over time can be conducted to assess the collective effects of a specific lesson.

4.5.2 Heart Rate

Heart Rate Signal Classification

One of the challenges in estimating the heart rate for very large PPG data sets is the extraction of valid heart rate signal regions within the greater PPG measurement recorded over multiple hours. The PPG can be highly affected by motion artifacts and noise that render the detection and estimation of the heart rate quite difficult to make. Due to the number of PPG readings taken over the course of a day [1], methods such as manual inspection and classification of signal regions by a researcher are extremely time consuming and error-prone, and reduce the automation quality of the system.

Instead, a classification algorithm was designed in order to automatically detect regions in the PPG that contain high quality heart rate components (i.e. high signal-to-noise ratio) for further signal processing and heart rate estimation. A portion of the overall PPG data collected in the school trial was used as a training and evaluation set for the algorithm.

Regions in the PPG signal training set were manually classified using binary labels indicating the presence or absence of heart rate content and their associated beginning and end timestamps. The regions were fed into a feature-extraction algorithm that applies a hampel filter to eliminate outlying readings, and extracts a limited number of discriminating features from segments in sequential 10 second windows of the filtered regions. The labels were then applied to the corresponding feature vectors.

One of the ways to distinguish heart rate signal from motion artifacts and noise in the PPG is by estimating the power spectral density (PSD) of the signal. The PSD represents the power, or magnitude, of the signal as a function of the frequency. Figure 4-6 shows the plots of the PSDs of a 35 different PPG segments with distinctly high or low heart rate signal-to-noise ratios, shown in blue and red respectively. The

[1] As described in subsection 2.1.2, the PPG reading frequency is 100 Hz, resulting in over 2 million readings recorded in each device over a single school day.

noisy segments demonstrate a greater magnitude of high frequency content than the heart rate signals, indicative of higher levels of noise, likely due to a lack of good contact of the PPG sensor and skin surface or ambient noise. Similarly, low frequency content (<1.5 Hz) was observed to be higher in these red regions, possibly because of large-scale device motion. The mean magnitude of the signal in the high frequency components (20-200Hz range) was chosen as a feature due to its relative success in separating the labeled segments, as can be visually observed in Figure 4-6.

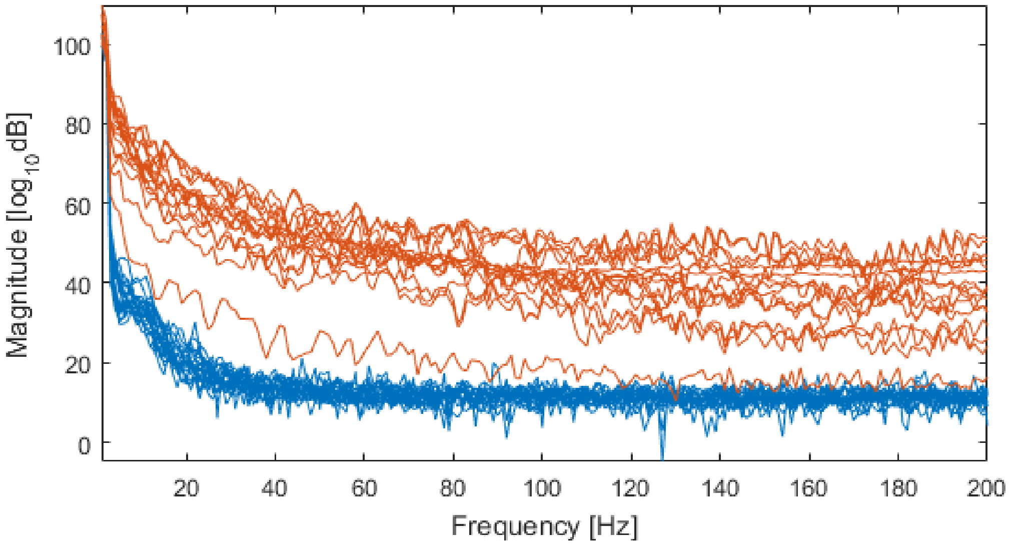

Figure 4-6: PPG signals' power spectral densities (PSD).

A second feature used is the slope of the power spectral density curves between the low and high frequency component ranges (LF/HF). A distinction of the heart rate signal is its relatively sharp transition, and high ratio, between magnitudes of signal components in the expected heart rate frequency range (1-2 Hz) and those in higher frequencies (>3 Hz), as opposed to signal segments showing a high degree of randomness due to motion artifacts and noise. Figure 4-7 demonstrates this quality in the slope of the PSD curves between 2 and 3 Hz.

A third feature implemented considers the assumption of regularity of local maxima in a heart rate signal. Prominent peaks are indicative of heart beats, and the distance between them is near-constant in short signal windows under normal physiological conditions. Thus, the regularity, or variance, of inter-peak distance can be

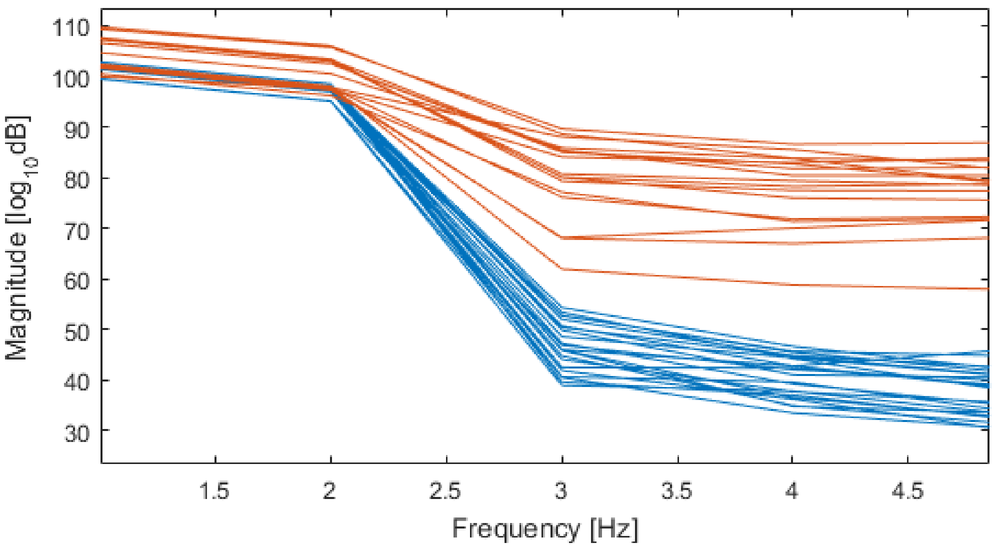

Figure 4-7: LF/HF slope in the power spectral densities plot.

used as a discriminating feature between high and low-HR content signal segments. To find the variance, a bandpass filter is applied and the peaks are extracted from the filtered signal, as later described in the heart rate estimation algorithm. The variance of the locations of the peaks can be calculated with the Matlab 'var' function. As per the initial assumption, low heart rate-content segments showed higher variance than those with high content, where the variance was often lower due to the regularity of the signal. It should be noted that the quality of this feature depends on the accuracy of the peak analysis, which can vary depending on the algorithm.

A training feature set of 882 entries, each with three features, was fed into Matlab's Classification Learner application, used to run the data against a large number of supervised classification algorithms. Table 4.1 lists the algorithms and statistics that obtained the best classification results with a labeled test set, showing a high correlation between distinctive features the heart rate presence/absence labels. The parallel coordinates plot in Figure 4-8 visualizes the results of the correlation between the three features used over the test set in the Ensemble algorithm.

Algorithm	**Accuracy (%)**	**True Positives (%)**	**True Negatives (%)**
Logistic Regression	94.1	90	95
Linear Support Vector Machine (SVM)	92.2	81	96
Simple Decision Tree	96.6	94	98
K-Nearest Neighbors (KNN)	92.9	85	96
Ensemble (Bagged Decision Trees)	97.1	93	99

Table 4.1: PPG algorithm classification accuracy results.

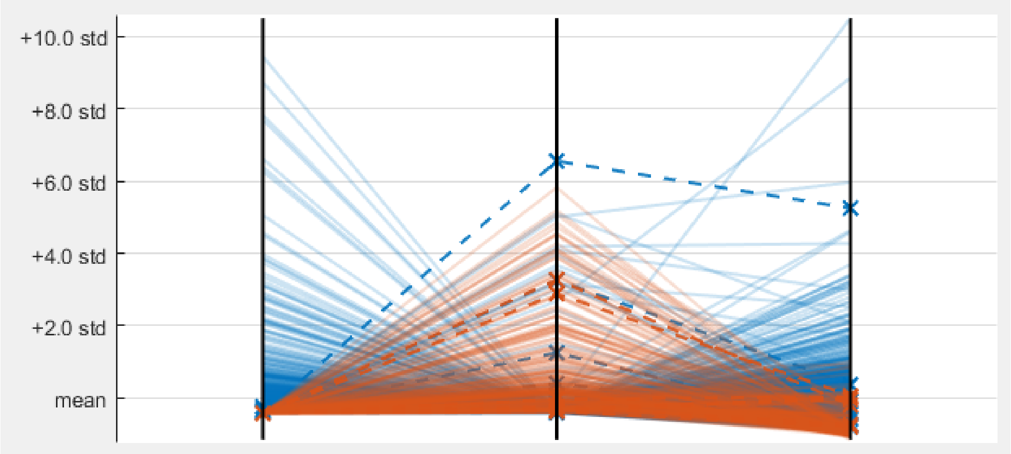

Figure 4-8: Parallel coordinates plot visualizing signal feature correlation. Left bar: mean magnitude of high frequency components, center bar: LF/HF magnitude ratio, right bar: peak distance variance. Correct predictions are marked with continuous connecting lines and incorrect ones are marked with dashed lines.

Heart Rate Estimation

Heart rate signal observed in the validated raw PPG measurement demonstrated low frequency components (<1Hz) due to external motion artifacts. The PPG was digitally applied a second-order high pass Butterworth filter with a corner frequency of 1Hz, the lowest expected frequency of the heart rate content (Figure 4-10a). A