Fusion*, vol. 83, pp. 19–52, 2022.

[46] J. Laiti, J. Donnelly, E. Byrne, and P. Dunne, “Co-creating wellby – a mobile app and wearable for student well-being management guided by a needs assessment and co-design,” *Frontiers in Digital Health*, vol. 7, 2025.

[47] S. Buerger, J. Holzer, T. Yanagida, B. Schober, and C. Spiel, “Measuring adolescents’ well-being in schools: The adaptation and translation of the epoch measure of adolescent well-being—a validation study,” *School Mental Health*, vol. 15, no. 2, pp. 611–626, 2023.

[48] W. Chen, Y. Liang, J. Cheng, E. Xie, Q. Zhu, and J. Liu, “Psychometric properties of the perceived stress scale short form (pss-10) in chinese adolescents: validity, measurement invariance across gender and left-behind status,” *Children and Youth Services Review*, vol. 156, p. 107252, 2024.

[49] R. De la Vega, C. Tomé-Pires, E. Solé, M. Racine, E. Castarlenas, M. P. Jensen, and J. Miró, “The pittsburgh sleep quality index: Validity and factor structure in young people,” *Psychological assessment*, vol. 27, no. 4, p. e22, 2015.

[50] N. Mitro, K. Argyri, L. Pavlopoulos, D. Kosyvas, L. Karagiannidis, M. Kostovasili, F. Misichroni, E. Ouzounoglou, and A. Amditis, “Ai-enabled smart wristband providing real-time vital signs and stress monitoring,” *Sensors*, vol. 23, no. 5, p. 2821, 2023.

[51] A. Sharma, C. W. Lo, C. Law, J. Ovia, M. Hia, P. Phongsopa, R. Horder, and R. Tang, “An evaluation of heart rate and pulse oximeter modules in capturing raw photoplethysmography signals in order to fabricate a wearable device to predict the onset of exertional heatstroke in athletes,” *Biomedical Journal of Scientific & Technical Research*, vol. 34, no. 5, pp. 27 147–27 160, 2021.

[52] M. Feli, K. Kazemi, I. Azimi, Y. Wang, A. M. Rahmani, and P. Liljeberg, “End-to-end ppg processing pipeline for wearables: From quality assessment and motion artifacts removal to hr/hrv feature extraction,” in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2023, Conference Proceedings, pp. 1895–1900.

[53] H. J. Han, S. Labbaf, J. L. Borelli, N. Dutt, and A. M. Rahmani, “Objective stress monitoring based on wearable sensors in everyday settings,” *Journal of Medical Engineering & Technology*, vol. 44, no. 4, pp. 177–189, 2020.

[54] S. Vadrevu and M. S. Manikandan, “A robust pulse onset and peak detection method for automated ppg signal analysis system,” *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 3, pp. 807–817, 2018.

[55] A. Tazarv, S. Labbaf, S. M. Reich, N. Dutt, A. M. Rahmani, and M. Levorato, “Personalized stress monitoring using wearable sensors in everyday settings,” in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, Conference Proceedings, pp. 7332–7335.

[56] F. Shaffer and J. Ginsberg, “An overview of heart rate variability metrics and norms.” *Frontiers in public health*, vol. 5, 2017.

[57] G. Vos, K. Trinh, Z. Sarnyai, and M. R. Azghadi, “Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review,” *International Journal of Medical Informatics*, vol. 173, p. 105026, 2023.

[58] R. Mukkamala, J. O. Hahn, O. T. Inan, L. K. Mestha, C. S. Kim, H. Töreyin, and S. Kyal, “Toward ubiquitous blood pressure monitoring via pulse transit time: Theory and practice,” *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 8, pp. 1879–1901, 2015.

[59] F. Peng, Z. Zhang, X. Gou, H. Liu, and W. Wang, “Motion artifact removal from photoplethysmographic signals by combining temporally constrained independent component analysis and adaptive filter,” *BioMedical Engineering OnLine*, vol. 13, no. 1, p. 50, 2014. [Online]. Available: https://doi.org/10.1186/1475-925X-13-50

[60] K. Vandecasteele, J. Lázaro, E. Cleeren, K. Claes, W. Van Paesschen, S. Van Huffel, and B. Hunyadi, “Artifact detection of wrist photoplethysmograph signals,” in *BIOSIGNALS*, 2018, Conference Proceedings, pp. 182–189.

[61] T. Pereira, C. Ding, K. Gadhoumi, N. Tran, R. A. Colorado, K. Meisel, and X. Hu, “Deep learning approaches for plethysmography signal quality assessment in the presence of atrial fibrillation,” *Physiological Measurement*, vol. 40, no. 12, p. 125002, 2019. [Online]. Available: https://dx.doi.org/10.1088/1361-6579/ab5b84

[62] H. S. Shin, C. Lee, and M. Lee, “Adaptive threshold method for the peak detection of photoplethysmographic waveform,” *Computers in biology and medicine*, vol. 39, no. 12, pp. 1145–1152, 2009.

[63] L. Breiman, “Random forests,” *Machine learning*, vol. 45, pp. 5–32, 2001.

[64] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[65] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[66] S. Balakrishnama and A. Ganapathiraju, “Linear discriminant analysis-a brief tutorial,” *Institute for Signal and information Processing*, vol. 18, no. 1998, pp. 1–8, 1998.

[67] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, pp. 273–297, 1995.

[68] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.

[69] A. Matuz, D. van der Linden, G. Darnai, and Á. Csathó, “Generalisable machine learning models trained on heart rate variability data to predict mental fatigue,” *Scientific Reports*, vol. 12, no. 1, p. 20023, 2022.

[70] K. Motaman, K. Alipour, B. Tarvirdizadeh, and M. Ghamari, “A dilated cnn-based model for stress detection using raw ppg signals,” *IET Wireless Sensor Systems*, vol. 15, no. 1, p. e70004, 2025.

[71] B. Coşkun, D. E. Barkana, Uzun, H. Bostancı, and D. Tarakçı, “Classification of stress and participation using physiological signals of children

during serious game-based therapy," in *2023 Medical Technologies Congress (TIPTEKNO)*. IEEE, 2023, Conference Proceedings, pp. 1–4.

[72] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.

[73] A. J. Camm, M. Malik, J. T. Bigger, G. Breithardt, S. Cerutti, R. J. Cohen, P. Coumel, E. L. Fallen, H. L. Kennedy, and R. E. Kleiger, "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," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[74] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and heart rate variability: a meta-analysis and review of the literature," *Psychiatry investigation*, vol. 15, no. 3, p. 235, 2018.

[75] A. Burlacu, C. Brinza, A. Brezulianu, and A. Covic, "Accurate and early detection of sleepiness, fatigue and stress levels in drivers through heart rate variability parameters: a systematic review," *Reviews in cardiovascular medicine*, vol. 22, no. 3, pp. 845–852, 2021.

[76] A. Ahmed, J. Ramesh, S. Ganguly, R. Aburukba, A. Sagahyroon, and F. Aloul, "Evaluating multimodal wearable sensors for quantifying affective states and depression with neural networks," *IEEE Sensors Journal*, vol. 23, no. 19, pp. 22 788–22 802, 2023.

[77] G. J. Martinez, T. Grover, S. M. Mattingly, G. Mark, S. D'Mello, T. Aledavood, F. Akbar, P. Robles-Granda, and A. Striegel, "Alignment between heart rate variability from fitness trackers and perceived stress: Perspectives from a large-scale in situ longitudinal study of information workers," *JMIR Hum Factors*, vol. 9, no. 3, p. e33754, 2022. [Online]. Available: https://humanfactors.jmir.org/2022/3/e33754https://doi.org/10.2196/33754http://www.ncbi.nlm.nih.gov/pubmed/35925662

[78] S. Solhjoo, M. C. Haigney, E. McBee, J. J. G. van Merrienboer, L. Schuwirth, A. R. Artino, A. Battista, T. A. Ratcliffe, H. D. Lee, and S. J. Durning, "Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load," *Scientific Reports*, vol. 9, no. 1, p. 14668, 2019. [Online]. Available: https://doi.org/10.1038/s41598-019-50280-3

[79] A. Siennicka, D. S. Quintana, P. Fedurek, A. Wijata, B. Paleczny, B. Ponikowska, and D. P. Danel, "Resting heart rate variability, attention and attention maintenance in young adults," *International Journal of Psychophysiology*, vol. 143, pp. 126–131, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167876019301898

[80] J. A. Suykens, "Nonlinear modelling and support vector machines," in *IMTC 2001. proceedings of the 18th IEEE instrumentation and measurement technology conference. Rediscovering measurement in the age of informatics (Cat. No. 01CH 37188)*, vol. 1. IEEE, 2001, Conference Proceedings, pp. 287–294.

[81] R. Udhayakumar, S. Rahman, D. Buxi, V. G. Macefield, T. Dawood, N. Mellor, and C. Karmakar, "Measurement of stress-induced sympathetic nervous activity using multi-wavelength ppg," *Royal Society Open Science*, vol. 10, no. 8, p. 221382, 2023.

[82] E. M. Arias, J. Parraga-Alava, and D. Z. Montenegro, "Stress detection among higher education students: A comprehensive systematic review of machine learning approaches," in *2024 Tenth International Conference on eDemocracy & eGovernment (ICEDEG)*, 2024, Conference Proceedings, pp. 1–8.

[83] M. Stojchevska, B. Steenwinckel, J. Van Der Donckt, M. De Brouwer, A. Goris, F. De Turck, S. Van Hoecke, and F. Ongenae, "Assessing the added value of context during stress detection from wearable data," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 268, 2022. [Online]. Available: https://doi.org/10.1186/s12911-022-02010-5

[84] L. K. Hill, D. D. Hu, J. Koenig, I. Sollers, John J., G. Kapuku, X. Wang, H. Snieder, and J. F. Thayer, "Ethnic differences in resting heart rate variability: A systematic review and meta-analysis," *Biopsychosocial Science and Medicine*, vol. 77, no. 1, 2015. [Online]. Available: https://journals.lww.com/bsam/fulltext/2015/01000/ethnic_differences_in_resting_heart_rate.4.aspx

[85] K. Jensen-Urstad, N. Storck, F. Bouvier, M. Ericson, L. Lindbland, and M. Jensen-Urstad, "Heart rate variability in healthy subjects is related to age and gender," *Acta Physiologica Scandinavica*, vol. 160, no. 3, pp. 235–241, 1997.

[86] S. Aeschbacher, S. Tobias, D. Laura, K. Rahel, N. Charlotte, S.-T. Arno, P.-H. N. M., R. Martin, R. Lorenz, , and D. Conen, "Heart rate, heart rate variability and inflammatory biomarkers among young and healthy adults," *Annals of Medicine*, vol. 49, no. 1, pp. 32–41, 2017, doi: 10.1080/07853890.2016.1226512. [Online]. Available: https://doi.org/10.1080/07853890.2016.1226512

[87] S.-C. Fang, Y.-L. Wu, and P.-S. Tsai, "Heart rate variability and risk of all-cause death and cardiovascular events in patients with cardiovascular disease: a meta-analysis of cohort studies," *Biological research for nursing*, vol. 22, no. 1, pp. 45–56, 2020.

[88] T. Benichou, B. Pereira, M. Mermillod, I. Tauveron, D. Pfabigan, S. Maqdasy, and F. Dutheil, "Heart rate variability in type 2 diabetes mellitus: A systematic review and meta–analysis," *PloS one*, vol. 13, no. 4, p. e0195166, 2018.

[89] S. F. M. Pizzoli, C. Marzorati, D. Gatti, D. Monzani, K. Mazzocco, and G. Pravettoni, "A meta-analysis on heart rate variability biofeedback and depressive symptoms," *Scientific Reports*, vol. 11, no. 1, p. 6650, 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-86149-7

[90] S. Karvekar, M. Abdollahi, and E. Rashedi, "A data-driven model to identify fatigue level based on the motion data from a smartphone," *BioRxiv*, p. 796854, 2019.

[91] S. K. D'Mello and B. M. Booth, "Affect detection from wearables in the "real" wild: Fact, fantasy, or somewhere in between?" *IEEE Intelligent Systems*, vol. 38, no. 1, pp. 76–84, 2023.

[92] S. Greene, H. Thapliyal, and A. Caban-Holt, "A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health," *IEEE Consumer Electronics Magazine*, vol. 5, no. 4, pp. 44–56, 2016.

## VII. Biography

**Justin Laiti** (Member, IEEE) received the B.S. degree in biomedical engineering from Virginia Tech, Blacksburg, VA, USA, in 2022. He was awarded a U.S. Fulbright grant to pursue a Ph.D. as part of the StAR PhD Program at the Royal College of Surgeons in Ireland (RCSI), where he is currently in his third year. His research focuses on the development of wearable-based tools for stress monitoring and well-being support in educational settings. In 2024, he received a Postgraduate Secondment Award to spend three months at the University of Oxford, working with Dr. Tingting Zhu on machine learning approaches for affect detection from photoplethysmography (PPG) data.

**Dr. Yu Liu** is a Postdoctoral Researcher in the Department of Engineering Science at the University of Oxford. He received his B.E. and Ph.D. (with honor) from Tsinghua University, China. His current research focuses on AI for Social Good, with particular emphasis on healthcare applications and urban computing.

**Dr. Elaine Byrne** received her PhD in Public Health from the University of the Western Cape, Cape Town, South Africa. She is currently a Senior Lecturer and Programme Director at the Centre for Positive Health Sciences, RCSI. Her research focuses on promoting health and well-being at individual, societal, and organisational levels, with an emphasis on inclusive and collaborative research methodologies. She has extensive experience in evaluating health interventions and developing strategies to enhance research accessibility. In addition to her academic work, Dr. Byrne serves on the board of directors of Beyond Stigma, an Irish charity dedicated to addressing internal stigma and fostering self-worth.

**Dr. Pádraic Dunne** received his PhD in Immunology from University College London, London, United Kingdom. He is a Senior Lecturer at the RCSI Centre for Positive Health Sciences, where he leads the Digital Health Research Group. His research focuses on the development of health coach-led health and well-being programmes for Irish communities, secondary school students, healthcare professionals, and individuals with chronic disease. Dr. Dunne is Co-Director of the Centre’s MSc in Positive Health Coaching and Director of the micro-credential course ”An Introduction to Lifestyle Medicine.”

**Dr. Tingting Zhu** graduated with the DPhil degree in information and biomedical engineering at Oxford University in 2016. This followed her MSc in Biomedical Engineering at University College London and BEng (Hons) in Electrical Engineering from the University of Malta. After DPhil, Tingting was awarded a Stipendiary Junior Research Fellowship at St. Hilda’s College, Oxford. In 2018, Tingting was appointed as the first Associate Member of Faculty at the Department of Engineering Science; in 2019, following the award of her Royal Academy of Engineering Research Fellowship, she was appointed to full Member of Faculty at the Department of Engineering Science.