Analyzing Human Poses for Emotion Recognition: Clustering and Supervised Learning Approaches

DENIS KRYLOV, University of Twente, The Netherlands

The accurate interpretation of emotions is crucial in enhancing various fields, such as assistive technologies and healthcare. People may mask their true emotions, as facial expressions are not always reliable indicators. This study explores the efficacy of using skeletal movements for emotion recognition. The research focuses on two primary questions. First, it evaluates the provided labels in the EiLA dataset by clustering skeletal movements into the seven basic emotions. Second, it examines the accuracy of different models in predicting emotions based on these movements. The methodology involves (1) extracting frames from video data, (2) using the PoseLandmarker algorithm to obtain normalized 3D coordinates of key skeletal points, (3) normalizing and truncating skeletal movements for consistency, and (4) converting them into feature vectors. These vectors are then clustered and used to train various models to determine their performance in emotion recognition. The average linkage method proved most effective for clustering skeletal movements into the seven basic emotions. However, qualitative analysis revealed challenges related to overlap and ambiguity in emotion labeling. Among the models evaluated, the Support Vector Machine (SVM) achieved the highest accuracy but exhibited moderate precision and recall, indicating difficulty in handling class imbalances. In contrast, the Random Forest model demonstrated more robust performance with the highest F1-Score, effectively identifying true positive emotions.

Additional Key Words and Phrases: emotion recognition, computer vision, skeletal movements, human-computer interaction, poselandmarker algorithm, hierarchical clustering, svm, random forest, neural network

1 INTRODUCTION

The capability to accurately interpret emotions can significantly impact various fields. For example, in the realm of assistive technologies, emotion-aware systems can be enhanced to be more empathetic and adaptive for individuals with disabilities. In the healthcare field, it can aid in the early diagnosis and monitoring of mental health conditions, providing a deeper understanding of patient well-being. Therefore, interpreting emotions is a pivotal task in the field of human-computer interaction.

Traditional methods of emotion recognition, which often rely on facial expressions, vocal intonations, or self-reported data, can be limiting or intrusive. There is a possibility that people may not reveal their emotions through facial expressions, as studied by Ekman et al. [9]. Additionally, individuals may feign emotions to mislead the observer. In such cases, cues from the experienced emotions can be extracted from other sources like body language [10].

The aim of this research is to determine the accuracy of defining emotions based on skeletal movements. To achieve this, the research focuses on answering the following questions:

- **RQ1**: *How effectively do the skeletal movements extracted from the EiLA dataset form clusters corresponding to the seven basic emotions?*
- **RQ2**: *How accurately can different models predict emotions using skeletal movements extracted from the EiLA dataset?*

The rest of the paper is organized as follows: Section 2 discusses the current state-of-the-art and describes technical background. Section 3 presents the methodology. Section 4 describes the experimental setup. Section 5 presents and describes the results. Section 6 analyzes and discusses the results, answers the research questions, and suggests future work. Finally, Section 7 presents the conclusions.

2 SCIENTIFIC BACKGROUND

2.1 Related works

Emotion recognition based on skeletal movements is relatively unexplored compared to methods based on facial expressions. However, the popularity of these approaches has increased in recent years. Costa et al. [4] summarized examples of correlations between body movements and emotions that were originally proposed by Darwin [6]. Additionally, Wallbott et al. [25] conducted a study on body language, specifically identifying behavioral cues related to body movements and language from six professional actors.

There are also works focused on emotion recognition from skeletal movements. For instance, Sapiński et al. [20] analyzed motion data captured under seven basic emotions using a Microsoft Kinect v2 sensor. Their performance measurements on CNN, RNN, and RNN-LSTM models resulted in an accuracy of 63%. Shichkina et al. [24] studied the correlation between emotions and body posture in a sitting position using a hardware-software system based on a posturometric armchair, achieving an overall accuracy of over 90% with various methods. Shi et al. [23] proposed an attention-based convolutional neural network and an attention-based fusion method to analyze emotions from videos, utilizing audio signals, skeletal data, and text information. Montepare et al. [19] identified emotions such as sadness, anger, happiness, and pride from gait information and found that specific cues could differentiate these emotions. For instance, angry movements were more heavyfooted, while sad movements had less arm swing comparing to the other gaits. In another study he concluded that negative emotions, particularly sadness, were more accurately recognized [18]. A similar study by Lima et al. [17] proposed the ST-Gait++ architecture to recognize four emotions (Anger, Happiness, Neutral, and Sadness), achieving an overall accuracy of 87.5%, which is an improvement of approximately 5% over the state-of-the-art. Kumar et al. [16] conducted a promising study using skeleton data obtained from a Microsoft Kinect v2 and a motion trajectory computation scheme using Fourier temporal features from the interpolation of skeleton joints, resulting in an overall accuracy of 95.32%.

TScIT 41, July 5, 2024, Enschede, The Netherlands

However, some of these studies require additional hardware (e.g., Microsoft Kinect v2 or a posturometric armchair) to obtain raw point cloud data, which can only be gathered under limited experimental conditions. Additionally, they may require supplementary inputs like voice or textual context to increase prediction accuracy [20, 23, 24]. In the current research, an analysis of raw skeletal movement data and its relation to primary emotions was proposed without the need for additional hardware or input data. The dataset used consists of video footage without audio signals from the TV show *MasterChef+ Brasil* (EiLA dataset).

2.2 Technical background

2.2.1 Pose Detection and Estimation. For pose estimation from videos, the PoseLandmarker was used. PoseLandmarker, developed by Google, detects and estimates poses from images or videos, returning normalized coordinates for each joint with estimated depth. This tool extracts 3D skeletal data from video frames. The model used is *PoseLandmarker (Full)* with an input size of $256 \times 256 \times 3$, returning 33 normalized points with (x, y, z) coordinates [1].

2.2.2 Clustering Analysis. Hierarchical clustering is a method of cluster analysis that builds a hierarchy of clusters, either agglomerative (bottom-up) or divisive (top-down). In this research, the bottom-up approach is used based on the dataset size. Agglomerative clustering helps analyze how well skeletal movements can be grouped, providing insights into the natural clustering of different emotions [13].

2.2.3 Classification Algorithms. Selecting appropriate classification algorithms is critical for accurately and reliably predicting emotions from 3D skeletal movements. Given the complexity of human emotions and the nature of skeletal movement data, a thorough comparison of different machine learning approaches is necessary. This research compares the performance of Support Vector Machines (SVM), Random Forest (RF), and Neural Networks (NN) in a multi-class classification setting.

Support Vector Machines. Support Vector Machines (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates different classes in the feature space. For non-linear classification, SVM uses kernel functions (e.g., RBF, polynomial) to project data into higher dimensions where linear separation is possible. SVM is suitable for this task due to its robustness with smaller datasets and its ability to handle multi-class classification [5, 22].

Random Forest. Random Forest is an ensemble learning method that constructs multiple decision trees during training and merges their results (either by averaging for regression or majority voting for classification). It is known for its simplicity and effectiveness in handling various types of data. Random Forest has high accuracy and robustness to over-fitting due to the averaging of multiple trees and can generalize well on smaller datasets, providing robust predictions [11, 21].

Neural Network. Neural Networks, especially deep learning models, are powerful tools for capturing complex patterns in data. They consist of multiple layers of neurons where each layer extracts higher-level features from the input data. Despite the small dataset, Neural Networks can be effective with proper data augmentation and regularization [2].

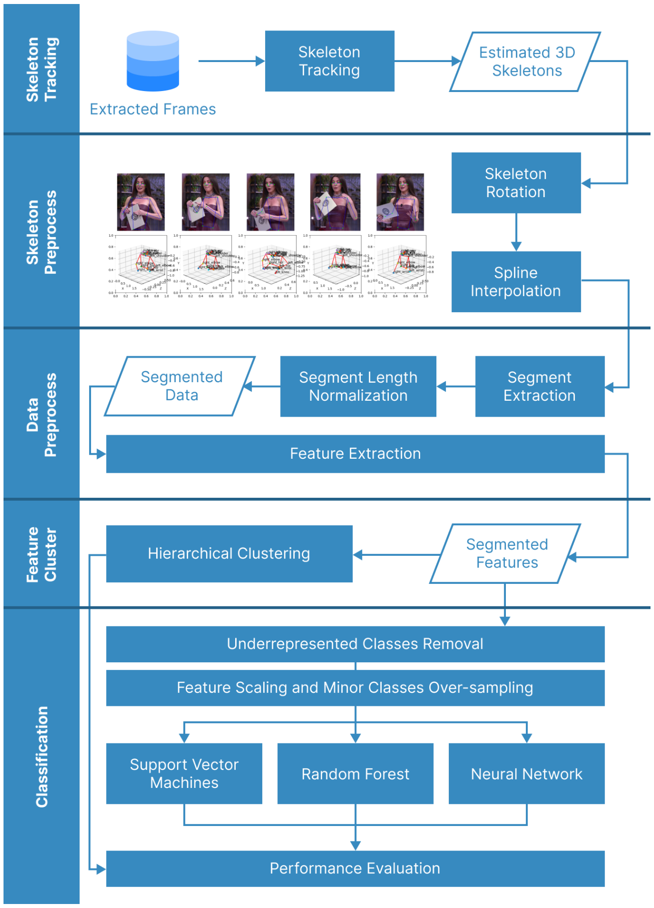


Fig. 1. Methodology Overview

3 METHODOLOGY

The overall methodology is depicted in Figure 1. A detailed description of each step follows below.

3.1 Skeleton Tracking

Prior to skeleton extraction, each frame underwent pre-processing steps including determining the boundary box for a person with *Person Id*, cropping the frame using this boundary box, resizing the image, and applying the PoseLandmarker algorithm for pose estimation, as illustrated in Figure 2.

For all videos provided in the EiLA Dataset, a total of 8087 frames were extracted, with the PoseLandmarker algorithm applied to each frame. The PoseLandmarker algorithm returns normalized detected points in the format (x, y, z), where x and y denote the coordinates of the joint on the frame, and z represents the estimated depth (Figure 3). Normalized points mitigate the influence of varying pose sizes. The PoseLandmarker returns 33 points for each frame, which

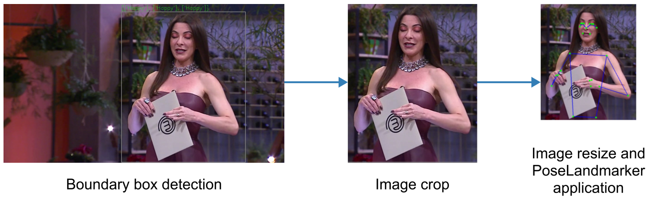


Fig. 2. Dataset Preparation Overview

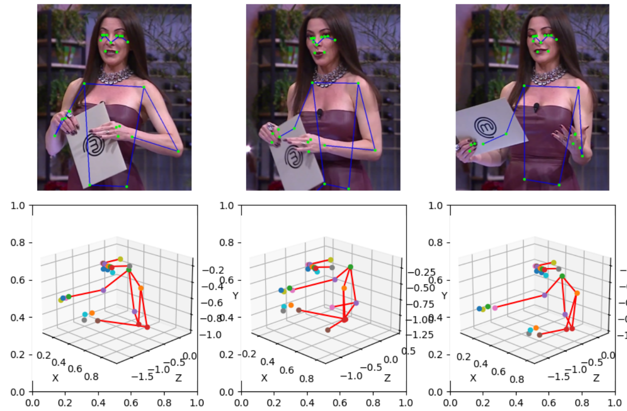


Fig. 3. Result of joints detection using PoseLandmarker

were reduced to 25 key points: nose (1), eyes (6), mouth (2), ears (2), shoulders (2), elbows (2), wrists (2), hips (2), and fingers (6) [12]. This combination of points, representing the human posture for each frame, is termed as the skeleton [7]. The resulting skeleton for each frame was integrated with original annotations, adding 75 additional columns for the X, Y and Z coordinates of each joint to the annotations file. An example of how these points are presented can be seen in Table 1.

Table 1. An example table of extracted skeleton points

nose_X	nose_Y	nose_Z	...
1.1883129256	0.18981633247	-0.04783265416	...
1.2084573548	0.19331540014	0.012150476112	...
1.356732110	0.18915118244	-0.22049692826	...
...	...	...	...

3.2 Skeleton Preprocess

Skeleton Rotation. Human actions are composed as a series of skeletons. To ensure consistency regardless of absolute body position and initial body orientation, all skeletons are transformed by rotating the coordinate system. Inspired by Jian et al.'s work [14], the skeleton is rotated using a rotation matrix R computed via Rodrigues' rotation formula. The rotation matrix was computed as following:

$$R = I + sin_{\theta} \cdot K + (1 - cos_{\theta}) \cdot K^2 \tag{1}$$

where:

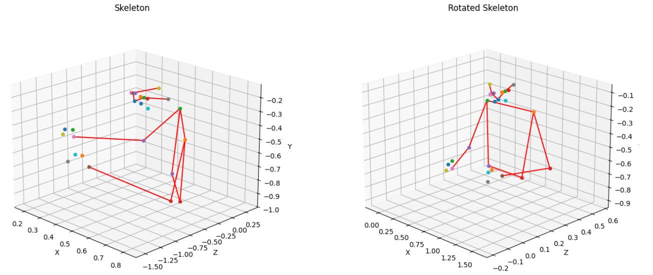


Fig. 4. Original skeleton (left) and rotated skeleton (right)

- R is the rotation matrix.
- I is the 3×3 identity matrix.
- θ is the angle of a counterclockwise rotation in radians.
- $\sin_{\theta}$ and $\cos_{\theta}$ are the sine and cosine of the angle θ, respectively.
- K is a unit vector, associated with Lie algebra $\mathfrak{so}(3)$.

The rotation angle θ is calculated such that the vector $\vec{H_R H_L}$ (from right hip H_R to left hip H_L) becomes parallel to the x-axis (Figure 4). Now the skeleton is independent from different viewpoints [14].

Spline Interpolation. Some joints may have missing values. To deal with that, the spline interpolation will be applied [16]. In the context of 3D skeleton joints, where the motion is continuous and smooth, spline interpolation can effectively fill in missing joint positions without introducing abrupt changes or discontinuities. In our research the interpolation used a cubic spline $S(x)$ with ($k = 3$) that consists of $m - 1$ polynomial segments, one for each interval $[x_i, x_{i+1}]$. Then all cubic polynomials were constructed for each interval with the form

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \tag{2}$$

where $x \in [x_i, x_{i+1}]$, a_i is a constant, b_i, c_i, d_i are linear, quadratic and cubic coefficients consecutively. The system of equations was solved to determine the missing values.

3.3 Data Preprocess

Segment Extraction and Normalization. In the EiLA dataset, the combination of *Video Tag*, *Clip Id*, and *Person Id* uniquely identifies a person. A set of frames representing a specific person is referred to as a segment. Each segment contains a sequence of frames capturing a particular motion, i.e., a series of skeletons related to the provided label. Before extracting features from a segment, its length is normalized to ensure all segments have an equal number of frames. During the normalization process, a target number of frames is defined. If the target is lower than the current number of frames, redundant frames are equally removed from the center. Conversely, if the target is higher, frames are duplicated in the same manner. As a result of this length normalization, all segments have an equal length matching the target length. The total number of normalized segments in this study is 326.

Feature Extraction. To represent a configuration of an entire skeleton, inspired by Ding et al [8] and Costa et al [4], a distance vectors were obtained from the most important skeleton points O [12]. O contains coordinates of a nose (1), shoulders (2), elbows (2), wrists (2), hips (2). O' is a set of all other coordinates in such a way that all distances from all elements $o \in O$ were obtained to all $p_n \in O'$ where $n \in [0, 23]$ and $o \neq p_n$. O' contains nose (1), eyes (6), mouth (2), ears (2), shoulders (2), elbows (2), wrists (2), hips (2) and fingers (6).

Let's take an arbitrary segment S_l, $o \in O$ as an original point where $i \in [0, 9]$ is a sequential number of a frame in S_l:

$$o = \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_i & y_i & z_i \end{pmatrix} \tag{3}$$

where $\{(x_i, y_i, z_i)\}_{i=1}^{k}$ are the coordinates of the original point o.

Then we identify all other points and their coordinates, creating the matrix of coordinates for every other points:

$$p_n = \begin{pmatrix} x'_1 & y'_1 & z'_1 \\ x'_2 & y'_2 & z'_2 \\ \vdots & \vdots & \vdots \\ x'_i & y'_i & z'_i \end{pmatrix} \tag{4}$$

where $\{(x'_i, y'_i, z'_i)\}_{i=1}^{k}$ are the coordinates of the other estimated point p_n.

For each frame i we compute the euclidean distance between the original point o and the every other point p_n:

$$d_{o,p_n,i} = \sqrt{(x_i - x'_i)^2 + (y_i - y'_i)^2 + (z_i - z'_i)^2} \tag{5}$$

At last, a 1680-dimensional vector of distances for each segment S_l was created (Equation 6). The distances ensure that during the classification the skeletons will be independent from their absolute joint positions.

$$d_{S_l} = [d_{o_0,p_0,0}, d_{o_0,p_0,1}, \ldots, d_{o_{10},p_{23},9}] \tag{6}$$

3.4 Feature Clustering

Before attempting to classify poses and their associated emotions, hierarchical clustering was applied to the set of feature vectors d_S (Equation 6). Specifically, *Agglomerative Clustering* was used to assess the distinctiveness of these poses independently. After each clustering iteration, Silhouette and Davies Bouldin scores were computed.

3.5 Classification

Data Preparation. Prior to training the models for classifying skeletons according to their emotional labels, it is crucial to ensure balanced data where all classes have sufficient samples for accurate classification. Therefore, underrepresented classes were removed from consideration. In the current study, the models were trained to classify only three emotions: *Neutral*, *Happy*, and *Anger*. After removing underrepresented classes and performing a stratified data split, the training data was further processed using SMOTE [3] to balance the class distribution, providing additional information to the models and mitigating bias in the classification process.

Classification Methods. The classification task utilized *Support Vector Machines* (SVM), *Random Forest*, and *Neural Network* algorithms from the Scikit-learn library [15]. Each method was applied to the set of feature vectors d_S (Equation 6). Finally, the performance of these models was evaluated with different training and testing splits.

4 EXPERIMENTAL SETUP

4.1 EiLA Dataset

EiLA (Emotions in LatAM) is a dataset provided by the University of Twente, designed for emotion recognition among individuals from Latin American backgrounds. The dataset consists of 9 videos from the MasterChef+ Brasil show. This dataset was chosen because individuals in these videos experience a wide range of emotions during the cooking process. The dataset annotations (see Table 3) cover 8087 frames from these videos and include the following annotations for each frame:

- **Video Tag**: Video identification present on YouTube.
- **Clip Id**: Identification for each clip from a source video, unique within the source video.
- **Labels**: Arrays containing labels assigned by each annotator of the dataset.
- **Frame Number**: Frame used for the annotation.
- **X**: Starting position of the bounding box on the x-axis.
- **Y**: Starting position of the bounding box on the y-axis.
- **Width**: Percentage of the video's width used as an offset for "X".
- **Height**: Percentage of the video's width used as an offset for "Y".
- **Pid**: Integer identifying a specific person for clips with the same "Video Tag" and "Clip Id".

The resulting sample counts per class (after segmentation) are presented in Table 2. Some classes are heavily underrepresented, specifically *Sad*, *Surprise*, *Disgust*, and *Fear*.

Table 2. Count of samples per class after segmentation

Neutral	Happy	Anger	Sad	Surprise	Disgust	Fear
161	87	42	11	10	8	7

4.2 Validation metrics

To assess the performance of the classifiers in predicting emotions from skeletal movements, a variety of validation metrics will be employed. These metrics will provide insights into the accuracy of our models, their effectiveness in handling multi-class classification tasks, and their robustness in various scenarios.

4.2.1 Accuracy. Accuracy represents the ratio of correctly predicted instances to the total instances in the dataset, providing an overview of the model's overall performance.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{7}$$

Table 3. An annotations for each frame for EiLA dataset

Video Tag	Cid	Labels	Frame Number	X	Y	Width	Height	Pid
aJKL0ahn1Dk	1	[['Happy'], ['Happy'], ['Happy']]	19532	41.965200	4.873195	44.216991	94.802684	0
aJKL0ahn1Dk	1	[['Happy'], ['Happy'], ['Happy']]	19538	41.564836	4.874640	44.216991	94.802684	0
aJKL0ahn1Dk	1	[['Happy'], ['Happy'], ['Happy']]	19544	41.164472	4.876086	44.216991	94.802684	0
aJKL0ahn1Dk	1	[['Happy'], ['Happy'], ['Happy']]	19550	40.764108	4.877532	44.216991	94.802684	0
aJKL0ahn1Dk	1	[['Happy'], ['Happy'], ['Happy']]	19556	39.646728	5.014136	44.216991	94.802684	0
...	...	...	...	...	...	...	...	...

4.2.2 *Precision, Recall and F1-Score.*

Precision indicates the proportion of true positive predictions among all positive predictions, showing the accuracy of the positive ones.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{8}$$

Recall measures the proportion of true positive predictions among all actual positive instances, indicating how well the model captures positive instances.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{9}$$

F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both precision and recall.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

4.2.3 *Confusion Matrix.*

The confusion matrix is a detailed table that presents the performance of the classification model by comparing actual vs. predicted classifications. It provides a visual representation of the model's performance across all classes, highlighting where the model misclassifies instances. The confusion matrix is crucial for understanding error distribution and identifying which classes are frequently confused with each other.

4.3 Implementation details

In the following section the implementation details will be discussed. The overall project was implemented using Python 3.12.3 version. The source code can be accessed in the GitHub repository.

4.3.1 *Data Split.*

The stratified data split was performed and the three distinct subsets were formed: training, validation and test.

- **Training set (80%)**.
- **Validation set (10%)**: used for tuning hyper-parameters and preventing over-fitting. Used only for NN validation.
- **Test set (20%)**: used to evaluate the models on unseen data.

4.3.2 *Hyper-parameters.*

To determine the optimal hyper-parameters, hyper-parameter tuning was conducted using the grid search technique. The selected hyper-parameters for each model are detailed in Table 4. The Neural Network architecture comprises 4 layers: (1) Input dense layer with 64 units, ReLU activation function, L2 regularization, and input shape corresponding to the number of features per sample; (2) Dropout layer, (3) Hidden layer with 35 units, ReLU activation function, and L2 regularization; (4) Output dense layer with 3 units and Softmax activation function.

Table 4. Best hyper-parameters for each model

Model	Hyper-parameter	Best Value
SVM	C	10
	Kernel	RBF
	Gamma	0.1
Random Forest	Number of Estimators	200
	Max Depth	None
	Min Samples Split	5
	Min Samples Leaf	1
Neural Network	Batch Size	64
	Epochs	50
	Dropout Rate	0.3
	L2 Regularization	0.01
	Optimizer	Adam

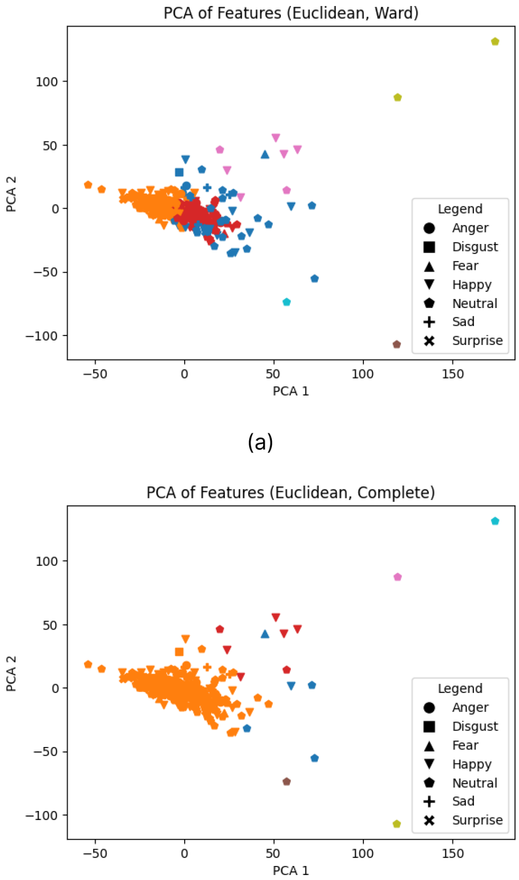
5 RESULTS

5.1 Clustering

Table 5 presents the cluster results for different numbers of target classes using various linkage methods and metrics. The table compares the clustering performance using the Silhouette coefficient and the Davies-Bouldin index. The clustering performed for the original dataset with 7 classes. The Silhouette score measures how similar an object is to its own cluster compared to other clusters. According to Table 5, the highest Silhouette score (0.7309) is achieved with 3 clusters using the Euclidean distance and Complete linkage method. The Silhouette score is also relatively high (0.7240) with the Average linkage method. The lowest Silhouette score (0.1507) was obtained with the Euclidean distance and Ward linkage method. The Davies-Bouldin index measures cluster separation, with lower

Table 5. Cluster results for different number of target classes

Clusters	Metric	Linkage	Silhouette	Davies-Bouldin
7	euclidean	ward	0.1507	1.5080
5	euclidean	ward	0.1663	1.4382
3	euclidean	ward	0.1603	1.5024
7	euclidean	average	0.5648	0.5397
5	euclidean	average	0.6225	0.6528
3	euclidean	average	0.7240	0.1723
7	euclidean	complete	0.5137	0.9460
5	euclidean	complete	0.5270	0.9960
3	euclidean	complete	0.7309	0.8466



(a)

(b)

Fig. 5. PCA clustered features

values indicating better-defined clusters. The lowest Davies-Bouldin index (0.1723) was obtained for 3 clusters with Euclidean metric and Average linkage. For 7 clusters with the same metric and linkage, the index was also relatively low. The highest value was obtained with 7 clusters, Euclidean metric, and Ward linkage.

The PCA plots visualize the clustering results in 2D space using the first two principal components for dimensionality reduction. Figure 5 (a) shows the PCA of features using the Ward linkage method, and (b) shows the PCA of features using the Complete linkage method. Each plot depicts the clustering of different emotions, with each color and shape representing a distinct emotion.

5.2 Classification

Classification results for different models are provided in Table 6. The classification was performed on the truncated dataset with 3 classes after SMOTE [3] resampling. Example confusion matrices are shown in Figure 7. According to the results in Table 6, SVM achieved an average accuracy of 0.5816 ± 0.0310 and an F1-Score of 0.4672 ± 0.0448. Random Forest attained an accuracy of 0.5539 ± 0.0563 and an F1-Score of 0.5114 ± 0.0608. Finally, the Neural Network showed an accuracy of 0.4916 ± 0.0758 and an F1-Score of 0.4174 ± 0.0300.

Table 6. Classification results for different models

Model	**Accuracy (Mean)**	**F1-Score (Mean)**
SVM	0.5816 ± 0.0310	0.4672 ± 0.0448
Random Forest	0.5539 ± 0.0563	0.5114 ± 0.0608
Neural Network	0.4916 ± 0.0758	0.4174 ± 0.0300

In the example in Figure 7, (a) SVM demonstrates exceptional performance in identifying Anger, achieving a 100% accuracy rate. However, it shows significant confusion in classifying Happy and Neutral emotions, correctly identifying only half of the Happy instances and misclassifying the rest as Neutral. Similarly, a substantial portion of Neutral instances is misclassified as Happy or Anger. In (b), Random Forest shows improved performance for Neutral emotions (67.74%), but decreased for Happy (33.33%) and Anger (14.29%). In (c), Neural Network exhibits balanced but lower performance across all emotions: Anger (14.29%), Happy (28.57%) and Neutral (56.67%).

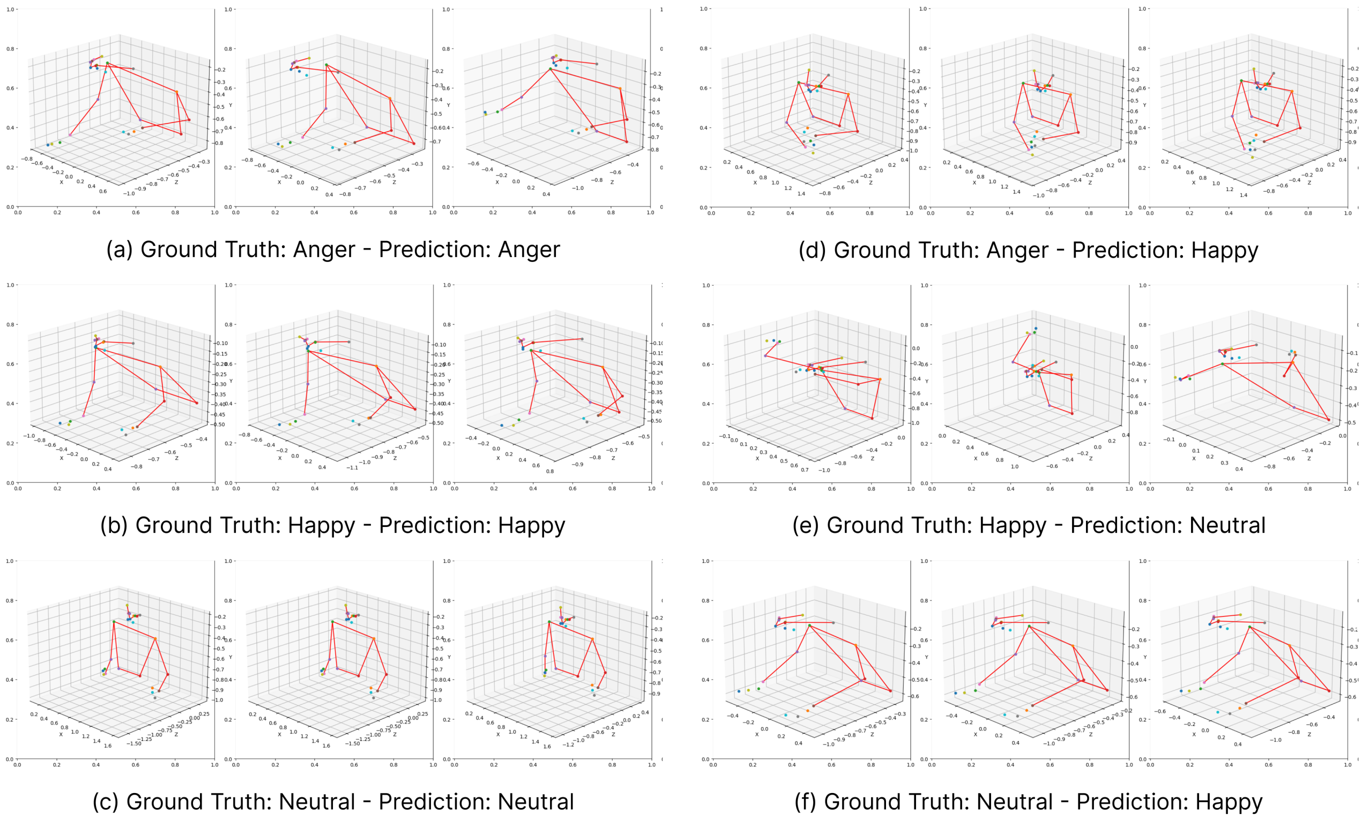
6 DISCUSSION

6.1 Answer to RQs

6.1.1 *RQ1: How effectively do the skeletal movements extracted from the EiLA dataset form clusters corresponding to the seven basic emotions?*

Quantitative analysis. Low Silhouette scores and high Davies-Bouldin scores for the Ward linkage method suggest that clusters are not well-separated and less distinct, indicating ineffectiveness in forming clear clusters corresponding to the seven basic emotions. Conversely, the Average linkage method demonstrates improved clustering effectiveness. The Silhouette scores are significantly higher, particularly for fewer clusters. The Davies-Bouldin index also shows marked improvement, suggesting more compact and distinct clusters. These results indicate that the Average linkage method forms clearer and more distinct clusters, making it more effective for clustering skeletal movements corresponding to emotions. Very similar results can be achieved with Complete linkage method.

Qualitative analysis. The PCA visualizations in Figure 5 cast doubt on the quality of classification using the Average method. In (a), the Ward linkage method shows more overlap and less distinct clusters. In contrast, (b) shows better-separated clusters with the Complete linkage method, although some overlap persists. However, upon closer inspection of the labels, it is evident that with both Complete and Average methods, most labels cluster into one group, with only a few assigned to different clusters. This occurs due to the similarity of skeletal movements in the EiLA dataset. Furthermore, as shown in Table 3, emotions labeled by different annotators can vary, adding ambiguity even for human identification of emotions. Ward linkage provides more consistent clustering resembling the original label distribution, but with many labels assigned to multiple clusters due to high feature density in the feature space, which aligns with lower Silhouette scores and higher Davies-Bouldin indices.



(a) Ground Truth: Anger - Prediction: Anger

(d) Ground Truth: Anger - Prediction: Happy

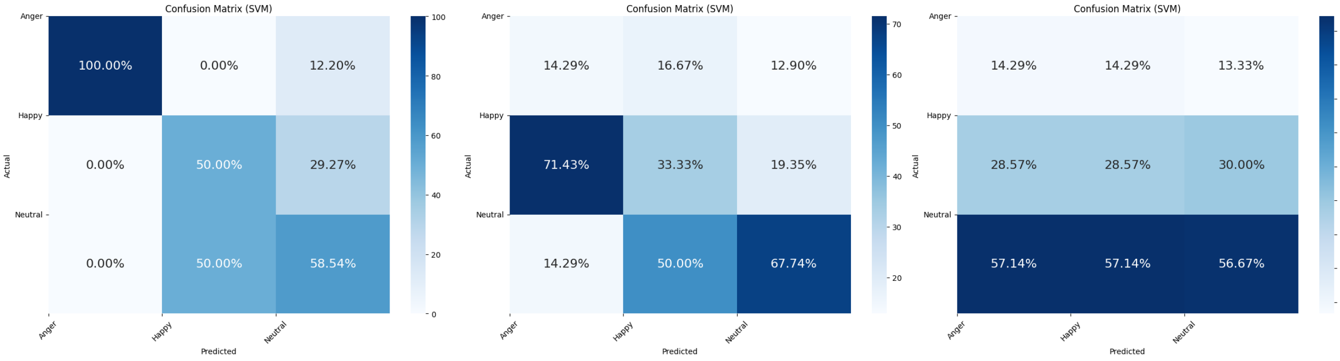
(b) Ground Truth: Happy - Prediction: Happy

(e) Ground Truth: Happy - Prediction: Neutral

(c) Ground Truth: Neutral - Prediction: Neutral

(f) Ground Truth: Neutral - Prediction: Happy

Fig. 6. Predictions of 3 basic emotions (first 3 frames, SVM)



(a)

(b)

(c)

Fig. 7. Example confusion matrices for SVM (a), Random Forest (b) and Neural Network (c)

6.1.2 RQ2: How accurately can different models predict emotions using skeletal movements extracted from the EiLA dataset?

Quantitative analysis. SVM achieved the highest accuracy among all three models, but its relatively low F1-Score suggests moderate precision and recall in classifying emotions. This indicates that SVM performs reasonably well in distinguishing between classes but may struggle with imbalanced classes or capturing nuanced data patterns. On the other hand, the Random Forest model resulted in the highest F1-Score, implying a better balance between precision and recall. This suggests that the Random Forest model is more effective in identifying true positive emotions and may handle class imbalances better (Equation 10). Finally, the Neural Network showed the lowest accuracy and F1-Score, likely due to its complex nature requiring more data to fully capture the nuances of skeletal movements. At last, all models showed low standard deviation for both accuracy and

F1-Score suggesting that the different data split didn't significantly affect the results.

Comparing to previous studies such as Sapiński et al. [20], who achieved 63% accuracy using CNN, RNN, and RNN-LSTM models on Kinect data, SVM and Random Forest models in the current study applied on the EiLA dataset show competitive performance given the different datasets and absence of additional hardware or contextual data. Other studies by Shichkina et al. [24] and Shi et al. [23] reported higher accuracies (over 90%), but they relied on additional hardware (e.g., Kinect v2, posturometric armchair) and multi-modal data (audio, text). Current approach using only video-derived skeletal data inherently faces more challenges, explaining the lower accuracies but also highlighting the potential of using skeletal data alone.

Qualitative analysis. In Figure 6, examples of correct and incorrect predictions made by SVM are illustrated. By looking at the skeleton's it is very hard to distinguish the emotions that people are experiencing. For instance, anger can be described as an experienced emotion, which should be followed by tremble, purposeless gestures, shaking fists and expanded chests, as observed in (a). In (b), wide arm movements might suggest Happiness. In (c), minimal changes in posture indicate a Neutral state [4].

Conversely, in (f), a pose similar to (b) results in an incorrect prediction of Happiness. In (e), despite purposeless gestures indicating Happiness, a depth estimation error by PoseLandmarker caused a distorted skeleton, leading to a Neutral prediction. In (d), subtle signs of Anger are hard to detect, and wide elbow positions might incorrectly predict Happiness, similar to (b). Overall, these examples underscore the uncertainty in the data, contributing to understandable prediction errors.

6.2 Limitations

Dataset Imbalance. One significant limitation encountered was the imbalance in the EiLA dataset, as indicated in Table 2. Classes such as *Sad*, *Surprise*, *Disgust*, and *Fear* were heavily underrepresented compared to more common classes like *Neutral*, *Happy*, and *Anger*. This imbalance necessitated the removal of these classes to train a model that would not be biased towards the most frequently occurring classes.

Size of Dataset. Another important limitation was the overall size of the dataset. When grouping 8087 frames into segments S_i representing motions, the number of samples decreased significantly (from 8087 samples to 645). This reduction in sample size posed challenges in training a model capable of generalizing well and robustly capturing nuanced motion patterns related to human emotions. In addition, EiLA dataset was originally designed to analyze human emotions based on facial expressions, leading to some frames containing only the facial parts of humans without their body parts. These frames, and consequently the segments they formed, had to be removed, further reducing the dataset size (from 645 initial segments to 326).

PoseLandmarker Depth Estimation. PoseLandmarker, used in this research, estimates the depth of joints. However, the accuracy of this estimation can vary, as illustrated in Figure 3. For instance, when a person holds their hands close to the body, the estimated 3D points may show the wrists as relatively far from the body. Additionally, the current method does not account for scenarios where a person may be holding objects, which can influence skeletal motion.

6.3 Future work

The future work should start by addressing the limitations of the current research. The larger dataset should be used with frames, containing the mostly upper body of humans, not only faces. In addition, dataset can be expanded (or another dataset should be chosen) that contains balanced classes for each of the basic emotions. Also, the methodology can be expanded by using Fourier temporal features from the interpolation of skeleton joints to identify actions of each human body part first [16], and then converted vector of features that represents these identified actions, can be used to train emotions classification model. In that case two models need to be trained: one will predict the actions of the human body parts, and the other one classify the emotions based on these actions. Finally, it would be advantageous to consider scenarios where individuals may be holding objects during skeletal movement analysis.

7 CONCLUSIONS

This study aimed to evaluate the effectiveness of clustering skeletal movements for emotion recognition and assess the accuracy of various models using the EiLA dataset. The findings provide detailed insights into both clustering techniques and model performances.

The Average linkage method emerged as the most effective for clustering skeletal movements into the seven basic emotions, showing higher Silhouette scores and lower Davies-Bouldin indices compared to Ward linkage and Complete linkage methods. However, qualitative analysis highlighted significant overlap and ambiguity despite improved clustering performance. This is partly due to inherent similarities in skeletal movements and subjective variations in emotion labeling.

Among the models evaluated, the Support Vector Machine (SVM) achieved the highest accuracy but exhibited moderate precision and recall, indicating challenges in handling class imbalances. The Random Forest model demonstrated a better balance with the highest F1-Score, showcasing robustness in identifying true positive emotions. In contrast, the Neural Network model performed the poorest, possibly due to its complexity and the dataset's limited size.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr. Estefanía Talavera Martínez, for her continuous support. The success of this research was significantly enhanced by our regular meetings and her invaluable guidance.

REFERENCES

[1] Google AI. 2024. MediaPipe Pose Landmarker. https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker. https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker Accessed: 2024-06-30.

[2] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. 2017. *Deep learning*. Vol. 1. MIT press Cambridge, MA, USA.

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[4] Willams Costa, Estefanía Talavera, Renato Oliveira, Lucas Figueiredo, João Marcelo Teixeira, João Paulo Lima, and Veronica Teichrieb. 2023. A survey on datasets for emotion recognition from vision: Limitations and in-the-wild applicability. *Applied Sciences* 13, 9 (2023), 5697.

[5] Nello Cristianini and John Shawe-Taylor. 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

[6] Charles Darwin. 1948. *The Expression of the Emotions in Man and Animals: The Expression of the Emotions in Man and Animals: Charles Darwin's Seminal Study on Emotional Expression*. Prabhat Prakashan.

[7] Prasad Dhore, Aparna Pande, Shital Mehta, and Saili Sable. 2022. Human Pose Estimation And Classification: A Review. *Neuroquantology* 20, 15 (2022), 3199.

[8] Weili Ding, Bo Hu, Han Liu, Xinming Wang, and Xiangsheng Huang. 2020. Human posture recognition based on multiple features and rule learning. *International Journal of Machine Learning and Cybernetics* 11 (2020), 2529–2540.

[9] Paul Ekman. 1993. Facial expression and emotion. *American psychologist* 48, 4 (1993), 384.

[10] Paul Ekman, Maureen O'Sullivan, Wallace V Friesen, and Klaus R Scherer. 1991. Invited article: Face, voice, and body in detecting deceit. *Journal of nonverbal behavior* 15, 2 (1991), 125–135.

[11] Trevor Hastie, Robert Tibshirani, Jerome Friedman, Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. Random forests. *The elements of statistical learning: Data mining, inference, and prediction* (2009), 587–604.

[12] Ruth M Holmes, Ellen Rushe, and Anthony Ventresque. 2024. The Key Points: Using Feature Importance to Identify Shortcomings in Sign Language Recognition Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 15970–15975.

[13] Anil K Jain and Richard C Dubes. 1988. *Algorithms for clustering data*. Prentice-Hall, Inc.

[14] Min Jiang, Jun Kong, George Bebis, and Hongtao Huo. 2015. Informative joints based human action recognition using skeleton contexts. *Signal Processing: Image Communication* 33 (2015), 29–40.

[15] Oliver Kramer and Oliver Kramer. 2016. Scikit-learn. *Machine learning for evolution strategies* (2016), 45–53.

[16] Naresh Kumar and Nagarajan Sukavanam. 2018. Motion trajectory for human action recognition using fourier temporal features of skeleton joints. *Journal of Image and Graphics* 6, 2 (2018), 174–180.

[17] Maria Luísa Lima, Willams De Lima Costa, Estefania Talavera Martínez, and Veronica Teichrieb. 2024. ST-Gait++: Leveraging spatio-temporal convolutions for gait-based emotion recognition on videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 302–310.

[18] Joann Montepare, Elissa Koff, Deborah Zaitchik, and Marilyn Albert. 1999. The use of body movements and gestures as cues to emotions in younger and older adults. *Journal of Nonverbal Behavior* 23 (1999), 133–152.

[19] Joann M Montepare, Sabra B Goldstein, and Annmarie Clausen. 1987. The identification of emotions from gait information. *Journal of Nonverbal Behavior* 11 (1987), 33–42.

[20] Tomasz Sapiński, Dorota Kamińska, Adam Pelikant, and Gholamreza Anbarjafari. 2019. Emotion recognition from skeletal movements. *Entropy* 21, 7 (2019), 646.

[21] scikit-learn developers. 2024. *RandomForestClassifier*. scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html Scikit-learn: Machine Learning in Python.

[22] scikit-learn developers. 2024. *Support Vector Machines*. https://scikit-learn.org/stable/modules/svm.html Accessed: 2024-06-30.

[23] Jiaqi Shi, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. 2021. 3D skeletal movement-enhanced emotion recognition networks. *APSIPA Transactions on Signal and Information Processing* 10 (2021), e12.

[24] Yulia Shichkina, Olga Bureneva, Evgenii Salaurov, and Ekaterina Syrtsova. 2023. Assessment of a Person's Emotional State Based on His or Her Posture Parameters. *Sensors* 23, 12 (2023), 5591.

[25] Harald G Wallbott and Klaus R Scherer. 1986. Cues and channels in emotion recognition. *Journal of personality and social psychology* 51, 4 (1986), 690.

A USE OF AI TOOLS

During the preparation of this work, the author used ChatGPT to review the grammatical correctness of sentences and ensure adherence to academic style. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the work.

B GITHUB REPOSITORY

https://github.com/denskrlv/PosEmotion