J. Vis. Commun. Image R. 91 (2023) 103743

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full length article

Classification networks for continuous automatic pain intensity monitoring in video using facial expression on the X-ITE Pain Database ☆

Ehsan Othman [a,*], Philipp Werner [a], Frerk Saxen [a], Ayoub Al-Hamadi [a], Sascha Gruss [b], Steffen Walter [b]

[a] Department of Neuro-Information Technology, Institute for Information Technology and Communications, Otto-von-Guericke University Magdeburg, Magdeburg, 39106, Germany
[b] Department of Medical Psychology, University Hospital of Ulm, ULM, 89081, Germany

ARTICLE INFO

Keywords:
Continuous pain intensity recognition
Random Forest classifier
Facial expression
Long-Short Term Memory
Sample weighting

ABSTRACT

So far, the current methods in the clinical application do not facilitate continuous monitoring for pain and are unreliable, especially for vulnerable patients. In contrast, several automated methods have been proposed for this task by using facial features that were extracted independently from every frame of a given sequence. However, the obtained results were poor due to the failure to represent movement dynamics. To solve this problem, this work introduces three distinct methods regarding classification to monitor continuous pain intensity: (1) A Random Forest classifier (RFc) baseline method, (2) Long-Short Term Memory (LSTM) method, and (3) LSTM using sample weighting method (LSTM-SW). In this study, we conducted experiments with 11 datasets regarding classification, then compared results to regression results in Othman et al. (2021). Experimental results showed that the LSTM & LSTM-SW methods for continuous automatic pain intensity recognition performed better than guessing and RFc except with small datasets such as the reduced tonic datasets.

1. Introduction

Automatic pain recognition has been attracting a lot of interest in the health domain in recent decades. Further, significant progress has been made in machine learning methods for such task [1]. Automatic systems can be more objective and robust pain diagnosis measures than human observers, since it has been shown that humans could not report pain with patients who are not able to self-report their pain, such as infants, unconscious patients in the intensive care unit, or adults with cognitive impairment [2]. Additionally, the human observer may be influenced by personal factors, such as the relationship to the sufferer [3] and the patient's attractiveness [4]. Othman et al. [5] reported that machines are much more adept at recognising pain intensity when analysing the frontal faces in videos than humans.

Facial expression analysis is very informative for pain detection [1, 6]. Several studies focused on automatic systems that use facial expression analysis to facilitate continuous monitoring of pain intensity. It has been shown that pain intensity recognition from single frames was outperformed by using the dynamic information available in a sequence of multiple frames [7]. Thus, the current work focuses on the temporal integration of frame-level facial features to recognise continuous pain intensity; it is good in describing the relevant dynamic information, such as speed, tendency, or overall variation. This paper introduces methods for monitoring continuous pain intensity in video using facial expression analysis. These methods were evaluated on the X-ITE Pain Database, which was recorded with healthy participants (subjects). Conducting experiments with healthy subjects have always played a vital role in a medical study to evaluate safety and tolerability without interference by concomitant pathological conditions [8]. Then, the efficacy of the represented model can be determined on patients (e.g., vulnerable groups).

The database is extremely imbalanced; that is time-windows without pain stimulus are the vast majority. To overcome the imbalanced

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail addresses: Ehsan.Othman@ovgu.de (E. Othman), Philipp.Werner@ovgu.de (P. Werner), Frerk.Saxen@ovgu.de (F. Saxen), Ayoub.Al-Hamadi@ovgu.de (A. Al-Hamadi), Sascha.Gruss@uni-ulm.de (S. Gruss), Steffen.Walter@uni-ulm.de (S. Walter).
URLs: http://www.nit.ovgu.de (E. Othman), http://www.nit.ovgu.de (P. Werner), http://www.nit.ovgu.de (F. Saxen), https://www.nit.ovgu.de (A. Al-Hamadi), https://scholar.google.com/citations?user=XDeOLZ4AAAAJ&hl=en (S. Gruss), https://www.uniklinik-ulm.de/psychosomatische-medizin-und-psychotherapie/forschung/sektion-medizinische-psychologie/team/steffen-walter.html (S. Walter).

https://doi.org/10.1016/j.jvcir.2022.103743
Received 5 October 2021; Received in revised form 6 August 2022; Accepted 28 December 2022
Available online 9 January 2023

database problem, we used the same datasets that were proposed by us in [9], which were utilised to reduce the impact of such problem: (1) Phasic Dataset (PD), (2) Heat Phasic Dataset (HPD), (3) Electrical Phasic Dataset (EPD), (4) Tonic Dataset (TD), (5) Heat Tonic Dataset (HTD), (6) Electrical Tonic Dataset (ETD), (7) Reduced Phasic Dataset (RPD), (8) Reduced Heat Phasic Dataset (RHPD), (9) Reduced Electrical Phasic Dataset (REPD), (10) Reduced Tonic Dataset (RTD), and (11) Reduced Electrical Tonic Dataset (RETD). The reduced datasets ignored some no pain samples that precede each sub-sequence of pain intensity samples. We preserved some no pain samples directly adjacent to each sub-sequence (samples) of pain intensity and ignored the rest of the no pain samples before. The number of preserved no pain samples varies based on the number of samples in each pain intensity sub-sequence. For example, for the sub-sequence of severe phasic electrical pain that contains five samples, the previous five no pain samples were kept, and the rest no pain samples before were deleted.

This work focuses on recognising continuous pain intensity regarding classification. According to the best of our knowledge, it is the first time that classification methods are used for continuous monitoring of pain intensity from frontal video in the X-ITE Pain Database. Our proposed methods were based on combining the temporal aspects of the preceding ten-seconds of Facial Activity Descriptor (FAD) [10] for 4 statistical measures (minimum, maximum, mean, and standard deviation). FAD was calculated from frame-level features to discriminate between no pain, three pain intensities (low, moderate, and severe) in four different qualities (phasic (short) and tonic (long) variants of each heat and electrical stimuli). Three classification methods were designed for predicting categorical pain intensity scales: Random Forest classifier (RFc) [11] as baseline method, Long-Short Term Memory (LSTM) [12,13], and LSTM using sample weighting method [5] (called LSTM-SW). We compared the difference between classification versus regression in recognising continuous pain intensity using the same datasets. The regression methods distinguish data into continuous real values, instead of using classes or discrete values. Such methods predict continuous labels and exploit their ordinal relationship. In contrast, classification methods categorise the data predicting discrete labels without using information about the order of classes. The regression results were provided in [9]; we used three regression methods for predicting continuous pain intensity: Random Forest regression (RFr) [11] as baseline method, Long-Short Term Memory (LSTM) [12,13], and LSTM using sample weighting method [5] (called LSTM-SW). The sample weighting method was used to reduce the impact of an imbalanced database problem to improve the performance of automatic models. We increased the weight of the training samples with more facial responses by duplicating the successfully predicted samples with prediction scores above 0.3 when using RFc models. Further, this work emphasises that the proposed 11 datasets from the X-ITE Pain Database that were introduced by Othman et al. [9] help to simplify the imbalanced database problem and improve the results.

The current work is organised as follows. Section 2 provides an overview of pain recognition methods based on facial expression then describes their relevance to this paper. Section 3 shows a description of the materials and methods for automatic monitoring of continuous pain intensity using facial expression: the X-ITE Pain Database preprocessing is shown in Section 3.1, feature extraction preprocessing is presented in Section 3.2, Random Forest (RF) method described in Section 3.3, Long-Short Term Memory (LSTM) and LSTM using sample weighting (called LSTM-SW) methods for classification and regression task is explained in Section 3.4, and the experimental setup is explained in Section 3.5. Section 4 presents a comparison between classification and regression models results with different measures, followed by a discussion in Section 5. Finally, we conclude the results and describe the potential future works in Section 6. All acronyms are listed in the abbreviation section.

2. Related work

Several databases have been designed and released for automatic pain recognition methods in computer vision and machine learning domains, which ranked from oldest to newest: COPE Database [14], UNBC-McMaster Shoulder Pain Database [15], BioVid Heat Pain Database [16], BP4D-Spontaneous Database [17], YouTube Database [18], BP4D+ Database [19], IIIT-S ICSD [20], SenseEmotion Database [21], Multimodal EmoPain Database [22], Mint PAIN Database [23], and X-ITE Pain Database [24]. Most methods of pain recognition used single modality: [7,25] used video, [26,27] used audio signal, and [28–30] used physiological signals. The recent methods used multiple modalities [31–33] that can improve the performance and flexibility of pain recognition. Further, Salekin et al. [34] introduced the first multimodal neonatal dataset that contains both behavioural and physiological responses to monitor and recognise neonatal pain.

Facial expression modality is the most commonly used in pain recognition [4], and many automatic systems have been suggested based on analysing individual facial Action Units (AUs) and their combinations. AUs were defined by the facial action coding system of Ekman and Friesen [35], which is the most commonly used method for coding facial expression. Many previous works have introduced methods for pain recognition using the extracted facial features (AUs) from video. Most recent studies in pain recognition have focused on features from video-level because it is more effective than features from frame-level for describing relevant dynamic information [7,36].

Several facial features descriptors have been proposed to analyse the spatio-temporal texture of facial videos: Local Binary Pattern (LBP) [37], Local Phase Quantisation (LPQ) [38], Binarized Statistical Image Features (BSIF) [39], LBP-TOP [40], LPQ-TOP [41], BSIF-TOP [42], HOG-TOP [43], and LGBP-TOP [25]. The LGBP-TOP are extended descriptors that use the Three Orthogonal Planes (TOP). Further, Werner et al. [7] and Kächele et al. [36] proposed the spatio-temporal descriptors based on appearance- & geometry-based facial features and head pose; the pain levels were classified using Random Forest (RF) [11] with those descriptors.

RF is widely used in pain recognition domain [5,31,36,44,45]. Alongside Werner et al. [31] method that has been used to classify pain intensity, we classified in [44] no pain and highest heat pain intensity using Random Forest classifier (RFc). We used the time series statistics descriptor [7] to represent face feature vector in the frontal video by calculating 16 statistical measures with their first and second derivatives per time series. In [23,46,47], the authors used various neural networks for pain recognition including Convolutional Neural Networks (CNN) [48] and Long-Short Term Memory (LSTM) [12]. Further, several hybrid deep learning networks have been proposed for pain recognition by combining CNN with LSTM [47,49,50], or CNN with Bidirectional LSTM [51,52].

In [5,44], we showed how RFc with Facial Activity Descriptor (FAD) performed well compared to reduced MobileNetV2 (using transfer learning with the first 5 inverted residual blocks), it performed similarly to simple Convolutional Neural Network (CNN). We improved the performance of CNN with frontal RGB images compared to RFc with FAD by about 1% when using the sample weighting method. Downweighting misclassified samples during training improved the performance, these samples often contain low or no facial responses to pain (see [53] for details of this phenomenon). We duplicated some training samples with more facial responses based on the classification score (score above 0.3). The performance improvement of CNN model was not very high to classify seven classes (no pain and three phasic pain intensities for heat and electrical modalities). Therefore, in this paper, we utilised RFc with FAD as a baseline approach to predict continuous phasic and tonic pain intensity and no pain from facial video sequences. Further, we used the previously mentioned FAD with LSTM for better handling time series prediction as an advanced continuous recognition method. So this paper advanced over [5] by investigating a more complex problem (classify phasic and tonic pain in video level), and it shows a comparison between classification and regression in continuous monitoring of pain intensity in the X-ITE Pain Database.

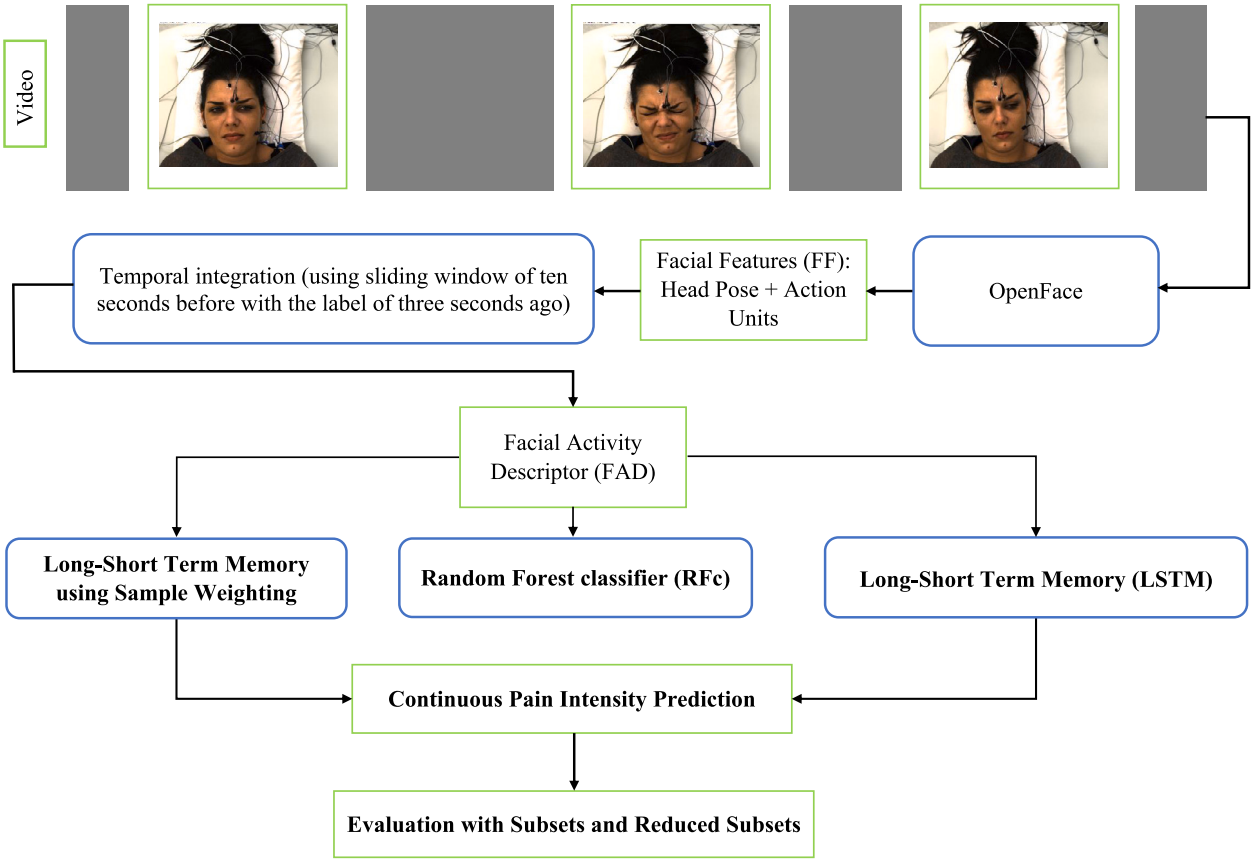

Fig. 1. Classification pipeline of pain intensity monitoring using RFc, LSTM, and LSTM-SW with FAD.

3. Material and methods

Fig. 1 shows an overview of the methodology of automatic facial expression analysis for the frontal faces in video to continuously monitor pain intensity. In this work, the Experimentally Induced Thermal and Electrical (X-ITE) Pain Database was used when the participants were stimulated with heat and electricity to induce pain in three intensities (low, medium, and high), see Section 3.1. First, we used OpenFace [54] (1) to detect the face from each frame for each participant and (2) to extract Facial Features (FF) and head pose. Second, the time window, which includes temporal integration of frame-level features, was represented by a time series statistics descriptor (called FAD); we moved the labels three seconds forward of each video (subject) and then used a sliding window with a time length of ten seconds ago. For more details, see Section 3.2. Third, we used the Random Forests classifier (RFc) with FAD as an automatic baseline method for continuous monitoring of pain intensity (see Section 3.3). Further, we introduced two deep learning methods for classification using FAD: (1) Long-Short Term Memory (LSTM) method (called LSTM, see Section 3.4.1) and (2) LSTM using the sample weighting method (called LSTM-SW, see Section 3.4.2). In Section 3.5 we give an overview of the conducted experiments and the database processing steps.

3.1. X-ITE pain databases

In this paper, we evaluated the performance of different continuous pain intensity monitoring methods on the X-ITE Pain Database [24]. In this database, a total of 134 healthy human participants aged between 18 and 50 years were stimulated with two types of pain (heat and electricity) in 3 intensities (low, medium, and high). The phasic and tonic stimuli were subjected to participants based on their personal pain tolerances using a thermal stimulator (Medoc PATHWAY Model ATS) and an electrical stimulator (Digitimer DS7A). The entire experiment took 3 h; approximately one and a half hours were the duration of the actual experiment for each participant (subject). For each phasic stimulus of both modalities (heat and electricity), the three pain intensities were repeated 30 times for 5 s duration; they were applied in randomised order with pauses of 8–12 s. Next to phasic stimuli, longer tonic pain stimuli for each modality were applied once per intensity for 60 s, followed by a pause of five minutes. There were three phases of how tonic heat and electrical pain intensity stimuli were applied: the two lower intensities were applied randomly during the phasic stimulus period, and the highest intensity was applied at the end of the experiment. For more details see Gruss et al. [24]. In line with Werner et al. [31] and Othman et al. [5,9], we selected the same 127 participants (subjects) subset, including samples only, for which data were available from all sensors (frontal RGB camera, audio, ECG: electrocardiogram, EMG: surface electromyography, EDA: electrodermal activity). In this work, we focused on analysing the facial expression data from frontal videos involving the phasic and tonic pain intensity during the application of the thermal and electrical pain stimuli and no pain.

3.2. Facial Activity Descriptor (FAD) and sliding time window

In this work, we extracted facial features (FF) from each frame for each video (subject) using OpenFace [54], each video duration of about one and a half hours. The OpenFace [54] tool can detect the face and facial landmarks, extract Action Units (AUs), and estimate head pose in this experiment. The FF we used include 21 features: 3 head pose (Yaw, Pitch, and Roll), AU1(binary occurrence output), and 17 AU intensity outputs of OpenFace, which are AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, and AU45. The FF contained 25 samples per second due to the video frame

Table 1
A summary of the LSTM architectures configurations (Regression and Classification). The meaning of acronyms is provided in the abbreviations section.

Layer type	Attribute	Architectures					
		Classification				Regression	
		A(c)	B(c)	C(c)	D(c)	A(r)	B(r)
Input	Size: Timesteps: Features:	10 × 252 10 252				10 × 252 10 252	
LSTM	No. of parameters: Activation: No. of units:	4112 ReLU 4	8352 ReLU 8	4112 ReLU 4	8352 ReLU 8	4112 ReLU 4	8352 ReLU 8
Dropout	with p:	0.5	0.5	0.5	0.5	0.5	0.5
Flatten	Output:	80	40	80	40	80	40
Dense1	No. of parameters: Activation: No. of units:	5248 ReLU 128	5184 ReLU 64	5248 ReLU 128	5184 ReLU 64	5248 ReLU 128	5184 ReLU 64
Dense2	No. of parameters: Activation: No. of units:	129 Softmax 7	65 Softmax 7	129 Softmax 4	65 Softmax 4	129 Linear/Sigmoid 1	65 1
Output	Continuous: Discrete:	– X 7 levels	– X 7 levels	– X 4 levels	– X 4 levels	X – –	X – –

Activation function		Loss function
Linear	$\hat{y}_{(r)i} = w_i^T x$	$MSE = \sum_{i=1}^{n} (\hat{y}_{(r)i} - y_{(r)i})^2 / n$
Sigmoid	$\hat{y}_{(r)i} = \frac{1}{1 + \exp^{(-w_i^T x)}}$	$BCE = -(\sum_{i=1}^{n} [y_{(r)i} \log \hat{y}_{(r)i} + (1 - y_{(r)i}) \log(1 - \hat{y}_{(r)i})])/n$
Softmax	$\hat{y}_{(c)i} = \frac{\exp^{(w_i^T x)}}{\sum_{j=1}^{C} \exp^{(w_j^T x)}}$	$CCE = -(\sum_{i=1}^{n} [y_{(c)i} \log \hat{y}_{(c)i}])/n$

rate of 25 frames per second. The temporal integration of frame-level features (called time series) was represented by a time series statistics descriptor [7,55] for describing the changes of FF, which is called Facial Activity Descriptor (FAD). FAD of each second was summarised by four statistics (minimum, maximum, mean, and standard deviation) of the time series itself and its first and second derivative, yielding a 12-dimensional descriptor per time series. We moved the labels of each subject 3 s after because the facial pain responses typically were delayed by 2–3 s compared to stimulus. Further, we applied a person-specific standardisation of the features [7] with FAD in order to focus on the within-subject response variation rather than the differences between subjects. The sliding time window with a length of ten seconds was applied once per second by combining the FAD of the preceding ten seconds to predict the next time step of the pain intensity label. The data of the first ten seconds was removed because there were no prior observations to use.

3.3. Random forest classifier (RFc) baseline method

Alongside Werner et al. [31] study, we trained RFc [11] with 100 trees using the sequence of samples (FAD sequences), FAD was described in Section 3.2. The suggested method performed well to predict pain intensity and no pain from the time windows of samples that were cut out from the continuous recording of the main stimulation phase [5,31]. Further, in our recent study (see Othman et al. [9]), we used similar Random Forest regression (RFr) structure and input format, but regarding regression for continuous pain intensity recognition in the X-ITE Pain Database. RFr showed good results compared to chance; it also outperformed LSTMs (LSTM and LSTM using sample weighting) with a small dataset. RFc was used to compare between classification and regression results. For more details about the comparison results, see Section 4.2.

3.4. The deep learning methods

This section describes the proposed deep learning methods that were applied to monitor continuous pain intensity from frontal faces in video using the Facial Activity Descriptor (FAD). These methods are Long-Short Term Memory (LSTM) and LSTM using the sample weighting method (called LSTM-SW), see Sections 3.4.1 and 3.4.2, respectively. We used LSTM [12] to learn long-term dependency among a longer time period by storing information from previous frames. It is an effective method for better handling time series prediction.

3.4.1. Long-Short Term Memory (LSTM)

After extracting the FAD from each video (subject) and after sliding time windows from time series data (FAD) (see Section 3.2), these FAD samples were fed into LSTM [12] one by one in sequence. In this work, four LSTM architectures were proposed for classification (see A(c), B(c), C(c), and D(c) in Table 1), which yielded the best results after being tested on all datasets. Further, we presented the two LSTM regression networks, which were proposed in [9], see A(r) and B(r) in Table 1. A(r), A(c), and C(c) comprised a single LSTM layer with 4 units activated by ReLU and followed by a flatten layer, and then one dense layer with 128 neurons activated by ReLU. B(r), B(c), and D(c) comprised a single LSTM layer with 8 units activated by ReLU and followed flatten layer, and then one dense layer with 64 neurons activated by ReLU. For recognising continuous pain intensity using regression architecture, the final dense output layer was activated using the linear or sigmoid function, and the used loss functions are the Mean Squared Error (MSE) or Binary Cross-Entropy (BCE), respectively in this sequence. With BCE, the sigmoid activation function was used to calculate the outputs between zero and one for multi-class classification. Further, the final dense output layer in classification architectures was Softmax activated and the Categorical Cross-Entropy (CCE) loss was used. The obtained models were trained for 2000 epochs with 10^{-4} or 10^{-5} or 10^{-6} learning rate. Our LSTM predicted one frame by using several adjacent frames, which kept the estimation line stable, smooth, and closed to the ground-truth labels. Table 1 shows the proposed LSTM architectures; it also presents the suggested activation and the loss functions.

$\hat{y}_{(r)i}$ is the continuous predicted value of the LSTM, and $\hat{y}_{(c)i}$ is the discrete predicted belonging to the i category, for $i = 1, 2, \ldots,$ C. x

Table 2
List of acronyms of pain stimuli type, modalities, intensities, and numerical class labels with the percentage samples distribution. E: Electrical pain stimulus, H: Heat pain stimulus. The meaning of acronyms is provided in the abbreviations section.

Type	Modality	Intensities			no Pain
		Severe	Moderate	Low	(77%)
Phasic	H E	PH3 = 3 (2.0%) PE3 = −3 (2.6%)	PH2 = 2 (2.1%) PE2 = −2 (2.6%)	PH1 = 1 (2.1%) PE1 = −1 (2.6%)	BL = 0
Tonic	H E	TH3 = 6 (1.0%) TE3 = −6 (1.0%)	TH2 = 5 (1.0%) TE2 = −5 (1.0%)	TH1 = 4 (1.0%) TE1 = −4 (1.0%)	BL = 0
−10 (0.5%) −11 (2.5%)	not used in the experiments				

is the frame-level features (time windows 10s-length) vector, n is the number of samples.

3.4.2. LSTM using sample weighting (LSTM-SW)

After observing the poor performance when using the highly imbalanced datasets (11 datasets), we decided to increase the weight of the training samples with more facial responses by duplicating the samples with a high score > 0.3. Table 2 summarises the class notation (no pain and pain intensity) in our experiments and shows the imbalance of the classes in the X-ITE Pain Database by giving the sample distribution with percentages.

In line with the regression LSTM-SW method in [9], the baseline RFc method was trained on the 11 datasets from the X-ITE Pain Database as mentioned in the Section 3.3. We identified the training samples with prediction scores above 0.3, which were correctly classified. We duplicated these samples once with each dataset, which increased the samples size. The validation and test data were kept unmodified in all experiments. See [5] for more details about the sample weighting method. Then, we trained the proposed LSTM in Section 3.4.1 on the same 11 datasets after duplicating the samples with more facial responses. This method is called LSTM-SW. All applied datasets were described in Section 3.5. We used all duplicate samples with Reduced Subsets, and we excluded the duplicate samples of no pain in Subsets to reduce their effect on increasing the imbalanced datasets problem. As a result, almost all LSTM-SW models increased the performance compared to RFc models in terms of classification and regression measures (see Section 4). Additionally, the computational cost of training RFc is negligible compared to the LSTM training.

Table 3
Subject's categories based their intensity expressions of pain for three database splits.

Training set (572 696 samples)	
Intensity	Subjects (100)
1	S002, S010, S019, S020, S022, S033, S042, S065, S082, S093, S048, S050, S051
2	S011, S016, S026, S027, S028, S037, S041, S044, S074, S077, S080, S089, S091, S092, S095, S102, S103, S118, S120, S122, S123, S124, S130, S132
3	S003, S004, S005, S006, S008, S012, S015, S017, S029, S032, S034, S035, S036, S038, S039, S040, S045, S052, S053, S054, S055, S057, S058, S061, S063, S064, S066, S067, S068, S069, S070, S072, S073, S075, S076, S078, S083, S084, S085, S086, S087, S090, S096, S098, S099, S101, S105, S106, S109, S110, S111, S112, S114, S116, S117, S119, S125, S129, S133
4	S060, S108, S115, S134
Validation set (75 537 samples)	
Intensity	Subjects (13)
1	S113
2	S046, S100, S128
3	S009, S013, S018, S043, S088, S094, S097, S126
4	S104
Test set (79 485 samples)	
Intensity	Subjects (14)
1	S079, S107
2	S007, S047, S056, S081
3	S031, S049, S058, S062, S071, S127, S131
4	S021

3.5. Experiments

This paper focuses on classification experiments that were applied on the frontal faces from the videos for continuous monitoring of pain intensity. The extremely imbalanced datasets problem (see Table 2) may be the reason why many automatic models perform poorly; all notations in Table 2 will be gradually explained in this section. Some problems with several samples in the experiments, such as samples with labels −10 or −11. −10 indicates samples with false start and restart of the stimuli, overlapping between heat or electrical stimulation, unbalanced phasic estimation, short pause, short tonic electrical stimulus, single heat stimulus in front, or additional stimulus. −11 indicates the samples when the subject speaks or interacts during the experiment (the beginning and after the first & second tonic stimuli of the experiment). In line with our study in [9], we used the same proposed four categories of subjects, which based on how intensely they expressed pain (see Table 3).

Table 3 shows the proposed splits for the database that use, which represent 80% of data for training, 10% for validation 10% for testing. Each split contains subjects from all intensity categories. The subjects were selected randomly from each category based on the proposed percentage.

Othman et al. [9] suggested different datasets from the imbalanced X-ITE Pain Database to reduce the impact of such problem, which was also used in this work. We processed the database: (1) We excluded all sequences of samples with labels −10, −11 and no pain samples sequence before and after these samples to simplify the problem and reduce the impact of the imbalance of these datasets; (2) We suggested to split the obtained dataset into 6 subsets to evaluate the proposed methods (see Subsets which are the first six datasets in Table 4); (3) We suggested to reduce each proposed dataset by reducing some no pain samples prior to pain intensity sequences in a time series for each subject (see Reduced Subsets which are the last six datasets in Table 4). All datasets were reduced to 50% except Tonic Subset (38%) and Electrical Tonic Subset (49%).

In Table 4, (1) PD: Excluded tonic samples (labelled 4, 5, 6, −4, −5, −6, −10, −11), no pain samples before these samples, and no pain samples after samples with −10 and −11 labelled, (2) HPD: Excluded electrical samples (labelled −1, −2, −3) from PD and no pain frames before these frames, (3) EPD: Excluded heat samples (labelled 1, 2, 3) from PD and no pain frames before these frames, (4) TD: Excluded phasic samples (labelled 1, 2, 3, −1, −2, −3, −10, −11), no pain samples before these samples and after samples with −10 and −11 labelled, (5) HTD: Excluded electrical samples (labelled −1, −2, −3) from TD and no pain frames before these frames, (6) ETD: Excluded heat samples (labelled 1, 2, 3) from TD and no pain frames before these frames, (7) RPD: Reduced the no pain frames in PD (reduced to about 50%), (8) RHPD: Reduced the no pain frames in HPD (reduced to about 50%),

Table 4
No. of samples in each dataset for each splits to evaluate proposed methods before & after applying sample weighting method. Red. Subsets: Reduced Subsets.

	Datasets	Description	Training set	Validation set	Test set	Apply sample weighting	
						Training set	Increased
Subsets	PD	Phasic Dataset	352 133	46 476	50 362	405 060	52,927 (20%)
	HPD	Heat Phasic Dataset	159 998	21 441	23 019	190 178	30,180 (20%)
	EPD	Electrical Phasic Dataset	316 939	41 794	45 325	35 3560	36,621 (10%)
	TD	Tonic Dataset	117 646	14 885	16 689	14 2667	25,021 (20%)
	HTD	Heat Tonic Dataset	21 198	2755	3103	37 087	15,889 (70%)
	ETD	Electrical Tonic Dataset	95 458	12 000	13 446	109 644	14,186 (10%)
Red. Subsets	RPD	Reduced Phasic Dataset	158 472	20 897	22 501	237 735	79,263 (50%)
	RHPD	Reduced Heat Phasic Dataset	69 390	9233	9933	119 780	50,390 (70%)
	REPD	Reduced Electrical Phasic Dataset	88 148	11 548	12 438	119 780	62,789 (70%)
	RTD	Reduced Tonic Dataset	55 804	7041	7983	99 455	43,651 (80%)
	RETD	Reduced Electrical Tonic Dataset	33 826	4156	4740	62 799	28,973 (90%)

(9) REPD: Reduced the no pain frames in EPD (reduced to about 50%), (10) RTD: Reduced the no pain frames in TD to about 38%, (11) RETD: Reduced the no pain frames in ETD to about 49%.

Our reduction strategy focused on reducing some no pain samples prior to each sequence of pain intensity by preserving different numbers of no pain samples that were directly adjacent to each pain intensity sequence; this number was assigned based on the number of samples in each pain intensity sub-sequence (e.g., for frame sequence of low tonic electrical pain that contained 60 samples, we kept the previous 60 no pain samples and deleted the rest before). Thus, we got five additional datasets (Reduced Subsets), and the Heat Tonic Dataset (HTD) was not reduced because it is nearly balanced.

The proposed RFc, LSTM, and LSTM-SW models were trained on all 11 datasets, which were also used to train regression models (RFr, LSTM, and LSTM) in [9]. The pain intensity labels were conditioned into the right format by: (1) converting the negative labels (−1, −2, −3) to positive (4, 5, 6) in Phasic Dataset (PD), the obtained labels were 1, 2, 3, 4, 5, 6, (2) converting the labels 4, 5, 6, −4, −5, −6 to 1, 2, 3, 4, 5, 6 in Tonic Dataset (TD), (3) converting the negative labels (−1, −2, −3) to positive (1, 2, 3) in Electrical Phasic Dataset (EPD), (4) converting labels 4, 5, 6 to 1, 2, 3 in Heat Tonic Dataset (HTD), and (5) converting the labels −4, −5, −6 to 1, 2, 3 in Electrical Tonic Dataset (ETD). The predictions from classification models were normalised to bring them in the range of [0,1].

Classification models were evaluated on the test set using different measures, which were useful when datasets varies in size (aggregate the contributions of all classes to compute the average metric). Those classification measures were Micro average precision (Micro avg. precision), Micro average recall (Micro avg. recall), and Micro average F1-score (Micro avg. F1-score). Further, we calculated accuracy. The Mean Squared Error (MSE) and the intraclass correlation coefficient (ICC) [56] were calculated to compare the classification versus regression models' performance. Table 5 shows both classification and regression measures that use in the conducted experiments; MSE and ICC were used to compare regression results to classification results. The evaluation results were presented in Section 4.

Table 5
The description and equation of experiment measures.

Classification
$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{n}$ Percentage of correctly classified samples. n: All samples.
$Micro\ avg.\ precision = \frac{Sum\ the\ TP\ of\ all\ classes}{Sum\ the\ TP\ of\ all\ classes + Sum\ the\ FP\ of\ all\ classes}$ An average per-class agreement of the data class labels with those of a classifier.
$Micro\ avg.\ recall = \frac{Sum\ the\ TP\ of\ all\ classes}{Sum\ the\ TP\ of\ all\ classes + Sum\ the\ FN\ of\ all\ classes}$ An average per-class effectiveness of a classifier to identify class labels.
$Micro\ avg.\ F1-score = 2 * \frac{Micro\ avg.\ precision * Micro\ avg.\ recall}{Micro\ avg.\ Precision + Micro\ avg.\ Recall}$ Relations between data of positive labels and those given by a classifier based on a per-class average.
Regression
$MSE = \frac{1}{n}\sum_{i=1}^{n}(Predicted\ values - Actual\ values)^2$ The average squared difference between the predicted values and the actual value.
$ICC = \frac{BMS - EMS}{BMS + (k-1)EMS}$ Intraclass correlation coefficient. ICC (3,1) [56] used to assess measurement reliability based on average of k measurements (conditions, raters). BMS: Between-targets means square, EMS: Within-targets means square.

Table 6
The defined classes in 7-Class and 4-Class datasets, The meaning of acronyms is provided in the abbreviations section.

Datasets	Classes
PD/RPD	BL, PH1, PH2, PH3, PE1, PE2, and PE3
HPD/RHPD	BL, PH1, PH2, and PH3
EPD/REPD	BL, PE1, PE2, and PE3
TD/RTD	BL, TH1, TH2, TH3, TE1, TE2, and TE3
HTD	BL, TH1, TH2, and TH3
ETD/RETD	BL, TE1, TE2, and TE3

4. Results

4.1. Classification

In this section, we compared the performance of the classification models that continuously monitor pain intensity using the proposed measurements that were shown in Table 5. Table 6 shows that 7 classes were obtained after applying continuous pain intensity classification. Those classes refer to each phasic and tonic stimuli, combining the heat and electrical pain stimuli that differ in pain intensity (no pain, low, moderate, and severe): BL, PH1, PH2, PH3, PE1, PE2, and PE3 for phasic pain recognition task, and BL, TH1, TH2, TH3, TE1, TE2, and TE3 for tonic pain recognition task. Further, 4 out of the 7 classes were extracted for each stimulation method (heat/electrical): BL, PH1, PH2, and PH3 for the phasic heat recognition task, BL, PE1, PE2, and PE3 for the phasic electrical recognition task, BL, TH1, TH2, and TH3 for tonic heat recognition task, and TE1, TE2, and TE3 for the tonic electrical recognition task.

Table 7
The increased sample size of training set for each class after applying sample weighting method. The meaning of acronyms is provided in the abbreviations section.

Datasets		Modalities with intensities						
Subsets	Phasic	BL	PH1	PH2	PH3	PE1	PE2	PE3
	PD	–	15%	15%	18%	16%	16%	21%
	HPD	–	32%	32.7%	35%	–	–	–
	EPD	–	–	–	–	32%	32%	37%
	Tonic	BL	TH1	TH2	TH3	TE1	TE2	TE3
	TD	–	17%	17%	19%	15%	15%	17%
	HTD	–	34%	33%	33%	–	–	–
	ETD	–	–	–	–	32%	33%	35%
Reduced Subsets	Phasic	BL	PH1	PH2	PH3	PE1	PE2	PE3
	PD	22%	12%	12%	13%	13%	13%	15%
	HPD	36%	21%	21%	22%	–	–	–
	EPD	29%	–	–	–	21%	21%	22%
	Tonic	BL	TH1	TH2	TH3	TE1	TE2	TE3
	TD	26%	12%	12%	13%	12%	12%	12%
	ETD	43%	–	–	–	19%	19%	19%

Table 4 shows that the samples in each pain intensity class from each dataset were increased after applying the sample weighting method to the RFc prediction score. The samples with a score higher than 0.3 were duplicated to use with LSTM-SW. No pain samples after sample weighting were not duplicated with all datasets in Subsets to avoid increasing imbalance. The highest pain intensity (severe) samples from Subsets were recognised better than moderate pain intensity samples; see the percentage of the increased samples with PH3, PE3, TH3, and TE3 in Table 7.

With Reduced Subsets, no pain samples were better in recognising because they were too many compared to pain intensity samples (about 50% of samples experience no pain with most datasets except Tonic Dataset (38%) and Electrical Tonic Subset (49%)).

Table 8 and Fig. 2 show how the proposed models successfully predicted discrete pain intensity level (7 and 4 classes) in sequences compared to Trivial (chance); the listed results are in terms of Micro avg. precision, Micro avg. recall, Micro avg. F1-score and accuracy. The paired t-test was used to calculate the p-value for evaluating the proposed models (LSTM and LSTM-SW) results if they are significantly better than the baseline model (RFc). Our models performed better than guessing (chance) and automatic RFc baseline models in terms of accuracy. The best results in terms of micro avg. precision were obtained from LSTM models except when using Heat Tonic Dataset (HTD) and Electrical Tonic Dataset (ETD) for which LSTM-SW performs better with about 0.34 and 0.24, respectively. LSTM-SW yielded the best Micro avg. recall results except when using Electrical Tonic Dataset (ETD) and Reduced Phasic Dataset (RPD), the LSTM performed better with 0.05 and 0.45. Additionally, RFc achieved the best Micro avg. recall results using Reduced Tonic Dataset (RTD) and Reduced Electrical Tonic Dataset (RETD) with about 0.12 and 0.15 due to their small data size. LSTM and LSTM-SW performed similarly in terms of Micro avg. F1-Score except when using Reduced Tonic Dataset (RTD) and Reduced Electrical Tonic Dataset (RETD), RFc performed the best with 0.13 and 0.17, respectively. The accuracy performance of LSTM models were significantly greater than RFc (p-value < 0.05) except with Heat Tonic Dataset (HTD) and Electrical Tonic Dataset (ETD), LSTM-SW performed well with these models about 33.7% and 82.4%, respectively.

Table 8
Comparison between Trivial, baseline (RFc), LSTM, and LSTM-SW methods to predict discrete pain level in sequences regarding classification task. no.C: no. of Classes, Red. Subsets: Reduced Subsets, Archit.: Architecture, lr: learning rate.

Datasets		no.C	Accuracy %				Micro avg. Precision				Archit.	lr
			Trivial	RFc	LSTM	LSTM-SW	Trivial	RFc	LSTM	LSTM-SW		
Subsets	PD	7	77.7	76.6	78.2[a]	77.5[a]	0	0.17	0.28[a]	0.25[a]	A(c)	10^{-5}
	HPD	4	78.5	77.8	78.9[a]	78.0	0	0.24	0.32[a]	0.27	C(c)	
	EPD	4	86.1	85.4	86.6[a]	85.4	0	0.28	0.39[a]	0.30	C(c)	
	TD	7	70.3	68.5	70.7[a]	70.0[a]	0	0.14	0.19	0.13	A(c)	10^{-6}
	HTD	4	20.0	29.1	32.8	33.7[a]	0	0.31	0.33	0.34	C(c)	
	ETD	4	82.0	80.7	81.7	82.4[a]	0	0.15	0.12	0.24	C(c)	
Red. Subsets	RPD	7	50.0	47.6	57.4[a]	54.7[a]	0	0.20	0.29[a]	0.27[a]	A(c)	10^{-4}
	RHPD	4	50.1	49.0	61.0[a]	58.7[a]	0	0.28	0.37[a]	0.34[a]	C(c)	
	REPD	4	50.0	49.2	61.5[a]	58.9[a]	0	0.32	0.41[a]	0.38[a]	C(c)	
	RTD	7	38.1	33.9	39.2[a]	37.4	0	0.16	0.31	0.28	B(c)	10^{-6}
	RETD	4	49.0	43.1	50.1[a]	47.5	0	0.20	0.28	0.19	D(c)	

Datasets		no.C	Micro avg. Recall				Micro avg. F1-Score				Archit.	lr
			Trivial	RFc	LSTM	LSTM-SW	Trivial	RFc	LSTM	LSTM-SW		
Subsets	PD	7	0	0.04	0.07[a]	0.09[a]	0	0.06	0.10[a]	0.12[a]	A(c)	10^{-5}
	HPD	4	0	0.07	0.12[a]	0.14[a]	0	0.10	0.15[a]	0.16[a]	C(c)	
	EPD	4	0	0.06	0.10[a]	0.12[a]	0	0.10	0.15[a]	0.16[a]	C(c)	
	TD	7	0	0.04	0.02	0.04	0	0.06	0.03	0.06	A(c)	10^{-6}
	HTD	4	0	0.49	0.92[a]	1.0[a]	0	0.38	0.47	0.48[a]	C(c)	
	ETD	4	0	0.03	0.05	0.04	0	0.04	0.06	0.06	C(c)	
Red. Subsets	RPD	7	0	0.12	0.45[a]	0.39[a]	0	0.15	0.35[a]	0.32[a]	A(c)	10^{-4}
	RHPD	4	0	0.22	0.58[a]	0.62[a]	0	0.24	0.45[a]	0.44[a]	C(c)	
	REPD	4	0	0.25	0.54[a]	0.59[a]	0	0.28	0.46[a]	0.46[a]	C(c)	
	RTD	7	0	0.12	0.06	0.09	0	0.13	0.08	0.12	B(c)	10^{-6}
	RETD	4	0	0.15	0.09	0.08	0	0.17	0.12	0.10	D(c)	

[a]p<0.05 when using paired t-test for comparing the results of the LSTM and LSTM-SW with RFc.

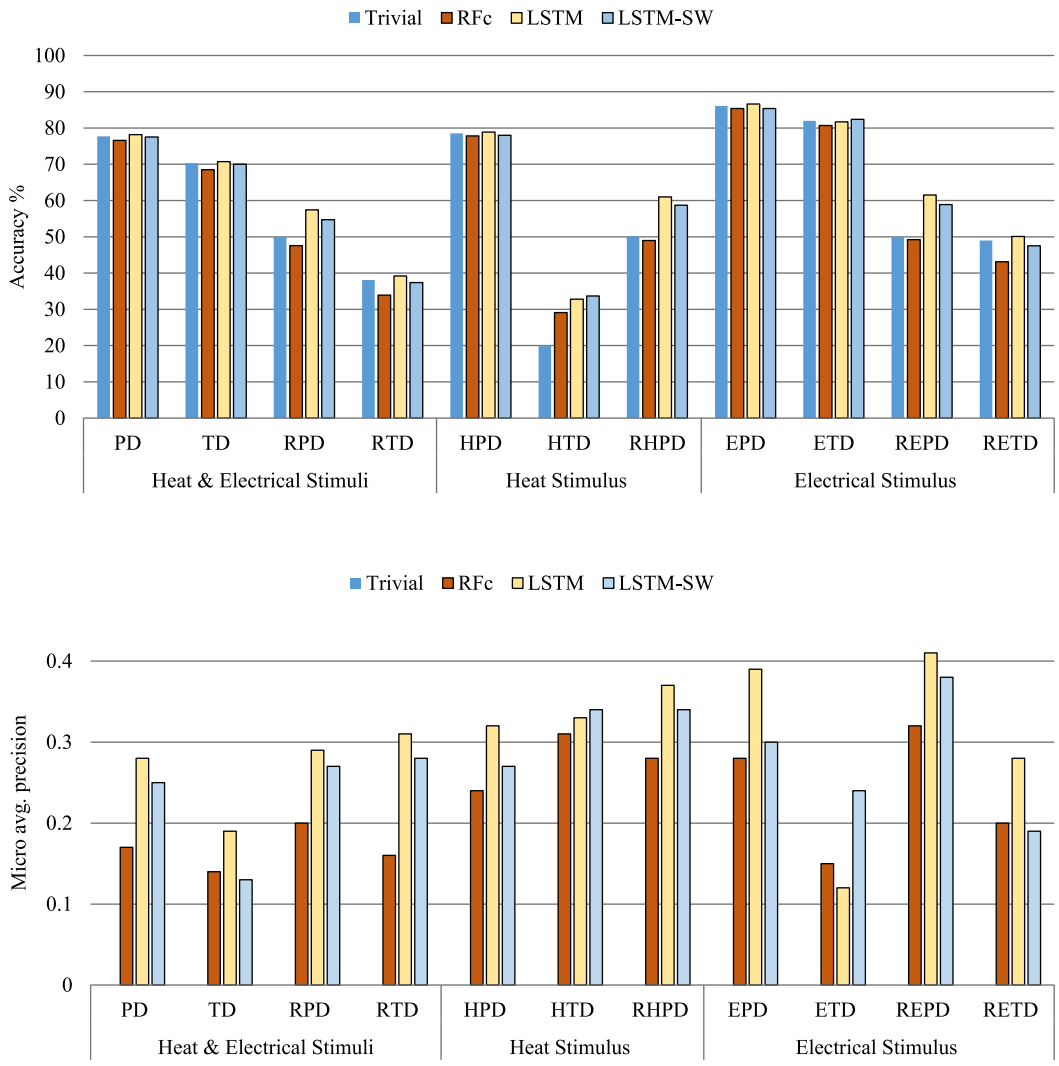

Fig. 2. Comparison of the Trivial and baseline method (RFc) with proposed LSTM and LSTM-SW methods to predict discrete pain level using classification measures.

4.2. Classification vs. regression

This section provides the comparison between Trivial and our methods with all datasets (see the 11 datasets in Section 3.5). To evaluate the discrete and continuous pain intensity monitoring models, Mean Squared Error (MSE) and the intraclass correlation coefficient (ICC) [56] were calculated from the test set. The results of regression models obtained from our recent study in [9]; before training the models, the labels were normalised [0,1]. Table 9 and Fig. 3 show that the Trivial failed to recognise pain intensity, whereas the baseline methods (RFc and RFr) are significantly better. After comparing regression methods to classification methods, LSTM using MSE loss function got the highest ICC values on Tonic Dataset (TD) and Electrical Tonic Dataset (ETD); they obtained the ICC of 0.14 and 0.13, respectively. The TD also got the smallest recognition error; the ETD obtained 0.08, whereas LSTM using BCE obtained the smallest recognition error (0.07). LSTM using BCE loss function got the highest ICC values and smallest recognition error on Heat Phasic Dataset (HPD) and Reduced Heat Phasic Dataset (RHPD); it obtained the MSE of 0.08 and 0.09 and the ICC of 0.28 and 0.62, respectively.

Further, LSTM-SW using BCE loss function got the highest ICC values and smallest recognition error on phasic Subset (PD and EPD), RHPD, and RETD. It obtained the MSE of 0.08, 0.06, 0.09, and 0.13 and the ICC of 0.22, 0.28, 0.62, and 0.20 respectively. Correspondingly, LSTM-SW and LSTM using the CCE loss function got the highest ICC values on HTD (0.20), RPD (0.57), and REPD (0.56), respectively in this sequence. RFc with Reduced Tonic Dataset (RTD) led to a higher ICC value (0.09); it performed better than LSTM methods, probably due to the small data size. The sample weighting method improved the performance of LSTM using BCE, but not significantly with most phasic and tonic datasets. Further, the models using A(r), A(c), and C(c) architectures performed better than models using B(r), B(c), and D(c) architectures.

5. Discussion

In this paper, we conducted several experiments in order to gain insights into both classification and regression for automatically monitoring continuous pain intensity using facial video on X-ITE Pain Database. First, we used the 11 datasets proposed in [9], which were also described in Section 3.5. In each dataset, we shifted the labels 3 s forward for each subject and combined the data (FAD) for ten seconds ago to predict pain intensity in the next time step. Second, we analysed the facial expression in proposed datasets using classification methods (RFc, LSTM and LSTM-SW). The results showed that it is challenging but possible to automatically predict the continuous pain intensities and qualities (heat and electrical stimuli) in phasic and tonic pain based

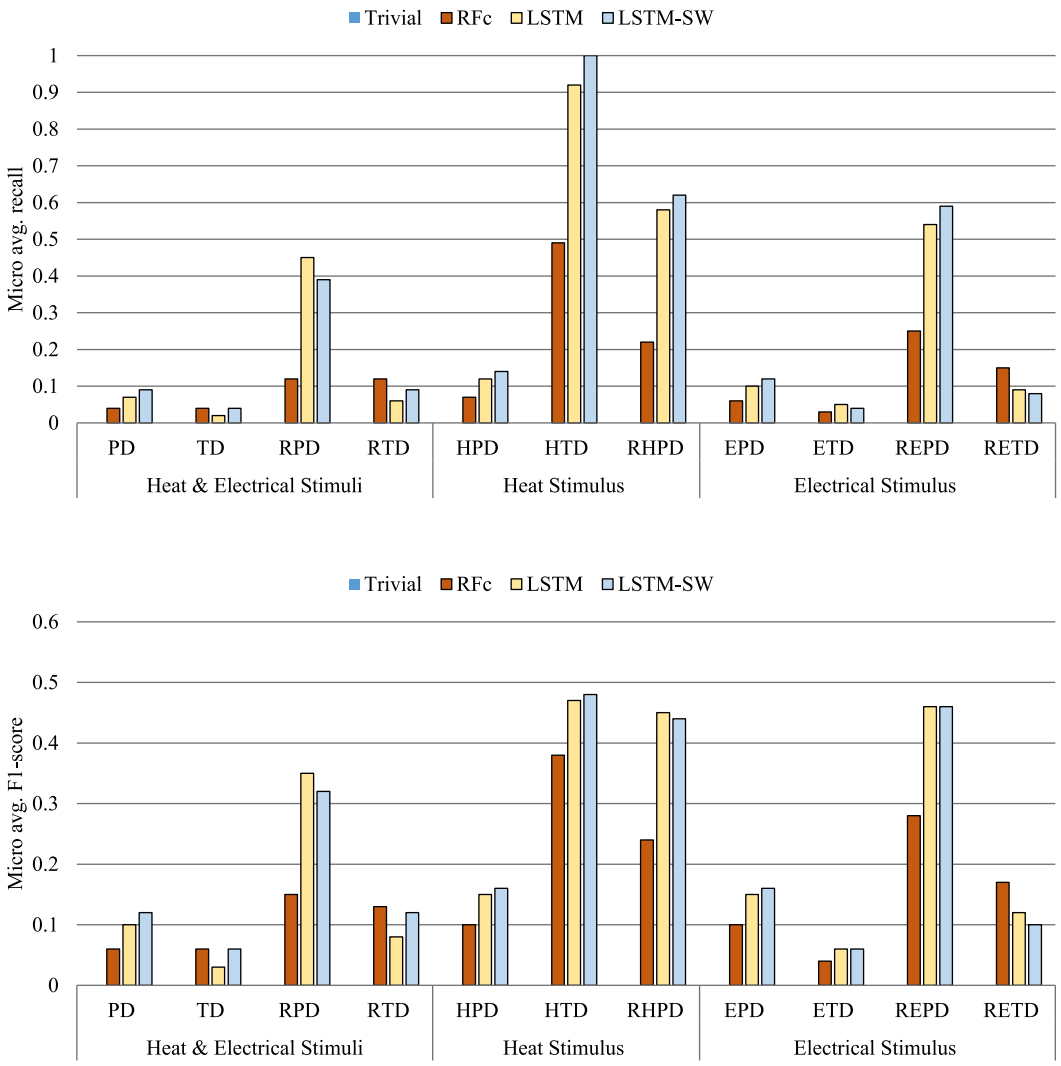

Fig. 2. (*continued*).

on facial expression analysis. Classification models obtained similar results to our study in [9].

Both classification and regression models were better than informed guessing (Trivial) because the trivial model always votes for the majority of the class (no pain in our experiment). This task is difficult due to: (1) Imbalanced datasets (see Table 2) and (2) A lack of facial responses to pain with some subjects (some people have low sensitivity to pain, see [5]) or vice versa (some people are extremely expressive of their pain). The reduced datasets showed significantly better performance (see Table 8 and Fig. 2), which leads to the hypothesis that the most confusions occur from outliers or label noise. Nevertheless, the results showed that it seems more important to reduce the noise in imbalanced data than to keep very hard samples. There is a large discrepancy between the size of the tonic samples and the phasic samples; the phasic data is about three times bigger than the tonic data. If we combine them to estimate pain intensity from both the tonic and phasic datasets, the model will fail to recognise the tonic samples; most of the samples' predictions will be no pain or phasic pain intensity. Although we separated tonic datasets, the performance was still poor due to the small size of the tonic datasets. In this work, we confirmed that continuous monitoring of pain intensity is possible using regression or classification methods based on the good results, see Table 9.

Further, recognition of pain stimuli intensity in the phasic datasets worked better than in the tonic datasets because of the large size of data and the deep learning models ability to benefit from huge datasets, see Tables 8, 9 and Figs. 2, 3. Regarding classification results, we closely looked at the correctly predicted samples when using RFc, LSTM, and LSTM-SW. We noticed that RFc performed the best with all phasic datasets in recognising intermediate pain (low and moderate).

In contrast, the LSTMs performed well in recognising no pain and the highest pain intensity, see Table 10. We concluded that LSTMs may have difficulty in recognising intermediate pain intensity stimulation in large imbalanced datasets. However, this finding cannot be generalised because further investigation is needed to understand exactly what properties are inherited by models that were trained using LSTM.

Additionally, training and testing LSTM on big imbalanced datasets will be helpful to confirm our hypothesis. With reduced phasic datasets (RPD, RHPD, and REPD) and Heat Tonic Dataset (HTD), the LSTMs obtained superior results in recognising baseline (no pain) versus pain intensity, possibly due to less noise of no pain data. HTD was almost balanced, and reduced phasic subsets were obtained after reducing the imbalance problem.

Table 8 and Fig. 2 show that most LSTM models were the best in recognising no pain versus pain intensity and in predicting more

Table 9
Comparison of the Trivial, baseline (RFc and RFr), LSTM, and LSTM-SW methods for classification and regression.

Meas.	Task			Classification				Regression					lr
	Loss		–	–		CCE		–		MSE	BCE		
	Models		Trivial	RFc	Archit.	LSTM	LSTM-SW	RFr	Archit.	LSTM	LSTM	LSTM-SW	
MSE	Subsets	PD	0.10	0.10	A(c)	0.09	0.10	0.09	A(r)	0.08	0.08	0.08	10^{-5}
		HPD	0.11	0.11	C(c)	0.11	0.11	0.09	C(r)	0.08	0.08	0.09	
		EPD	0.07	0.07	C(c)	0.07	0.08	0.06	C(r)	0.05	0.05	0.06	
		TD	0.12	0.12	A(c)	0.12	0.12	0.10	A(r)	0.09	0.09	0.09	10^{-6}
		HTD	0.41	0.25	C(c)	0.18	0.16	0.13	C(r)	0.13	0.12	0.16	
		ETD	0.09	0.09	C(c)	0.09	0.08	0.09	C(r)	0.08	0.07	0.08	
	Red. Subsets	RPD	0.23	0.20	A(c)	0.12	0.14	0.12	A(r)	0.10	0.08	0.09	10^{-4}
		RHPD	0.26	0.23	C(c)	0.13	0.14	0.13	C(r)	0.10	0.09	0.09	
		REPD	0.26	0.21	C(c)	0.13	0.15	0.12	C(r)	0.10	0.10	0.11	
		RTD	0.25	0.23	B(c)	0.25	0.24	0.14	B(r)	0.13	0.13	0.13	10^{-6}
		RETD	0.25	0.24	D(c)	0.24	0.26	0.15	D(r)	0.14	0.14	0.13	
ICC	Subsets	PD	0	0.10	A(c)	0.18	0.20	0.13	A(r)	0.18	0.20	0.22	10^{-5}
		HPD	0	0.16	C(c)	0.26	0.26	0.19	C(r)	0.27	0.28	0.27	
		EPD	0	0.16	C(c)	0.25	0.24	0.18	C(r)	0.24	0.27	0.28	
		TD	0	0.08	A(c)	0.07	0.08	0.10	A(r)	0.14	0.11	0.12	10^{-6}
		HTD	0	0.11	C(c)	0.19	0.20	0.17	C(r)	0.13	0.15	0.15	
		ETD	0	0.07	C(c)	0.09	0.11	0.09	C(r)	0.13	0.09	0.11	
	Red. Subsets	RPD	0	0.19	A(c)	0.57	0.49	0.23	A(r)	0.49	0.56	0.54	10^{-4}
		RHPD	0	0.21	C(c)	0.56	0.55	0.26	C(r)	0.58	0.62	0.62	
		REPD	0	0.27	C(c)	0.56	0.52	0.32	C(r)	0.50	0.52	0.51	
		RTD	0	0.09	B(c)	0.05	0.08	0.05	B(r)	0.08	0.04	0.07	10^{-6}
		RETD	0	0.12	D(c)	0.15	0.10	0.15	D(r)	0.14	0.09	0.20	

Table 10
Recall % result of 7-Class and 4-Class pain intensity recognition tasks on testing set.

Subsets															
Phasic	PD (7-Class)							HPD (4-Class)				EPD (4-Class)			
Model	BL	PH1	PH2	PH3	PE1	PE2	PE3	BL	PH1	PH2	PH3	BL	PE1	PE2	PE3
Trivial	100	0	0	0	0	0	0	100	0	0	0	100	0	0	0
RFc	97.7	0.1	0.2	6.2	0.2	0.2	11.9	97.5	0.6	1.3	16.2	98.2	0.8	0.4	17.0
LSTM	98.9	0	0	17.7	0	0.1	19.2	97.8	0.8	1.3	27.7	99.1	0	0	27.6
LSTM-SW	97.5	0	0	20.5	0.1	0.3	25.5	96.4	0.5	0.9	31.2	97.4	0.1	10.5	31.9
Subsets															
Tonic	TD (7-Class)							HTD (4-Class)				ETD (4-Class)			
Model	BL	TH1	TH2	PH3	TE1	TE2	TE3	BL	TH1	TH2	TH3	BL	TE1	TE2	TE3
Trivial	100	0	0	0	0	0	0	100	0	0	0	100	0	0	0
RFc	96.0	0.7	0.2	13.4	0.2	0.6	3.4	32.5	21.3	21.7	42.5	97.9	0.7	0.4	7.0
LSTM	99.7	0	0	9.6	0	0.1	0.5	1.8	49.3	26.6	46.5	98.6	0	0	14
LSTM-SW	98.1	0	0	15.4	0	0	3.0	0	51.4	32.7	42.1	99.6	0	0	12.6
Reduced Subsets															
Phasic	RPD (7-Class)							RHPD (4-Class)				REPD (4-Class)			
Model	BL	PH1	PH2	PH3	PE1	PE2	PE3	BL	PH1	PH2	PH3	BL	PE1	PE2	PE3
Trivial	100	0	0	0	0	0	0	100	0	0	0	100	0	0	0
RFc	85.9	3.0	3.7	11.7	4.4	3.9	27.1	81.2	11.2	11.3	27.1	78.5	11.7	10.5	38.1
LSTM	91.5	8.9	13.4	34.8	14.6	15.0	50.2	89.0	31.7	20.8	45.9	88.9	28.3	31.1	50.8
LSTM-SW	87.9	9.2	9.0	37.1	12.7	13.6	47.9	84.7	25.2	23.0	49.2	82.7	28.9	24.6	52.1
Reduced Subsets															
Tonic	RTD (7-Class)							RHTD (4-Class)				RETD (4-Class)			
Model	BL	TH1	TH2	TH3	TE1	TE2	TE3	BL	TH1	TH2	TH3	BL	TE1	TE2	TE3
Trivial	100	0	0	0	0	0	0	–	–	–	–	100	0	0	0
RFc	75.6	8.3	4.2	20.6	2.6	3.6	10	–	–	–	–	75.8	5.6	8.7	21.9
LSTM	95.1	0	0.1	21.4	0	0	5.8	–	–	–	–	94.3	0.6	2.0	22.4
LSTM-SW	85.9	0	3.8	24.0	0	0.1	18.0	–	–	–	–	89.3	0	0.5	32.1

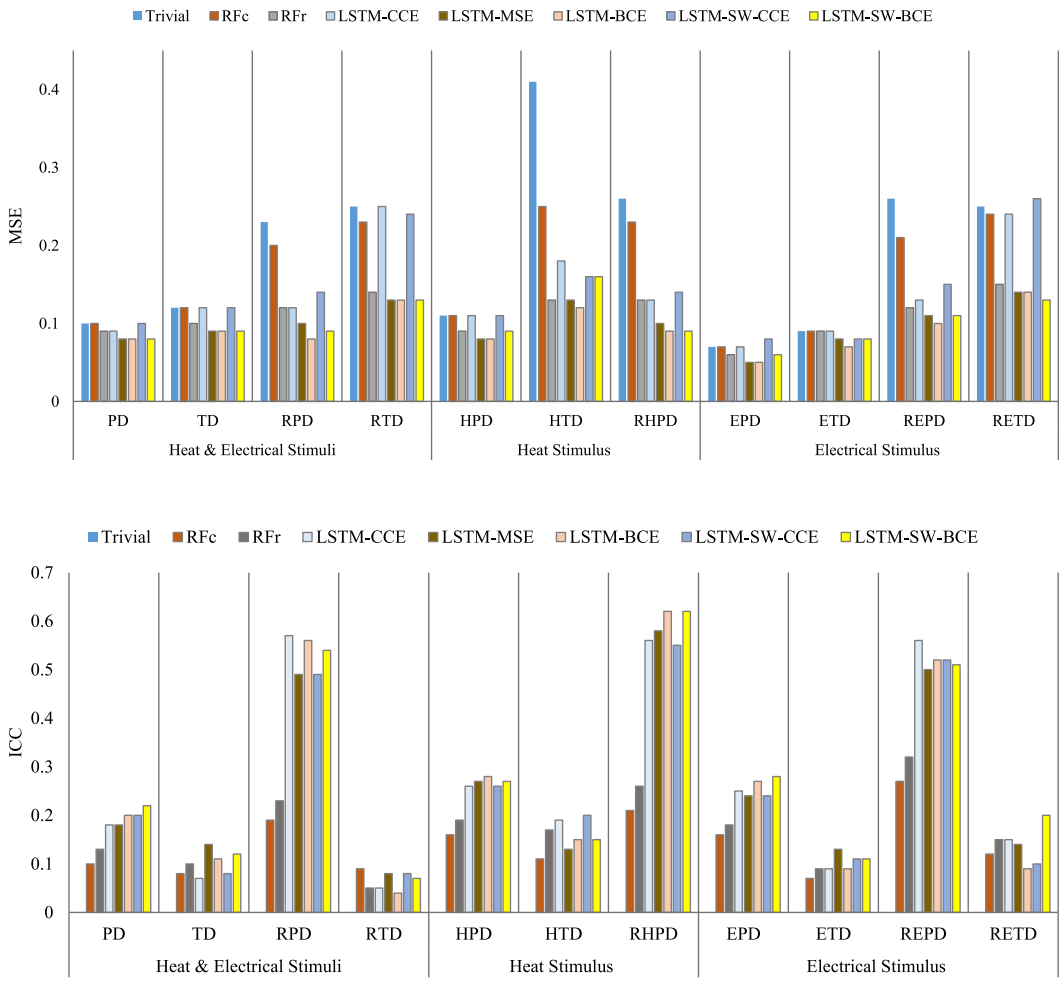

Fig. 3. Comparison of the Trivial and baseline method (RFc) with proposed LSTM and LSTM-SW methods to predict discrete and continuous pain level using regression measures.

pain intensity from sample sequences based on accuracy and precision measurements. Further, most LSTM-SW models were the best in continuously recognising pain intensity based on recall and F1 score measurements. The sample weighting method [5] was good with some datasets to reduce the influence of unimportant samples (noise) and increased pain intensity recognition performance in sequences. RFc was the best in predicting correct pain intensity from reduced tonic datasets (RTD and RETD) because it performs well with very small datasets.

6. Conclusion

As showed previously in this paper [Table 2], the database was extremely imbalanced; moreover, there were a lot of outliers or label noise because of the inconsistencies between the label and the video when some samples show a lack or extreme facial responses. Thus, we reduced the impact of the imbalanced datasets problem by reducing no pain samples using our strategy (see Section 3.5) and [9], the obtained results were good and indicate that most of the deleted samples are probably outliers. Fig. 4 shows how the proposed models performed. In regards to classification and regression, this paper reports the comparison between three continuous pain intensity recognition methods using Facial Activity Descriptors (FAD) on 11 datasets proposed from the X-ITE Pain Database.

The suggested methods were the Random Forest (RF), Long-Short Term Memory (LSTM), and LSTM using the sample weighting method (called LSTM-SW). We obtained good results indicating that most of the deleted samples were outliers. All methods were significantly better than guessing. LSTM and LSTM-SW outperformed RF (Random Forest classifier (RFc) and Random Forest regression (RFr)) with most datasets. RF models were the best with small datasets such as Reduced Tonic Dataset (RTD). Further, LSTM classification models performed the best with almost balanced datasets such as Heat Tonic Dataset (HTD) and reduced phasic datasets (RPD and REPD) because the huge imbalance tends to decrease the performance. Both classification and regression methods performed well. The presented results in this paper were promising.

The results showed that it is possible to provide automatic and continuous monitoring of pain intensity in patients. We believe that a larger dataset with more pain intensities is necessary for more reliable automatic monitoring of continuous pain intensity. Thus, we assume that the LSTM or other deep learning methods will surpass our proposed methods. In addition, for future research, continuous multimodal monitoring models of pain intensity are likely to be more reliable due to literature studies [29], such as combining facial expression modality with Electromyogram [EMG] and Electrodermal Activity [EDA] modalities.

To further improve the results, we plan to investigate the use of weighted cross-entropy, oversampling technique [57,58], and undersampling techniques in [59,60] to address the problem of imbalanced datasets. The obtained results will be compared with those with LSTM-SW in this work to introduce the best method.

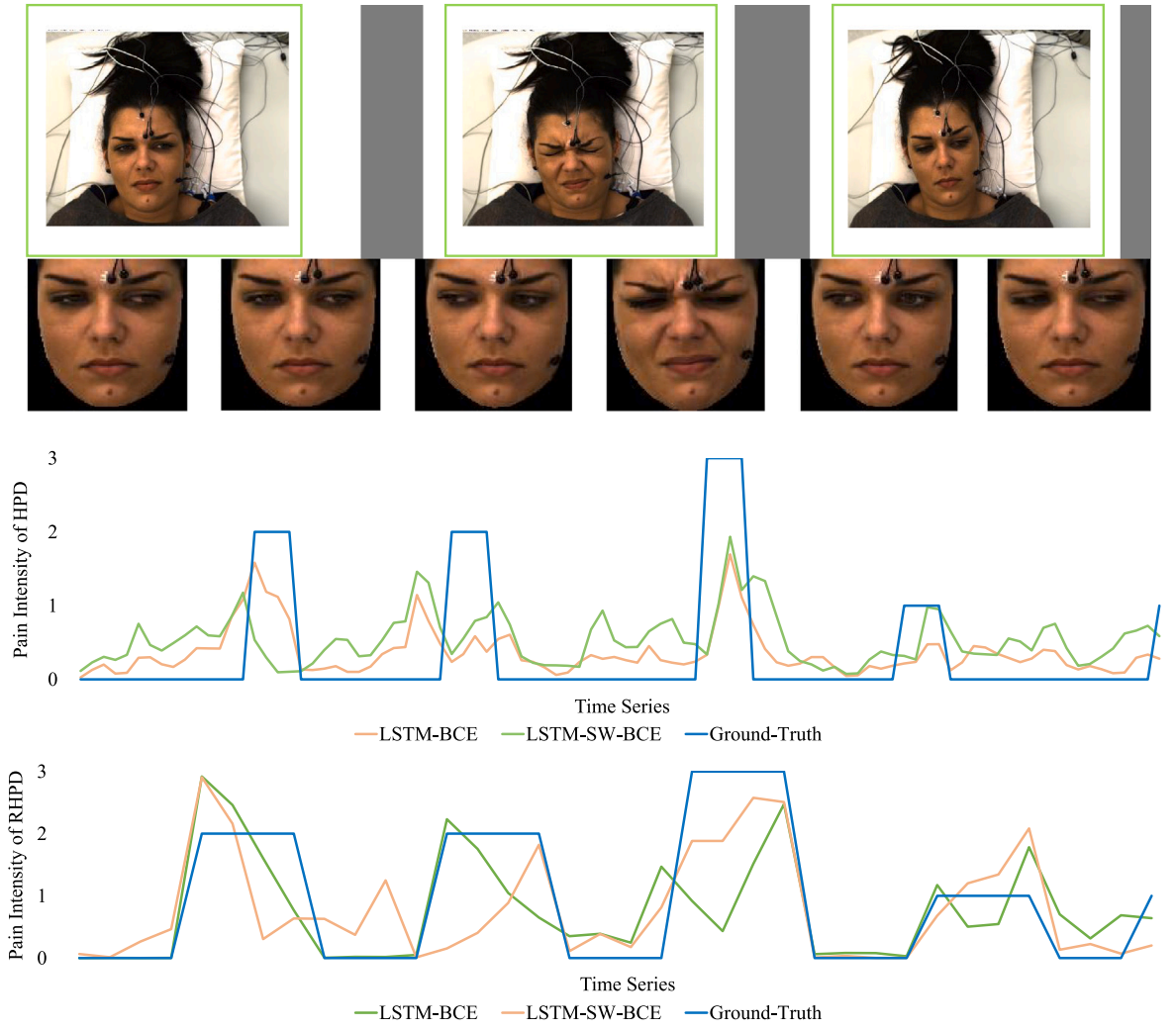

Fig. 4. An example sequence of pain intensity recognition for LSTM-BCE, LSTM-SW-BCE, and Ground-Truth.

Abbreviations:

The following abbreviations are used in this manuscript:

Archit.	Architecture
BCE	Binary Cross-Entropy loss function
CCE	Categorical Cross-Entropy loss function
CNN	simple Convolutional Neural Network
ECG	Electrocardiogram
EDA	Electrodermal activity
EMG	Electromyography
EPD	Heat Phasic Dataset
ETD	Electrical Tonic Dataset
FAD	Facial Activity Descriptor
HPD	Heat Phasic Dataset
HTD	Heat Tonic Dataset
ICC	Intraclass Correlation Coefficient
LSTM	Long-Short Term Memory
LSTM-SW	Long-Short Term Memory using sample weighting
Loss	Loss function
lr	learning rate
Meas.	Measures
MSE	Mean Squared Error
PD	Phasic Dataset
PE1	Phasic electrical stimulus with low pain intensity
PE2	Phasic electrical stimulus with moderate pain intensity
PE3	Phasic electrical stimulus with severe pain intensity
PH1	Phasic heat stimulus with low pain intensity
PH2	Phasic heat stimulus with moderate pain intensity
PH3	Phasic heat stimulus with severe pain intensity
Red. Subsets	Reduced Subsets
REPD	Reduced Electrical Phasic Dataset
RETD	Reduced Electrical Tonic Dataset
RHPD	Reduced Heat Phasic Dataset
RF	Random Forest
RFc	Random Forest classifier
RFr	Random Forest regression
RPD	Reduced Phasic Dataset
RTD	Reduced Tonic Dataset
TD	Tonic Dataset
TE1	Tonic electrical stimulus with low pain intensity
TE2	Tonic electrical stimulus with moderate pain intensity
TE3	Tonic electrical stimulus with severe pain intensity
TH1	Tonic heat stimulus with low pain intensity
TH2	Tonic sic heat stimulus with moderate pain intensity
TH3	Tonic heat stimulus with severe pain intensity

CRediT authorship contribution statement

Ehsan Othman: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Philipp Werner:** Conceptualization, Writing – original draft, Writing – review & editing. **Frerk Saxen:** Conceptualization, Writing – review & editing. **Ayoub Al-Hamadi:** Conceptualization, Supervision, Writing – review & editing. **Sascha Gruss:** Conceptualization, Data gathering, Writing – review & editing. **Steffen Walter:** Conceptualization, Data gathering, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request from [Sascha.Gruss@uni-ulm.de] or [Steffen.Walter@uni-ulm.de].

Acknowledgements

This research was supported by German Academic Exchange Service (DAAD) and the Federal Ministry of Education and Research of Germany (BMBF); projects Robo-Lab no. 03ZZ04X02B and German Research Foundation (DFG) under grants AI 638/13-1 and AI 638/15-1.

Institutional review board statement

The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of the Ulm University (protocol code: 372/16, date of approval: 5 January 2017).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jvcir.2022.103743.

References

[1] P. Werner, D. Lopez-Martinez, S. Walter, A. Al-Hamadi, S. Gruss, R.W. Picard, Automatic recognition methods supporting pain assessment: A survey, Trans. Affect. Comput. (2019) http://dx.doi.org/10.1109/TAFFC.2019.2946774.

[2] K. Herr, P.J. Coyne, M. McCaffery, R. Manworren, S. Merkel, Pain assessment in the patient unable to self-report: Position statement with clinical practice recommendations, Pain Manage. Nurs.: Off. J. Am. Soc. Pain Manage. Nurses 12 (4) (2011) 230–250, http://dx.doi.org/10.1016/j.pmn.2011.10.002.

[3] K.D. Craig, The social communication model of pain, Canad. Psychol. 50 (1) (2009) 22–32, http://dx.doi.org/10.1016/j.pmn.2011.10.002.

[4] K.D. Craig, The facial expression of pain better than a thousand words? APS J. 1 (3) (1992) 153–162, http://dx.doi.org/10.1016/1058-9139(92)90001-S.

[5] E. Othman, P. Werner, F. Saxen, A. Al-Hamadi, S. Gruss, S. Walter, Automatic. vs, Human recognition of pain intensity from facial expression on the X-ITE pain database, Sensors 21 (9) (2021) http://dx.doi.org/10.3390/s21093273.

[6] A.C. d. C Williams, Facial expression of pain: An evolutionary account, Behav. Brain Sci. 25 (4) (2002) 439–455, http://dx.doi.org/10.1017/S0140525X02000080.

[7] P.Werner A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, H.C. Traue, Automatic pain assessment with facial activity descriptors, Trans. Affect. Comput. 8 (3) (2017) 286–299, http://dx.doi.org/10.1109/TAFFC.2016.2537327.

[8] G. Pasqualetti, G. Gori, C. Blandizzi, M.D. Tacca, Healthy volunteers and early phases of clinical experimentation, Eur. J. Clin. Pharmacol. 66 (7) (2010) 647–653, http://dx.doi.org/10.1007/s00228-010-0827-0.

[9] E. Othman, P. Werner, F. Saxen, A. Al-Hamadi, S. Walter, Regression networks for automatic pain intensity recognition in video using facial expression on the X-ITE pain database, in: The 25th Int'L Conf on Image Processing, Computer Vision & Pattern Recognition (IPCV'21), Las vegas, USA, 2021.

[10] P. Werner, A. Al-Hamadi, S. Niese, S. Walter, H.C. Gruss, H.C. Traue, Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges, in: Proceedings of the British Machine Vision Conference, UK, 2013, http://dx.doi.org/10.5244/C.27.119.

[11] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.

[12] S. Zhou, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[13] F. Gers, F. Cummins, J. Schmidhuber, Learning to forget: Continual prediction with LSTM, Neural Comput. 12 (10) (2000) http://dx.doi.org/10.1162/089976600300015015.

[14] S. Brahnam, C.-F. Chuang, F.Y. Shih, M.R. Slack, SVM classification of neonatal facial imagesof pain, in: The Fuzzy Logic and Applications, 6th International Workshop, Crema, Italy, 15-17 September, 2005, http://dx.doi.org/10.1007/11676935_15.

[15] P. Lucey, J.F. Cohn, K.M. Prkachin, P.E. Solomon, I. Matthews, Painful data: The UNBC-mcmaster shoulder pain expression archive database, in: International Conference on Automatic Face & Gesture Recognition (FG) Pain, Santa Barbara, CA, USA, 2011, pp. 57–64, http://dx.doi.org/10.1109/FG.2011.5771462.

[16] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H.C. Traue, P. Werner, A. Al-Hamadi, S. Crawcour, A.O. Andrade, G.M. d. Silva, The BioVid heat pain database: Data for the advancement and systematic validation of an automated pain recognition system, in: The Cybernetics (CYBCONF), Lausanne, Switzerland, 2013, http://dx.doi.org/10.1109/CYBConf.2013.6617456.

[17] X. Zhanga, L. Yin, J.F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, J.M. Girard, BP4D-spontaneous: A high-resolution spontaneous 3D dynamic facial expression database, Image Vis. Comput. 32 (10) (2014) http://dx.doi.org/10.1016/j.imavis.

[18] D. Harrison, M. Sampson, J. Reszel, K. Abdulla, N. Barrowman, J. Cumber, A. Fuller, C. Li, S. Nicholls, C.M. Pound, Too many crying babies: A systematic review of pain management practices during immunizations on YouTube, BMC Pediatr. 14 (134) (2014) http://dx.doi.org/10.1186/1471-2431-14-134.

[19] Z. Zhang, J.M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J.F. Cohn, Q. Ji, L. Yin, Multimodal spontaneous emotion corpus for human behavior analysis, in: The Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, 2016, http://dx.doi.org/10.1109/CVPR.2016.374.

[20] V.K. Mittal, Discriminating the infant cry sounds due to pain vs. Discomfort towards assisted clinical diagnosis, in: The SLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies, San Francisco, USA, 2016, http://dx.doi.org/10.21437/SLPAT.2016-7.

[21] M. Velana, S. Gruss, G. Layher, P. Thiam, Y. Zhang, D. Schork, V. Kessler, S. Meudt, H. Neumann, J. Kim, F. Schwenker, E. André, H.C. Traue, S. Walter, The SenseEmotion database: A multimodal database for the development and systematic validation of an automatic pain- and emotion-recognition system, in: The Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction, Vol. 4, Cancun, Mexico, 2016, http://dx.doi.org/10.1007/978-3-319-59259-6_11.

[22] M.S.H. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, A.C. Elkins, N. Kanakam, A. d. Rothschild, N. Tyler, P.J. Watson, A.C. d. C. Williams, M. Pantic, N.Bianchi-Berthouze, The automatic detection of chronic pain-related expression: Requirements, challenges and the multimodal EmoPain dataset, IEEE Trans. Affect. Comput. 7 (4) (2016) 435–451, http://dx.doi.org/10.1109/TAFFC.2015.2462830.

[23] M.A. Haque, R.B. Bautista, F. Noroozi, K. Kulkarni, C.B. Laursen, R. Irani, M. Bellantonio, S. Escalera, G. Anbarjafari, K. Nasrollahi, O.K. Andersen, E.G. Spaich, T.B. Moeslund, Deep multimodal pain recognition: A database and comparison of spatio-temporal visual modalities, in: The International Conference on Automatic Face & Gesture Recognition, Xi'an, China, 2018, http://dx.doi.org/10.1109/FG.2018.00044.

[24] S. Gruss, M. Geiger, P. Werner, O. Wilhelm, H.C. Traue, A. Al-Hamadi, S. Walter, Multi-modal signals for analyzing pain responses to thermal and electrical stimuli, J. Vis. Exp. 146 (2019) e59057, http://dx.doi.org/10.3791/59057.

[25] P. Thiam, V. Kessler, F. Schwenker, Hierarchical combination of video features for personalised pain level recognition, in: European Symposium on Artificial Neural Networks, ESANN, Bruges, Belgium, 2017, pp. 465–470.

[26] P. Thiam, F. Schwenker, Combining deep and hand-crafted features for audio-based pain intensity classification, in: Multimodal Pattern Recognition of Social Signals in Human–Computer-Interaction, Bejing, China, 2018, pp. 49–58.

[27] F.-S. Tsai, Y.-L. Hsu, W.-C. Chen, Y.-M. Weng, C.-J. Ng, C.-C. Lee, Toward development and evaluation of pain level-rating scale for emergency triage based on vocal characteristics and facial expressions, in: The Interspeech, San Francisco, CA, USA, 2016, pp. 49–58, http://dx.doi.org/10.21437/Interspeech.2016-408.

[28] Y. Chu, X. Zhao, J. Han, Y. Su, Physiological signal-based method for measurement of pain intensity, Front. Neurosci. 11 (279) (2017) http://dx.doi.org/10.3389/fnins.2017.00279.

[29] S. Walter, S. Gruss, K. Limbrecht-Ecklundt, H.C. Traue, Automatic pain quantification using autonomic parameters, Psychol. Neurosci. 7 (3) (2017) 363–380, http://dx.doi.org/10.3922/j.psns.2014.041.

[30] D. Lopez-Martinez, R. Picard, Continuous pain intensity estimation from autonomic signals with recurrent neural networks, in: Presented at the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, Hawaii, USA, 2018, pp. 5624–5627, http://dx.doi.org/10.1109/EMBC.2018.8513575.

[31] P. Werner, A. Al-Hamadi, S. Gruss, S. Walter, Twofold-multimodal pain recognition with the X-ITE pain database, in: Workshops and Demos (ACIIW) International Conference on Affective Computing and Intelligent Interaction, Cambridge, UK, 2019, http://dx.doi.org/10.1109/ACIIW.2019.8925061.

[32] P. Thiam, F. Schwenker, Multi-modal data fusion for pain intensity assessment and classification, in: Image Processing Theory, Tools and Applications, IPTA, Montreal, QC, Canada, 2017, http://dx.doi.org/10.1109/IPTA.2017.8310115.

[33] P. Thiam, P. Bellmann, H.A. Kestler, F. Schwenker, Exploring deep physiological models for nociceptive pain recognition, Sensors 19 (20) (2017) http://dx.doi.org/10.3390/s19204503.

[34] M.S. Salekin, G. Zamzmi, J. Hausmann, D. Goldgof, R. Kasturi, M. Kneusel, T. Ashmeade, T. Ho, Y.Suna, Multimodal neonatal procedural and postoperative pain assessment dataset, Comput. Biol. Med. 129 (2021) 104150, http://dx.doi.org/10.1016/j.compbiomed.2020.104150.

[35] P. Ekman, W.V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, Palo Alto, CA, USA, 1978, http://dx.doi.org/10.2217/nmt-2015-0006.

[36] M. Kächele, P. Thiam, M. Amirian, P. Werner, S. Walter, F. Schwenker, G. Palm, Multimodal data fusion for person-independent, continuous estimation of pain intensity, in: The 16th International Engineering Applications of Neural Networks Conference, Rhodes, Greece, 2015, http://dx.doi.org/10.1007/978-3-319-23983-5_26.

[37] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: Application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 2037–2041, http://dx.doi.org/10.1109/TPAMI.2006.244.

[38] V. Ojansivu, J. Heikkilä, Blur insensitive texture classification using local phase quantization, in: The Image and Signal Processing - 3rd International Conference, Octeville, France, 2008, http://dx.doi.org/10.1007/978-3-540-69905-7_27.

[39] J. Kannala, E. Rahtu, BSIF: Binarized statistical image features, in: The Proceedings of the 21st International Conference on Pattern Recognition, Tsukuba, Japan, 2012, pp. 1363–1366.

[40] R. Yang, S. Tong, M. Bordallo, E. Boutellaa, J. Peng, X. Feng, A. Hadid, On pain assessment from facial videos using spatio-temporal local descriptors, in: The 6th International Conference on Image Processing Theory, Tools and Applications, IPTA, Oulu, Finland, 2016, http://dx.doi.org/10.1109/IPTA.2016.7820930.

[41] B. YJiang, M.F. Valstar, M. Pantic, Action unit detection using sparse appearance descriptors in space–time video volumes, in: The Automatic Face & Gesture Recognition and Workshops, FG, Santa Barbara, CA, USA, 2011, pp. 314–321, http://dx.doi.org/10.1109/FG.2011.5771416.

[42] S.R. Arashloo, J. Kittler, Dynamic texture recognition using multiscale binarized statistical image features, IEEE Trans. Multimed. 16 (8) (2014) 2099–2109, http://dx.doi.org/10.1109/TMM.2014.2362855.

[43] J. Chen, Z. Chi, H. Fu, A new framework with multiple tasks for detecting and locating pain events in video, Comput. Vis. Image Understand. 155 (2017) 113–123, http://dx.doi.org/10.1016/j.cviu.2016.11.003.

[44] E. Othman, P. Werner, F. Saxen, A. Al-Hamadi, S. Walter, Cross-database evaluation of pain recognition from facial video, in: International Symposium on Image and Signal Processing and Analysis, ISPA, Dubrovnik, Croatia, 2019, http://dx.doi.org/10.1109/ISPA.2019.8868562.

[45] P. Thiam, V. Kessler, M. Amirian, P. Bellmann, G. Layher, Y. Zhang, M. Velana, S. Gruss, S. Walter, H.C. Traue, J. Kim, D. Schork, E. Andre, H. Neumann, F. Schwenker, Multi-modal pain intensity recognition based on the SenseEmotion database, IEEE Trans. Affect. Comput. (2019) http://dx.doi.org/10.1109/TAFFC.2019.2892090.

[46] F. Wang, X. Xiang, C. Liu, T.D. Tran, A. Reiter, G.D. Hager, H. Quon, J. Cheng, A.L. Yuille, Regularizing face verification nets for pain intensity regression, in: At the IEEE International Conference on Image Processing, Beijing, China, 2017, http://dx.doi.org/10.1109/ICIP.2017.8296449.

[47] P. Rodriguez, G. Cucurull, J. Gonz'alez, J.M. Gonfaus, Deep pain: Exploiting long short-term memory networks for facial expression classification, IEEE Trans. Cybern. PP (99) (2017) 1–11, http://dx.doi.org/10.1109/TCYB.2017.2662199.

[48] Y. Lecun, K. Kavukcuoglu, C. Farabet, Regularizing face verification nets for pain intensity regression, in: The IEEE International Conference on Image Processing, Paris, France, 2010, http://dx.doi.org/10.1109/ISCAS.2010.5537907.

[49] N. Kalischek, P. Thiam, P. Bellmann, F. Schwenker, Deep domain adaptation for facial expression analysis, in: The 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, Cambridge, United Kingdom, UK, 2019, pp. 317–323, http://dx.doi.org/10.1109/ACIIW.2019.8925055.

[50] G. Bargshady, J. Soar, X. Zhou, R.C. Deo, F. Whittaker, H. Wang, A joint deep neural network model for pain recognition from face, in: The 4th International Conference on Computer and Communication Systems, Singapore, 2019, http://dx.doi.org/10.1109/CCOMS.2019.8821779.

[51] J. Soar, G. Bargshady, X. Zhou, F. Whittaker, Deep learning model for detection of pain intensity from facial expression, in: The International Conference on Smart Homes and Health Telematics, Singapore, 2018, pp. 249–254, http://dx.doi.org/10.1007/978-3-319-94523-1_22.

[52] P. Thiam, H.A. Kestler, F. Schwenker, Two-stream attention network for pain recognition from video sequences, Sensors 20 (3) (2021) http://dx.doi.org/10.3390/s20030839.

[53] P. Werner, A. Al-Hamadi, S. Walter, Analysis of facial expressiveness during experimentally induced heat pain, in: Presented at the Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW, San Antonio, TX, USA, 2017, http://dx.doi.org/10.1109/ACIIW.2017.8272610.

[54] T. Baltrusaitis, P. Robinson, L.-P. Morency, OpenFace: An open source facial behavior analysis toolkit, in: Winter Conference on Applications of Computer Vision, WACV, Lake Placid, NY, USA, 2016, http://dx.doi.org/10.1109/WACV.2016.7477553.

[55] E. Othman, F. Saxen, D. Bershadskyy, P. Werner, A. Al-Hamadi, J. Weimann, Predicting the group contribution behaviour in a public goods game from Face-to-Face communication, Sensors 19 (12) (2019) http://dx.doi.org/10.3390/s19122786.

[56] P.E. Shrout, J.L. Fleiss, Intraclass correlations: Uses in assessing rater reliability, Psychol. Bull. 86 (2) (1979) 420–428, http://dx.doi.org/10.1037//0033-2909.86.2.420.

[57] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357, http://dx.doi.org/10.1613/jair.953.

[58] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, in: IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 2008, pp. 1322–1328.

[59] S.J. Yen, Y.S. Lee, Cluster-based under-sampling approaches for imbalanced data distributions, Expert Syst. Appl. 36 (3) (2009) 5718–5727, http://dx.doi.org/10.1016/j.eswa.2008.06.108.

[60] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, in: Icml, Vol. 97, 1997, p. 179, (1).