Applied Intelligence
https://doi.org/10.1007/s10489-021-02458-4

An experimental study of objective pain measurement using pupillary response based on genetic algorithm and artificial neural network

Li Wang[1] · Yikang Guo[1] · Biren Dalip[1] · Yan Xiao[2] · Richard D. Urman[3] · Yingzi Lin[1]

Accepted: 20 April 2021

Abstract
Obtaining an objective measurement of the pain level of a patient has always been challenging for health care providers. The most common method of pain assessment in the hospital setting is asking the patients' verbal ratings, which is considered to be a subjective approach. In order to get an objective pain level of a patient, we propose measuring pain level objectively using the pupillary response and machine learning algorithms. Thirty-two healthy subjects were enrolled in this study at Northeastern University. A painful stimulus was applied to healthy subjects by asking them to place their hands inside a bucket filled with iced water. We extracted 11 features from the pupil diameter data. To get the optimal subset of the features, a genetic algorithm (GA) was used to select features for the artificial neural network (ANN) classifier. Before feature selection, the f1-score of ANN was $54.0 \pm 0.25\%$ with all 11 features. After feature selection, ANN had the best performance with an accuracy of 81.0% using the selected feature subset, namely the Mean, the Root Mean Square (RMS), and the Pupillary Area Under Curve (PAUC). The experimental results suggested that pupillary response together with machine learning algorithms could be a promising method of objective pain level assessment. The outcomes of this study could improve patients' experience of pain measurement in telehealthcare, especially during a pandemic when most people had to stay at home.

Keywords Objective pain measurement · Pupillary response · Genetic algorithm · Artificial neural network · Machine learning algorithm

1 Introduction

Pain is an unpleasant and harmful feeling for human beings. It can be divided into chronic pain (long-lasting) and acute pain (short-lasting) [1]. Knowing the pain level of patients is crucial. Thus, a patient's pain level is the fundamental information observed by the medical staff before providing medical treatment. Most hospitals in the United States obtain a patient's pain level through a survey based on patients' perception of their pain. Several common surveys are the numeric rating scale (NRS), the visual analogue scale (VAS), the verbal rating scale (VRS), among others [2–4]. Obviously, the pain level derived from a survey is subjective. There has been significant research showing the drawbacks of these pain measurement methods. Ekblom and Hansson pointed out that VRS had a low correlation with other pain measurement methods and it might be less sensitive [5]. Bergh et al. concluded that the probability of accomplishing pain rating decreased for patients with higher age, especially with VAS [6]. NRS and VAS were compared in Carpenter's study, demonstrating that more than three quarters of pain ratings derived from NRS and VAS were not equivalent. It concluded

This work has been financially supported by a collaborative National Science Foundation project entitled "Novel Computational Methods for Continuous Objective Multimodal Pain Assessment Sensing System (COMPASS)" under the award #1838796, 1838650 and 1838621.

✉ Yingzi Lin
yi.lin@northeastern.edu

[1] Intelligent Human-Machine Systems Laboratory, Northeastern University, Boston, MA, USA

[2] College of Nursing and Health Innovation, University of Texas at Arlington, Arlington, TX, USA

[3] Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Published online: 17 May 2021

that scale ratings varied considerably for both patients' pain level reports and the nurses' usage of medications depending on the pain scale used [7]. Another disadvantage of taking pain measurements via surveys is that patients need to be conscious, cognitively intact, and able to speak to report their pain levels. It's impossible to get the pain levels of unconscious patients and new-born babies by using the survey methods. There have been studies trying to use physiological signals and machine learning algorithms to assess pain states [8–10]. To overcome these challenges, we propose to use the pupillary response to predict a patient's pain level based on machine learning algorithms.

1.1 Pupillary response

As an effective physiological signal, the pupillary response has been used extensively among a variety of fields [11–13]. Pain measurement by pupillary response has been researched actively for decades since the 1950s [14–16]. Rubin et al. [14] conducted a study on children with abdominal pain in 1967. In their study, pupil dilation in darkness was measured from both healthy children and children with abdominal pain. This study found that there was a significant reduction of pupillary dilation in children with abdominal pain after being exposed to stress. In the study of Constant et al. [15], pupillary diameter (PD) was recorded from children undergoing sevoflurane anesthesia. PD increased significantly in all children after noxious stimulation. The study concluded that, for noxious stimulation, PD was more sensitive than the commonly used factors, such as heart rate, arterial blood pressure, etc. In 2012, Aissou et al. [16] measured pupillary dilatation reflex (PDR) from postoperative patients. The results showed that PDR was highly correlated to patients' pain levels. PDR was proved to be valuable for guiding morphine administration for patients in the immediate postoperative period. A few commonly used pupillary features are PD, PDR, variation coefficient of the pupillary diameter (VCPD), pupillary light reflex (PLR), etc [17, 18]. Table 1 shows some studies using pupillary features to assess the pain level of different kinds of patients.

1.2 Machine learning algorithms

The genetic algorithm (GA) was first developed by John Holland in the 1960s [19]. After GA was introduced, numerous studies have utilized it as a feature selection tool when there were quantities of features to analyze. Recently, Nakisa et al. [20] extracted 30 optimal features out of 1440 features from 32 channels of brain wave signals. Goswami et al. [21] achieved an improvement of 12% in classification performance by using GA. Among all of these studies, a few of them were about pain measurement [22, 23]. Brahnam et al. [22] conducted research to diagnose neonatal pain from neonatal facial images. In their study, GA was employed to search for a parsimonious network, which was shown to be successful in finding the optimal solutions for the neural network. In a childhood abdominal pain estimation study [23], GA was used to prune the artificial neural network (ANN) architecture and minimize the number of diagnostic factors. GA is shown to be a promising tool for feature selection and improving classification performance. ANN [22–24] has been actively used as a machine learning classifier for pain measurement. The ANN had a good performance as a pain measurement classifier. To our knowledge, no research team has used pupillary response coupled with GA and ANN to measure the pain level.

1.3 Scope

The motivation of this work was to overcome the challenges of survey-based pain measurement methods. This study aimed to find a way to measure the pain level objectively through pupillary response and machine learning algorithms. Towards this objective, we first filtered the noisy pupillary response data using pupil velocity. Eleven features were extracted from the cleaned pupillary diameter data. ANN was utilized to classify the pupillary features extracted by GA. The main contribution of this study was integrating machine learning techniques to achieve an objective pain level assessment. The remaining parts of this paper are organized as follows: Section 2 represents the methodology of this study. Section 3 shows

Table 1 Studies using pupillary features to predict the pain level of different types of patients

Year	Subject population	Pain trigger	Pupillary feature	Pain assessment method
2019 [18]	345 postoperative patients	Postoperative pain	PLR, VCPD, PD	Statistical analysis
2017 [17]	40 patients during labor	Obstetrical labor pain	VCPD	Statistical analysis
2012 [16]	100 postoperative patients (42 males, 58 females)	Postoperative pain	PDR	Statistical analysis
2006 [15]	24 children	Standardized skin incision	PD	Statistical analysis
1967 [14]	25 children	Cold water	PD	Statistical analysis

the experimental results in the present study. In Section 4, we discuss the findings and results of this paper. Section 5 concludes the paper and states the future work for this topic.

2 Methodology

2.1 Participants

Thirty-two subjects (6 females and 26 males), aged from 18 to 24 (mean = 21.25, SD = 1.64) took part in this experiment in the Intelligent Human-Machine System(IHMS) laboratory at Northeastern University. Among all the subjects, 87.5% (n=28) reported they were Caucasian, 6.25% (n=2) Hispanic/Latino, 3.125% (n=1) Asian/ Pacific Islander, 3.125% (n=1) African American. All subjects were healthy, fluent in English, and none of them had any experience of chronic pain. This study was approved by the Northeastern University Institutional Review Board (IRB #: 17-01-25). All subjects read and signed the consent form before the experiment. The subjects were informed that they could stop this experiment at any time.

2.2 Apparatus

Tobii Pro Glasses 2 (Tobii Technology, Danderyd, Sweden) was used to measure subjects' pupillary response at a sampling rate of 50 Hz. Robust pupil diameter measurement was ensured by Tobii Glasses' four embedded infrared cameras and a unique 3D eye model. The Tobii Pro Glasses 2 data recording system is shown in Fig. 1. A bucket filled with iced water (temperature $\approx 0\,^{\circ}C$) was used to trigger pain from the cold temperature. A Dell monitor was used to display guidelines during the experiment.

Fig. 1 Tobii Pro Glasses 2. (Glasses and data recorder)

2.3 Experimental design and procedures

The experimental design of this study was based on the cold pressor test [25–27], a widely used protocol for quantitative sensory testing (QST). The subjects were asked to read and sign the consent form upon their arrival at the experiment site. Prior to experimental data collection, the subjects were asked to place their hands inside the iced water bucket for 5 to 10 seconds to get familiar with the stimulation of cold pain. Data collection was conducted in the following steps: *Step 1,* the subject sat in a comfortable armchair and wore the Tobii Pro Glasses 2. Tobii Glasses calibration was completed. The environmental light condition during the experiment was controlled at a constant level. *Step 2,* to minimize the eye movements, the subject was asked to look at a green dot on a monitor in front of him/her at a distance of about 30 cm. *Step 3,* 20 seconds of baseline data were collected. *Step 4,* the subject was asked to put his/her right hand in a bucket filled with iced water. The subject was asked to report his/her pain level on a scale from 0 (No pain) to 10 (Excruciating pain) every 20 seconds, according to the NRS. The reported pain level was recorded. *Step 5,* for each subject, the experiment ended after 10 pain ratings were obtained or when the subject asked to stop the experiment. Fig. 2 shows the experimental setting of the present study.

2.4 Data description

The pupillary diameter data of both eyes were recorded in Tobii Pro Glasses 2 with a sampling rate of 50 Hz.

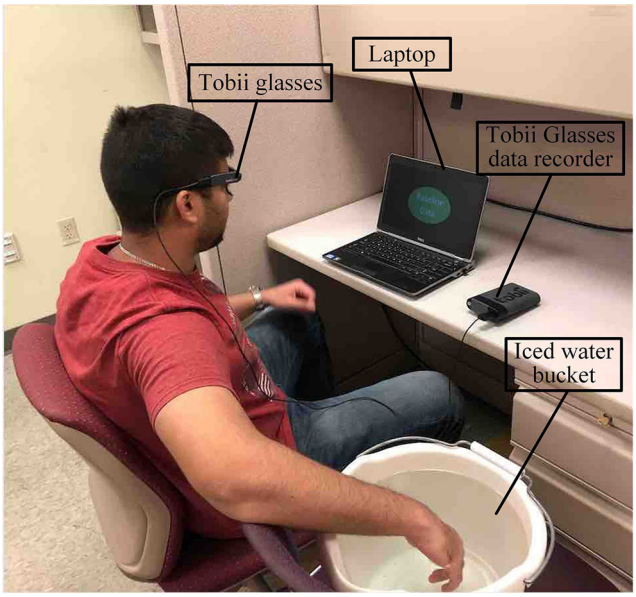

Fig. 2 Schematic diagram of the experimental settings (a subject wearing the Tobii Glasses, sitting in front of a laptop and dipping his hand into iced water)

The raw data were saved in .csv files, containing two columns (left pupil diameter and right pupil diameter). The pupillary data were segmented into 10-second epochs. There were 660 data samples/epochs in total since some subjects withdrew from the experiment before 10 pain levels were recorded. The data samples were labeled based on their corresponding reported pain ratings in the following way: Level 0 (No-pain/Baseline), Level 1-5 (Low-pain), and Level 6-10 (High-pain). The 660 pupil data samples were randomly assigned to training data (80%) and test data (20%).

2.5 Data preprocessing

After visual inspection, it was found that a small proportion of the raw data had missing values and they were contaminated by the subject's eye blinks during the experiment, as shown in Fig. 3a. The artifacts in the pupillary data were removed by using the pupil diameter velocity method.

Pupil diameter signal was designated by $x(i)$ and corresponding time stamp was $t(i)$, where $i = 1, 2, 3..., N$ and N was the data length. Pupil diameter velocity (mm/s) meant the changing rate of pupil diameter [28]. Absolute pupil diameter velocity, $x'(i)$, was used to detect blinking noise:

$$x'(i) = \left| \frac{x(i) - x(i-1)}{t(i) - t(i-1)} \right| \tag{1}$$

Absolute pupil diameter velocity is shown in Fig. 3b. Absolute pupil velocity threshold was set at 2 mm/s, which means the pupil diameter data with a corresponding velocity greater than 2 mm/s were considered as blinking noise and were removed from the raw data. The detected blinking noise was marked in red in Fig. 3c. After blinking noise was removed, an interpolation package in SciPy [29] was exploited to fill the missing data points in the raw pupil diameter data. The filtered pupil diameter is shown in Fig. 3d. Since normal pupil diameter varies among people in different ages and genders [30], the pupil diameter data were scaled to a 0-1 range:

$$x_s(i) = \frac{x(i) - min(\mathbf{X})}{max(\mathbf{X}) - min(\mathbf{X})} \tag{2}$$

Studies [31, 32] suggested that the left and right pupil diameters were highly correlated. The mean of the filtered left and right pupillary data was utilized in the following computation.

2.6 Pupillary features

Eleven features were extracted from each of the 660 pupillary data epochs, as follows:

(I) the maximum value of pupil diameter data:

$$Max = max\left(\mathbf{X}\right) \tag{3}$$

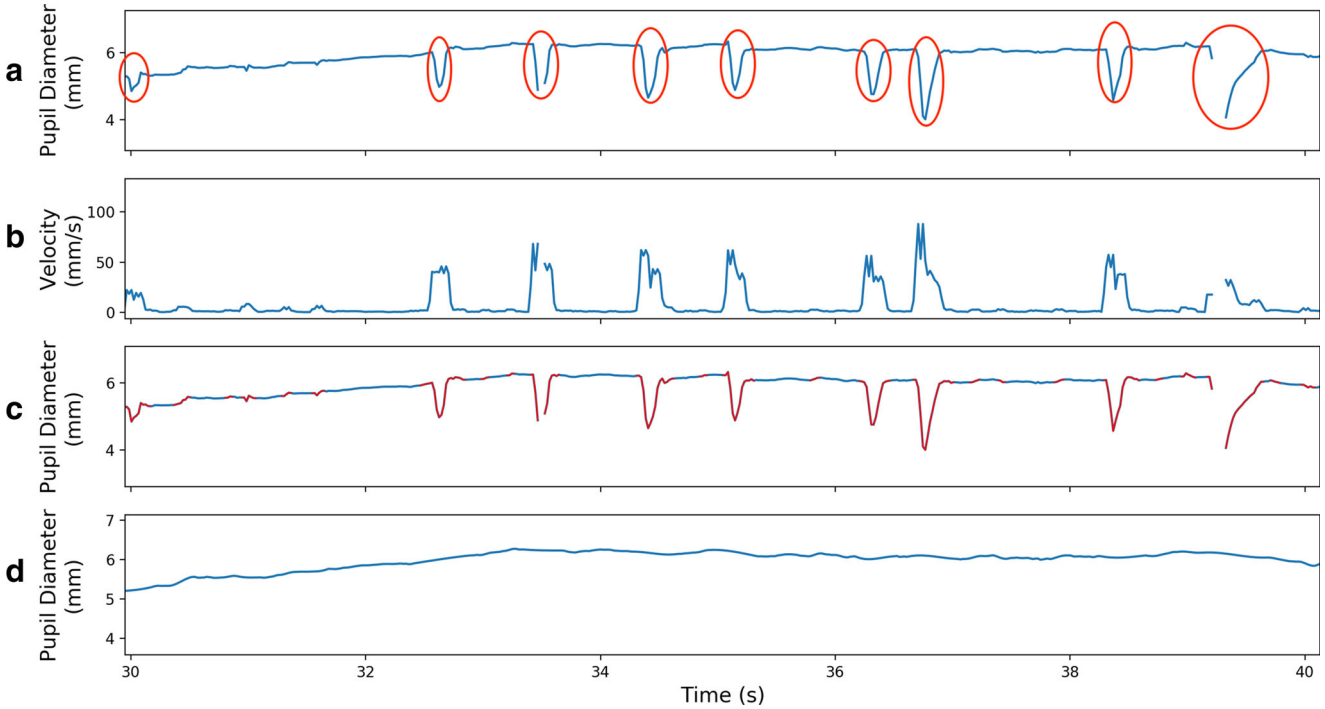

Fig. 3 Filtering process for pupil diameter data. **a** Raw pupil diameter data with blinking noise, marked in red circles. **b** Pupil diameter velocity. **c** Raw pupil diameter data with detected blinking noise, highlighted in red. **d** Pupil diameter data filtered by a velocity filter

(II) the minimum value of pupil diameter data:

$$Min = min\left(\mathbf{X}\right) \tag{4}$$

(III) the range of pupil diameter data:

$$Range = Max - Min \tag{5}$$

(IV) the mean value of pupil diameter data:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x(i) \tag{6}$$

(V) the standard deviation (SD) of pupil diameter data:

$$\sigma = \left(\frac{1}{N-1}\sum_{i=1}^{N}\left(x(i)-u\right)^2\right)^{1/2} \tag{7}$$

(VI) the interquartile range (IQR) value of pupil diameter data (Q_3 and Q_1 mean the third and first quartile respectively):

$$IQR = Q_3\left(\mathbf{X}\right) - Q_1\left(\mathbf{X}\right) \tag{8}$$

(VII) the root mean square (RMS) of pupil diameter data:

$$RMS = \left(\frac{1}{N}\sum_{i=1}^{N} x(i)^2\right)^{1/2} \tag{9}$$

(VIII) the skewness of pupil diameter data:

$$Skewness = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x(i)-\mu}{\sigma}\right)^3 \tag{10}$$

(IX) the kurtosis of pupil diameter data:

$$Kurtosis = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x(i)-\mu}{\sigma}\right)^4 \tag{11}$$

(X) the variation coefficient of the pupillary diameter (VCPD):

$$VCPD = \frac{\frac{1}{N}\sum_{i=1}^{N}|x(i)-median|}{median} \tag{12}$$

(XI) the pupillary area under the curve (PAUC):
PAUC was calculated by transforming the pupillary data from time-domain to frequency-domain using the multitaper method [33]. As suggested in [34], the Area Under Curve (AUC) between 0.3 Hz and 3 Hz was effective to detect the pain states on patients. PAUC is the AUC of the pupillary spectrum between 0.3 Hz and 3 Hz.

$$PAUC = \int_{0.3}^{3} multitaper\left(x(t)\right)d\omega \tag{13}$$

where *multitaper* was the power spectral density estimation method and ω represented the frequency domain of the transformed data.

2.7 Machine learning modeling

Our pain assessment system consisted of three major parts, namely data acquisition, feature extraction, and GA & Grid search. The overview of the system framework is shown in Fig. 4. The data acquisition part was conducted by using a wearable eye-tracking device with a sampling rate of 50 Hz. The feature extraction part was achieved by extracting 11 features (Section 2.6) from the pre-processed pupil diameter data. The GA & Grid search part was accomplished by integrating GA and grid search to search for the optimal feature set and hyperparameters. The first step was to initialize a set of random features for the first iteration of GA. Next, GA evaluated the fitness score of the system using the ANN hyperparameters in Table 2. GA proceeded to the Selection step, the Crossover step, and the Mutation step. And the second iteration of GA started. The grid search of hyperparameters updated each time when GA finished one full cycle. The stop criteria were when all ANN hyperparameters were searched and the max generation number of GA was reached. When the stop criteria were satisfied, the system output the best classification accuracy with the optimal feature set and ANN hyperparameters.

2.7.1 Feature selection

As previously mentioned, eleven features (*Max*, *Min*, *Range*, *Mean*, *SD*, *IQR*, *RMS*, *Skewness*, *Kurtosis*, *VCPD*,

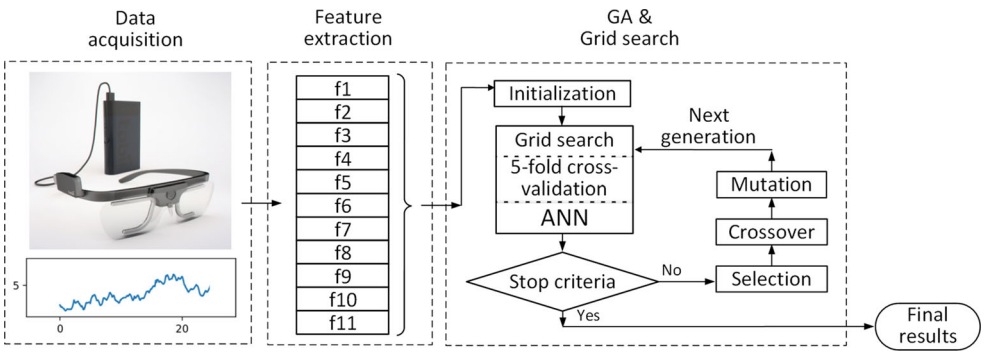

Fig. 4 The overview of the pain assessment system framework, including data acquisition, feature extraction, and GA & Grid search

Table 2 The settings for the ANN hyperparameter tuning, including activation function, alpha penalty, and tolerance

Hyperparameter	Settings
Activation function	'logistic', 'tanh', 'relu'
Alpha penalty	0.0001, 0.001, 0.01
Tolerance	0.0001, 0.001, 0.01

and *PAUC*) were extracted from the pre-processed raw data. To get the optimal combination of features for the machine algorithms, GA was used to select features. GA was inspired by Darwin's Theory of Evolution by Natural Selection [35]. GA mimicked the evolution theory by considering individuals as chromosomes. To use GA as a feature selection tool, a string of binary digits represented an individual (chromosome), in which 0 meant the feature at that digit was not selected and 1 meant the feature was selected. For example, in an eight-digit binary string, 00111011, the first, second, and sixth features were not selected, while the third, fourth, fifth, seventh, and eighth features were selected. The crossover and mutation processes of GA are shown in Fig. 5. The evaluation process of GA was performed by using ANN with grid search.

2.7.2 Hyperparameter tuning

ANN is a non-linear classifier that is inspired by the biological neural network of human beings [36]. ANN was used as the classifier for pain assessment in this study. A grid search method was utilized to tune the hyperparameters for ANN, as shown in *'GA & Grid search'* section of Fig. 4. Three kinds of hyperparameters (activation function, alpha penalty, and tolerance) were tuned in this study, as shown in Table 2.

2.7.3 Evaluation

Since our data set had an unbalanced number of samples in each class, a stratified 5-fold cross-validation was utilized to evaluate the classifier. In the stratified 5-fold cross-validation, each training or testing set contained approximately the same proportion of samples of each class as the entire data set (class 'B': 62 samples, class 'L': 236 samples, and class 'H': 362 samples). The stratified 5-fold cross-validation on our data set is shown in Fig. 6. The

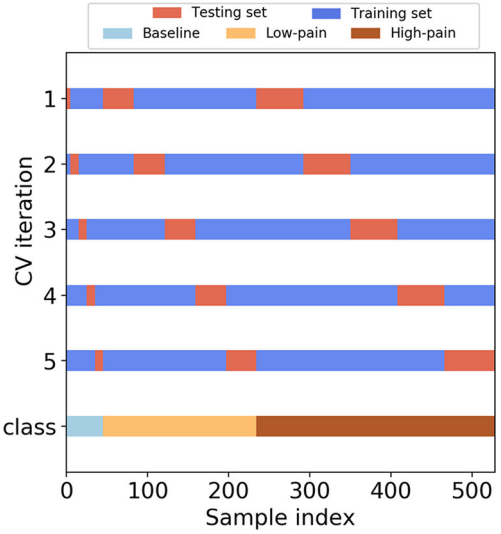

Fig. 6 A stratified 5-fold cross-validation. The testing set and the training set of each iteration are shown with different colors in the first five rows. The size of each class (Baseline, Low-pain, High-pain) is shown in the bottom row

performance of the classifier was measured by f1-score, as shown in equation (16).

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{14}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegtive} \tag{15}$$

$$F1 = \frac{2}{Precision^{-1} + Recall^{-1}} \tag{16}$$

3 Experimental results

3.1 Subjective pain level ratings

Iced water was used in numerous pain measurement studies [27, 37–39], and it proved to be an effective tool to simulate pain. In the present study, a bucket of iced water was used to introduce cold pain for the subjects. As mentioned before, the subjects could terminate the experiment whenever they wished. Thirty subjects out of thirty-two made it to the end of the experiment. Based on the survey, none of the subjects suffered from acute or chronic pain at the time of

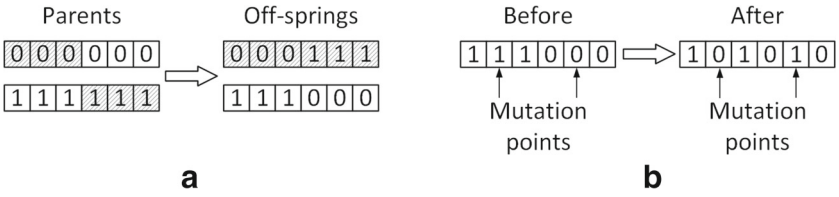

Fig. 5 Simple illustration of GA's crossover and mutation processes. **a** crossover **b** mutation

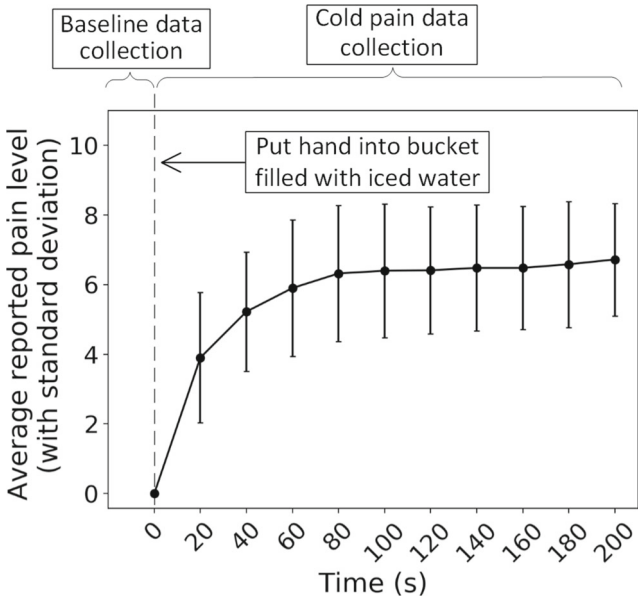

Fig. 7 Average reported pain ratings with standard deviation from 32 subjects. Subjects put their hands in the iced water at the time of 0 s

this experiment. The baseline pain level of all subjects was zero. The average reported pain levels of all subjects are shown in Fig. 7. The pain ratings rose very quickly in the first minute of the experiment and kept at a constant level after around the middle of the experiment.

3.2 Frequency domain analysis

Pupillary unrest is the fluctuation of the subjects' pupillary diameter. It has been utilized to detect patients' pain states in prior studies [34, 40]. We utilized the multitaper method to transform the pupillary responses from time-domain to frequency-domain. After the transformation, the width of the frequency bin was 0.1 Hz. The PAUC was obtained by calculating the AUC of the pupillary responses spectrum from 0.3 Hz to 3 Hz [34]. Figure 8a shows three samples of pupillary responses in three different pain states (Baseline, Low pain, and High pain). The data sample in the high pain state demonstrated more fierce oscillation than the data samples in low pain and baseline states. Figure 8b illustrates the power spectra of their corresponding pupillary responses. The power spectrum of the high pain data had higher values than the low pain and baseline data in the lower frequency range, which suggested that there was more pupillary unrest in the high pain state. A post-hoc analysis was carried out by using the Tukey's Honestly Significant Difference (HSD) test on the PAUC from all the subjects. Figure 8c shows the statistical results of the PAUC in three pain states from all participants. The PAUC differences were significant in Baseline vs. High pain ($p < 0.01$) and Low vs. High pain ($p < 0.01$).

3.3 Classification performance

Eleven pupillary features were extracted from the pupillary responses. GA was used to select an optimal subset of features for the ANN. Grid search was utilized to tune the hyperparameters of the ANN. The final parameter settings for GA and ANN are as follows:

- GA: crossover possibility (0.5), mutation possibility (0.4), and population size (20). GA was evaluated with a generation number of 50, 100, 200 respectively.
- ANN: activation function (relu), solver (lbfgs), alpha penalty parameter (0.0001), tolerance (0.0001). Two hidden layers were used in this Multi-layer Perceptron classifier.

Table 3 presents the performance of the ANN based on training data using the above parameters. The ANN was

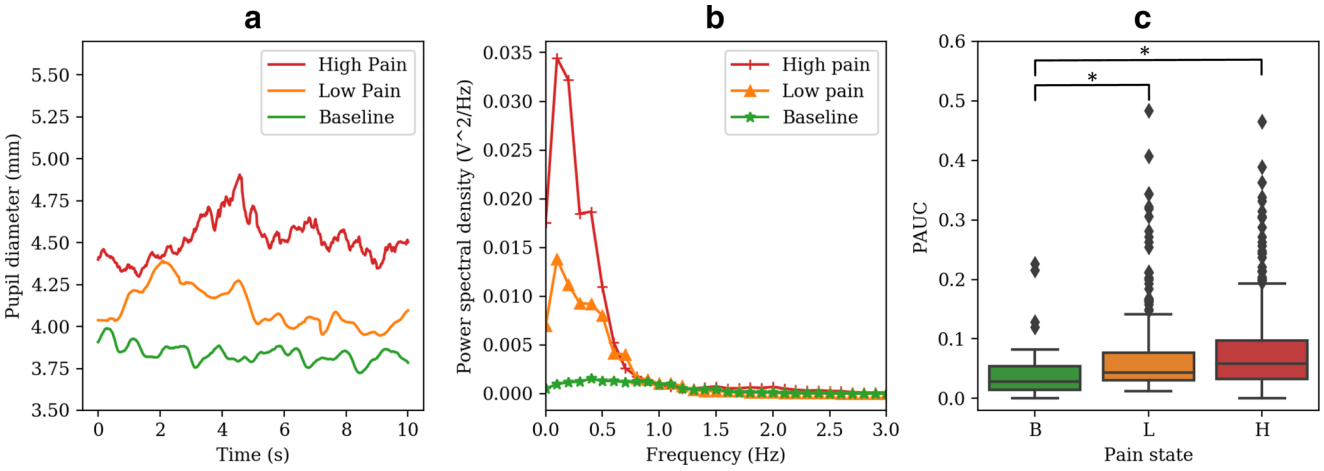

Fig. 8 **a** Pupillary responses in three pain states. **b** The power spectra of the pupillary response from 0 Hz to 3 Hz. **c** The boxplot of the PAUC in three pain states from all participants (* $p < 0.01$, One-way ANOVA with Tukey's HSD)

Table 3 Classification results on the training data. Features in "GA feature selection" column follow the order: Max, Min, Range, Mean, SD, IQR, RMS, Skewness, Kurtosis, VCPD, and PAUC (1: the feature is selected; 0: the feature is not selected)

Classifier	F1-score with all features (%)	No. of generation	GA feature selection	F1-score (%)
ANN	54.0 ± 0.25	50	1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1	75.6 ± 2.6
		100	0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1	82.57 ± 3.1
		200	0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1	82.57 ± 3.1

firstly evaluated with all 11 features using a stratified five-fold cross-validation. The f1-score with all features is shown in the second column of Table 3. With all 11 features, the ANN achieved an accuracy of $54.0 \pm 0.25\%$. After feature selection by GA, the performance of ANN climbed from $54.0 \pm 0.25\%$ to $82.57 \pm 3.1\%$. ANN's performance achieved the optimal rate with a generation number of 100. ANN's performance increased from $75.6 \pm 2.6\%$ to $82.57 \pm 3.1\%$ with the growing of generation number from 50 to 200. For ANN, the optimal feature set was Mean, RMS, and PAUC.

The final accuracy of the ANN using selected features on the test data set was 81.0%. Table 4 shows the precision, recall, and f1-score of ANN's performance on the test data set. The f1-scores of Baseline, Low pain, and High pain classes were 0.88, 0.65, and 0.87, respectively.

The Receiver Operating Characteristic (ROC) curves of three pain states using ANN classifier are shown in Fig. 9. The straight dashed diagonal line from $(0, 0)$ to $(1, 1)$ represents the ROC curve of a random-guessing classifier. The ROC curve of class B was the closest to the upper left corner, which meant the classifier had both good sensitivity and good specificity over class 'B'. The ROC curve of class 'H' was closer to the straight diagonal line, which meant the performance was worse on class 'H' than on class 'B'. The performance on class 'L' was worse than random guessing, as the ROC curve of class 'L' was under the straight diagonal line. Class 'B' had the highest AUC of 0.94. The AUCs of class 'L' and class 'H' were 0.40 and 0.75, respectively.

The confusion matrix of classification results of the ANN using selected features is shown in Fig. 10. The ANN classifier had a good performance for data labeled as "B" (baseline) and "H" (high-pain). Among class 'B', 79% were classified correctly and 21% were classified as class 'L'. Among class 'H', 95% were classified correctly and only 5% were classified as class 'L'. However, the ANN had a bad performance for "L" (low-pain) data. Among class 'L', only about half were classified correctly and 44% were classified as class 'H', which meant the classifier was not sensitive to the "L" data. It could also be seen in Fig. 9, the ROC curve of class 'L' was below the random-classifier ROC curve. The confusion matrix also suggested that the type 1 error (false positive) and type 2 error (false negative) for class 'B' were 0 and 3 respectively. The type 1 and type 2 errors for class 'L' were 7 and 18 respectively. The type 1 and type 2 errors for class 'H' were 18 and 4 respectively.

Table 4 Classification performance using selected features on the test data, showing precision, recall, f1-score, and number of samples in each class

	Precision	Recall	F1-score	Number of samples
Baseline	1.00	0.79	0.88	14
Low pain	0.77	0.56	0.65	41
High pain	0.8	0.95	0.87	77

4 Discussions

This study presented a method to predict the pain level of healthy subjects using pupillary responses based on machine learning algorithms. The subjects' pain levels were categorized into three classes (No-pain/baseline, Low-pain, High-pain). GA and ANN were utilized to classify the pupillary response data.

Iced water was widely used in cold pressor pain tests. In our experiment, the most common feedback from the subjects was that their hands became numb after a while when placed in the iced water bucket [41]. The reason that

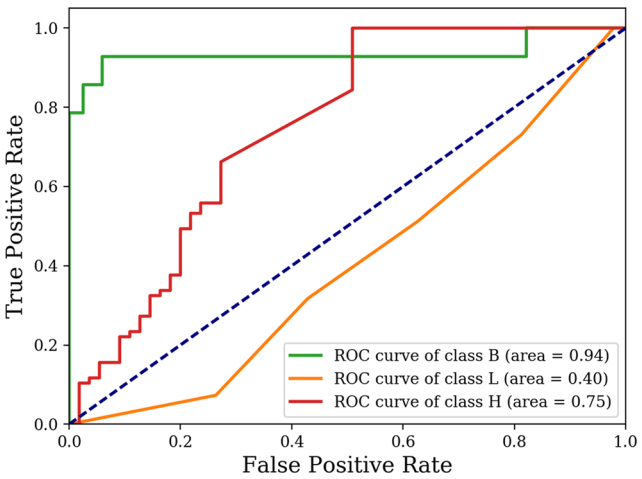

Fig. 9 The ROC curve for all three classes in ANN (the AUC of each class shown in the lower-right corner)

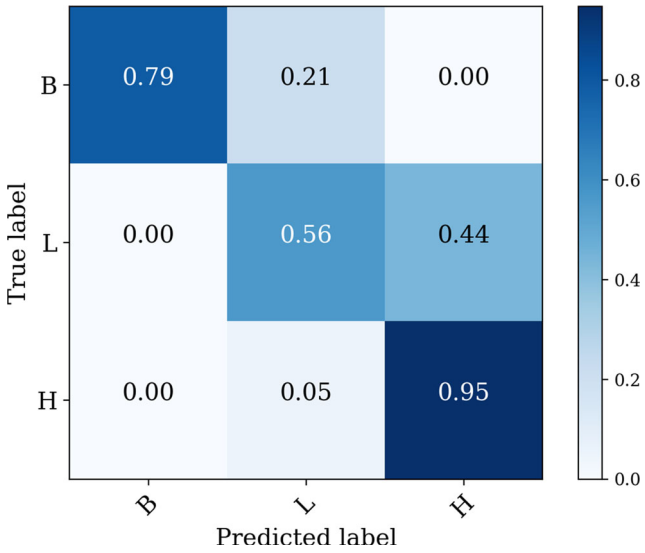

Fig. 10 Confusion matrix of classification results from ANN with selected features (Mean, RMS, and PAUC). (B: baseline, L: low-pain, H: high-pain)

subjects' hands became numb was probably that pain was introduced to subjects by using the mixture of ice and water, in which the temperature was technically 0 °C and 0 °C might be too cold for some of our subjects. In some other studies, a higher temperature of cold water was utilized to trigger pain for subjects. In a study by Dowman et al. the temperature of cold water was 4.3 ± 0.8 °C [42]. In future experiments, it's recommended to use cold water with a higher temperature to introduce pain.

Since we had a small number of data samples (660 samples in total), we used "lbfgs" as ANN's solver for better performance. We used two hidden layers for ANN and each hidden layer had 5 and 3 nodes respectively. ANN had an f1-score of 54.0% while using all 11 features. However, we found that ANN classified all data samples into one class (High). While using the selected features (Mean, RMS, and PAUC) by GA, ANN's performance improved from 54.0% to 82.57%. The problem that ANN classifies all instances into the same class was called "dying ReLU". During the training process of ANN, parts of the training data went to a hard zero zone of ReLU [43]. Using the features selected by GA, ANN's classification performance improved by 28.57% and the "dying ReLU" problem was solved.

Some pain assessment studies used machine learning techniques with other physiological signals, but without pupillary response [44–46]. A few other common physiological parameters for pain measurement are Electrocardiogram (ECG) [44], Electromyography (EMG) [44], Skin Conductance (SC) [44, 45], Functional Magnetic Resonance Imaging (fMRI) [46], etc. Table 5 compares different studies using bio-signals to predict pain states. It is worth noting that the compared studies performed two-class classification tasks, while our study did three-class classification tasks. It suggested that Random Forest (RF) achieved the best accuracy of 79% by classifying baseline and level 1 pain data using ECG, EMG, and SC [44]. In a study by Lopez and Picard [45], ECG and SC were utilized to predict heat pain using multi-task neural networks (NN). The best performance was 82.75% by classifying the baseline and the highest pain data in this study. Brown et al. used Support Vector Machine (SVM) and fMRI to classify pain triggered by thermal stimuli [46]. They obtained an accuracy of 84% from SVM on classifying painful and non-painful data. Some of the parameters need at least one electrode attached to the patients/subjects' skin, like ECG, SC, etc. Some of the parameters need overly complex devices to complete the signal collection, like fMRI. However, assessing pain levels through the pupil diameter has two advantages over the other signals. First, measuring pupil response can be non-invasive by using a camera. There is not any invasive contact between the data collection device and the subjects, which minimizes the uncomfortable feeling of the subjects. Second, it can be accomplished by using a simple and portal device (a camera), which allows pain level measurement to be taken in an easy and fast manner.

There have been some studies using pupillary response for pain assessment with statistical analysis methods instead of machine algorithms [17, 47, 48]. All of the aforementioned studies proved the correlation between pupillary response and the pain level of patients. However, one limitation of their methods is that it's impossible to build a real-time pain measurement system using statistical methods. Our work can be extended to build a real-time system for objective pain measurement.

Table 5 The comparison of similar studies using bio-signals to predict pain states

	Method	Signal	# of classes	Accuracy
This study	GA, ANN	Pupillary responses	3	81.0%
[44] (2019)	RF	ECG, EMG, and SC	2	87%
[45] (2017)	Multi-task NN	ECG and SC	2	82.75%
[46] (2011)	SVM	fMRI	2	84%

5 Conclusion and future work

In this study, we proposed an objective way to measure pain levels based on pupillary response using machine learning algorithms. Pupillary response data were collected from 32 subjects and preprocessed using a "pupil velocity" method. Eleven features were extracted from raw pupillary data. GA was used to select the optimal subset of features. We used ANN to perform data classification. ANN achieved the best performance using the three selected features (Mean, RMS, and PAUC) with an accuracy of 81.0%. GA improved the performance of the ANN and reduced the amount of data for ANN to deal with, proving itself to be a valuable tool for feature selection. As a non-invasive measurement, the pupillary response was implied to be an effective way for objective pain assessment.

A major limitation of this work was the limited number of data samples that we collected in this experiment. With more data samples, we could have a more balanced number of samples in each class. Although we received promising experimental results, there were still a few improvements that need to be accomplished in the future. For instance, the next step of this work should target real patients and collect clinical pain data. Also, the current study was based on a dataset we collected already. In the future, a real-time objective pain assessment system should be developed.

Acknowledgements This work has been financially supported by a National Science Foundation project entitled "Collaborative: Novel Computational Methods for Continuous Objective Multimodal Pain Assessment Sensing System (COMPASS)" under the awards #1838796, 1838650 and, 1838621. The opinions are those of the authors and do not necessarily reflect the official positions of the sponsor.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Loeser J. D. et al (2001) Bonica's management of pain, vol 3. Lippincott Williams & Wilkins, Philadelphia
2. Scott J., Huskisson E. (1976) Graphic representation of pain. Pain 2(2):175
3. Downie W., Leatham P., Rhind V., Wright V., Branco J., Anderson J. (1978) Studies with pain rating scales. Ann. Rheum. Dis. 37(4):378
4. Breivik H., Borchgrevink P., Allen S., Rosseland L., Romundstad L., Breivik Hals E., Kvarstein G., Stubhaug A. (2008) Assessment of pain. BJA:, British Journal of Anaesthesia 101(1):17
5. Ekblom A., Hansson P. (1988) Pain intensity measurements in patients with acute pain receiving afferent stimulation. J. Neurol. Neurosurg. Psychiatry 51(4):481
6. Bergh I., Sjöström B, Odén A, Steen B (2000) An application of pain rating scales in geriatric patients. Aging Clin. Exp. Res. 12(5):380
7. Carpenter J. S., Brockopp D. (1995) Comparison of patients' ratings and examination of nurses' responses to pain intensity rating scales. Cancer Nursing 18(4):292
8. Wang L., Xiao Y., Urman R. D., Lin Y. (2020) Cold pressor pain assessment based on EEG power spectrum. SN Applied Sciences 2(12):1
9. Lin Y., Wang L., Xiao Y., Urman R. D., Dutton R., Ramsay M. (2018) Objective pain measurement based on physiological signals. In: Proceedings of the international symposium on human factors and ergonomics in health care, vol 7. SAGE Publications Sage, Los Angeles, pp 240–247
10. Yu M., Sun Y., Zhu B., Zhu L., Lin Y., Tang X., Guo Y., Sun G., Dong M. (2020) Diverse frequency band-based convolutional neural networks for tonic cold pain assessment using EEG. Neurocomputing 378:270
11. Lin Y., Zhang W., Watson L. G. (2003) Using eye movement parameters for evaluating human–machine interface frameworks under normal control operation and fault detection situations. Int. J. Hum. Comput. Stud 59(6):837
12. Lin Y., Zhang W. J., Wu C., Yang G., Dy J. (2009) A fuzzy logics clustering approach to computing human attention allocation using eyegaze movement cue. Int. J. Hum. Comput. Stud 67(5):455
13. Cheng B., Zhang W., Lin Y., Feng R., Zhang X. (2012) Driver drowsiness detection based on multisource information. Human Factors and Ergonomics in Manufacturing & Service Industries 22(5):450
14. Rubin L. S., Barbero G. J., Sibinga M. S. (1967) Pupillary reactivity in children with recurrent abdominal pain. Psychosom. Med. 29(2):111
15. Constant I., Nghe M. C., Boudet L., Berniere J., Schrayer S., Seeman R., Murat I. (2006) Reflex pupillary dilatation in response to skin incision and alfentanil in children anaesthetized with sevoflurane: a more sensitive measure of noxious stimulation than the commonly used variables. BJA:, British Journal of Anaesthesia 96(5):614
16. Aissou M., Snauwaert A., Dupuis C., Atchabahian A., Aubrun F., Beaussier M. (2012) Objective assessment of the immediate postoperative analgesia using pupillary reflex measurement: a prospective and observational study. Anesthesiology: J Am Soc Anesthesiologists 116(5):1006
17. Charier D. J., Zantour D., Pichot V., Chouchou F., Barthelemy J. C. M., Roche F., Molliex S. B. (2017) Assessing pain using the variation coefficient of pupillary diameter. J. Pain 18(11):1346
18. Charier D., Vogler M. C., Zantour D., Pichot V., Martins-Baltar A., Courbon M., Roche F., Vassal F., Molliex S. (2019) Assessing pain in the postoperative period: Analgesia Nociception Index^{TM} vs pupillometry. British Journal of Anaesthesia
19. Holland J. H. (1992) Genetic algorithms. Sci. Am 267(1):66
20. Nakisa B., Rastgoo M. N., Tjondronegoro D., Chandran V. (2018) Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors. Expert Syst. Appl. 93:143
21. Goswami S., Chakrabarti A., Chakraborty B. (2018) An empirical study of feature selection for classification using genetic algorithm. Int J Adv Intell Paradigms 10(3):305
22. Brahnam S., Chuang C. F., Sexton R. S., Shih F. Y. (2007) Machine assessment of neonatal facial expressions of acute pain. Decis. Support. Syst. 43(4):1242
23. Mantzaris D., Anastassopoulos G., Adamopoulos A., Stephanakis I., Kambouri K., Gardikis S. (2007) Selective clinical estimation of childhood abdominal pain based on pruned artificial neural

networks. In: Proceedings of the 3rd WSEAS international conference on cellular and molecular biology, biophysics and bioengineering, pp 50–55
24. Sanders N. W., Mann III N. H. (2000) Automated scoring of patient pain drawings using artificial neural networks: efforts toward a low back pain triage application. Comput Biology Med 30(5):287
25. Wood D. L., Sheps S. G., Elveback L. R., Schirger A. (1984) Cold pressor test as a predictor of hypertension. Hypertension 6(3):301
26. Walsh N. E., Schoenfeld L., Ramamurthy S., Hoffman J. (1989) Normative model for cold pressor test. Am J Phys Med Rehabil 68(1):6
27. Littlewort G. C., Bartlett M. S., Lee K. (2007) Faces of pain: automated measurement of spontaneousallfacial expressions of genuine and posed pain. In: Proceedings of the 9th international conference on Multimodal interfaces. ACM, pp 15–21
28. Bergamin O., Schoetzau A., Sugimoto K., Zulauf M. (1998) The influence of iris color on the pupillary light reflex. Graefe's Archive for Clinical and Experimental Ophthalmology 236(8):567
29. Jones E., Oliphant T., Peterson P., et al. (2001) SciPy: Open source scientific tools for Python. http://www.scipy.org/. Online Accessed 2018-11-20
30. MacLachlan C., Howland H. C. (2002) Normal values and standard deviations for pupil diameter and interpupillary distance in subjects aged 1 month to 19 years. Ophthalmic Physiol Opt 22(3):175
31. Jackson I., Sirois S. (2009) Infant cognition: going full factorial with pupil dilation. Developmental Science 12(4):670
32. Kret M. E., Sjak-Shie E. E. (2019) Preprocessing pupil size data: Guidelines and code. Behav. Res. Methods 51(3):1336
33. Thomson D. J. (1982) Spectrum estimation and harmonic analysis. Proc. IEEE 70(9):1055
34. Neice A. E., Behrends M., Bokoch M. P., Seligman K. M., Conrad N. M., Larson M. D. (2017) Prediction of opioid analgesic efficacy by measurement of pupillary unrest. Anesthesia & Analgesia 124(3):915
35. Mitchell M (1998) An introduction to genetic algorithms. MIT Press, Cambridge
36. McCulloch W. S., Pitts W. (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5(4):115
37. Hilgard E. R. (1967) A quantitative study of pain and its reduction through hypnotic suggestion. Proc. Natl. Acad. Sci. 57(6): 1581
38. Nayak S., Shiflett S. C., Eshun S., Levine F. M. (2000) Culture and gender effects in pain beliefs and the prediction of pain tolerance. Cross-cultural Research 34(2):135
39. Ferreira-Valente M. A., Pais-Ribeiro J. L., Jensen M. P. (2011) Validity of four pain intensity rating scales. Pain® 152(10): 2399
40. Bokoch M. P., Behrends M., Neice A., Larson M. D. (2015) Fentanyl, an agonist at the mu opioid receptor, depresses pupillary unrest. Auton. Neurosci. 189:68
41. Kunkle E. C. (1949) Phasic pains induced by cold. J. Appl. Physiol. 1(12):811
42. Dowman R., Rissacher D., Schuckers S. (2008) EEG indices of tonic pain-related activity in the somatosensory cortices. Clin. Neurophysiol. 119(5):1201
43. Maas A. L., Hannun A. Y., Ng A. Y. (2013) Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml, vol 30, p 3
44. Naeini E. K., Shahhosseini S., Subramanian A., Yin T., Rahmani A. M., Dutt N. (2019) An edge-assisted and smart system for real-time pain monitoring. In: 2019 IEEE/ACM International conference on connected health: applications, systems and engineering technologies (CHASE). IEEE, pp 47–52
45. Lopez-Martinez D., Picard R. (2017) Multi-task neural networks for personalized pain recognition from physiological signals. In: 2017 Seventh international conference on affective computing and intelligent interaction workshops and demos (ACIIW). IEEE, pp 181–184
46. Brown J. E., Chatterjee N., Younger J., Mackey S. (2011) Towards a physiology-based measure of pain: patterns of human brain activity distinguish painful from non-painful thermal stimulation. PloS one 6(9):e24124
47. Wildemeersch D., Peeters N., Saldien V., Vercauteren M., Hans G. (2018) Pain assessment by pupil dilation reflex in response to noxious stimulation in anaesthetized adults. Acta Anaesthesiol. Scand. 62(8):1050
48. Connelly M. A., Brown J. T., Kearns G. L., Anderson R. A., St Peter S. D., Neville K. A. (2014) Pupillometry: a non-invasive technique for pain assessment in paediatric patients. Archives of Disease in Childhood 99(12):1125

Li Wang received his B.E. in Beijing Institute of Technology in Beijing, China, in 2015. He is currently a Ph.D. candidate at Northeastern University in Boston, MA. His concentration is Human-Machine interaction. He is working in the Intelligent Human-Machine Systems (IHMS) Laboratory as a research assistant. His research interests include machine learning algorithm design, brain-computer interface, and physiological signal processing.

Yikang Guo received his B.S. and M.Sc. degree from School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China, and School of Automation, Beijing Institute of Technology, Beijing, China, respectively. Currently he is pursing toward the doctoral degree in Northeastern University, Boston, USA. His research interests include neural networks, pattern recognitions, and human-brain interaction.

Biren Dalip received his B.Sc. in Electronic Engineering Technology from Wentworth Institute of Technology in Boston, MA, in 2016. He received his M.Sc. in Industrial Engineering from Northeastern University in Boston, MA in 2018. He is currently a master student at Northeastern University in Boston, MA. His concentration is Human Factors Engineering. He is working in the Intelligent Human-Machine Systems (IHMS) Laboratory as a research assistant. His research interests include human machine system and pain study for clinical applications.

Dr. Richard D. Urman received his MD from Harvard Medical School and MBA from Harvard Business School, and completed a residency in anesthesia. His research interests are in patient safety, novel anesthetic drugs, and operating room efficiency. He has lectured nationally and internationally. His other interests are medical education and curriculum design, and he was recently named the Morgan-Zinsser Fellow of the Harvard Medical School Academy. He has authored several books, including Pocket Anesthesia, Essential Clinical Anesthesia, Operating Room Leadership and Management, Essential Regional Anesthesia, Physicians' Pathways to Non-Traditional Careers and Leadership Opportunities, Ambulatory Anesthesia, and several others.

Dr. Yingzi Lin is a Professor of the Department of Mechanical and Industrial Engineering, College of Engineering, Northeastern University, Boston, MA, USA, where she directs the Intelligent Human-Machine Systems (IHMS) Laboratory. She has published hundreds of technical papers in referred journals and conference proceedings. Her area of expertise includes: intelligent human-machine systems, driver-vehicle systems, smart structures and systems, sensors and sensing systems, multimodality information fusion, human machine interface design, and human friendly mechatronics, human state detection and monitoring, human factors and patient safety. Dr. Lin was the Chair of the Virtual Environments Technical Group of the Human Factors and Ergonomics Society (HFES). She was on the committees of the Transportation Research Board (TRB) of the National Academy of Sciences. She served as an Associate Editor of the IEEE Trans. on Systems, Man and Cybernetics - Part A: Systems and Humans. She has also been on the organizing committee of a number of professional meetings in the areas of Advanced Sensors, Mechatronic Systems, Dynamic Systems and Control, Advanced Smart Materials and Smart Structures, and human-machine interaction.