ML-based Pain Recognition Model Using Mixup Data Augmentation

Raghu Mysore Shantharam and Freidhelm Schwenker *

Posted Date: 7 August 2024

doi: 10.20944/preprints202408.0493.v1

Keywords: machine learning; pain recognition; support vector machine; mixup; electrodermal activity

Article

ML-Based Pain Recognition Model Using Mixup Data Augmentation

Raghu M. Shantharam and Freidhelm Schwenker *

Institute of Neural Information Processing, Ulm University, James-Franck-Ring, 89081 Ulm, Germany; raghu.mysore@uni-ulm.de

* Correspondence: friedhelm.schwenker@uni-ulm.de

Abstract: Machine Learning (ML) has revolutionized healthcare by enhancing diagnostic capabilities because of its ability to analyze large datasets and detect minor patterns often overlooked by humans. This is beneficial, especially in pain recognition, where patient communication may be limited. However, ML models often face challenges such as memorization and sensitivity to adversarial examples. Regularization techniques like mixup, which trains models on convex combinations of data pairs, address these issues by enhancing model generalization. While mixup has proven effective in image, speech, and text datasets, its application to time-series signals like electrodermal activity (EDA) is less explored. This research uses ML for pain recognition with EDA signals from the BioVid Heat Pain Database to distinguish pain by applying mixup regularization to manually extracted EDA features and using a Support Vector Machine (SVM) for classification. Results showed that this approach achieved an average accuracy of 75.87% using leave-one-subject-out cross-validation (LOSOCV) compared to 74.61% without mixup. This demonstrates the mixup's efficacy in improving ML model accuracy for pain recognition from EDA signals. This study highlights the potential of mixup in ML as a promising approach to enhance pain assessment in healthcare.

Keywords: machine learning; pain recognition; support vector machine; mixup; electrodermal activity

1. Introduction

Artificial Intelligence (AI) dates back to 1950 as described in [1], where computers were used to simulate intelligent behavior. John McCarthy described this act of simulating intelligence as AI. What began as a simple software program, AI is now advanced includes several complex algorithms, and can even perform functions similar to a human brain. AI has evolved significantly over the last few years, especially with Machine Learning (ML) and Deep Learning (DL) advent. According to [2] ML is an evolving branch of computational algorithms designed to emulate human intelligence. ML algorithms can learn from past data and improve their predictive accuracy over time. ML has paved its way into the healthcare industry, making its applications seamless. Its applications include diagnosing diseases and predicting the therapeutic response as mentioned in [1]. Several literatures have also stated that it is possible to detect early stages of cancer [3] which was almost impossible in the past. Although AI was already studied in the mid-90s, its applications in medicine were limited because of the problem of "overfitting" and this limitation was overcome in the 2000s with the availability of larger datasets and several regularization methods. Because of ML's ability to analyze large datasets and identify minor patterns that are not possible through conventional techniques, it can be used to identify pain in humans. Pain is a complex phenomenon, and according to [4] pain is defined as "an unpleasant sensory and emotional experience". Despite knowledge and technology, pain is still often poorly managed [5,6], and this mostly affects patients with limited communication abilities, who cannot report their pain experience. Such vulnerable groups are infants and elderly people making it difficult for medical professionals to come to a concrete conclusion or solution. ML techniques driven by data hold significant potential in automating pain recognition, thereby assisting healthcare professionals in delivering appropriate medical interventions and treatments. This automatic pain recognition could be done in several ways such as through facial expressions, speech, body movements, and/or by using biophysical signals such as Electromyography (EMG), Electrocardiogram (ECG), Electrodermal

 doi:10.20944/preprints202408.0493.v1

activity (EDA/GSR). Numerous challenges such as memorization, sensitivity to adversarial examples, and overfitting hinder the effectiveness of machine learning models. Several techniques are aimed to overcome these obstacles. One such approach is the mixup [7] that has been proven to enhance the performance. Despite its success, the mixup has yet to be applied to physiological signals. Motivated by this methodology, this literature aims to incorporate the mixup strategy to investigate whether it enhances the performance of pain recognition models.

Either ML or DL could be used for pain recognition, in this literature the conventional ML approach is used wherein relevant features are extracted manually with the help of a feature extractor. This high-level input data representation is subsequently used to optimize an inference model. Such approaches have proven to be effective. Pain recognition can be done using numerous approaches by utilizing physiological signals, some of them by extracting hand-crafted features and then using them to train the ML model, others using a DL architecture. The authors of [8] have proposed a classification architecture based on Deep Belief Networks (DBNs) to classify pain using photoplethysmography signals. Here the DBNs were trained using a set of manually engineered features, so a permutation and combination of different methodologies are possible. While ML models have enabled significant breakthroughs, they also exhibit undesirable behaviors such as memorization and sensitivity to adversarial samples, reducing the performance and generalization capabilities of these models. One solution to overcome this unexpected behavior is data augmentation. [9]Data augmentation is the method of choice to train on similar but different examples to the training data. Data augmentation improves the generalization and performance of ML algorithms by applying transformations such as rotations, translations, and flips. However, the procedure is dataset-dependent and thus requires the use of expert knowledge. Data augmentation also assumes that the examples in the vicinity share the same class, and does not model the vicinity relatation across samples of different classes. These issues motivated the authors of [7] to introduce a simple and data-agnostic data augmentation technique called "mixup". Mixup has gained significant attention and has been shown to enhance the robustness and accuracy of ML and DL models. Mixup involves blending pairs of samples and their corresponding labels to create new synthetic training examples. Mixup takes two samples, x_i, and x_j, and their associated labels, y_i, and y_j, and creates new examples $\hat{x}$ and $\hat{y}$ as weighted linear combinations.

The equations can be represented as below:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \tag{1}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \tag{2}$$

where (x_i,y_i) and (x_j, y_j) are two samples drawn at random from the extracted features as mentioned in [7]. Here, $\lambda \in [0, 1]$ is a random value drawn from a beta distribution with a user-defined parameter α, typically set between 0.1 and 0.4. The generated synthetic example $\hat{x}$ and the label $\hat{y}$ are used as input during training. While previous works have demonstrated the effectiveness of mixup, they have been limited to images, sound, and text data. As part of this literature, Mixup is applied to physiological signals that are then used to train conventional machine learning algorithms/models to recognize pain. A comparison is drawn between machine learning architectures that integrate mixup and those that do not utilize mixup. It is possible to apply mixup on time-series data because the two samples, x_i, and x_j are raw input vectors [7] not limited to images. At the end of this work, it is proved that mixup could also be done on time-series data like EDA signals, eventually leading to a better performance.

This paper is organized as follows. Section 2 summarizes the latest techniques available for pain recognition and the current methodology used to perform mixup. Section 3 explains the proposed method and the datasets and pre-processing in detail. The results are discussed in Section 4 and the paper is concluded with future recommendations in Section 5.

 doi:10.20944/preprints202408.0493.v1

2. Literature Review

As stated in the previous sections, it is difficult to interpret pain, especially when the patient cannot communicate his or her pain experience. Detecting pain has always been tricky because there is no easy method to measure and detect the presence of pain directly, most of the methods available in the past relied solely on techniques such as verbal communication and tend to lack reliability. However, advancements in artificial intelligence have eased the process of pain recognition, by offering automated detection of medical symptoms. Artificial intelligence requires a large amount of data with which it learns to be able then to make predictions on unseen data. The authors in [10] present a multimodal data set, in which pain is induced at different levels. The data set is called the Biopotential and Video (BioVid) Heat Pain Database and this dataset is used for research purposes. This section presents some of the well-established machine learning approaches, some of them making use of the BioVid dataset for pain recognition, additionally, mixup a data augmentation technique, and its prior applications in domains like speech and images have been covered.

There is constantly growing work focusing on pain recognition using physiological signals as also stated in [11]. Categorization is then based on the nature of pain elicitations. Many authors have proposed traditional machine learning approaches such as decision trees or support vector machines (SVM) that involve using a set of carefully engineered features extracted from raw data. The extracted features are then used to optimize an inference model. Some of them are worth noting. In [12], a pain detection approach based on Electroencephalogram (EEG) signals is proposed. Choi-Williams quadratic time-frequency distribution was used to extract necessary features from EEG signals, and then the extracted features were used to train an SVM which was then able to recognize pain based on the classification task. In [13] camera Photoplethysmography (PPG) signals were combined with ECG and EMG signals, and the dataset that was then obtained was used for pain intensity classification based on a fusion architecture at the feature level with SVMS and Random Forests as base classifiers. Authors in [14,15] have used the BVDB [16] to perform different pain intensity classification experiments based on a carefully selected set of features extracted from both physiological and video signals, classification task was then performed based on Random Forest and SVM classifiers. In [17] a user-independent pain intensity classification evaluation based on physiological signals using the same dataset has been done. An adaptive confidence learning approach for personalized pain estimation was also proposed by the same authors based on both physiological and video modalities, the fusion was applied at the feature level. Here Random Forests were used to solve the regression task.

Authors of [8] have proposed a classification approach based on Deep Belief Networks (DBNs). The work mainly focuses on the assessment of patient's pain levels during surgery, using photoplethysmography (PPG) signals. However, it is to be noted that in [8] the proposed architecture has been built on a set of manually engineered features extracted from PPG signals.

Now that we have seen the benefits of machine learning techniques, especially in pain detection, the efficiency of these techniques might be limited because they exhibit undesirable behaviors on adversarial examples and unseen data that impact the overall performance. Several solutions have been proposed to overcome these undesirable behaviors, mixup [7] a data augmentation technique, is one such solution to overcome these limitations. Mixup [7] trains a neural network on convex combinations of pairs of examples and their labels. Authors of [7] have conducted experiments on the ImageNet-2012, CIFAR-10, CIFAR-100, and speech data showing that the mixup improves the generalization of state-of-the-art neural network architectures. It is also found that mixup reduces the memorization of corrupt labels, additionally, mixup also increases the robustness to adversarial examples and along with it stabilizes the training of generative adversarial networks.

While mixup proposed by the authors of [7] works by linearly interpolating inputs and their corresponding labels, it is shown in [18] that the samples can also be mixed in a non-linear fashion, the authors of [18] have named this technique "nonlinear Mixup". Unlike linear mixup where the input and output label pairs share the same, linear, scalar mixing policy, the nonlinear mixup approach embraces nonlinear interpolation policy for both the input and their corresponding labels, the mixing

 doi:10.20944/preprints202408.0493.v1

policy for the labels is adaptively learned based on the mixed input. Experiments have been done on benchmark sentence classification datasets and it is indicated that nonlinear mixup plays a vital role in improving the performance of the classification model. The study conducted by [19] demonstrates the effectiveness of mixup data augmentation in natural language processing (NLP).

Mixup technique is used extensively in the field of health care, especially for medical image classification. In [20] a technique called Balanced-Mixup has been proposed for highly imbalanced medical image classification. In this approach, a balanced mixup simultaneously performs regular and balanced sampling of the training data. The resulting two sets of samples are then mixed up to create a more balanced training distribution from which a neural network can effectively learn without incurring heavily under-fitting the minority classes. Authors of [21] examined the effectiveness of mixup in enhancing the performance of a COVID-19 image classification model. They conducted a comparative analysis between traditional methods and those methods that incorporated mixup. Their findings consistently indicated that the models using mixup augmentation surpassed those relying solely on conventional methods, underscoring the superiority of mixup in enhancing classification accuracy.

So far we have seen various methods that have been proposed and put into practice for pain recognition. These methods involve accessing physiological signals and employing machine learning algorithms to automatically detect pain in patients without relying on their verbal communication. We also saw a popular data augmentation technique called mixup and its ability to enhance the efficiency of deep learning architectures on unfamiliar and adversarial examples, however, it is important to note that, the input data to machine learning or deep learning models that are designed for medical applications are not only images, but they can also be other forms of data such as physiological signals as discussed. It is necessary for such models working on raw physiological signals to perform well on test data and adversarial instances failing which the pain recognition model overfits and becomes unreliable for pain recognition. However, though there are many proposals explaining the benefits of mixup most of them focus mainly on images as input data but for time-series input data like EDA, ECG, and EMG the behavior of mixup is unknown.

In our current work, we focus on a novel approach that employs the mixup strategy on features derived from time-series signals like EDA. Subsequently, we utilize the augmented data to train our machine learning model after which the efficiency of the machine learning model is seen to be better than the models without mixup. These findings illustrate a significant enhancement in machine learning models when incorporating mixup, unlike previous methods that do not utilize mixup.

3. Methodology

3.1. The BioVid Heat Pain Database (BVDB)

Physiological signals play a vital role in pain assessment, we now need a methodology to collect these signals and form a database to be used as an input to the AI model. The BioVid Heat Pain Database(BVDB) [16] is used to meet these requirements. The BVDB was collected in collaboration with the Neuro-Information Technology group of the University of Magdeburg and the Medical Psychology group of the University of Ulm. Here, multimodal data recordings from healthy subjects wherein the subjects are subjected to different levels of artificial pain stimuli. The pain elicitation in the form of heat was conducted through the professionally designed PATHWAY (http://www.medoc-web/products/pathway) thermode attached to the participants' right forearm. Three physiological signals EDA, ECG, and EMG were recorded during this experiment. Along with the physiological signals, video signals were also collected, however, as part of this work, only physiological signals are considered. Figure 1 from [16] shows a rough experimental setup that was used.

 doi:10.20944/preprints202408.0493.v1

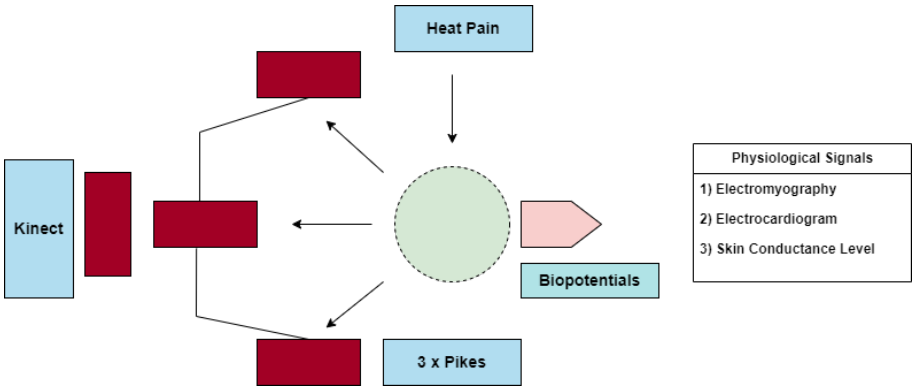

Figure 1. Experimental setting with heat stimulation and recording of biopotentials and video.

The database consists of five parts among which only part A is used in this literature. The starting temperature for all the subjects was 32° Celsius, this starting level is also called the global pain-free level BL1, and then to stimulate pain artificially the temperature was slowly increased. The temperature was increased until the participants felt a change from heat to pain(pain threshold PA1), likewise, the temperature was incremented up to a maximum threshold of PA4 and the pain became hardly bearable. Additionally, two in-between pain elicitation levels PA2 and PA3 were calculated, making the four individual pain levels PA1, PA2, PA3, and PA4 equidistant [11,16], this can be observed in Figure 2.

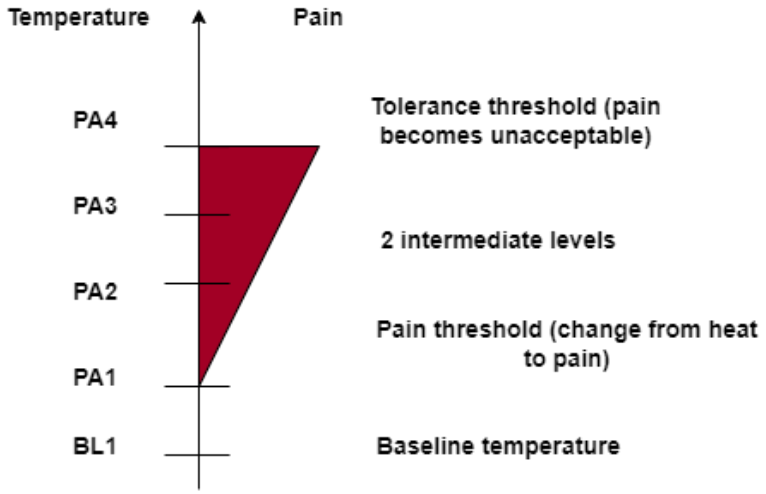

Figure 2. Elicited pain levels.

Each of the pain levels was held for a total of 4s. Each pain stimulation was followed by a rest period during which the baseline temperature was held for a random duration of 8-12s. There were a total of 90 participants in this experiment and the age groups were 18-35, 36-50, and 51-65 years respectively. There were 15 male participants and 15 female participants in each of these age groups. The database is available for public usage, due to technical issues data from 3 participants could not be collected, hence the database consists of data from the remaining 87 participants.

Figure 3 from [16] shows a temperature plot of a stimulus and the following pause. As part of this work, EDA signals play a vital role in pain assessment as stated in [11] and EDA signals are primarily considered as part of this research work. Two important pain levels are BL1(when there is no pain present) and PA4 (the highest level of perceived pain). The AI model is trained to be a binary classifier with two labels BL1 indicating no pain and PA4 indicating the presence of pain.

 doi:10.20944/preprints202408.0493.v1

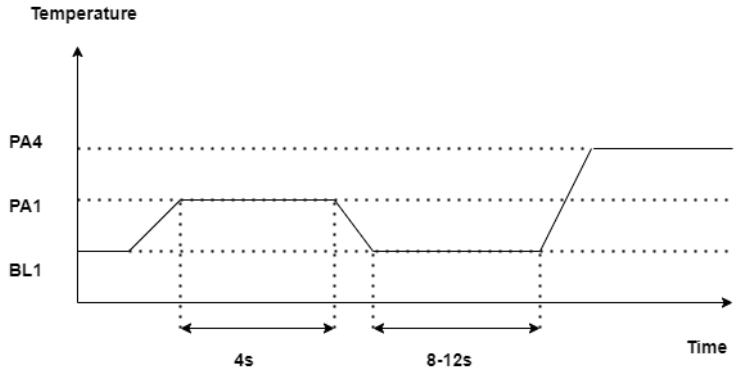

Figure 3. Heat stimulus and pause between stimuli.

3.2. Preprocessing

The physiological signal EDA was preprocessed before performing the classification tasks doing so reduces the computational requirements. The signal's sampling rate was reduced to 256 Hz following [11]. The amount of noise in the recorded data was also significantly reduced by applying different signal preprocessing techniques. A third-order low-pass Butterworth filter with a cut-off frequency of 0.2 Hz was applied to the EDA signals. Additionally, instead of using 5.5s windows with a shift of 3s from the elicitation's onset, the preprocessed signals were segmented into windows of length 4.5s, shifting from the elicitation's onset of 4s. In [11] the other physiological signals EMG and ECG have also been preprocessed, however, as part of this work we are only focussing on the EDA signals and hence ECG and EMG remain unprocessed.

3.3. Feature Extraction

The EDA signal is preprocessed as discussed in the previous section and a cleaned signal is extracted from the EDA signal, the EDA signal is further decomposed into phasic and tonic components, more specifically Phasic Skin Conductance Response (SCR) and Tonic Skin Conductance Level (SCL). The phasic component represents the stimulus-dependent fast-changing signal and the tonic component represents the slow-changing and continuous signal. The cleaned, phasic, and tonic components of an EDA signal could be pictorially represented as shown in Figure 4.

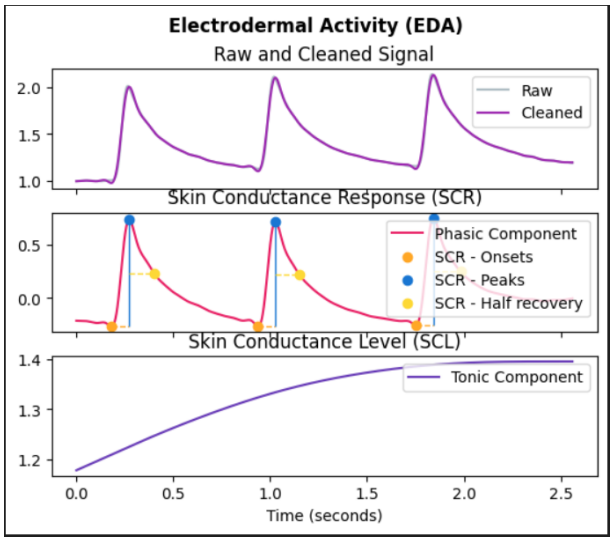

Figure 4. EDA processed signal.

Thirty-six features, comprising twelve statistical features extracted from each of the three signals mentioned above were utilized to train the machine learning model. The twelve statistical features are:

 doi:10.20944/preprints202408.0493.v1

1. Maximum of the signal
2. Minimum of the signal
3. Mean value of the signal
4. Standard deviation of the signal
5. Variance of the signal
6. Root Mean Square of the signal (RMS)
7. Peak-to-Peak (maximum of the signal - minimum of the signal)
8. Skewness of the signal
9. Kurtosis of the signal
10. Mean value of the first difference
11. Mean absolute value of the first difference
12. Mean absolute value of the second difference

As mentioned in [23], the statistical features can be calculated for each of the signal channels in the following way. Let the three signals EDA cleaned, phasic, and tonic be designated by X_i, $i = 1, 2, 3$ respectively. Let X_{in} represent the value of the nth sample of the ith signal, where $n = 1, 2, \ldots, N$ with N in the range of $3,840$ to $7,680$. Let f_{ij} represent the jth feature of the ith signal.

The mean of the signals: f_{i1}:

$$\mu_i = \frac{1}{N}\sum_{n=1}^{N} X_{in}, \quad i = 1, 2, 3 \tag{3}$$

The standard deviation of the signals: f_{i2}:

$$\sigma_i = \left(\frac{1}{N-1}\sum_{n=1}^{N}(X_{in} - \mu_i)^2\right)^{\frac{1}{2}}, \quad i = 1, 2, 3 \tag{4}$$

The mean of the absolute values of the first differences of the signals: f_{i3}:

$$\theta_{i1} = \frac{1}{N-1}\sum_{n=1}^{N-1}|X_{in+1} - X_{in}|, \quad i = 1, 2, 3 \tag{5}$$

The mean of the absolute values of the second differences of the signals: f_{i4}:

$$\theta_{i2} = \frac{1}{N-2}\sum_{n=1}^{N-2}|X_{in+2} - X_{in}|, \quad i = 1, 2, 3 \tag{6}$$

The minimum: f_{i5}:

$$Min = minimum(X_{in}), \quad i = 1, 2, 3 \tag{7}$$

The maximum: f_{i6}:

$$Max = maximum(X_{in}), \quad i = 1, 2, 3 \tag{8}$$

The variance of the signal: f_{i7}:

$$(\sigma_i)^2 = \left(\frac{1}{N-1}\sum_{n=1}^{N}(X_{in} - \mu_i)^2\right), \quad i = 1, 2, 3 \tag{9}$$

The root mean square (RMS) of the signal: f_{i8}:

$$\mathrm{RMS}_i = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(X_{in})^2}, \quad i = 1, 2, 3 \tag{10}$$

 doi:10.20944/preprints202408.0493.v1

The peak-to-peak of the signal also called range of the signal: f_{i9}:

$$range = Max - Min \tag{11}$$

The skewness of the signal : f_{i10}:

$$\text{Skewness}_i = \frac{\frac{1}{N}\sum_{n=1}^{N}(X_{in} - \mu_i)^3}{\left(\frac{1}{N}\sum_{n=1}^{N}(X_{in} - \mu_i)^2\right)^{\frac{3}{2}}}, \quad i = 1,2,3 \tag{12}$$

The kurtosis of the signal : f_{i11}:

$$\text{Kurtosis}_i = \frac{\frac{1}{N}\sum_{n=1}^{N}(X_{in} - \mu_i)^4}{\left(\frac{1}{N}\sum_{n=1}^{N}(X_{in} - \mu_i)^2\right)^{2}}, \quad i = 1,2,3 \tag{13}$$

The mean value of the first difference of the signal: f_{i12}:

$$\theta_{i3} = \frac{1}{N-1}\sum_{n=1}^{N-1}(X_{in+1} - X_{in}), \quad i = 1,2,3 \tag{14}$$

3.4. ML Pain Recognition Model Design

The given dataset is multi-model, where three physiological signals were captured, namely EDA, ECG, and EMG. Based on the experiments and analysis from [11], using an EDA signal significantly outperforms both EMG and ECG in all the classification experiments. Hence as part of the current work EDA signal has been considered for the classification task because it has been proven to be the best-performing single modality[11]. Furthermore, two approaches have been implemented one is a conventional machine learning approach which is built upon a set of carefully engineered features that are manually extracted from the raw EDA signal. The extracted features are subsequently utilized to fine-tune an inference model, essentially categorizing them as either indicative of pain or not, the extracted features are not augmented using mixup. The second approach is based on data augmentation using mixup before training the inference model. The performance of both these models is then compared.

The classifier used here is a binary classifier and the type of learning employed is supervised learning which means the training is supervised with labels. Since this is a binary classification, two labels are used BL1 (baseline temperature) representing no-pain and PA4 (pain tolerance temperature) representing pain. In this experiment, only BL1 and PA4 have been considered again based on the observations by [11] wherein the discrimination between the baseline temperature BL1 and the pain threshold temperature PA1, as well as the two intermediate temperatures PA2 and PA3, constitute a very difficult classification and the modalities that were employed to perform classification were unable to perform classification successfully. On the other hand in [11] it was stated that classification using BL1 vs PA4 as well as PA1 and PA4 yielded successful results.

3.4.1. ML Based Pain Recognition Model with and without Mixup

Figure 5, shows the workflow pipeline of the pain recognition model incorporated using machine learning.

 doi:10.20944/preprints202408.0493.v1

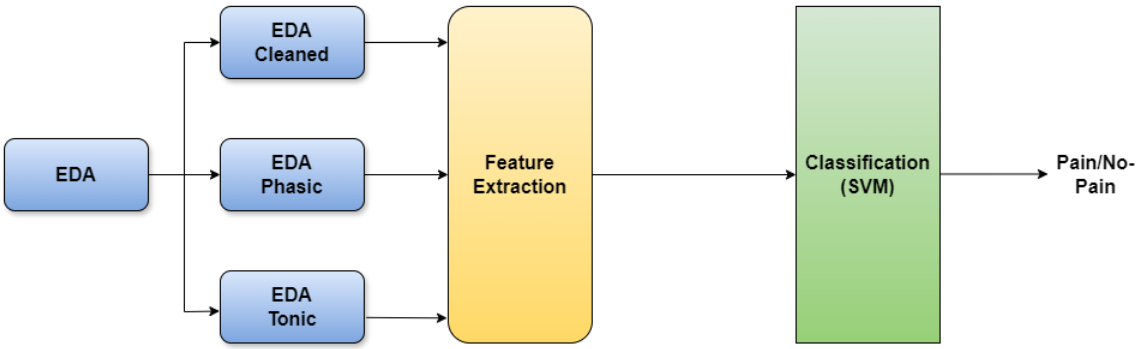

Figure 5. Machine learning model for pain recognition.

The functionality of the model can be illustrated in a few simple steps:

1. Physiological signals are captured from the subjects, following the methodology outlined in Section 3.1 utilizing the BioVid heat pain database.
2. EDA is preprocessed and further decomposed into EDA cleaned, EDA phasic, and EDA tonic. The model is designed by making use of EDA signal only aligning with the findings of [11] given its efficiency and success in pain detection.
3. From each signal, twelve statistical features are extracted, resulting in thirty-six features.
4. In the case of an ML model without mixup, the thirty-six features extracted using the above step are used to train a conventional machine learning classifier. In our study, a support vector machine(SVM) is used which then categorizes the data into either pain or no-pain states.
5. On the other hand, for ML model with mixup the features extracted using step 3 are augmented linearly using mixup [7] where two random samples of extracted features and their corresponding labels are selected and interpolated using equations (1) and (2) where (x_i, y_i) and (x_j, y_j) are two samples drawn at random from the extracted features as mentioned in [7] and $\lambda \in [0, 1]$. This augmentation is called mixup, we get a new sample and corresponding label.
6. This augmented data and the original features are then used to train the support vector machine model which categorizes the data into pain and no-pain states. Mixup is applied iteratively within a loop to ensure that all data samples are covered.

4. Results

Validating a machine learning model is an important aspect as it not only helps us understand the model better but also estimates an unbiased generalization performance, it plays an important role in identifying and correcting overfitting enabling us to select the best model for a given task. There is no single validation method that works in all scenarios. This section discusses a few validation techniques comparing the pain recognition models employed with and without Mixup.

4.1. Leave One Subject Out Cross Validation (LOSOCV)

Cross-validation is a technique used to train and test machine learning models. In cross-validation, the machine learning model is trained on a subset of data points and tested with the remaining data and this is repeated in a loop for n times, n being the number of cross folds. A common approach when multiple subjects are involved is to use leave-one-subject cross-validation (LOSOCV). LOSOCV involves iterating through each subject in a dataset of n subjects. During each iteration, one subject is left out as a test set while the model is trained on the remaining n-1 subjects. This process is repeated until all the subjects have been used as test sets exactly once. When a subject is dropped out, all observations from the subject are left out of the training set. In our work, the value of n is 87, since we have 87 subjects involved in the experiment.

The pain recognition model with mixup performed better in this experimental setup than the model without mixup. The mixup was performed using equations (1) and (2) as mentioned in [7].

 doi:10.20944/preprints202408.0493.v1

where (x_i,y_i) and (x_j, y_j) are two samples drawn at random from the extracted features and $\lambda \in [0,1]$. The λ values are sampled from the Beta distribution.

The Beta distribution is utilized because it effectively models random variables that are constrained within a specific range, typically between 0 and 1, making it suitable for representing probabilities. Its shape parameter, alpha, often shows a pronounced peak near the extremes of 0 or 1. This indicates a high likelihood of observing values close to these limits, suggesting a strong tendency toward extreme outcomes, where the data sample is categorized as one class. Meanwhile, the distribution maintains variability within the range, with a somewhat constant probability of selecting values in the middle (0.33 as likely as 0.5 for instance). This implies that while extreme outcomes are more probable, values closer to the center are still feasible. The parameter $\alpha \in [0.1, 0.4]$ leads to improved performance, smaller α creates less mixup effect, whereas, for larger α mixup leads to underfitting.

A performance evaluation based on the average accuracy of the designed architectures in a binary classification task consisting of the discrimination between the baseline temperature BL1 and the pain tolerance temperature PA4 is reported in Table 1.

Table 1. Performance comparison of SVM with and without mixup for the binary classification task BL1 vs PA4 on EDA signals from BioVid database (PartA).

Method	**SVM**
No mixup	74.61%
Mixup (α = 0.1)	75.87%
Mixup (α = 0.2)	75.04%
Mixup (α = 0.3)	75.61%
Mixup (α = 0.4)	75.12%

At a glance, the designed models with Mixup performed better than the models without Mixup. The average accuracy of the classical machine learning model trained as a binary classifier was 74.61 % without mixup data augmentation. However, when the data was augmented with mixup the performance turned out to be better, λ is the hyperparameter that can be tuned, where $\lambda \in$ [0,1]. Further, λ is sampled from the Beta distribution, the classifier was trained with different values of α and the performance of the classical machine learning model was the best when α = 0.1 where the average accuracy of the pain recognition model showed up to be 75.87%.

4.2. Wilcoxon Signed Rank Test

Wilcoxon signed-rank test is a non-parametric hypothesis test that compares the median of two paired groups and tells if they are identically distributed or not. Wilcoxon signed-rank test is also known as the Wilcoxon matched pair test. It can be used when the differences between the pairs of data are non-normally distributed. This test is used for comparing two machine learning models. The working of a Wilcoxon test can be briefly explained in a few steps:

- The null (h0) and alternate (h1) hypothesis are determined. Null Hypothesis (ho): The median difference between the paired samples is zero, or there is no difference between the two related populations. Alternate Hypothesis (h1): The median difference between the paired samples is not zero, indicating a difference between the two related populations.
- A difference (D) is calculated between each pair of observations.
- The pairs with zero differences are discarded.
- We find out the absolute differences and assign ranks to them starting from the lowest to the highest.
- The Sum of ranks for positive and negative differences are calculated separately.
- The smaller sum of ranks(W) is used as the test statistic.
- Obtained value W is compared with the critical value from the Wilcoxon signed-rank distribution table or the p-value is calculated.

 doi:10.20944/preprints202408.0493.v1

The obtained p-value is compared with a so-called significance level which is typically set to 0.05, the null hypothesis is rejected if the p-value is less than the significance level and we can conclude that there is a statistically significant difference between the two models/ populations. On the contrary, if the obtained p-value is greater than the significance level, it is an indication that there is no statistically significant difference between the populations/models. In our experiment, the p-values were less than 0.05 for all values of $\alpha \in [0.1, 0.4]$ indicating that the ML models incorporating mixup performed better than the ones without mixup.

As part of our experiment, we compared the machine learning models with and without mixup by using the Wilcoxon-signed-rank test, and the p-value is calculated for two lists of accuracy scores: one obtained by performing a LOSOCV on a machine learning model without mixup and the other obtained by performing a LOSOCV on a machine learning model with mixup. The p-value obtained was compared with a significance value and it was much smaller than 0.05 indicating statistical significance which means that the null hypothesis could be rejected hence proving that the machine learning model with mixup performed better than the model without mixup.

4.3. Confusion Matrix

In machine learning, to measure the performance of a model, a confusion matrix is used where exact allocation is possible as it displays the accurate and inaccurate instances based on the model's predictions. The matrix displays the number of instances produced by the model on the test data. A comparison is made within the realm of pain recognition, assessing a conventional machine learning model trained on manually extracted statistical features, with and without the utilization of mixup, and the results are summarized in Figure 6 and Figure 7 respectively.

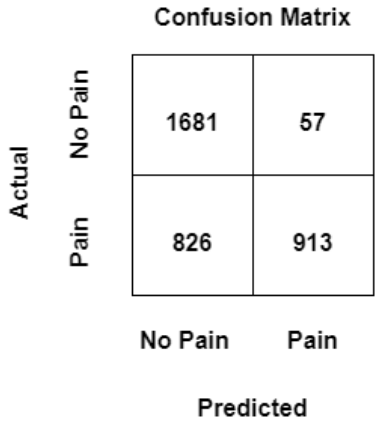

Figure 6. Confusion matrix without mixup.

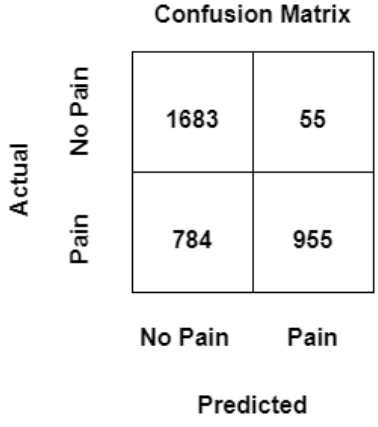

Figure 7. Confusion matrix with mixup.

 doi:10.20944/preprints202408.0493.v1

- True Positive (TP): Number of events correctly predicted as "pain" by the ML model.
- True Negative (TN): Number of events correctly predicted as "no pain" by the ML model.
- False Positive (FP): It is the total count of events predicted as "pain" by the ML model when the actual value is "no pain".
- False Negative (FN): It is the total count of events predicted as "no pain" by the model when the actual value is "pain".

It can be noticed that values that have been previously wrongly predicted as "no pain" by the machine learning model without mixup are corrected by the machine learning model with a mixup with a hyperparameter value set to 0.1, and the model can classify these events as "pain" correctly, likewise, two instances that have been wrongly classified as "pain" by the machine learning model without mixup has been correctly classified as "no pain" when the same model is trained with augmented data. It can be observed that the architectures can consistently discriminate between the baseline temperature BL1 and pain tolerance temperature PA4, the architectures that use augmented data through Mixup perform better than their counterparts without Mixup.

4.4. ROC Curve

A ROC curve or receiver operating characteristic curve has also been illustrated which is nothing but a graphical plot demonstrating the performance of a binary classifier at varying threshold values.

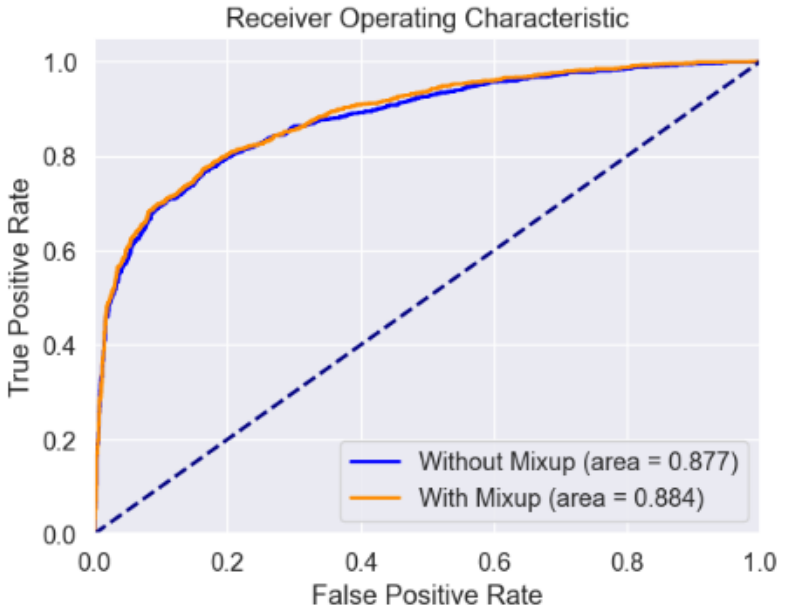

Figure 8. ROC curve of an ML pain recognition model with and without mixup.

In our study, we evaluated the performance of an ML model with and without mixup using a receiver operating characteristic (ROC) curve. It was observed that the ROC curve for the ML model trained without mixup resulted in an Area Under the Curve (AUC) of 0.877. When the model was trained with mixup we obtained an AUC of 0.884 suggesting an improvement in the model's performance compared to the model that was trained without mixup. This increase demonstrates the effectiveness of the mixup in enhancing the model's performance which has been designed to recognize pain in humans.

5. Conclusions and Future Work

This section includes a summary of the research conducted and its future scope. This literature explored the application of mixup, a data-agnostic and straightforward data augmentation principle

 doi:10.20944/preprints202408.0493.v1

for pain intensity classification based on physiological data focusing on EDA. We explored machine-learning architectures that could effectively be used for this application with and without mixup. We used the BioVid Heat Pain Database (PART A) and designed a binary classifier to be able to distinguish between BL1("no pain") and PA4("pain"). An SVM was used for this purpose that was trained with and without augmented data. The data augmentation that was used is called mixup. Incorporating mixup into the training pipeline required only a few lines of code and introduced minimal computational overhead. comparison between architectures employing mixup and those that were trained without augmented data was made and it was observed that the performance of the models using mixup was better than the ones without mixup. All designed architectures were trained comprehensively, within the constraints of the computational resources available for model development. Therefore, refining different components of the models through additional adjustments is believed to enhance the system's performance potentially. The efficacy of mixup augmentation relies heavily on the parameter α, which determines the extent of augmentation and thus requires a careful investigation to achieve a better performance. Several other additional experiments were also conducted such as nonlinear mixup and Nonlinear mixup proved to have a higher potential for further research.

Mixup also opened up several possibilities for further exploration. First, as part of this work, the primary focus of this study was not on achieving the highest accuracy for the pain recognition model. Rather, the aim was to validate the concept of mixup by applying it to a time-series EDA signal. Basic machine learning architectures were employed for this purpose, which may result in lower accuracy compared to other works of literature utilizing more advanced architectures. There is a potential to explore the suggested mixup approach further by implementing it on a neural network-based architecture, in [22] the authors have used a random forest classifier instead of a support vector machine to achieve an accuracy of 73.80% by using only EDA signals and an average accuracy of 74.10% using an early fusion technique. In literature [17] the average accuracy of the machine learning model with more features extracted is reported to be 81.10% using only EDA signals and 82.73% by using all three signals ECG, EDA, and EMG. It is possible to apply mixup on the architectures proposed in other literature to observe the impact on the overall performance of the model. Another approach worth exploring for the EDA signals is the application of nonlinear mixup, as proposed in [18]. Further, the application of linear interpolation was showcased using two randomly selected samples, additionally, it would also be possible to explore mixing varying numbers of samples for future investigation. Lastly, the possibility of applying the mixup technique to other learning problems such as unsupervised learning can also be explored.

Funding: This research did not receive any external funding.

Institutional Review Board Statement: Not applicable since the study did not involve any clinical trials.

Informed Consent Statement: The validation and testing of the developed machine learning model was carried out on a publicly available dataset; therefore, informed consent is not applicable.

Data Availability Statement: The dataset and the code used for implementation will be available based on the consent of the authors.

Acknowledgments: The BVDB was collected in collaboration with the Neuro-Information Technology group of the University of Magdeburg and the Medical Psychology group of the University of Ulm.

Conflicts of Interest: The authors declare no conflict of interest.

 doi:10.20944/preprints202408.0493.v1

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
EDA	Electrodermal Activity
GSR	Galvanic Skin Response
ECG	Electrocardiogram
EMG	Electromyography
EEG	Electroencephalogram
PPG	Photoplethysmography
SVM	Support Vector Machine
RVC	Random Forest Classifier
DBN	Deep Belief Networks
NLP	Natural Language Processing
BVDB	BioVid Heat Pain Database
SCL	Skin Conductance Level
RMS	Root Mean Square
LOSOCV	Leave-One-Subject-Out-Cross-Validation
ROC	Receiver Operating Characteristic

References

1. Kaul, V.; Enslin, S.; Gross, S.A. History of artificial intelligence in medicine. *Gastrointest. Endosc.* **2020**, *92*, 807–812.
2. El Naqa, I.; Murphy, M.J. What is machine learning? In *What is machine learning?*; Springer: New York, NY, USA, 2015; pp. 1–100.
3. Hunter, B.; Hindocha, S.; Lee, R.W. The role of artificial intelligence in early cancer diagnosis. *Cancers* **2022**, *14*, 1524.
4. Werner, P.; Lopez-Martinez, D.; Walter, S.; Al-Hamadi, A.; Gruss, S.; Picard, R.W. Automatic recognition methods supporting pain assessment: A survey. *IEEE Trans. Affect. Comput.* **2019**, *13*, 530–552.
5. Lynch, M.E. The need for a Canadian pain strategy. *Pain Res. Manag.* **2011**, *16*(2), 77.
6. Turk, D.C.; Melzack, R. The measurement of pain and the assessment of people experiencing pain. *The Guilford Press: New York, NY, USA*, 2011.
7. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* **2017**.
8. Lim, H.; Kim, B.; Noh, G.-J.; Yoo, S.K. A deep neural network-based pain classifier using a photoplethysmography signal. *Sensors* **2019**, *19*(2), 384.
9. Simard, P.Y.; LeCun, Y.A.; Denker, J.S.; Victorri, B. Transformation invariance in pattern recognition—tangent distance and tangent propagation. In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 239–274.
10. Simard, P.Y.; LeCun, Y.A.; Denker, J.S.; Victorri, B. Transformation invariance in pattern recognition—tangent distance and tangent propagation. In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 239–274.
11. Thiam, P.; Bellmann, P.; Kestler, H.A.; Schwenker, F. Exploring deep physiological models for nociceptive pain recognition. *Sensors* **2019**, *19*(20), 4503.
12. Alazrai, R.; Al-Rawi, S.; Alwanni, H.; Daoud, M.I. Tonic cold pain detection using Choi–Williams time-frequency distribution analysis of EEG signals: a feasibility study. *Appl. Sci.* **2019**, *9*(16), 3433.
13. Kessler, V.; Thiam, P.; Amirian, M.; Schwenker, F. Pain recognition with camera photoplethysmography. In Proceedings of the 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), Montreal, QC, Canada, 28-30 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–5.

 doi:10.20944/preprints202408.0493.v1

14. Gruss, S.; Treister, R.; Werner, P.; Traue, H.C.; Crawcour, S.; Andrade, A.; Walter, S. Pain intensity recognition rates via biopotential feature patterns with support vector machines. *PLoS One* **2015**, *10*(10), e0140330.
15. Werner, P.; Al-Hamadi, A.; Niese, R.; Walter, S.; Gruss, S.; Traue, H.C. Automatic pain recognition from video and biomedical signals. In Proceedings of the 2014 22nd International Conference on Pattern Recognition (ICPR), Stockholm, Sweden, 24-28 August 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 4582–4587.
16. Walter, S.; Gruss, S.; Ehleiter, H.; Tan, J.; Traue, H.C.; Werner, P.; Al-Hamadi, A.; Crawcour, S.; Andrade, A.O.; da Silva, G.M. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In Proceedings of the 2013 IEEE International Conference on Cybernetics (CYBCO), Lausanne, Switzerland, 13-15 June 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 128–131.
17. Kächele, M.; Thiam, P.; Amirian, M.; Schwenker, F.; Palm, G. Methods for person-centered continuous pain intensity assessment from bio-physiological channels. *IEEE J. Sel. Top. Signal Process.* **2016**, *10*, 854–864.
18. Guo, H. Nonlinear mixup: Out-of-manifold data augmentation for text classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7-12 February 2020; Volume 34, Number 04, pp. 4044–4051.
19. Sun, L.; Xia, C.; Yin, W.; Liang, T.; Yu, P.S.; He, L. Mixup-transformer: Dynamic data augmentation for NLP tasks. *arXiv* **2020**, arXiv:2010.02394.
20. Galdran, A.; Carneiro, G.; González Ballester, M.A. Balanced-mixup for highly imbalanced medical image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V*; Springer: Cham, Switzerland, 2021; pp. 323–333.
21. Özdemir, Ö.; Sönmez, E.B. Attention mechanism and mixup data augmentation for classification of COVID-19 Computed Tomography images. *J. King Saud Univ. - Comput. Inf. Sci.* **2022**, *34*(8), 6199–6207.
22. Werner, P.; Al-Hamadi, A.; Niese, R.; Walter, S.; Gruss, S.; Traue, H.C. Automatic pain recognition from video and biomedical signals. In Proceedings of the 2014 22nd International Conference on Pattern Recognition (ICPR), Stockholm, Sweden, 24-28 August 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 4582–4587.
23. Chu, Y.; Zhao, X.; Han, J.; Su, Y. Physiological signal-based method for measurement of pain intensity. *Front. Neurosci.* **2017**, *11*, 279.