EmoMusic Dataset and EmoPain Modelling

Group 6

BKRY7 CPVV0 BYQD5 YRDR8 ZKMZ7 CKLM4 WXFG3

***Abstract*—The first part focuses on creating and labelling a dataset for an Emotion Automatic Recognition System when listening to pure music, called EmoMusic. The second part focuses on detecting chronic-pain protective behaviour for EmoPain Dataset via fusion of multiple modalities with three levels. (Code could be found in https://github.com/981526092/COMP0053_Group6)**

I. Part 1

A. Introduction

1) Aim of the Emotion Recognition System: Existing music recommendation engines rely on historical preferences or music content, but neglect the influence of users' emotional states on their music preferences [1]. Thus, our project aims to develop an Emotion Automatic Recognition System (EARS) for data labelling and core recommendation stages of a pure music recommendation system. We create a dataset called EmoMusic. We will collect and analyze users' emotional data in real-time by conducting experiments using pre-selected pure music stimuli to annotate facial expressions and physiological signals during music listening.

2) Affective States: The dimensional model of emotions [2] is a commonly used emotion classification model that quantifies emotions using a valence-arousal scale. However, the model may experience ambiguity when processing data, according to Lin et al [3]. Therefore, we discretized valence's strength into 5 common affective states (disgust, fear, neutral, surprise, and happiness) for our study, as these emotions are easily elicited by sound, especially in pure music listening scenarios.

3) Sensors and Modalities: Jung et al. [4] found that affective classification models using multimodal biological information data have advantages over single-modal models. Therefore, for our study, we used a wide range of wearable sensors to achieve better identification and collection of participants' physiological states without requiring physical activity. These sensors directly respond to changes in biological systems caused by emotions. In addition, facial expression recognition is another modality we will use to expand our dataset. The corresponding sensors for measuring data in these two modalities are Galvanic Skin Response (GSR), photoplethysmography (PPG), the respiration sensor, the thermal camera and a RGB camera for OpenFace data.

4) Elicitation Methods: In the experiment, pre-selected pure music is played to evoke intended affective states in participants. The music is chosen in advance to ensure clear emotional characters and is alternated to elicit a range of emotional responses. The team collected 30-second music clips corresponding to the 5 basic emotions and 7 clips for each affective state to create a reliable dataset of fair size. The order of music playing also contributes to the balance of data.

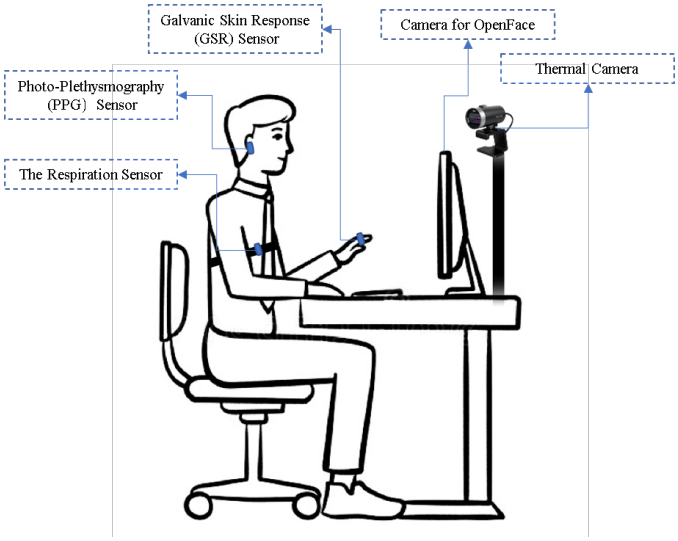


Fig. 1. An overview of the 5 sensors used for our dataset and where they were placed

B. Data Collection

In this study, six members of the project team participated after obtaining UCL-GDPR certification and being fully informed about the procedures, objectives, and potential risks. To comply with ethical standards, the experiment lasted for a maximum of 20 minutes, and the environment was controlled for lighting and acoustics to prevent noise interference. Participants were required to minimize body and head movements during the testing process to ensure accurate measurement by the sensors. The system was kept offline to prevent data leakage.

1) Baseline: Our study required two types of baseline data: sensor measurements (e.g. heart rate, facial temperature, and respiration rate) and self-reported emotional states using a five-point scale. The baseline data was collected before the experiment, after participants sat quietly for 5 minutes to ensure stable physiological readings. One minute of data was selected from this period to serve as the baseline.

TABLE I
AFFECTIVE STATES

Affective state	Original	Now
Most Negative	Disgust	Sad, Disgust and Angry
Less Negative	Fear	Fear, Scared and Nervous
Neutral	Neutral	Neutral, Peace and Calm
Less Positive	Surprise	Surprise, Excited, Thrill
Most Positive	Happiness	Happiness, Delighted, Enjoy

2) Specifics about Sensor: Multiple wearable sensors, including a PPG ear sensor for heart rate, a GSR sensor for skin conductance, and a respiration sensor for monitoring respiration, were used in the study. A thermal camera was also utilized to detect facial temperature changes, and an RGB camera captured facial expression videos for emotion extraction using the OpenFace model. Each sensor was monitored during the experiment, and data was recorded after each music clip. An overview of sensor placement is shown in Fig. 1.

3) Refinement of the Protocol: After running an initial experiment, we modified our protocol to improve it. We expanded our emotional categorization scheme to include three subcategories within each of the five primary affective states, which is presented in Table. I. We also adjusted the sequencing of musical stimuli by presenting consecutive tracks with the same emotional label before introducing opposing affective states.

C. labelling

Existing research [5] has repeatedly demonstrated a strong correlation between facial expressions and human emotions. Therefore, our research team used both self-reported and sensor-validation methods to derive affective labels. Participants reported their affective states after listening to each music clip, and we employed an RGB camera to validate the labelling process by capturing facial expressions. We utilized five affective states (disgust, fear, neutral, surprise, and happiness) for self-reported labelling, and a pre-trained emotion classification model based on facial expressions(code could be found in https://github.com/WuJie1010/Facial-Expression-Recognition.Pytorch) to validate the data labelling accuracy. This model trained on FER2013 Dataset, which consists of many facial images, and the output is one of six emotional categories. We modified this model by aggregating the result into three categories (positive, negative and neutral). Specifically, we capture the raw video from RGB camera as a picture of each frame, which would fed into the pre-trained model. The results of pre-trained classification model can further evidence the reliability of self-reported labelling and prevent the experiment participants report an imprecise result.

For implementation, participants were instructed on the five affective states and were asked to select the corresponding label after listening to each music clip. They had 30 seconds to complete this process. Facial expression validation was conducted after the experiment, and we found that the self-reported labels were largely consistent with the validation labels obtained.

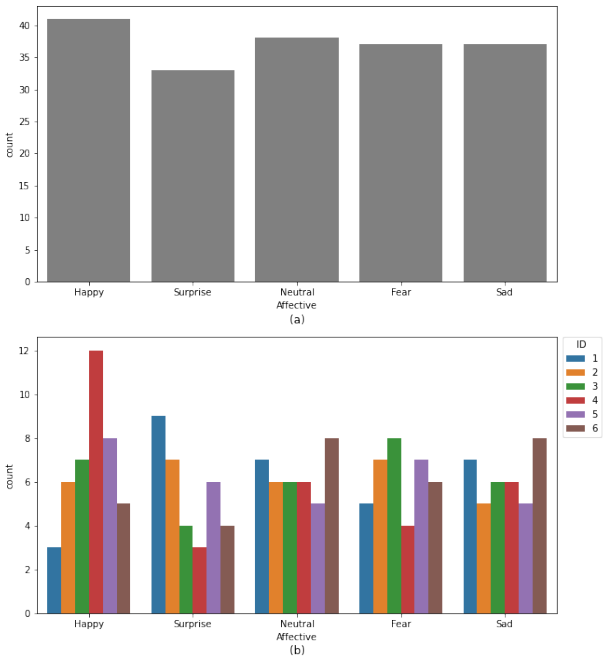


Fig. 2. The bar charts for the labelling. (a) shows the distribution of five affective states. (b) illustrates the count of affective states for each participant

D. Analysis

1) Distribution of Labelling: The labelling variable distribution (shown in Fig. 2) indicates that each affective state has a similar proportion of the total data, although there is variability across participants. This variance is acceptable as different feelings may manifest among the participants, resulting in a balanced dataset for further analysis.

2) PPG Data: We used the 'neurokit2' package in Python to analyze physiological signals from photoplethysmography (PPG). Our analysis included the mean and variance of heart rate and RR interval, and we visualized the data using box plots and swarm plots, presented in Fig. 3. To reduce systematic errors, we use the difference between the observed data and the baseline data as the analysis data. Our analysis aimed to identify relationships between these variables and the five affective states. According to the box plots, we have observed that the means of the aforementioned variables across the five affective states exhibit a similar range and are comparable in magnitude. Moreover, our analysis of the swarm plots has revealed that there is a significant degree of individual variability in the data, as evidenced by the substantial deviations observed between different individuals.

3) Thermal Camera Data: The Thermal Camera Data were processed using the 'fnv.reduce' package in Python to obtain maximum and mean temperature values for the face and mean temperature for the nose. To minimize systematic errors,

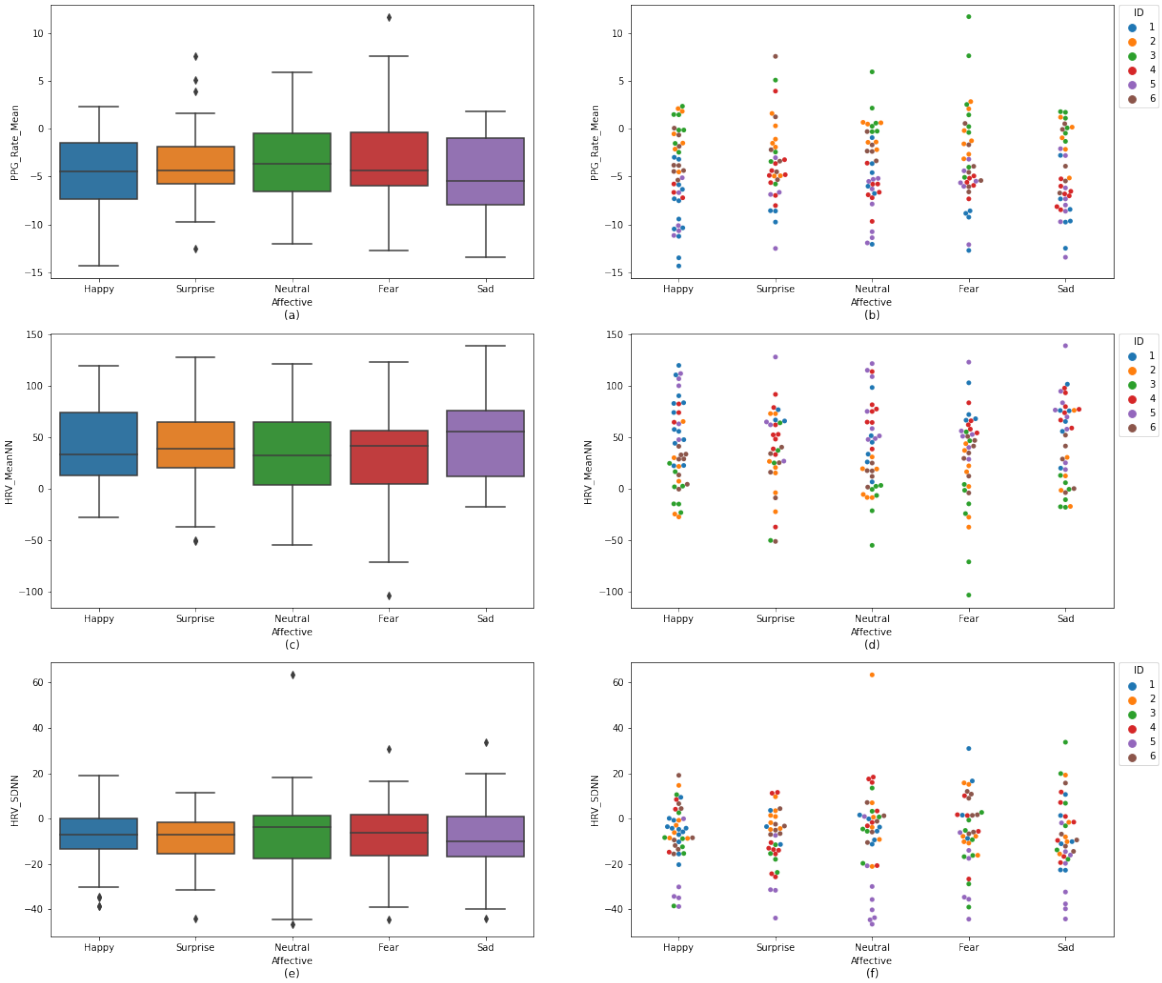


Fig. 3. Six charts show the relationship between interest of affective states and data. All data are changes compared to baseline data. (a) and (b) illustrate the relationship between affective states and the mean of heart rate by the box plot and swarm plot respectively; (c) and (d) reveal the relationship between labelling and the mean of RR interval with box plot and swarm chart; (e) and (f) utilize the same kind of chart to visualize the relationship between affective states and standard deviation of RR interval.

we utilized the difference between the observed data and the baseline data as the analysis data. We visualized the associations between affective states and the three identified thermal camera variables in Fig. 4, which consists of six plots. Our analysis of the data revealed that the mean values of the three variables across the five affective states were largely comparable, as indicated by the three box plots. However, it is worth noting that the range of maximum facial temperature specifically for the affective state of surprise differed from that of the other affective states. Additionally, we observed a significant degree of inter-individual variability, as evidenced by the considerable deviations observed between different individuals.

4) Statistical Analysis for data and labelling: In order to investigate the relationship between the sensor data and the affective states which are known as labels, we built up a regression model where affective states are the dependent variables, normalised PPG rate and normalised facial temperature are the independent variables, then, we conducted hypothesis testing to find out the contributions of each independent variable. We selected multinomial regression as our model since our labels follow multinomial distribution. The p-value for mean

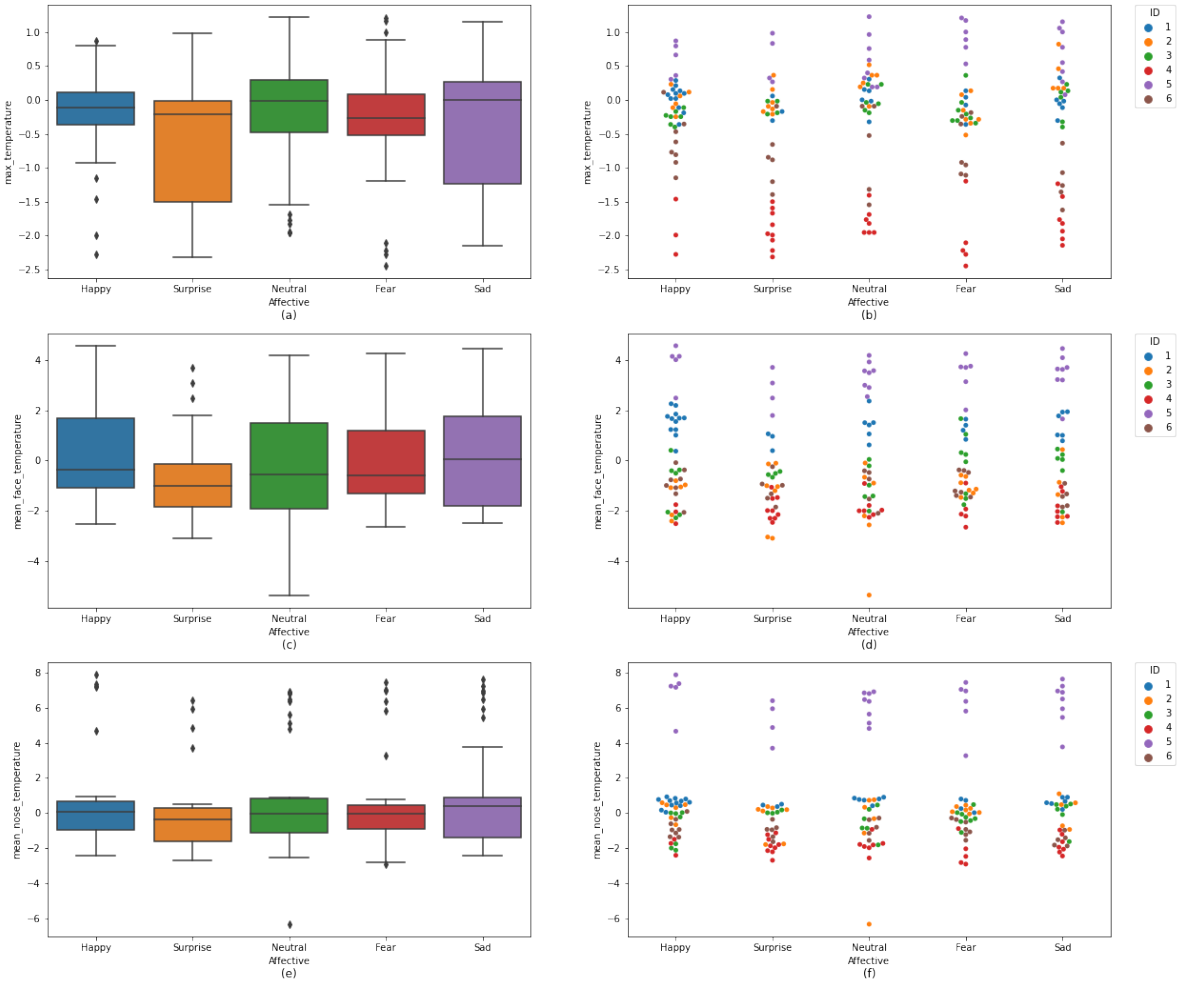


Fig. 4. Six charts show the relationship between interest of affective states and data. All data are changes compared to baseline data. (a) and (b) illustrate the relationship between affective states and the maximum temperature of face by the box plot and swarm plot respectively; (c) and (d) reveal the relationship between labelling and the mean temperature of face with box plot and swarm chart; (e) and (f) utilize the same kind of chart to visualize the relationship between affective states and the mean temperature around nose area.

temperature of face are lower than 0.05 when the affective state is labeled as surprised. We then conclude that the mean temperature of face have significant contribution to the model, we are over 95% confident that there is a relationship between the facial temperature and the affective states. However, the p-values for normalised PPG-rate are larger than 0.05 for each affective state, which means there is no significant relationship between PPG-rate and the affective states.

II. Part2

A. Introduction

The display of protective behavior by individuals with chronic pain (CP) while engaging in physical activities provides valuable insights into their physical and emotional conditions [6]. This project aims to develop an Automatic Recognition System (ARS) using various fusion techniques and neural network architectures to detect protective behavior in individuals with chronic pain during physical activities. The EmoPain dataset [7] was used, which includes body movement data for 14 individuals with chronic pain and

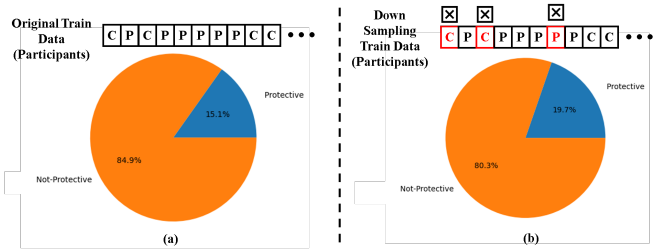


Fig. 5. (a) shows the distribution of labels before downsampling. (b) shows the distribution of labels after downsampling

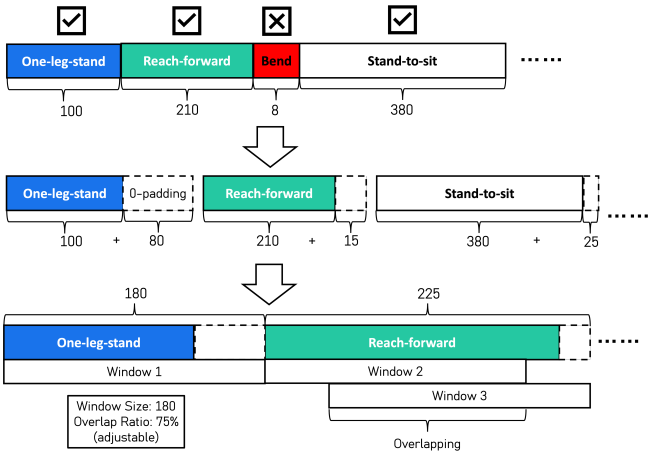


Fig. 6. A flow chart of separate exercise instances and sliding window segmentation

9 healthy participants, along with affective behavior labels indicating protective and non-protective behavior. The dataset includes two modalities: position of body joints and back electrophysiology measured by IMU and sEMG sensors.

B. Data Preprocessing

1) Data DownSampling: In the segmented training dataset, we observed a substantial class imbalance, potentially causing classification bias and overfitting, particularly for the Not-Protective class. Due to the small window instances and time series nature of the data, we avoided traditional resampling methods such as SMOTE [8] and ADASYN [9], as they introduce noise and disrupt temporal relationships. We employed a downsampling strategy, selectively removing participants' data containing only Not-Protective labels. This approach provided a more balanced class distribution, reducing risks associated with class imbalance while preserving the data's temporal structure. Fig. 5 depicts the downsampling process and label distribution before and after downsampling.

2) Separate Exercise Instances: To separate data into exercise instances, we identified continuous actions of the same exercise type. When encountering a different exercise type, we added the instance to a list if its frame count exceeded 12(adjustable), indicating a duration over 0.2 seconds. We checked if the new instance matched the last one in the list and concatenated them if they did. This approach addressed

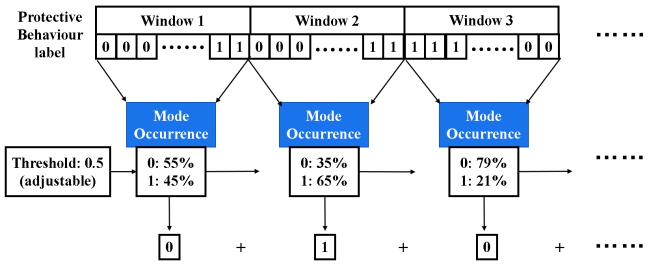


Fig. 7. An overview of target label preprocessing

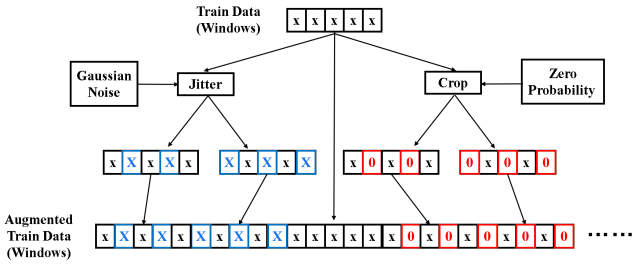


Fig. 8. A flow chart for data augmentation process

instances where continuous actions were interrupted by single frames of different exercise types, likely due to measurement errors. Excluding instances with few frames resulted in the removal of only a small number of instances.The process can be found in Fig. 6.

3) Sliding Window Segmentation: To effectively extract information, we applied sliding window segmentation to each exercise instance. We first zero-padded each instance, setting the default window size to 180 frames and an overlap ratio of 0.75, both of which are adjustable. For example, a 225-frame exercise instance generated two windows: one from frame 1 to 180 and another from frame 46 to 225. This approach yielded a greater number of windows instances compared to exercise instances. The process can be found in Fig. 6.

4) Target Label Preprocessing: For target binary label preprocessing, we aimed to determine the most representative label for each window. We calculated the occurrences of 0s and 1s and considered a 50% threshold, corresponding to the mode. Fig. 7 illustrates an overview of the process. Varying thresholds can impact the dataset, especially with imbalanced label distributions. An optimal threshold help achieve a balanced dataset.

5) Data Augmentation: To enhance model performance and diversify the dataset, we employed jittering and cropping, inspired by Wang et al. [10]. Jittering introduces random noise to input data, generating two jittered datasets while preserving the original structure. This assists in pattern recognition despite input variations. Cropping randomly zeroes portions of input data, resulting in two partially masked datasets. After combining these with the original dataset, we obtained an augmented dataset five times larger than the original. The data

Angle Description

	Angle Name	Respective Anatomical points
1	Crown-Hip-LeftFoot	26-1-4
2	Crown-Hip-RightFoot	26-1-9
3	Spine-Hip-LeftLeg	12-1-3
4	Spine-Hip-RightLeg	12-1-8
5	LeftKnee	2-3-4
6	RightKnee	7-8-9
7	LeftElbow	15-16-17
8	RightElbow	20-21-22
9	LeftShoulder	24-14-15
10	Rightshoulder	24-19-20
11	Leftshoulder-LeftUpperLeg-LeftLowerLeg	16-14-2
12	RightShoulder-RightUpperLeg-RightLowerLeg	21-19-7
13	Neck	12-24-26

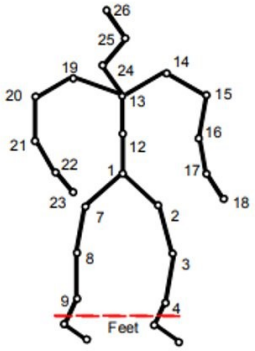


Fig. 9. The arrangement of the 22 body joints and angle description [6]

augmentation process is shown in Fig8.

6) Modality Transformation: Modality transformation was implemented to compare inputs and optimize models. In addition to 22 body joint XYZ coordinates, we used 13 joint angles and corresponding energies as inputs, based on a previous study [6]. Fig. 9 shows the angle description. Each angle was calculated from three relevant anatomical points. For example, the left knee angle is formed by joining joints 2, 3, and 4. Mathematically, let A be a spatial vector between body points 3 and 2, and B be a vector between 3 and 4. Then the left knee angle can be calculated using the following formula:

$$\cos\theta = \frac{\vec{A}\cdot\vec{B}}{|\vec{A}||\vec{B}|} \tag{1}$$

The angular velocity is proportional to the angle difference, as the time between each frame in the coordinate dataset is constant:

$$\omega = \frac{\Delta\theta}{\Delta t} \propto \Delta\theta \tag{2}$$

The kinetic energy can be calculated using the following formula:

$$Energy = 2\pi m\omega^2 \propto \omega^2 \tag{3}$$

Since the weight of each subject remains constant, we simply use the square of the angular velocity as the energy.

This angle-energy modality reduced input dimensionality and removed noise, integrating spatial and temporal information for model training along with sEMG.

C. Early-level Fusion of Modalities

Early fusion combines modalities at the initial stage before feeding them into the model. Fig. 10 provides an overview of models employing early-level fusion, excluding SVM.

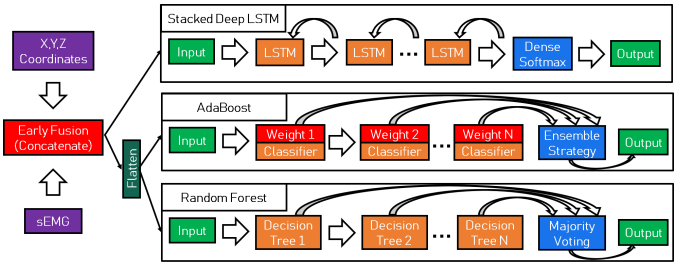


Fig. 10. The architectures of models in early-level fusion of modalities [6]

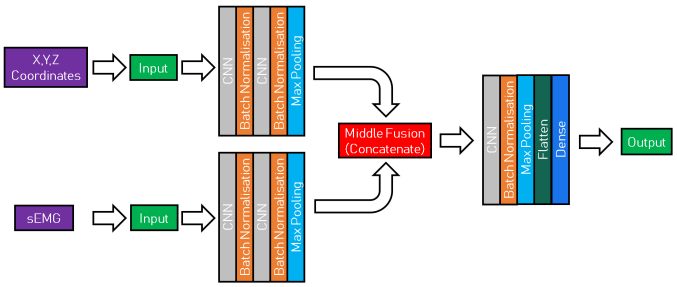


Fig. 11. The architecture of Middle-level fusion Convolutional Neural Network

1) Machine Learning Models: We investigated three ML models—random forest, Adaboost, and SVM—by concatenating features from all 180 frames within each window, resulting in 12,600 features per window. Upon flattening and training, the models displayed high accuracy but failed to predict protective behavior, predominantly classifying instances as non-protective. Notably, SVM trained exclusively on sEMG data produced remarkable results, comparable to late-level fusion models, implying that coordinate information might be less impactful during early fusion stages.

2) Stacked Deep LSTM: Following concatenation of coordinate data (or energy-angle data) and sEMG data, we implemented a two-layer LSTM model with dropout layers and a ReLU-activated dense layer. This model, tested with both coordinate and energy-angle data, combats overfitting using dropout layers and includes a ReLU-activated dense layer preceding the softmax output layer.

D. Middle-level Fusion of Modalities

Middle-level fusion extracts pertinent features from each modality and combines them into a unified representation. We utilized CNNs for feature extraction and label detection, and proposed BANet(LSTM) and BANet(CNN) based on Wang et al.'s work [10].

1) Middle-level Fusion Convolutional Neural Network: Fig. 11 presents the proposed dual-branch architecture, designed for feature extraction from distinct input modalities. Each branch includes two convolutional blocks with Conv1D layers, batch normalization, and ReLU activation, capturing local patterns and reducing overfitting. Outputs from both branches are concatenated and processed through Conv1D,

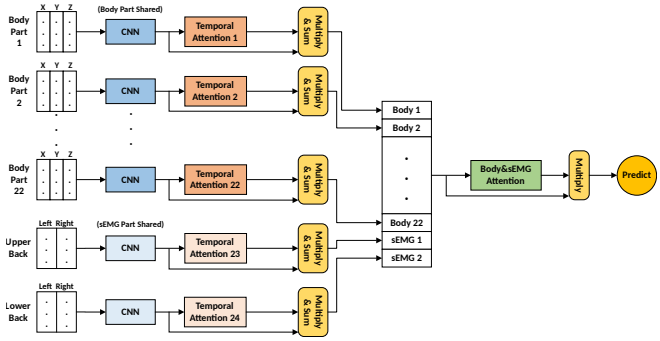


Fig. 12. The Architecture of BANet(LSTM)

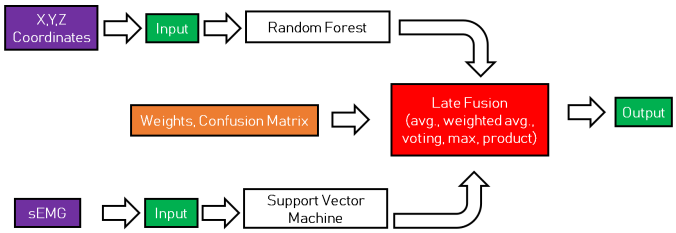


Fig. 13. The architecture of late-level fusion using Ensemble Methods

batch normalization, and ReLU layers. The flattened output is input to a dense layer for binary classification.

2) Body Attention Network using LSTM: BaNet, proposed by Wang et al. [10], used LSTM networks for single modality feature extraction. To reduce overfitting, the original BANet incorporated three LSTM layers and a Dropout layer, which deactivated 50% of the network nodes. We adapted this approach, employing two BANet architectures with three LSTM layers each to independently extract features from different modalities. The data was then fed into an Attention Neural Network for prediction.

3) Body Attention Network using CNN: Despite the improvements, BANet suffered from overfitting, yielding a macro F1-score of about 50%. We addressed this by replacing the original three-layer BANet with a two-layer CNN and adding a Dropout layer after each convolutional layer. We also added a CNN for sEMG processing. Two parallel CNNs extracted features from different modalities, and the data was fed into an Attention Neural Network for prediction, as shown in Fig. 12.

E. Late-level Fusion of Modalities

Late fusion involves merging the outcomes of distinct models, each trained on separate modalities, at a later stage in the pipeline.

1) Ensemble Method: We experimented with providing coordinate data to a random forest model and feeding sEMG data to an SVM model (Fig. 13). We investigated late-fusion/ensemble strategies, including maximum rule, mean calculation, weighted averages with distinct metrics, and mode-based voting. However, the random forest model's poor predictions hindered the SVM model's performance when fused.

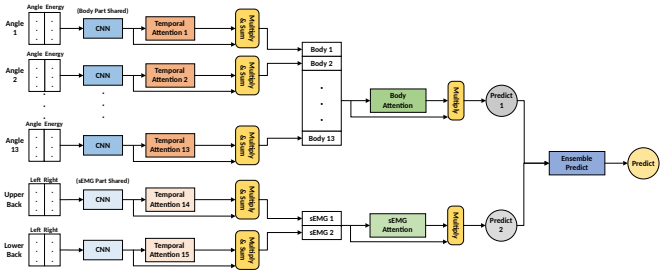


Fig. 14. The Architecture of Bi-BA-CNN

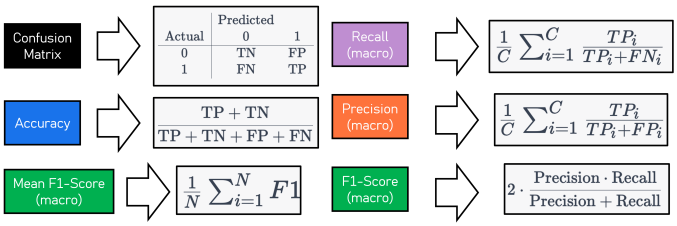


Fig. 15. The details for six evaluation metrics

We then trained an Adaboost classifier using coordinate data and fused the results with the SVM using max rule and F1-score weighted approaches. Both methods showed only a slight improvement over using the SVM model alone.

We also explored feeding features from two modalities into parallel Attention Neural Networks for prediction, averaging the two prediction values as the final result.

2) Bi Body Attention Convolutional Neural Network(Bi-BA-CNN): BANet can also be used to fuse multiple modalities at a later stage. The detailed structure of the model is shown in Fig. 14. This model can be seen as using a dual-layer parallel CNN to implement BANet.

F. Evaluation

1) Methods:

a) Leave-P-Participants-Out-Cross-Validation: Given the availability of multiple participant data, we selected LPPOCV. This method respects subject-wise organization and prevents information leakage, making it suitable for the dataset and yielding accurate performance estimates for unseen participants.

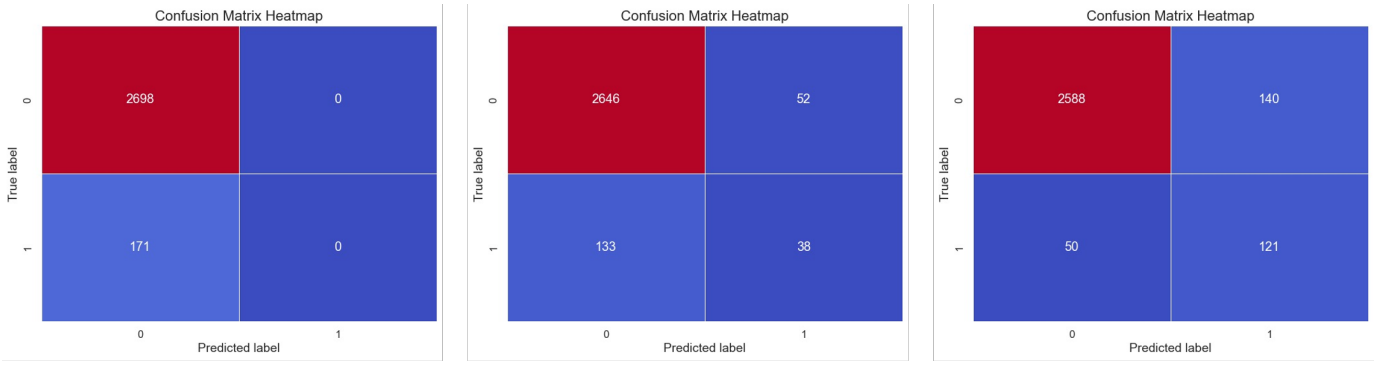
b) Time-Series-Split-Cross-Validation: TSSCV, designed for time-dependent data, reveals potential data leakage from overlapping windows. While it shows potential, TSSCV is less consistent compared to LPPOCV.

c) Hold Out Test: Due to the computational expense of both cross-validations, we opted to perform cross-validation on the best-performing model and employ a holdout test for the remaining models. This approach maintains evaluation efficiency and provides performance insights.

2) Metrics: Binary classification performance is evaluated using the confusion matrix and metrics such as precision, accuracy, recall, and F1 score. The choice of metric depends on data balance and misclassification tolerance. Imbalanced data

TABLE II
MODEL PERFORMANCE COMPARISON

Level of fusion	Model	Dataset	F1 Score	Recall	Precision	Accuracy
Early-level Fusion	Random Forest(Baseline)	Coordinate + sEMG	0.49	0.54	0.51	0.76
	Support Vector Machine	Coordinate + sEMG	0.49	0.5	0.47	0.94
	Adaboost	Coordinate + sEMG	0.58	0.58	0.58	0.9
	Stacked-Deep-LSTM	Coordinate + sEMG	0.48	0.5	0.47	0.94
	Stacked-Deep-LSTM	Angle&Enegy + sEMG	0.58	0.66	0.57	0.84
Middle-level Fusion	CNN-Normal	Coordinate + sEMG	0.63	0.6	0.69	0.94
	BANet(LSTM)	Coordinate + sEMG	0.48	0.5	0.47	0.94
	BANet(CNN)	Coordinate + sEMG	0.57	0.55	0.71	0.94
Late-level Fusion	RF-SVM-Ensemble	Coordinate + sEMG	0.67	0.67	0.67	0.93
	Adaboost-SVM-Ensemble	Coordinate + sEMG	0.73	0.72	**0.74**	0.94
	Bi-BA-CNN	Coordinate + sEMG	0.73	**0.83**	0.69	0.92
	Bi-BA-CNN	Angle&Enegy + sEMG	**0.76**	**0.83**	0.72	0.93



(a) Early-level Fusion (b) Middle-level Fusion (c) Late-level Fusion

Fig. 16. Confusion Matrixes for three levels of fusion. (a): Stacked-Deep-LSTM(Coordinate) model. (b):CNN-Normal(Coordinate) model. (c):BI-BA-CNN(Angle&Energy).

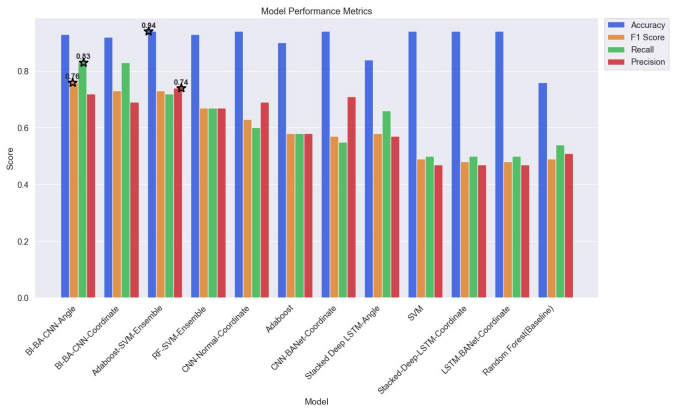


Fig. 17. The grouped bar chart for model performance

makes precision, recall, and F1 score more informative than accuracy, which fails to consider class distribution. Besides, the impact of misclassification on evaluation metrics varies. Precision is suitable for minimizing false positives (Type I error), while high recall values minimize false negatives (Type II error). Moreover, the F1 score provides a balanced view of precision and recall, which is our core metric, with Mean Macro F1-Score used for LPSOCV and TSSCV evaluations. Other metrics were used for the hold-out test.

3) Model Performance: Late-level fusion models, such as Bi-BA-CNN(AngleEnergy + sEMG), generally outperformed early and middle-level fusion models in terms of F1 score, precision, and recall. However, some models, such as Random Forest(Baseline) and Stacked Deep LSTM(AngleEnergy + sEMG), had significantly lower F1 scores despite high accuracy values due to overfitting to predict all 0 labels. Analyzing the updated confusion matrix for the Bi-CNN-BANet-Ensemble-Angle model, it had a high false positive rate (140 out of 272) for the majority class (0), leading to a lower precision score despite having a high recall score for the minority class (1). In contrast, CNN-Normal-Coordinate had a higher precision score for the majority class but a lower recall score for the minority class. Stacked-Deep-LSTM-Coordinate performed poorly for both the majority and minority classes. These results suggest trade-offs between precision and recall for the majority and minority classes.

4) Ablation Study: Our preliminary experiments revealed that using flattened data after sliding window segmentation with random forest and Adaboost classifiers yielded significantly better results compared to directly using frame-by-frame data for model training. Furthermore, we observed that downsampling improved the prediction accuracy for protective behavior, but resulted in a slight decrease in the prediction accuracy for non-protective behavior.

REFERENCES

[1] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion based music recommendation system using wearable physiological sensors," *IEEE transactions on consumer electronics*, vol. 64, no. 2, pp. 196–203, 2018.

[2] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[3] C. Lin, M. Liu, W. Hsiung, and J. Jhang, "Music emotion recognition based on two-level support vector classification," in *2016 International conference on machine learning and cybernetics (ICMLC)*, vol. 1, pp. 375–389, IEEE, 2016.

[4] T.-P. Jung, T. J. Sejnowski, *et al.*, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 96–107, 2019.

[5] P. Ekman, "Facial expression and emotion.," *American psychologist*, vol. 48, no. 4, p. 384, 1993.

[6] C. Wang, Y. Gao, A. Mathur, A. C. De C. Williams, N. D. Lane, and N. Bianchi-Berthouze, "Leveraging activity recognition to enable protective behavior detection in continuous data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 2, pp. 1–27, 2021.

[7] M. S. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, *et al.*, "The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal emopain dataset," *IEEE transactions on affective computing*, vol. 7, no. 4, pp. 435–451, 2015.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[9] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 1322–1328, IEEE, 2008.

[10] C. Wang, M. Peng, T. A. Olugbade, N. D. Lane, A. C. D. C. Williams, and N. Bianchi-Berthouze, "Learning temporal and bodily attention in protective movement behavior detection," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 324–330, IEEE, 2019.