Ensemble of multimodal deep learning autoencoder for infant cry and pain detection [version 2; peer review: 2 approved with reservations]

Yosi Kristian1, Natanael Simogiarto1, Mahendra Tri Arif Sampurna2, Elizeus Hanindito3, Visuddho Visuddho4

1Department of Informatics, Institut Sains dan Teknologi Terpadu, Surabaya, Jawa Timur, 60284, Indonesia
2Department of Pediatrics, Faculty of Medicine, Universitas Airlangga, Surabaya, Jawa Timur, 60132, Indonesia
3Department of Anesthesiology, Faculty of Medicine, Universitas Airlangga, Surabaya, Jawa Timur, 60132, Indonesia
4Faculty of Medicine, Universitas Airlangga, Surabaya, Jawa Timur, 60132, Indonesia

Abstract

Background: Babies cannot communicate their pain properly. Several pain scores are developed, but they are subjective and have high variability inter-observer agreement. The aim of this study was to construct models that use both facial expression and infant voice in classifying pain levels and cry detection.

Methods: The study included a total of 23 infants below 12-months who were treated at Dr Soetomo General Hospital. The the Face Leg Activity Cry and Consolability (FLACC) pain scale and recordings of the baby's cries were taken in the video format. A machine-learning-based system was created to detect infant cries and pain levels. Spectrograms with the Short-Time Fourier Transform were used to convert the audio data into a time-frequency representation. Facial features combined with voice features extracted by using the Deep Learning Autoencoders was used for the classification of infant pain levels. Two types of autoencoders: Convolutional Autoencoder and Variational Autoencoder were used for both faces and voices.

Result: The goal of the autoencoder was to produce a latent-vector with much smaller dimensions that was still able to recreate the data with minor losses. From the latent-vectors, a multimodal data representation for Convolutional Neural Network (CNN) was used for producing a relatively high F1 score, higher than single data modal such as the voice or facial expressions alone. Two major parts of the experiment were: 1. Building the three autoencoder models, which were autoencoder for the infant's face, amplitude spectrogram, and dB-scaled spectrogram of infant's voices. 2. Utilising the latent-vector result from the autoencoders to build the cry detection and pain
Conclusion: In this paper, four pain classifier models with a relatively good F1 score was developed. These models were combined by using ensemble methods to improve performance, which resulted in a better F1 score.

Keywords
Infant facial pain classification, infant cry detection, autoencoder, audio frequency features, deep learning.
List of abbreviations
ANN: Artificial Neural Networks
ASM: Active Shape Model
BCE: binary cross-entropy
CAE: convolutional autoencoder
FACS: Facial Action Coding System
FLACC: Face, Legs, Activity, Cry, Consolability
IFPaLVD: Infant FLACC Pain Level Video Dataset
LSTM: Long-Short Term Memory
MFCCs: Mel-Frequency Cepstrum Coefficients
MSE: Mean Squared Error
NIPS: Neonatal Infant Pain Scale
ROI: region of interest
SSIM: Structural Similarity
STFT: Short-Time Fourier Transform
VAE: Variational Autoencoder

Background
Babies cannot communicate their feelings properly, as such they cry to show hunger, discomfort (wet diaper and itchy skin), and pain. Methods such as Face, Legs, Activity, Cry, Consolability (FLACC),1 Neonatal Infant Pain Scale (NIPS),2 Wong-Baker FACES Rating Scale,3 and others,4 are used to detect pain experienced by a baby. Based on these methods, the main factors for detecting pain are the baby’s facial expression, sound of crying, and body movements. Most of the pain scores are subjective, and they have high variability inter-observer agreement. Therefore, there is a need for a pain measurement tool to objectively measure and classify the pain.

Hospitals mainly have more patients than medical staff, therefore, manually monitoring all their patients would be very time consuming and difficult. Normally, neonates are kept inside incubators, which can reduce the volume of their voice, resulting in difficulty for the medical staff to notice their discomfort. Therefore, an automation program with an artificial intelligence system that can identify the baby’s condition is needed. It can also be used to recognize and monitor the infant’s pain scale’s progress and increase or reduce pain medication. Automatization will allow a doctor to monitor infant conditions remotely with input from the nurses. This invention can significantly improve the health care services, especially in developing countries where doctors are not always on-site.

M. Petroni et al.5 used Artificial Neural Networks (ANN) to classify infant voices to assess whether they’re in pain, fear, or anger. J. O. Garcia and C. A. R. Garcia6 also used infant voices to classify a normal or a pathological cry. This research used Mel-Frequency Cepstrum Coefficients (MFCCs)7 to encode speech signals before the signals were forwarded into the neural network. One study8 also used infant voices with MFCC representation to predict respiratory diseases such as wheeze, asthma, and crackles. Besides pain classification, MFCC has also been used to detect cries using audio features.9 This research divided an audio sequence into segments with 10 seconds duration. Instead of using infant voices, some studies also used facial expression features. K. Sikka et al.10 used facial expression feature to classify pain by using a Facial Action Coding System (FACS). Y. Kristian et al.11 tried hand-crafted extraction of facial expression features by using the Active Shape Model (ASM), based on the infant facial landmark detection and cry detection results from their previous research.12

These studies were further developed by using the deep learning autoencoder method to build a cry detection and pain classifier system.13 In 1986, D. E. Rumelhart et al.14 introduced a method called autoencoder. Autoencoder is a neural network consisting of two parts, encoder, and decoder. The encoder maps the input to the latent-vector with a smaller dimension, while the decoder maps the latent-vector back into the original input.14 An autoencoder for image representation with the use of deep networks was used by Vincent et al.15 for denoising images. That research inspired S. Gao et al.16 to use a deep autoencoder for extracting features from human faces, which showed promising results.
The deep learning extracts facial expression features automatically with the help of facial landmark detection as demonstrated by Kristian et al. J. Egede et al. utilized facial expression by trying to combine hand-crafted feature extraction and deep learning feature extraction. All these studies applied a facial expression as the only feature for classifying pain levels and crying detection. P. Werner et al. used a facial expression as a feature for one of their models. However, this research built other models by using the Biomedical Signals features while applying the Random Forest algorithm to combine these features.

The aim of this paper was to use both facial expression and infant voice as features for classifying pain levels and cry detection. As such deep learning autoencoders to automatically extract features for both facial and audio features were used. Additionally, the results for three features: only facial expression, only voice, and both facial expression and voice were compared in this study.

The key contributions of this paper are:

1) Automated three frame selection from videos dataset. Since several frames consist of an infant’s face, it was not possible to use all the frames. However, automated frame selection can also be a step to building an automated system.

2) Autoencoders were constructed to build a latent-vector with a smaller dimension than the original dimension but capable of representing the original data for both facial and audio data.

3) Introduction of the classifier models using the Convolutional Neural Network (CNN) for the classification of pain level and cry detection.

4) Combination of the best four pain classifier models that uses the ensemble methods to produce a higher F1 score.

Methods

Study setting and participants

This study was an observational study enrolled at Dr. Soetomo General Hospital (Surabaya City, East Java Province, Indonesia). The study included 23 infants below 12-months who were treated for Hernia at Dr Soetomo General Hospital from November 2011 until December 2012. These infants were selected using a consecutive sampling. The FLACC pain scale and audio visual recordings of the baby's cries were taken in the video format and then analyzed using the software.

Ethics approval and consent

This study was approved by Clinical Research Unit at Dr. Soetomo General Hospital (248/Panke.KKE/III/2016). Due to the nature of the research, the visual data documenting facial expressions of observed infants will not be disclosed under the ethical considerations and regulations. Informed written consent was obtained from all patient's parents, confirming that the results and any accompanying images including direct or indirect identifiers can be published.

Dataset

Hanindito has taken 46 videos from 23 infants. Those videos were obtained before and after surgery (mostly hernia surgery) for each infant. The videos were labeled using FLACC measurement methods which took place in Dr Soetomo General Hospital. From this scoring method, all the data were concluded into three classification labels, which were no pain, moderate pain, and severe pain (See Underlying data). In 2017, Kristian et al. added ten more videos using the same FLACC ground truth labels and method.

For this research, these videos were divided into smaller segments with a 10 second duration to generate more data. At the same time, it provided the opportunity to avoid videos with high noise for both video frames and audio waves. Some videos could only be divided into one segment, while others into more than ten segments, depending on the number of frames that contain an extractable infant frontal face. After this process, 253 videos with an average of 10 second duration was obtained. Unfortunately, 64 videos were discarded as they did not contain any extractable frontal face frames. This dataset is called Infant FLACC Pain Level Video Dataset (IFPaLVD), which can be downloaded.

The 189 videos were used to apply the data augmentation process to increase the number of data collection. The augmentation process included both image and audio augmentation. For image data, horizontal flip and image rotation was used. The angles used for image rotation were -10°, 10°, -20°, 20°, and 15°. The output example of this face augmentation can be seen in Figure 1.
There is the time-shifting process (0.25 seconds) during spectrogram conversion. Therefore, the duration of 9.75 seconds was used to obtain 10 seconds for audio data. Further audio augmentation process was tuning the volume $2\times$, $3\times$, and $4\times$ higher or $2\times$ and $3\times$ lower. Therefore, each video was generated into 12 unique videos, resulting in the final data to consist of 2,668 videos. Visually, the augmentation result for the audio data did not look any different.

**Experimental setup**

In this paper, the facial expression and infant’s voice to build a cry detection and pain classifier system was used. From the video frames, face detection and face landmarks were applied to gain the face region. From the audio frames, the wave signal was transformed into amplitude spectrograms and dB-scaled spectrograms by using the Short-Time Fourier Transform (STFT). To gain the features from the infant’s face and voices spectrograms, autoencoders were used for the production of the latent-vector. These latent-vectors were used for CNN classifiers to ultimately predict crying condition and pain level. Four pain classifier models were used to form an Ensemble model. An overview of the system is shown in Figure 2.

**Frame selection and preprocessing**

Every 10 second video had approximately 259 frames depending on the Frames Per Second (FPS) from each video. Therefore, in order to save resources, instead of using all the frames three selected frames were chosen. The first frame was at the time which the baby cried the loudest. This was made possible by selecting the highest value of audio frames along with mapping the index value from audio frames to match the video frames’ index value. Following this, two more frames were chosen from one second before and after the first selected frame. However, there was a possibility that the chosen frame might have not had an extractable frontal face. Therefore, with the selected frame as the starting point, each frame was evaluated until a frame with an extractable frontal face was reached.

![Figure 1. Face augmentation from original image (a) using 15° rotation (b) and horizontal-flip (c).](image)

**Figure 1.** Face augmentation from original image (a) using 15° rotation (b) and horizontal-flip (c).

**Figure 2. Architecture system.** The face region will be extracted from the selected frames using face landmark. The amplitude and dB-scaled spectrograms are obtained from STFT process. These data are forwarded into autoencoders resulting five latent-vector which will be the input for the cry detection and pain classifier system.
Upon selecting three desired frames, the remaining frames which included the non-face region was discarded. Since grayscale images were used in this process it was important to convert the image color. The region of interest (ROI), which is the face region was selected in the grayscale images. This process used the face detection model from a previous study.\(^{21}\) In this model, a rectangle area that consisted of the baby’s face was selected, and facial mapping (landmark) with the use of DLIB library was applied. This technique resulted in 64 points for the face landmarks. These points were used to draw a filled convex polygon and eliminate all the values outside the polygon. Finally, the frame was resized into the shape of 200 × 200 to uniform all the data dimensions.

**Audio preprocessing**

Sounds are continuous analog signals, and their waveforms can take any value.\(^{22}\) Therefore, to digitalize and homogenize an audio signal, computers convert the value into specific sample rate. Similar to FPS for video, audio has a variety of sample rates. The sample rates needed to be uniform in the dataset. The sample rate used was 22,050 Hz, resulting in an array with the length of 21,500 for 9.75 seconds of audio, which represented the audio wave. However, it was not possible to extract any information from this audio wave. Therefore, the wave representation was changed into the time-frequency representation. The method to change the audio representation was by STFT\(^{21}\) (Equation 1). This equation contains two parameters, hop counter (\(h\)) and frequency index (\(f\)).

\[
\text{STFT}(h,f) = \sum_{n=0}^{N-1} x_n \times \text{Hann} \times e^{j2\pi fn/N}
\]  

(Equation 1)

In the STFT process a 2,048 length “Hann” window with 512 hops was used. The total of 215000 samples, produced 420 hop iterations, with each iteration having a 2,048 window size (\(N\)). The result was a two-dimensional array with the shape of 1025 × 420. The two-dimensional array still had a complex number (\(i\)), therefore an absolute function to remove that complex number was done (Amplitude spectrogram). This amplitude spectrogram has many near-zero values as babies cry for a certain frequency range. Therefore, row 50 until 350 for this spectrogram was used and the rest was discarded, resulting in the final shape of 300 × 420. The amplitude spectrogram in its linear scale was transformed into its power log scale or dB scale\(^{23}\) (Equation 2), and the result was called dB-scaled spectrogram. The value of 1.0 was used for the power reference (\(P_0\)), and \(A\) represented the amplitude spectrogram.

\[
\text{dB} = 10 \log_{10} \left( \frac{A^2}{P_0} \right)
\]  

(Equation 2)

**Autoencoder**

Variational Autoencoder (VAE) is a new variation of an autoencoder that was proposed in 2013.\(^{24}\) The VAE can control the latent-vector distribution, due to its structure that contains two latent-vectors (Figure 3). As shown in previous research,\(^{25}\) VAE can generate new data from its latent-vector, and build a face generation model to create a completely new face image. As shown in Figure 3, the two latent-vectors are combined into a single latent-vector using a reparameterization trick (Equation 3). The combination of latent-vector is important as it enables the backpropagation process for VAE. In this equation, \(\mu\) represents mean (first latent-vector), \(\sigma\) represents variance (second latent-vector), and \(\epsilon\) represents random normal. Moreover, VAE also adds the reconstruction loss with its latent-loss, which is usually called KL Divergence Loss (Equation 4) with \(N\) representing the size of the latent-neuron.

![Figure 3. Architecture of Variational Autoencoder (VAE) consist of encoder and decoder with two latent-vector, which usually called Mean and Standard Deviation. These latent-vector were combined using reparameterization trick.](image)
\[ z = \mu + e^{\frac{y}{2}} \epsilon \tag{3} \]

\[ \text{KL}(x, y) = \frac{1}{2} \sum_i N_i (e^{x_i} + y_i^2 - 1 - x_i) \tag{4} \]

Apart from autoencoders for image data, a study in 2017 applied autoencoder for denoising a single-channel audio source. The autoencoder type used in this study was a convolutional autoencoder (CAE). They attempted to recreate the magnitude spectrograms. The spectrogram was a time-frequency representation of an audio signal. Figure 4 shows the example of dB-scaled spectrograms from this study data. This research indicates that using CAE from denoising audio produced promising results.

Autoencoders are unsupervised learning neural network algorithm with a backpropagation process. This algorithm is unsupervised learning because it will try to produce an output that has a close similarity to the input, which can lead to both the output and the input to have similar shapes. Therefore, eliminating the use of data labels.

After being forwarded into several hidden layers, autoencoders will reduce the data dimension. At a certain point, the layers will start to increase the data dimension. Finally, the last layer will have the same size as the input. The main idea of autoencoders is to reduce the original data dimension without losing important details so that it can be reconstructed back into the original data.

Based on this explanation, autoencoders consist of four main parts. The first part is the encoder, where the model will keep reducing the data dimension. The second part is called bottle-neck, or latent-vector. Latent-vector is the result of the encoder, which has the smallest size in the model. The third part is the decoder, which will try to recreate the original data from the latent-vector. The last part is the reconstruction loss, which is a function to measure the similarity between input and output. From all these parts, the encoder and latent-vector is needed. The purpose of the decoder and the reconstruction are to make sure there is an excellent encoder and latent-vector. Finally, only the encoder is used to get the latent-vector, and this latent-vector will be forwarded into the classifier.

In this paper, the autoencoders were used as the feature learning process. As mentioned, this paper has used autoencoders for both face and audio data. Both CAE and VAE were used, and the results were compared. Additionally, the results were analyzed with and without using a dense latent-vector. Mostly, the autoencoders consist of three kinds of layers, which were convolutional, max pooling, and upsampling layer. For the models that used a dense latent-vector, a fully connected layer was added. Rectified Linear Unit (ReLU) was used as the activation function of the convolutional layers and the sigmoid function for the fully connected layers. Other activation functions did not increase the performances.

\[ f(x) = \max (0, x) \tag{5} \]
Both amplitude and dB-scaled spectrograms used the same structure, but they were trained separately. Although layers with variant parameters were used, all convolutional layers only used one stride and the same padding so that the convolutional layers don’t reduce the dimension. The max-pooling layer for the dimensionality reduction task used $2 \times 2$ kernels to minimize the detail loss. The structure for the best model for CAE without dense latent-vector can be seen in Table 1 for face autoencoder, and Table 2 for spectrogram autoencoder. For the models that used dense latent-vector, a similar architecture was used with a different bottle-neck part (yellow color). The bottle-neck part was changed into a fully connected layer with 1,250 neurons, and the next layer with 8,250 neurons for the face autoencoder. For the spectrogram autoencoder layers were not changed, however, a convolutional layer with ten filters and a $3 \times 3$ kernel before the bottle-neck, and a fully connected layer with 7,875 neurons after the bottle-neck part was added. For the VAE models, the same architecture for both with and without dense latent-vector was used with the difference that the bottle-neck parts were duplicated, and an extra layer was added to implement the combination of the two latent-vector by using the reparameterization trick.

Based on the autoencoders structure without dense latent-vector, the latent features will have the shape of $25 \times 25 \times 2$ for the face data and the shape of $75 \times 105 \times 1$ for spectrograms data. This shape is the smallest that could be produced which still gave a good reconstruction result. In the next part, we’ll present that autoencoders without dense latent-vector provide a better reconstruction result.

Autoencoders need a loss function to measure their performance. The most common and simple function is the Mean Squared Error (MSE) function. However, using this function won’t produce the best result for this study therefore, Structural Similarity index (SSIM) was used as the loss function for the face autoencoder. In addition, binary cross-entropy (BCE) instead of MSE was used in the spectrogram autoencoder. The benefits of using SSIM and BCE compared to MSE has been shown in in Figures 5 and 6.

Equations 7 and 8 for SSIM and BCE loss functions were used with $h$ representing model prediction, and $y$ representing the desired output. For the SSIM loss function, 0.01 for $k_1$ and 0.03 for $k_2$, while $L$ represents the maximum value was used (This value is one since the image has already been normalized).

### Table 1. Face CAE architecture.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Model</th>
<th>Layer</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Conv 3×3, 30</td>
<td>9</td>
<td>Conv 3×3, 10</td>
</tr>
<tr>
<td>2</td>
<td>Max Pool 2×2</td>
<td>10</td>
<td>UpSampling 2×2</td>
</tr>
<tr>
<td>3</td>
<td>Conv 3×3, 30</td>
<td>11</td>
<td>Conv 5×5, 20</td>
</tr>
<tr>
<td>4</td>
<td>Max Pool 2×2</td>
<td>12</td>
<td>UpSampling 2×2</td>
</tr>
<tr>
<td>5</td>
<td>Conv 3×3, 20</td>
<td>13</td>
<td>Conv 7×7, 20</td>
</tr>
<tr>
<td>6</td>
<td>Max Pool 2×2</td>
<td>14</td>
<td>UpSampling 2×2</td>
</tr>
<tr>
<td>7</td>
<td>Conv 3×3, 10</td>
<td>15</td>
<td>Conv 3×3, 30</td>
</tr>
<tr>
<td>8</td>
<td>Conv 3×3, 2</td>
<td>16</td>
<td>Conv 3×3, 1</td>
</tr>
</tbody>
</table>

### Table 2. Spectrogram CAE architecture.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Model</th>
<th>Layer</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Conv 3×3, 50</td>
<td>7</td>
<td>Conv 3×3, 30</td>
</tr>
<tr>
<td>2</td>
<td>Max Pool 2×2</td>
<td>8</td>
<td>UpSampling 2×2</td>
</tr>
<tr>
<td>3</td>
<td>Conv 3×3, 40</td>
<td>9</td>
<td>Conv 3×3, 40</td>
</tr>
<tr>
<td>4</td>
<td>Max Pool 2×2</td>
<td>10</td>
<td>UpSampling 2×2</td>
</tr>
<tr>
<td>5</td>
<td>Conv 3×3, 30</td>
<td>11</td>
<td>Conv 3×3, 50</td>
</tr>
<tr>
<td>6</td>
<td>Conv 3×3, 1</td>
<td>12</td>
<td>Conv 3×3, 1</td>
</tr>
</tbody>
</table>
Figure 5. **Spectrogram autoencoders output result.** (a) Original spectrogram, (b) BCE reconstruction, and (c) MSE reconstruction. MSE can’t reconstruct the original data.

$$SSIM(h, y) = \frac{2\mu_h \mu_y + (k_1 L)^2}{\mu_h^2 + \mu_y^2 + (k_1 L)^2} \frac{2\sigma_{hy} + (k_2 L)^2}{\sigma_h^2 + \sigma_y^2 + (k_2 L)^2}$$ \hspace{1cm} (7)

$$BCE(h, y) = -\frac{1}{N} \sum_{i=0}^{N} (y_i \log h_i + (1 - y_i) \log (1 - h_i))$$ \hspace{1cm} (8)

Figure 6. **Face autoencoders output result.** (a) Original image, (b) SSIM reconstruction, and (c) MSE reconstruction.
Multimodal pain and cry detection

In the present study, we used joint representation to co-learning multimodal data together at the same time. We also use a pre-train representation of unsupervised learning, such as autoencoders to overcome the limited labeled data in this study. This research used the ASM model from previous research\(^1\) to crop the face region and discarded unimportant data. We transforms the ASM model into Kazemi et al. face landmark algorithm\(^3\) (Figure 7). DLIB library supports this algorithm. The extracted features were forwarded into the Long-Short Term Memory (LSTM) model to classify pain and for cry detection from a sequence of video frames.

In the present study classifiers for both cry detection and pain classifier were used. Throughout this study, from all the autoencoders tested, both CAE and VAE without dense latent-vector had better performances (Results shown in the next section). Therefore, the latent-vector from autoencoders without a dense-latent vector was used as the input for the classifiers. Following this, it was noticed that autoencoder models with similar performances may have completely different latent features from each other. Some latent features may be able to provide good classification results while this might not be possible for others. Due to this reason, the use of a neural network classifier consisting of just fully connected layers and dropout layers could not be used in this study. Therefore, at least one convolutional layer to boost the classifier performances was needed. In addition to the convolutional layer, dropout layers to prevent overfitting were used considering the small amount of data in this study.

In totals, five latent-vectors from the autoencoders were used. All these features were not used at once, however, models were built by using a single or combination of features. There were five feature parameters for the cry detection and pain classifier system: faces, dB-Scaled, amplitude, faces + dB-Scaled, and faces + amplitude. For a model that used more than one feature, the convolutional layer’s output was concatenated. Therefore, based on the multimodal models, there were four convolutional layers, three for face data and one for spectrogram, before concatenation process.

For the cry detection model, there were only two target classes. Therefore, one neuron output, and the sigmoid activation function was used. However, as there were three target classes for the pain classifier, three output neurons were needed. Additionally, the softmax activation function was used\(^3\) (Equation 9) to make the neuron’s output as the prediction probability for each class. For the cost function, binary cross-entropy for the detection system, and for the pain classifier, a categorical cross-entropy (Equation 10) were used. In these equations, \(N\) represents the number of output neurons.

\[
f(x_i) = \frac{e^{-x_i}}{\sum_j e^{-s_j}} \tag{9}
\]

\[
\text{CCE}(y, h) = -\sum_{i}^{N} (y_i \log h_i) \tag{10}
\]

As this study is a classification problem, accuracy as the performance measurement was used. On the other hand, the most widely used metric to measure model performances for classification problem is the F1 score (Equation 11). F1 score is the harmonic mean of precision (Equation 12) and recall (Equation 13). The final measurement is the Macro F1 Score. This measurement is the average of F1 scores for each class.


\[ F1 = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \]  
(11)

\[ \text{precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \]  
(12)

\[ \text{recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \]  
(13)

**Ensemble model**

This study presents one model that was the best for the cry detection system. However, as a significant high F1 score for the pain classifier model could not be achieved, four models that provided a significant high F1 score were used instead. Each model had different parameters: face + dB-Scaled, face + amplitude, only dB-scaled, and only amplitude. From these results, it was decided to use all the models rather than choosing the best models. The prediction of the pain class was by using four models, and the majority votes method as the ensembling method was used. This method voted for all classifier results and used the majority class label as the final predicted class. In the case of tied votes, the class with higher pain labels were chosen. For example, if the severe pain and moderate pain classes had the same voting points, severe pain as the final prediction was selected.

**Results**

In this paper, two major parts of the experiment were presented. In the first part, three autoencoder models were built, which were autoencoder for the infant’s face, amplitude spectrogram, and dB-scaled spectrogram of infant’s voices. In the second part, the latent-vector result from the autoencoders were used to build the cry detection and pain classification models.

**Face autoencoder**

In total, 20% of the 16,575 face dataset was used as the testing data, which resulted in 3,093 data. The results for the facial autoencoder can be seen in Table 3, and the image output sample can be seen in Figure 8. These results indicate that both CAE and VAE can recreate the data with a small SSIM loss value. From the reconstruction loss and the sample output, it can be seen that the best model was the use of CAE without dense latent-vector.

A dense latent-vector had reduced the model performance for both CAE and VAE. Although, the performance was only slightly reduced, and there was sufficient reconstruction of the original face, the dense latent-vector used much more weight parameters because of its fully connected layers. Therefore, it can be concluded that autoencoders without dense latent-vector were the best option in this case.

Even though CAE and VAE cannot be compared as they are different methods, it can be seen in Figure 8 that CAE gave a much better reconstruction result. It was assumed that VAE tried to build a more general model, as a result the model was forced to miss the small details.

**Spectrogram autoencoders**

Two spectrogram autoencoder models were built, one for dB-scaled spectrogram and one for amplitude spectrogram and the same structure and loss function for both spectrograms were used. Moreover, the same data training and testing distribution was used. The testing data was 506 data as it represents 20% of the total data (2530 short videos). The use of the dense latent-vector for the spectrogram autoencoder produced the same results as the face autoencoder (Results not

<table>
<thead>
<tr>
<th>Type</th>
<th>E</th>
<th>Weight param</th>
<th>SSIM loss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Train</td>
</tr>
<tr>
<td>CAE</td>
<td>25</td>
<td>46,373</td>
<td>0,052239</td>
</tr>
<tr>
<td>CAE-d</td>
<td>20</td>
<td>7,552,851</td>
<td>0,083503</td>
</tr>
<tr>
<td>VAE</td>
<td>80</td>
<td>46,555</td>
<td>0,123958</td>
</tr>
<tr>
<td>VAE-d</td>
<td>100</td>
<td>29,493,161</td>
<td>0,175157</td>
</tr>
</tbody>
</table>

E: Epoch.
The results of the spectrograms CAEs and VAEs can be seen in Table 4, and the output samples were shown in Figure 9 for the dB-scaled spectrogram and Figure 10 for the amplitude spectrogram.

From the output samples, it can be seen that VAEs’ reconstructions were not good for both spectrograms. Based on the results in Figure 10, it can be agreed that the amplitude spectrogram cannot be reconstructed by using VAE. For the dB-scaled spectrogram, the reconstruction may be similar to the original data. However, few spectrogram differences

---

**Table 4.** Spectrogram autoencoder on various type.

<table>
<thead>
<tr>
<th>Type</th>
<th>E</th>
<th>Weight param</th>
<th>BCE loss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Train</td>
</tr>
<tr>
<td><strong>dB-scaled spectrogram</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAE</td>
<td>20</td>
<td>59.553</td>
<td>0.007848</td>
</tr>
<tr>
<td>VAE</td>
<td>20</td>
<td>59.553</td>
<td>0.007848</td>
</tr>
<tr>
<td><strong>Amplitude spectrogram</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAE</td>
<td>20</td>
<td>59.282</td>
<td>0.004835</td>
</tr>
<tr>
<td>VAE</td>
<td>20</td>
<td>59.553</td>
<td>0.007848</td>
</tr>
</tbody>
</table>

E: Epoch.

shown). The results of the spectrograms CAEs and VAEs can be seen in Table 4, and the output samples were shown in Figure 9 for the dB-scaled spectrogram and Figure 10 for the amplitude spectrogram.
might give a huge difference in their audio signal. Therefore, the present study did not test with the classifier by using the VAE latent-vector for the spectrograms. It was also assumed that VAEs were unable to recreate the perfect original data due to their ability to generalize.

Data label distribution
Two data training and testing distribution was tested. In the first distribution, seven original videos, which were three pre-surgery, and four post-surgery videos were excluded to prevent overfitting of the system. This resulted in the use of 1,836 training data and 432 testing data (Table 5). For the second distribution, the data order was mixed, and 25% of the data for the purpose of testing was excluded. Therefore, 1,701 training data and 567 data for testing was obtained. Further details about this distribution can be seen in Table 6. This distribution was provided to compare the result with the first distribution.

Cry detection
In this study several feature combinations by using both data distributions were tested (Table 7). The results show that the use of the second distribution gave a much higher results on all parameters. It is possible that using the second dataset distribution couldn’t detect an overfit model because of the high similarities between the training and testing data.

Table 5. First dataset distribution.

<table>
<thead>
<tr>
<th>Labels</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe</td>
<td>276</td>
<td>132</td>
</tr>
<tr>
<td>Moderate</td>
<td>492</td>
<td>132</td>
</tr>
<tr>
<td>No pain</td>
<td>1068</td>
<td>168</td>
</tr>
<tr>
<td>Cry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>744</td>
<td>240</td>
</tr>
<tr>
<td>No</td>
<td>1092</td>
<td>192</td>
</tr>
</tbody>
</table>

Table 6. Second dataset distribution.

<table>
<thead>
<tr>
<th>Labels</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe</td>
<td>301</td>
<td>107</td>
</tr>
<tr>
<td>Moderate</td>
<td>469</td>
<td>155</td>
</tr>
<tr>
<td>No pain</td>
<td>931</td>
<td>305</td>
</tr>
<tr>
<td>Cry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>735</td>
<td>249</td>
</tr>
<tr>
<td>No</td>
<td>966</td>
<td>318</td>
</tr>
</tbody>
</table>

Figure 10. Sample of amplitude spectrogram autoencoder results using BCE loss function with original spectrogram (a), output from CAE (b), and output from VAE (c).
features. Therefore, it can be concluded that using the voice feature was more relevant and in need of less data than the facial feature. The use of a dB-scaled spectrogram had a better outcome than an amplitude spectrogram, even though the amplitude spectrogram also had better result than the facial features. However, the best result was obtained by the multimodal feature, which used the dB-scaled spectrogram as the audio representation, and CAE as the facial feature extraction. The multimodal feature slightly increased the accuracy even though it used more resources.

According to these results, it can be said that with the existence of enough resources, it’s recommended to use multimodal features. However, in case of limited resources, it is best to use only the voice feature as it can still produce an acceptable

| Table 7. Cry detection using various features. |
|---|---|
| **Input** | **Face autoencoder** | **Spectrogram autoencoder** | **Weight param** | **First distribution** | **Second distribution** |
| | | | | **Acc** | **Macro F1** | **Acc** | **Macro F1** |
| 1 | Multimodal | CAE | dB-Scaled CAE | 439,680 | **0.875000** | **0.871008** | 0.934744 | 0.933486 |
| 2 | Multimodal | VAE | dB-Scaled CAE | 439,680 | 0.791666 | 0.782214 | 0.906525 | 0.903647 |
| 3 | Multimodal | CAE | Amplitude CAE | 439,680 | 0.835648 | 0.834004 | 0.943562 | 0.942811 |
| 4 | Multimodal | VAE | Amplitude CAE | 439,680 | 0.756944 | 0.754735 | 0.888888 | 0.885610 |
| 5 | Face | CAE | - | 283,834 | 0.444444 | 0.307692 | 0.945326 | 0.944099 |
| 6 | Face | VAE | - | 283,834 | 0.604166 | 0.598118 | 0.899470 | 0.898254 |
| 7 | Audio | - | dB-Scaled CAE | 237.787 | 0.872685 | 0.868943 | **0.980599** | **0.980245** |
| 8 | Audio | - | Amplitude CAE | 237.787 | 0.824075 | 0.820374 | 0.938271 | 0.937316 |

Table 8. Confusion matrix of cry detection.

<table>
<thead>
<tr>
<th>Multimodal</th>
<th>No</th>
<th>dB-scaled</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cry</td>
<td>151</td>
<td>41</td>
<td>Cry</td>
</tr>
<tr>
<td>159</td>
<td>33</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>227</td>
<td>39</td>
<td>202</td>
</tr>
</tbody>
</table>

Table 9. Pain classification using various features.

<table>
<thead>
<tr>
<th>Input</th>
<th>Face autoencoder</th>
<th>Spectrogram autoencoder</th>
<th>Weight param</th>
<th>First distribution</th>
<th>Second distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Acc</strong></td>
<td><strong>Macro F1</strong></td>
</tr>
<tr>
<td>1</td>
<td>Multimodal</td>
<td>CAE</td>
<td>dB-Scaled CAE</td>
<td>439,712</td>
<td>0.717593</td>
</tr>
<tr>
<td>2</td>
<td>Multimodal</td>
<td>VAE</td>
<td>dB-Scaled CAE</td>
<td>439,712</td>
<td>0.61574</td>
</tr>
<tr>
<td>3</td>
<td>Multimodal</td>
<td>CAE</td>
<td>Amplitude CAE</td>
<td>439,712</td>
<td>0.708333</td>
</tr>
<tr>
<td>4</td>
<td>Multimodal</td>
<td>VAE</td>
<td>Amplitude CAE</td>
<td>439,712</td>
<td>0.518518</td>
</tr>
<tr>
<td>5</td>
<td>Face</td>
<td>CAE</td>
<td>-</td>
<td>113,179</td>
<td>0.372685</td>
</tr>
<tr>
<td>6</td>
<td>Face</td>
<td>VAE</td>
<td>-</td>
<td>113,179</td>
<td>0.446759</td>
</tr>
<tr>
<td>7</td>
<td>Audio</td>
<td>-</td>
<td>dB-Scaled CAE</td>
<td>118,359</td>
<td>0.631944</td>
</tr>
<tr>
<td>8</td>
<td>Audio</td>
<td>-</td>
<td>Amplitude CAE</td>
<td>118,359</td>
<td>0.655092</td>
</tr>
<tr>
<td><strong>Ensemble model 1, 3, 7, 8</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>0.756944</strong></td>
<td><strong>0.739956</strong></td>
</tr>
</tbody>
</table>
result. The confusion matrix of using the multimodal and voice feature by using the dB-scaled spectrogram audio representation has been provided in Table 8.

Pain classification
The result for pain classification problem is shown in Table 9. It provided extremely high accuracy and F1 score for most cases, even when the first distribution gave a very low result. Therefore, the best method and feature from the first distribution was selected. The highest result came from the model that used the multimodal feature with CAE for face autoencoder, and the dB-scaled spectrogram as its audio representation. However, this parameter only produced a slightly better result than other parameters. This result was also not good enough to be titled as the best model for pain classification. Therefore, the top-four models were ranked based on their F1 scores. Ensemble of these four models were tested with the use of several methods, and the best result was produced using the majority votes technique and the tiebreaker of the sum of each model’s probability from the four models. This ensemble model had a better result based on the accuracy and F1 score. The confusion matrix for this ensemble model can be seen in Table 10. The input and output example in has also been provided in Figure 11.

**Table 10. Confusion matrix of ensemble pain classification.**

<table>
<thead>
<tr>
<th></th>
<th>No pain</th>
<th>Moderate</th>
<th>Severe</th>
</tr>
</thead>
<tbody>
<tr>
<td>No pain</td>
<td>153</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>Moderate</td>
<td>7</td>
<td>104</td>
<td>21</td>
</tr>
<tr>
<td>Severe</td>
<td>5</td>
<td>57</td>
<td>70</td>
</tr>
</tbody>
</table>

![Figure 11. The example of data input with their results.](image-url)
Discussion
In this paper, several methods have been tested for building a deep learning model for pain and cry detection systems. It is evident that a voice feature was more relevant than a facial expression for pain detection. However, the combination of both features can achieve a higher score. The multimodal deep learning autoencoder can assist the detection of infant cry and pain.

Table 11 shows comparison of this work with others related works. Since 2010, researchers have tried to use image as a modality to classify the pain in infants. They found an excellent accuracy of using images as modality of pain detection. Our research which proposed a multimodal deep learning autoencoder for infant cry and pain detection can provide a greater excellent accuracy to the F1 score 0.991.

Apart from studies using facial expression, several studies tried to classify infant pain using acoustic features from the infant voices data. In 1995, Petroni et al. tried to classify the infant’s cry by using the Mel-cepstrum coefficients representation and filter-band energies with the help of ANN. This research used three classification targets: anger, fear, and pain. The results provided good accuracy for predicting anger and pain. However, this research had many prediction errors for the fear label.

In 2003, J. O. Garcia and C. A. R. Garcia set out to classify normal and pathological infant’s cries. This research provided a high accuracy by using 506 data to classify two target classes. Instead of using the audio wave signal, MFCC as the audio representation was used in this study, which is the same as the spectrogram but adapted to resemble the human ear. This research also used the help of a neural network to classify their final prediction.

In 2011, M. A. Nicolaou et al. used a multimodal data without deep learning methods. Since the features were hand-crafted, autoencoders were not used. The same joint representation method along with three fusion methods were used by that study. The methods were feature-level (used in this study), model-level, and output-level fusion. The feature-level fusion is combining the feature to be predicted using one model. While model-level and output-level fusions predict each representation individually, and combine them using another model.

In 2018, T. Baltrušaitis et al. summarized methods that have been used for multimodal data representation. At present, there are only two methods for co-learning multimodal data at the same time: joint and coordinated representations. The basic concept of the joint representation method is to fuse the data representation by concatenating its individual features. On the other hand, the coordinated representation let each representation to be learned separately while still in coordination. Joint representation was used in the present study as it is the most used method. However, limited labeled data has been the main obstacle faced in this case. The solution to this problem was to use a pre-train representation of unsupervised learning, such as autoencoders. The method of using autoencoder for unsupervised learning has been used previously by another study.

<table>
<thead>
<tr>
<th>No</th>
<th>Author, year</th>
<th>Dataset</th>
<th>Modality</th>
<th>Techniques</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nanni et al., 2010</td>
<td>200 Still Image</td>
<td>Image</td>
<td>LBP/ELTP</td>
<td>0.926 (AUC)</td>
</tr>
<tr>
<td>2</td>
<td>Kristian et al., 2017</td>
<td>132 Still Images</td>
<td>Image</td>
<td>Geometrical Features + LBP + SVM</td>
<td>0.952 (F1 Score)</td>
</tr>
<tr>
<td>3</td>
<td>Kristian et al., 2018</td>
<td>2.983 Sequence of Images</td>
<td>Sequence of image</td>
<td>DCNN Autoencoder + LSTM</td>
<td>0.949 (F1 Score)</td>
</tr>
<tr>
<td>4</td>
<td>Sun et al., 2019</td>
<td>22 Videos</td>
<td>Sequence of image</td>
<td>LBP + HOG + Template Matching</td>
<td>0.870 (Accuracy)</td>
</tr>
<tr>
<td>5</td>
<td>Brahanm et al., 2020</td>
<td>234 Videos</td>
<td>Sequence of image</td>
<td>Bag-of-Features (BOF) and descriptors based on Gaussian of Local Descriptors (GOLD) + SVM</td>
<td>0.790 (AUC)</td>
</tr>
<tr>
<td>6</td>
<td>This Research</td>
<td>2268 Short Videos</td>
<td>Multimodal (Sequence of images + Voice)</td>
<td>Convolutional Autoencoder + Spectrogram + Ensemble Learning</td>
<td>0.991 (F1 Score)</td>
</tr>
</tbody>
</table>
In our previous study\textsuperscript{11,12} a cry detection and pain classification system was created. This study used 46 videos taken by E. Hanindito et al.\textsuperscript{18} and added ten more videos with multiple duration. These studies introduced an ASM for landmarking infant’s face. The ASM was built using the frontal face on the videos as their data to extract hand-crafted geometric features. From these features, they’re able to achieve high accuracy for the infant facial cry detection and pain classifier.

In 2018, while deep learning had already been widely used as a feature learning method, Y. Kristian, et al.\textsuperscript{39} removed all the hand-crafted feature extraction, and switched to deep learning by using autoencoders. This research still used the same ASM from their previous research\textsuperscript{12} to crop the face region and discarded unimportant data. This paper transforms the ASM model into Kazemi et al. face landmark algorithm.\textsuperscript{33} DLIB library supports this algorithm. The extracted features were forwarded into the Long-Short Term Memory (LSTM) model to classify pain and for cry detection from a sequence of video frames. The face landmark algorithm from this research was used in this paper by using the DLIB library.

This research reported a high accuracy for both cry detection and pain classification. Based on this finding, it can be suggested that multimodal representation such as combination with a physiological parameter of vital signs should be considered more in future work. Severe stimuli resulting vital sign parameters such as increase in heartbeat, changes in breathing patterns, oxygenation instead of a change of face and cry pattern should be explored.

However, several limitations exist in our study. The limited amount of data may introduce bias, however this paper was able to achieve a relatively good result from the use of the proposed method. To improve the performances even further, increasing the data amount would have a significant effect.

A patient monitoring system is very important, as most of the pain scores are subjective. These findings support the hospital monitoring system and increase awareness of health care providers for the infant’s pain. If the pain persists, modifying the pain medication may be needed.

Conclusion

This study has introduced a method for cry and pain detection by using the infant’s facial expression and voice. In addition, an alternative method in cases with limited resources have been provided. It has been shown that the infant’s voice is a more relevant feature than the facial expression, but the two features combined can obtain a better result.

For feature extraction, it has been demonstrated that CAE performs better than VAE for facial expression and voice spectrograms. This is because VAE is used more for generating new data rather than extracting features. Moreover, it has been shown that using dense latent-vector in the autoencoder increased the reconstruction loss.

Data availability

Underlying data

Figshare: Ensemble of multimodal deep learning autoencoder for infant cry and pain detection. https://doi.org/10.6084/m9.figshare.16910299.v1.\textsuperscript{19}

This project contains the following underlying data:

The data file contains the labels for video dataset for multi modal infant pain and cry classification.

Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 4.0 Public domain dedication).

Author contributions

References


Open Peer Review

Current Peer Review Status:  ?  ?

Version 1

Reviewer Report 02 August 2022

https://doi.org/10.5256/f1000research.76734.r144993

© 2022 Saleh Ba Wazir A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abdulaziz Saleh Ba Wazir

Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia

The paper presents a bright idea of deep learning usage for infants cry and pain detection. The paper used multimodal and utilized both visual and audio features of the dataset for training and testing purposes. The use of both features added more value in the field of infants cry researches using deep learning. Here is a few comments to boost the paper contribution and organization as well.

1. Many of the literature on autoencoders was highlighted in details in the methods section. However, it is proper to be discussed as part of the literature in introduction section or as a separate literature section. And only refer to it in the methods section.

2. Many of the literature on infants cry and pain detection as references 11, 13, and 15 was highlighted in details in the methods section. However, it is proper to be further discussed as part of the literature in introduction section or as a separate literature section. And only refer to it in methods section.

3. It is highly recommended to separate methods and experimental settings into sections. Evaluation and metrics can be part of the experimental settings.

4. Discussion either can be extended and detailed or added to results section and have one section (results and discussion) and detail the discussions within the results if combined.

5. Although several literature was cited on the same dataset, the paper lacks comparison with those researches. As parts of the results and discussion a comparison should be tabulated and discussed, even if the size and type of the dataset are different. The table should highlight the dataset, approach, deep learning models, and achieved metrics. At least few literature can be compared such as 11, 12, 13, 14, and 15.

6. For the comparison table as in comment number 5, the table should highlight at least one or two common metrics. If common metrics is missing, adding one or two metrics is still acceptable than not having a comparison at all.
7. The comparison table shall be discussed to highlight the novelty and contributions of this paper over other methods, such as usage of audio and images, or out performance of current work over others in terms of metrics, or weights. One or two difference is enough, but the more the better.

8. The novelty and contributions based on the comparison table can be used to further detail the contributions in background and conclusion sections.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Deep Learning, Image processing, Computer Vision, Acoustics using deep learning, Audio and Signal Processing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 03 Jan 2023
Mahendra Tri Arif Sampurna, Universitas Airlangga, Surabaya, Indonesia

We would like to thank the reviewer for reading and commenting on our submission. We will attempt to answer each question and suggestion as clearly as possible. Hopefully, revisions made on the manuscript would allow further considerations on indexing the article.

Thank you for reviewing our manuscript. We hereby list the responses and revisions made on the original manuscript:
Reviewer 2

The paper presents a bright idea of deep learning usage for infants cry and pain detection. The paper used multimodal and utilized both visual and audio features of the dataset for training and testing purposes. The use of both features added more value in the field of infants cry researches using deep learning. Here is a few comments to boost the paper contribution and organization as well.

1. Many of the literature on autoencoders was highlighted in details in the methods section. However, it is proper to be discussed as part of the literature in introduction section or as a separate literature section. And only refer to it in the methods section.

Response: Thank you for your valuable comment. We agree that this can lead to more effective reading, therefore we have move part of method to introduction.

2. Many of the literature on infants cry and pain detection as references 11, 13, and 15 was highlighted in details in the methods section. However, it is proper to be further discussed as part of the literature in introduction section or as a separate literature section. And only refer to it in methods section.

Response: Thank you for your suggestion. We also think it will be better if there is “Related Work” section before method. However, to conform with other paper published in this journal we maintain several previous work on the method section which have the basis of our method and move several literature to the discussion section.

3. It is highly recommended to separate methods and experimental settings into sections. evaluation and metrics can be part of the experimental settings.

Response: Thank you for your suggestion. We now have separate the method. The first three subsection (study setting, ethics approval, and dataset) is related to study settings and data collection and continue with part of experimental settings.

4. Discussion either can be extended and detailed or added to results section and have one section (results and discussion) and detail the discussions within the results if combined.

Response: Thank you for you valuable suggestion. We have added several related in discussion and hope that there will provide more understanding for our study.

5. Although several literature was cited on the same dataset, the paper lacks comparison with those researches. As parts of the results and discussion a comparison should be tabulated and discussed, even if the size and type of the dataset are different. The table should highlight the dataset, approach, deep learning models, and achieved metrics. At least few literature can be compared such as 11, 12, 13, 14, and 15.
Response: Thank you for your valuable suggestion. We found that this is very brilliant and have added the table for comparison other works.

6. For the comparison table as in comment number 5, the table should highlight at least one or two common metrics. If common metrics is missing, adding one or two metrics is still acceptable than not having a comparison at all.

Response: Thank you for your valuable suggestion. We have added the different in one or two common metric as suggested

7. The comparison table shall be discussed to highlight the novelty and contributions of this paper over other methods, such as usage of audio and images, or out performance of current work over others in terms of metrics, or weights. One or two difference is enough, but the more the better.

Response: Thank you for your valuable suggestion. We have added the difference in one or two common metric as suggested

8. The novelty and contributions based on the comparison table can be used to further detail the contributions in background and conclusion sections.

Response: Thank you for your valuable suggestion. We have added the table comparison of other work to assist this findings.

Competing Interests: We declare that we have no competing interest
comparison with other works.

3. In method section, it is recommended to separate experimental setup such as database and experimental details in a separate section.

4. It is better to highlight more contributions in Introduction and discussion.

5. In the references list, I can see the author have not mentioned the other related works. There are so many important research related to deep learning and pain detection that have been done before by others and published in high quality journals which is required to bring in the reference list and compare the results with them.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Affective computing, deep learning, computer vision

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 03 Jan 2023

Mahendra Tri Arif Sampurna, Universitas Airlangga, Surabaya, Indonesia

We would like to thank the reviewer for reading and commenting on our submission. We will attempt to answer each question and suggestion as clearly as possible. Hopefully, revisions made on the manuscript would allow further considerations on indexing the article.
Thank you for reviewing our manuscript. We hereby list the responses and revisions made on the original manuscript:

Reviewer 1

The paper is good, the model and novelty is explained well, however there are some feedbacks which the authors need to address before indexing.

1. Discussion should discuss more details, explain well, and talk more about future works, please discuss the achievement and compare with other work. This section need more work.

   **Response:** Thank you for your valuable suggestion. We acknowledge our paper lack on this part and now have added the table for comparison other works.

2. The comparison with other work is not discussed well. It is better the authors use table for comparison with other works.

   **Response:** Thank you for your valuable suggestion. We acknowledge our paper lack on this part and now have added the table for comparison other works.

3. In method section, it is recommended to separate experimental setup such as database and experimental details in a separate section.

   **Response:** Thank you for your suggestion. We now have separate the method. The first three subsection (study setting, ethics approval, and dataset) is related to study settings and data collection and continue with part of experimental settings.

4. It is better to highlight more contributions in Introduction and discussion.

   **Response:** Thank you for your valuable suggestion. We agree with you comment and now have highlighted the contributions in several point on the last part of introduction and conclusion.

5. In the references list, I can see the author have not mentioned the other related works. There are so many important research related to deep learning and pain detection that have been done before by others and published in high quality journals which is required to bring in the reference list and compare the results with them.

   **Response:** Thank you for your valuable suggestion. We have added other high quality journal to prove more understanding on our study.

**Competing Interests:** We declare that we have no competing interest
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com