

Subject-independent Pain Recognition using Physiological Signals and Para-linguistic Vocalizations

Nadeen Shoukry, Omar Elkilany, Patrick Thiam, Viktor Kessler and Friedhelm Schwenker^a
Ulm University, Institute of Neural Information Processing, 89081 Ulm, Germany

Keywords: Machine Learning, Pain Recognition, Affective Computing, Biopotential Signals, Para-linguistic Data.

Abstract: Pain is the result of a complex interaction among the various parts of the human nervous system. It plays an important role in the diagnosis and treatment of patients. The standard method for pain recognition is self-report; however, not all patients can communicate pain effectively. In this work, the task of automated pain recognition is addressed using para-linguistic and physiological data. Hand-crafted and automatically generated features are extracted and evaluated independently. Several state-of-the-art machine learning algorithms are applied to perform subject-independent binary classification. The *SenseEmotion* dataset is used for evaluation and comparison. Random forests trained on hand-crafted features from the physiological modalities achieved an accuracy of 82.61%, while support vector machines trained on hand-crafted features from the para-linguistic data achieved an accuracy of 63.86%. Hand-crafted features outperformed automatically generated features.

1 INTRODUCTION


Pain is the result of a complex interaction among the various parts of the nervous system and plays an important role in the diagnosis and treatment of patients. It varies in duration, intensity, and meaning. For example, the sudden pain associated with syndromes, such as causalgia, can be contrasted with the progressive pain related to some types of cancer (Turk et al., 1983). Furthermore, there are physical and psychological causes of pain. While physical pain has concrete causes, psychogenic pain occurs without external stimuli (Engel, 1959). These factors complicate the pain assessment task.

The standard method for pain assessment is self-reporting, which assumes that the subjects experiencing the pain are capable of comprehending the task and communicating their feelings effectively (Zwakhlen et al., 2006). This assumption, however, does not hold with elderly patients suffering from dementia or cognitive disabilities. For these patients, pain may become a barrier to social inclusion (Velana et al., 2016). Moreover, if it is not managed correctly, it may induce pathological effects, such as higher blood pressure and rapid heart rates, that may lead to further complications (Turk and Gatchel, 2018). With

the growth in the proportion of the population above the age of 65 in modern industrial societies, more reliable pain recognition and assessment methods are required, possibly not to replace but to aid self-reports (Velana et al., 2016).

In this paper, two different approaches for automatic pain recognition are explored. The first approach relies on physiological signals, collected through sensors attached to the bodies of patients, while the second approach depends on para-linguistic vocalizations, such as moans and similar sounds, which do not belong to a standardized language. State-of-the-art machine learning algorithms are applied to perform binary, subject-independent classification, and their accuracies are compared briefly. Leave-one-subject-out cross-validation is used as the evaluation approach. Assessment is performed on the recently recorded *SenseEmotion* dataset (Velana et al., 2016).

The rest of this work is organized as follows: In section 2, an overview of related work is presented. In section 3, the dataset itself is described with more focus on the physiological and para-linguistic modalities. The data preprocessing and preparation steps are described in section 4. The conducted experiments and their results are provided in section 5. Finally, this work is concluded in section 6.

^a  <https://orcid.org/0000-0001-5118-0812>

2 RELATED WORK

This section provides an overview of the research related to the pain recognition task with more focus on the two relevant modalities: physiological signals and para-linguistic vocalizations.

Accurate pain recognition requires complex analysis and a consideration of the environment where the pain occurs (Hammal and Cohn, 2014). With this fact in mind, several databases have been developed to support research on pain recognition. One of the first of these databases is the *UNBC-McMaster Shoulder Pain Expression Archive Database*, which comprises data from patients suffering from shoulder pain (Lucey et al., 2011). To collect the data, patients were asked to perform exercises using their limbs, and their facial expressions were captured and annotated. The database was limited to this modality.

The *EmoPain Dataset* is another pain recognition database, which focused on chronic pain and included data from both patients and healthy individuals (Aung et al., 2016). To collect the data, patients performed exercises in a rehabilitation setting. This database contained multi-modal data, including multiple-view face videos, audio signals, 3D motion capture data, and electromyographic signals.

The *BioVid Heat Pain Database* provided visual and biopotential data to advance the recognition of acute pain (Walter et al., 2013). Similarly, the *SenseEmotion* dataset provided a multi-modal database on acute pain, which was artificially induced using heat stimulation in healthy patients (Velana et al., 2016).

2.1 Para-linguistic Data

Most of the previous machine learning projects done in automatic pain recognition relied on facial expressions, bio-potential signals, and in some cases the fusion of both data modalities (Thiam and Schwenker, 2017; Thiam et al., 2017; Sellner et al., 2018; Schwenker, 2018). In a single study, para-linguistic audio signals were used along with other data modalities to train pain recognition models (Thiam et al., 2019). However, para-linguistic audio signals were not used on their own to recognize pain before.

Environmental and background sound classification projects used hand-crafted features of audio signals, such as Mel Frequency Cepstral Coefficients (MFCC). Support vector machines and random forests were trained using the audio features to perform the classification task (Saki and Kehtarnavaz, 2014; Wang et al., 2008; Lu et al., 2003).

There were previous attempts to use infant para-

linguistic vocalizations (cries) to recognize pain, non-pain, fear, and hunger. Instead of manually extracting audio features, deep neural networks were trained on the audio signals to extract features automatically and perform the classification (Chang and Li, 2016). Other studies used hand-crafted features to train deep neural networks (Abdulaziz and Ahmad, 2010; Petroni et al., 1995). Another study used raw audio data to train a convolutional neural network (CNN) and manually extracted audio features to train a fuzzy support vector machine (FSVM); in this last study, the FSVM outperforms the CNN and requires significantly less amount of data to achieve better results (Barajas-Montiel and Reyes-García, 2006).

2.2 Physiological Signals

Several medical studies have established correlations between physiological responses produced by the autonomic nervous system and the experience of pain (Ledowski et al., 2009; Colloca et al., 2006; Loggia et al., 2011). Many of these responses have been measured and used as potential indicators of affective state. They include heart rate, diastolic and systolic blood pressure, pupil dilation, pulse, respiration, skin conductance and color, and temperature (Picard, 2000).

Several studies have used electrocardiography, electromyography, and electrodermal activity to address the pain recognition task (Gruss et al., 2015; Thiam et al., 2019; Kächele et al., 2016). Signal processing and extensive feature extraction were used to produce low-dimensional representations of the information contained in these signals, and the extracted features were fed to support vector machines or random forests for classification. These three signals were also used to develop real-time pain recognition systems, which are capable of outputting personalized, continuous estimations of pain intensity (Kächele et al., 2016). Furthermore, unaffected by differences among the individuals experiencing pain, electrodermal activity has been found to be the best performing single modality for pain recognition (Thiam et al., 2019).

3 THE DATASET

As described by Velana et al., the *SenseEmotion* dataset comprises multi-modal data from 45 healthy participants. Pain was elicited in these participants using thermal stimuli ranging from 32 to 50.5 degrees Celsius. A baseline temperature T_0 of 32 degrees Celsius was used commonly in all of the experiments. At

the start of an experimental session, calibration was conducted to determine the pain threshold T_1 and the tolerance threshold T_3 of each individual participant, where the pain threshold marks the point where the participant perceives pain and the tolerance threshold marks the point where the participant can no longer bear the pain. The mean of T_1 and T_3 was used to define another level labelled T_2 . The resulting dataset consists of 120 events per participant, 30 for each heat level (Velana et al., 2016).

In this paper, the no-pain state corresponds to the baseline temperature T_0 , while the pain state corresponds to the pain tolerance threshold T_3 .

3.1 The Para-linguistic Modality

The audio signals used consist of para-linguistic data, which are non-verbal signals of speech, including breathing, moaning, and sighing sounds, beyond the pure transcriptional contents of spoken speech (Cai et al., 2017).

The recording of the audio signals relied primarily on a digital wireless headset microphone (Line6 XD-V75HS) combined with a directional microphone (Rode M3). The wireless headset was crucial to allow free head movements and record any sounds produced by the participants. The directional microphone was used to record the ambient acoustic noises surrounding the participants. Another audio signal was recorded from the Microsoft Kinect V2 integrated microphone, which was able to capture ambient noises as well. All recordings were sampled at a rate of 48kHz and synchronized with both the video and bio-physiological streams using the SSI framework (Wagner et al., 2013).

3.2 The Physiological Modalities

The physiological modalities were recorded using the Social Signal Interpretation (SSI) framework synchronously in real-time (Wagner et al., 2013). The following subsections provide a brief description of each signal.

3.2.1 Electrocardiography

Electrocardiography is used to trace the electric current that the heart muscle generates during a heartbeat. It provides information about the condition and performance of the heart. This modality can be used to extract features such as heart rate, inter-beat interval, and heart rate variability. Heart rate has been used to differentiate between positive and negative emotions, with finer differentiation obtainable using

a measure of finger temperature (Kim and André, 2008).

3.2.2 Electromyography

Electromyography is the graphing of the electrical characteristics of muscles. Three electrodes were used to measure the activity of the upper-right trapezius muscle and collect data for the *SenseEmotion* dataset. Electrical muscle activity is a general indicator of arousal. Increased muscle activity corresponds to increased sympathetic nervous system activity. Furthermore, high muscle tension is expected to occur in the case of pain (Velana et al., 2016).

3.2.3 Electrodermal Activity

The psychogalvanic reflex, or galvanic skin response, is an alteration in the electrical properties of the skin that follows harmful or alerting stimuli. Because sweat glands, which affect skin conductivity, are controlled by the sympathetic branch of the autonomic nervous system, the galvanic skin response is more sensitive as an indicator of emotional arousal than other physiological responses (Öhman et al., 1993).

3.2.4 Respiration

The respiration signal was captured using an elastic belt, which was worn by the participants over their clothing in the thorax area. It is suggested that respiration patterns reflect relaxation and tension (Boiten et al., 1994). Acute, cutaneous pain stimuli have been shown to cause increases in inspiratory flow, which affects inspiration intensity, even under general anaesthesia (Jafari et al., 2017). Therefore, the respiration signal may provide valuable and consistent data for the pain recognition task.

4 DATA PREPARATION

From the data of the 45 participants of the *SenseEmotion* dataset, the data of 40 participants could be used. The other five participants were excluded because they had erroneous raw data, which were missing one or more sensor information in certain time segments. In this section, an overview of the data preparation steps is presented.

4.1 Para-linguistic Data

The audio signals that were preprocessed and used came from the wireless headset microphone alone because it was the only device able to capture sounds

produced by the participants at a satisfactory level; it was placed at the nasolabial area, nearer than the integrated Kinect V2 microphone and the directional microphone, which were placed at a distance of about one meter from the participants and captured ambient noises only (Thiam et al., 2019).

There are several representations of the audio signals: the raw audio recordings, the audio spectrograms, and the hand-crafted features. The audio spectrogram is a visual representation of the spectrum of frequencies of the audio signals as they vary with time (Mallik et al., 2019). The spectrogram is able to represent many useful acoustic features, including frequency, pitch, and dB (Chang and Li, 2016).

Each participant experienced 60 stimuli on each forearm throughout the duration of the experiment. During the application of each pain/non-pain stimulus, 10 frames of spectrograms were recorded. Samples of the data spectrograms are presented in figures 1 and 2.

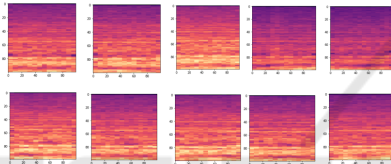


Figure 1: Sample of the data spectrograms (using the left forearm recorded data of participant 20160404_14_w during the application of stimulus 1 (pain stimulus)).

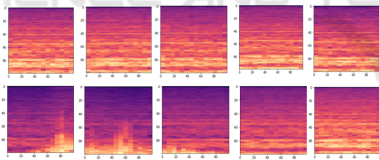


Figure 2: Sample of the data spectrograms (using the left forearm recorded data of participant 20160404_14_w during the application of stimulus 3 (non-pain stimulus)).

Obtaining hand-crafted features started with the processing of the data. Several low-level descriptors were extracted from the raw audio signal. Low-Level Descriptors (LLDs) are parameters computed from short time frames of a whole signal to significantly reduce the processed amount of data. LLDs were obtained using the openSMILE feature extraction toolkit (Thiam et al., 2019). The most commonly used LLDs in speech processing are the Mel Frequency Cepstral Coefficients (MFCCs).

MFCCs have been shown to work well for structured sounds, such as speech and music, but their performance tends to drop in the presence of noise; therefore, other LLDs are analyzed along with the MFCCs.

Obtaining MFCCs is basically the process of converting audio in the time domain to the frequency domain to be easily interpreted, analyzed, and processed. One can think of it as mimicking the human ear cochlea, which has few filters at high frequencies and more filters at low frequencies to allow for the detection of the most quiet sounds possible (Dave, 2013).

MFCCs are known to provide clear insights about sound data and outstanding results in speech recognition, emotion recognition, and speaker identification tasks (Dave, 2013). For the present work, 13 MFCCs were extracted, each combined with its first and second order temporal derivatives, resulting in a total of 39 MFCC-based LLDs. In addition to the MFCCs, another set of LLDs was computed using the Relative Spectral Perceptual Linear Predictive Coding (RASTA-PLP), which is an extension to the perceptual linear predictive analysis that improves the robustness of the computed coefficients against distortion. For the present work, 6 RASTA-PLP coefficients were extracted, each in combination with its first and second order temporal derivatives, resulting in a total of 18 RASTAPLP-based LLDs. Finally, a third set of LLDs was computed which involved the root mean square signal energy and logarithmic signal energy, in combination with their first and second order temporal derivatives. Additionally, the following descriptors were extracted: the loudness contour, the zero-crossing rate, the mean-crossing rate, the maximum absolute sample value, the minimum and maximum sample values, and the arithmetic mean of the sample values. This last set represents a total of 13 LLDs (Thiam et al., 2019).

Then the resulting signals were further processed to substantially reduce the noise using band-pass filtering, signal smoothing, and detrending. After these operations, a set of high-level descriptors (HLDs) was extracted from the previously processed signals. HLDs are extracted from segmenting the LLDs based on a fixed window. In the current work, the following set of 14 statistical functions is applied on the segmented LLD signals for the extraction of HLDs: mean, median, standard deviation, maximum, minimum, range, skewness, kurtosis, first and second quartiles, inter-quartile, 1%-percentile, 99%-percentile, and range from 1%- to 99%-percentile (Thiam et al., 2019).

The MFCC-based feature vectors have a total dimensionality of $14 \times 39 = 546$. The RASTA-PLP-based feature vectors have a total dimensionality of $14 \times 18 = 252$, and the last set of feature vectors from the temporal domain has a total dimensionality of $14 \times 13 = 182$. In the following step of preprocessing, the HLDs were standardised individually and per

participant using the z-score (Thiam et al., 2019).

4.2 Physiological Data

Both hand-crafted and automatic features were used to perform pain recognition using the physiological modalities.

The hand-crafted features were engineered using the same methodology presented in (Thiam et al., 2019). The work presented in this paper has made use of 335 hand-crafted features after they were provided by the authors. Out of these 335 features, 10 features were removed because they had zero variance among the dataset samples. Furthermore, 14 features were dropped because more than 25% of their values were undefined, under the assumption that they would not be beneficial to the learning task. The median value was calculated independently for each feature and used to replace what remained of missing or undefined values. Finally, before they were fed to support vector machines and neural networks, the remaining 311 features were standardized. The standardization step was skipped for random forests.

Standardization for feature x was done by subtracting the mean μ and dividing by the standard deviation σ to produce the standardized feature z as follows:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

For the automatic extraction of features, Butterworth filters were used to process the raw signals of the four modalities before the pain events were isolated. A fifth-order Butterworth bandpass filter was applied to the electrocardiography signals with a frequency range of [0.4, 35] Hz. A third-order Butterworth bandpass filter was applied to the respiration signals with a frequency range of [0.2, 0.8] Hz. A third-order Butterworth bandpass filter was applied to the electromyographic signals with a frequency range of [0.05, 25] Hz, and a third-order Butterworth low-pass filter was applied to the electrodermal activity signals with a cut-off frequency of 0.2 Hz.

To mitigate inter-subject variance, all the signals were normalized to fit in the range $[-1, 1]$. For the training set signals, the normalization was done as follows:

$$z_i = 2 \cdot \frac{x_i - x_{min}}{x_{max} - x_{min}} - 1 \quad (2)$$

where z is the output signal, x is the input signal, x_{min} is the minimum value of the input signal, x_{max} is the maximum value of the input signal, and the subscript i is used to index signal values.

The average maximum and minimum over all the training samples were calculated for every modality

as follows:

$$max_{avg} = \frac{1}{m} \sum_{j=1}^m x_{max,j} \quad (3)$$

$$min_{avg} = \frac{1}{m} \sum_{j=1}^m x_{min,j} \quad (4)$$

where m is the number of samples in the training set, $x_{max,j}$ is the maximum of the j^{th} sample, and $x_{min,j}$ is the minimum of the j^{th} sample.

Finally, the test data was normalized as follows:

$$z_i = 2 \cdot \frac{x_i - min_{avg}}{max_{avg} - min_{avg}} - 1 \quad (5)$$

where z is the output signal, x is the input signal, the subscript i is used to index signal values, and max_{avg} and min_{avg} are obtained from equations 3 and 4 respectively.

Figure 3 shows a sample of the electrodermal activity signal before any preprocessing, while figure 4 shows the same sample after the previously described preprocessing and normalization methods are applied.

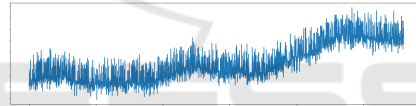


Figure 3: A sample from the electrodermal activity signal of one of the dataset samples before any preprocessing.

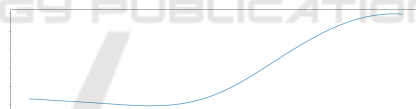


Figure 4: The same sample from figure 3 after it went through the preprocessing pipeline.

5 EXPERIMENTS AND RESULTS

In this section, the different experiments that were conducted are presented along with their results. All the experiments were conducted using the combined data from the left and right forearms from the *SenseEmotion* dataset.

Several algorithms were applied to the problem. Support vector machines work by maximizing the margin between the training instances and the classification boundary (Boser et al., 1992). A random forest is an ensemble classification model, where decision tree classifiers are trained on different subsets of the training data. The predictions of the individual decision tree classifiers are then combined together through a voting scheme to produce the final

output of the random forest (Breiman, 2001). Neural networks are nonlinear models that are trained using back-propagation and that output predictions using the feed-forward operation (Schwenker, 2018).

5.1 Para-linguistic Data

The para-linguistic data were used to train a support vector machine, a random forest, a naive Bayes classifier, a voting classifier, and a deep neural network. Randomized search cross-validation was used to tune the hyperparameters of the different models. The data was evaluated using 10-fold cross-validation and leave-one-subject-out cross-validation.

5.1.1 Support Vector Machines

The three hand-crafted features from the audio signals were combined and fed to a support vector machine. Randomized cross-validation search was used to tune three hyperparameters. The resulting hyperparameters from this search were $kernel = RBF$, $C = 2$ and $\gamma = 1/3$. These hyperparameters achieved a mean accuracy of 64.94% over the 10 cross-validation folds. The same model was evaluated using leave-one-subject-out cross-validation and achieved a mean accuracy of 63.86% over the 40 subjects.

5.1.2 Random Forests

The three hand-crafted features from the audio signals were combined and fed to a random forest. Randomized cross-validation search was used to tune the hyperparameters. The resulting hyperparameters from this search were number of estimators (trees) = 200, maximum features = $1/\sqrt{3}$, maximum depth = 17, minimum samples to split = 40, and minimum samples per leaf = 2. This combination achieved a mean accuracy of 63.52% over the 10 cross-validation folds. This model achieved a mean accuracy of 61.20% with the same combination of hyperparameters in the leave-one-subject-out cross-validation.

5.1.3 Naive Bayes Classifiers

The three hand-crafted features from the audio signals were combined and fed to a naive Bayes classifier. It achieved a mean accuracy of 61.97% over the 10 cross-validation folds and a mean accuracy of 61.68% over the 40 subjects.

5.1.4 Voting Classifier

A voting classifier is a classifier that formulates the classification decision based on the aggregation of the

votes from several classifiers. The voting classifier consisted of a support vector machine, a logistic regression classifier, a Gaussian naive Bayes classifier, and a random forest. The voting classifier achieved a mean accuracy of 63.87% over the 10 cross-validation folds and a mean accuracy of 63.28% in the leave-one-subject-out cross-validation.

5.1.5 Neural Network

A pre-trained model was used to do automatic feature extraction from the audio spectrograms. The pre-trained model used was VGG16, a 16-layer model originally trained on the ImageNet database (Simonyan and Zisserman, 2014). The classification was done using a bidirectional long short-term memory neural network (LSTM). A bidirectional LSTM uses previous and upcoming information from the sequence while learning (Graves and Schmidhuber, 2005). Dropout with a keep probability of 0.5 was used to reduce overfitting (Srivastava et al., 2014). The output layer used the sigmoid activation function. The model was trained using binary cross-entropy and the Adam optimizer (Kingma and Ba, 2014). It was first trained for 20 epochs with a batch size of 64. It achieved 49.47% mean accuracy over the 10 cross-validation folds and 49.01% mean accuracy in the leave-one-subject-out cross validation. After the same model was trained for 50 epochs, it achieved a mean accuracy of 50.35% over the 10 cross-validation folds and 50.12% in the leave-one-subject-out cross-validation.

5.2 Physiological Data

The physiological data were used to train random forests, support vector machines, neural networks, autoencoders, and convolutional neural networks. 5-fold cross-validation was used to tune the hyperparameters of the different models. Then Leave-one-subject-out cross-validation was used as the final evaluation metric. Table 1 provides a summary of the achieved results.

5.2.1 Random Forests

The hand-crafted features from the four physiological modalities were concatenated and fed to random forests. Two searches of the hyperparameter space were conducted. In both, entropy was used to grow the trees. 240 combinations of hyperparameters were tried in total. The best setting had a maximum depth of 20 levels per tree. All features were examined at every node, and a minimum of 10 samples was needed

to create child nodes in the trees. A total of 100 decision trees were trained. This setting achieved a mean accuracy of 83.36% on the 5 cross-validation folds and a mean accuracy of 82.61% on the 40 subjects of the leave-one-subject-out cross-validation.

5.2.2 Support Vector Machines

The hand-crafted features from the four physiological modalities were concatenated and fed to support vector machines with both polynomial and Radial-Basis Function (RBF) kernels.

With the RBF kernel, 195 hyperparameter combinations were evaluated in total. The best combination was $C = 2000$ and $\gamma = 2 \times 10^{-5}$. This combination achieved a mean accuracy of 81.94% over the 5 cross-validation folds and a mean accuracy of 81.34% over the 40 subjects.

With the polynomial kernel, 297 hyperparameter combinations were evaluated in total. The best combination of hyperparameters was $degree = 3$, $C = 20$, and $\gamma = 2 \times 10^{-3}$. This combination achieved a mean accuracy of 78.78% over the 5 cross-validation folds and a mean accuracy of 78.04% over the 40 subjects.

5.2.3 Neural Networks

Several architectures of neural networks with fully connected layers were tested. The number of hidden layers ranged from 2 to 4. The Exponential Linear Unit (ELU) function was used as the activation function for the hidden layers. The sigmoid function was used as the activation function for the output layer. Batch normalization and He initialization were used (Ioffe and Szegedy, 2015; He et al., 2015). Dropout with a keep probability of 0.5 was applied to the outputs of all the hidden layers in the networks (Srivastava et al., 2014). Binary cross-entropy was minimized using the Adam optimizer to train the networks (Kingma and Ba, 2014).

The best performing architecture had 3 hidden layers with 32, 16, and 4 units respectively. It achieved a mean accuracy of 81.96% on the 5 cross-validation folds and a mean accuracy of 81.57% on the 40 subjects.

5.2.4 Autoencoders

A single autoencoder was trained on all four modalities at once to perform automatic feature extraction. The input to this autoencoder consisted of 10240 normalized signal values. The autoencoder consisted of three hidden layers with 128, 64, and 128 units respectively. The 64-unit layer was used as the encoding layer. Batch normalization and He initialization

were used. The ELU function was used as the activation for the hidden layers of the autoencoder. A small neural network consisting of two layers with 32 and 1 units respectively was used on top of the encoding layer to perform the classification. This architecture achieved a mean accuracy of 70.00% on the 40 subjects in leave-one-subject-out cross-validation.

5.2.5 Convolutional Neural Networks

A 1D convolutional neural network with 3 convolutional layers was trained on the normalized signals of the four modalities to perform automatic feature extraction and classification (Kiranyaz et al., 2019). Each layer had 96 filters, and each filter was of size 3. The three layers were followed by a global averaging layer, which computes the average over each filter output independently. Finally, a dense layer with one unit was used to output the classification results. The binary cross-entropy function was used as the cost function to train the network. The Rectified Linear Unit (ReLU) function was used as the activation function for the convolutional layers. The sigmoid function was used as the activation function for the output layer. The model was trained using stochastic gradient descent. This architecture achieved a mean accuracy of 76.90% on the 40 subjects in leave-one-subject-out cross-validation.

Table 1: A summary of the results of the six algorithms that were applied to the physiological modalities to perform pain recognition. The mean accuracy of leave-one-subject-out cross-validation is presented. Random forests, support vector machines (SVM), and neural networks used hand-crafted features, while autoencoders and 1D convolutional neural networks (CNN) performed automatic feature extraction. For the autoencoders approach, a 2-layer neural network was used as the classifier.

Algorithm	Accuracy
Random Forests	82.61 ± 10.74%
Neural Network	81.57 ± 10.58%
RBF Kernel SVM	81.34 ± 10.62%
Polynomial Kernel SVM	78.04 ± 10.57%
1D CNN	76.90 ± 12.60%
Autoencoders	70.00 ± 12.00%

6 CONCLUSION

In this paper, several machine learning algorithms have been evaluated and compared on the pain recognition task using data from the *SenseEmotion* dataset.

On para-linguistic data, support vector machines were the best performing models in the leave-one-subject-out cross-validation. The naive Bayes clas-

sifer was also able to achieve comparable results with less processing time. Analysis of the trained models shows that the MFCC values were the most important features for the classifiers.

The physiological signals achieved higher accuracies. This high performance can be attributed to the use of the electrodermal activity signal, which has been confirmed as an effective indicator of pain. The random forests achieved the highest performance on the hand-crafted features. Support vector machines and neural networks followed closely. Error analysis has shown that the four models that used hand-crafted features made similar mistakes in their predictions. Automatic feature extraction achieved lower accuracies but produced different classification errors. Therefore, a combination of both hand-crafted and automatic features can be tested in the future and may lead to improvements in accuracy.

In both approaches, the use of hand-crafted features outperformed automatic feature extraction. The difference is especially clear in the case of the para-linguistic data, where the VGG-16 bidirectional model performed like a random classifier. This disparity can be attributed to the data size; the dataset does not contain enough data samples to train automatic feature extractors that can generalize well. Furthermore, the inter-subject variance in the values of the signals further complicates automatic extraction. Data augmentation can be tested in the future as one method to deal with these problems.

Pain recognition remains a difficult task. The future collection of more data, especially from a realistic clinical setting, can help improve the performance of the state-of-the-art machine learning algorithms and can allow for the use of deep learning for feature extraction. Data collection should be coupled with the development of new methods to deal with inter-subject variance, which stands in the way of generalizable results.

REFERENCES

- Abdulaziz, Y. and Ahmad, S. M. S. (2010). Infant cry recognition system: A comparison of system performance based on mel frequency and linear prediction cepstral coefficients. In *2010 International Conference on Information Retrieval & Knowledge Management (CAMP)*, pages 260–263. IEEE.
- Aung, M. S. H., Kaltwang, S., Romera-Paredes, B., Martinez, B., Singh, A., Cella, M., Valstar, M., Meng, H., Kemp, A., Shafizadeh, M., Elkins, A. C., Kanakam, N., de Rothschild, A., Tyler, N., Watson, P. J., d. C. Williams, A. C., Pantic, M., and Bianchi-Berthouze, N. (2016). The automatic detection of chronic pain-related expression: Requirements, challenges and the multimodal emopain dataset. *IEEE Transactions on Affective Computing*, 7(4):435–451.
- Barajas-Montiel, S. E. and Reyes-García, C. A. (2006). Fuzzy support vector machines for automatic infant cry recognition. In *Intelligent Computing in Signal Processing and Pattern Recognition*, pages 876–881. Springer.
- Boiten, F. A., Frijda, N. H., and Wientjes, C. J. E. (1994). Emotions and respiratory patterns: review and critical analysis. *International journal of psychophysiology*, 17(2):103–128.
- Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- Cai, D., Ni, Z., Liu, W., Cai, W., Li, G., Li, M., Cai, D., Ni, Z., Liu, W., and Cai, W. (2017). End-to-end deep learning framework for speech paralinguistics detection based on perception aware spectrum. In *INTER-SPEECH*, pages 3452–3456.
- Chang, C.-Y. and Li, J.-J. (2016). Application of deep learning for recognizing infant cries. In *2016 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pages 1–2. IEEE.
- Colloca, L., Benedetti, F., and Pollo, A. (2006). Repeatability of autonomic responses to pain anticipation and pain stimulation. *European Journal of Pain*, 10(7):659–665.
- Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. *International journal for advance research in engineering and technology*, 1(6):1–4.
- Engel, G. L. (1959). psychogenic pain and the pain-prone patient. *The American journal of medicine*, 26(6):899–918.
- Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.
- Gruss, S., Treister, R., Werner, P., Traue, H. C., Crawcour, S., Andrade, A., and Walter, S. (2015). Pain intensity recognition rates via biopotential feature patterns with support vector machines. *PLoS one*, 10(10):e0140330.
- Hammal, Z. and Cohn, J. F. (2014). Towards multimodal pain assessment for research and clinical use. In *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*, pages 13–17. ACM.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034.
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

- Jafari, H., Courtois, I., Van den Bergh, O., Vlaeyen, J. W., and Van Diest, I. (2017). Pain and respiration: a systematic review. *Pain*, 158(6):995–1006.
- Kächele, M., Thiam, P., Amirian, M., Schwenker, F., and Palm, G. (2016). Methods for person-centered continuous pain intensity assessment from bio-physiological channels. *IEEE Journal of Selected Topics in Signal Processing*, 10(5):854–864.
- Kim, J. and André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083.
- Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
- Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., and Inman, D. J. (2019). 1d convolutional neural networks and applications: A survey.
- Ledowski, T., Ang, B., Schmarbeck, T., and Rhodes, J. (2009). Monitoring of sympathetic tone to assess post-operative pain: skin conductance vs surgical stress index. *Anaesthesia*, 64(7):727–731.
- Loggia, M. L., Juneau, M., and Bushnell, M. C. (2011). Autonomic responses to heat pain: Heart rate, skin conductance, and their relation to verbal ratings and stimulus intensity. *PAIN*, 152(3):592 – 598.
- Lu, L., Zhang, H.-J., and Li, S. Z. (2003). Content-based audio classification and segmentation by using support vector machines. *Multimedia systems*, 8(6):482–492.
- Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., and Matthews, I. (2011). Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Face and Gesture 2011*, pages 57–64.
- Mallik, S., Chowdhury, D., and Chittopadhyay, M. (2019). Development and performance analysis of a low-cost mems microphone-based hearing aid with three different audio amplifiers. *Innovations in Systems and Software Engineering*, 15(1):17–25.
- Öhman, A., Esteves, F., Flykt, A., and Soares, J. J. F. (1993). *Gateways to Consciousness: Emotion, Attention, and Electrodermal Activity*, pages 137–157. Springer US, Boston, MA.
- Petroni, M., Malowany, A. S., Johnston, C. C., and Stevens, B. J. (1995). Identification of pain from infant cry vocalizations using artificial neural networks (anns). In *Applications and Science of Artificial Neural Networks*, volume 2492, pages 729–739. International Society for Optics and Photonics.
- Picard, R. W. (2000). *Emotions Are Physical and Cognitive*, pages 21–45. MIT press, Cambridge, MA.
- Saki, F. and Kehtarnavaz, N. (2014). Background noise classification using random forest tree classifier for cochlear implant applications. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3591–3595. IEEE.
- Schwenker, F. (2018). Multimodal affect classification using deep neural networks. *Studentexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018*, pages 240–246.
- Sellner, J., Thiam, P., and Schwenker, F. (2018). Visualizing facial expression features of pain and emotion data. In *IAPR Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction*, pages 101–115. Springer.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.
- Thiam, P., Kessler, V., Amirian, M., Bellmann, P., Layher, G., Zhang, Y., Velana, M., Gruss, S., Walter, S., Traue, H. C., Kim, J., Schork, D., Andre, E., Neumann, H., and Schwenker, F. (2019). Multi-modal pain intensity recognition based on the senseemotion database. *IEEE Transactions on Affective Computing*.
- Thiam, P., Kessler, V., and Schwenker, F. (2017). Hierarchical combination of video features for personalised pain level recognition. In *ESANN*.
- Thiam, P. and Schwenker, F. (2017). Multi-modal data fusion for pain intensity assessment and classification. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE.
- Turk, D. C. and Gatchel, R. J. (2018). *Psychological approaches to pain management: A practitioner's handbook*. Guilford publications.
- Turk, D. C., Meichenbaum, D., and Genest, M. (1983). *Pain and behavioral medicine: A cognitive-behavioral perspective*, volume 1. Guilford Press.
- Velana, M., Gruss, S., Layher, G., Thiam, P., Zhang, Y., Schork, D., Kessler, V., Meudt, S., Neumann, H., Kim, J., Schwenker, F., André, E., Traue, H. C., and Walter, S. (2016). The senseemotion database: A multimodal database for the development and systematic validation of an automatic pain- and emotion-recognition system. In *MPRSS 2016*, pages 127–139.
- Wagner, J., Lingenfelder, F., Baur, T., Damian, I., Kistler, F., and André, E. (2013). The social signal interpretation (ssi) framework: Multimodal signal processing and recognition in real-time. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 831–834, New York, NY, USA. ACM.
- Walter, S., Gruss, S., Ehleiter, H., Junwen Tan, Traue, H. C., Werner, P., Al-Hamadi, A., Crawcour, S., Andrade, A. O., and Moreira da Silva, G. (2013). The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *2013 IEEE International Conference on Cybernetics (CYBCO)*, pages 128–131.
- Wang, J.-C., Lee, H.-P., Wang, J.-F., and Lin, C.-B. (2008). Robust environmental sound recognition for home automation. *IEEE Transactions on Automation Science and Engineering*, 5(1):25–31.
- Zwakhalen, S. M., Hamers, J. P., Abu-Saad, H. H., and Berger, M. P. (2006). Pain in elderly people with severe dementia: a systematic review of behavioural pain assessment tools. *BMC geriatrics*, 6(1):3.