

1. Introduction

Emotion recognition is a core problem in affective computing, with applications in human-computer interaction, healthcare, education, and surveillance. While facial expressions and speech signals are the most commonly studied features, body language and posture also convey meaningful and important emotional information. Importantly, body-based approaches offer advantages in scenarios where facial data are unavailable or privacy concerns limit video or audio usage.

Studies have demonstrated that body movements can effectively reflect changes in affective state, even among primates. People pay more attention to body expressions than facial expressions or voices when dealing with affective states such as information in high intensity, perceptual ambiguity conditions, or when information from these channels is incongruent. As increasing psychological studies indicated the significant role of body movement in transmitting information and emotional states, artificial intelligence for emotion recognition is changing from facial expression system or body expression system to a multi-channel information combination.

In this project, we study emotion recognition from body language using pose coordinates only. The task is formulated as a multi-class classification problem in which skeletal joint coordinates are mapped to discrete emotion categories. The primary objective is not only to achieve high classification accuracy but also to analyze the limitations and generalization behavior of pose-based emotion recognition models under realistic evaluation settings.

Dataset Overview

The dataset consists of short clips consisting of 97 frames represented by 2D human pose landmarks. Each clip contains a sequence of frames with joint coordinates obtained from pose estimation systems. For each frame, the dataset provides:

- X and Y coordinates of body joints.
- Confidence scores indicating landmark reliability.
- Subject identifiers, enabling subject-based data splitting.
- Emotion labels corresponding to the clip.

There are seven emotion classes available in the dataset: Anger (A), Disgust (D), Fear (F), Happy (H), Neutral (N), Sad (SA), and Surprise (SU).

This structure allows for modeling both spatial posture and temporal motion patterns while enforcing subject-independent evaluation.

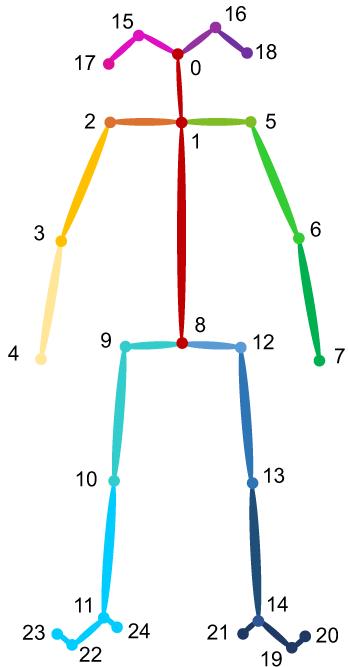
Evaluation Metrics

To assess model performance, we use:

Accuracy: measuring overall classification correctness

Macro-averaged F1-score: which equally weights all emotion classes and is robust to class imbalance

Macro F1-score is selected as the primary metric for model comparison.



0 - Nose	13 - Left knee
1 - Neck	14 - Left ankle
2 - Right shoulder	15 – Right eye
3 - Right elbow	16 – Left eye
4 - Right wrist	17 – Right ear
5 - Left shoulder	18 – Left ear
6 - Left elbow	19 - Left big-toe
7 - Left wrist	20 - Left small-toe
8 - Mid hip	21 - Left heel
9 - Right hip	22 - Right big-toe
10 - Right knee	23 - Right small-toe
11 - Right ankle	24 - Right heel
12 - Left hip	

Figure 1

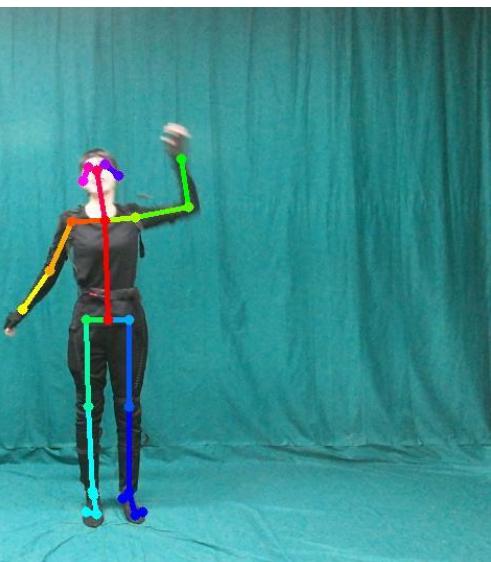


Figure 2



Figure 3

2. Related Work

Accurate recognition of human emotions from non-verbal cues has long been a central topic in affective computing and human-computer interaction. While facial expressions and speech signals have been extensively studied, recognition based on body language and postural movement has received comparatively less attention, in part due to the limited availability of high-quality annotated datasets.

Zhang *et al.* introduced the Multi-view Emotional Expressions Dataset (MEED), a large 2D pose dataset specifically designed to support research in bodily emotion recognition. MEED consists of 4,102 video recordings capturing six basic emotional expressions—anger, disgust, fear, happiness, sadness, and surprise—as well as neutral body movements, from three viewpoints (left, front, and right). The dataset includes both pose estimation results (e.g., coordinates and confidence scores for 25 body keypoints generated by OpenPose) and corresponding video frames, enabling comprehensive analysis of pose dynamics from multiple perspectives.

Importantly, MEED addresses critical gaps in earlier datasets by providing multi-view 2D pose annotations at scale, with nearly 400,000 pose frames and associated metadata, facilitating research that bridges controlled laboratory motion capture and more naturalistic body expressions. The dataset has been applied in various fields including affective computing, human–computer interaction, social neuroscience, and psychology due to its combination of emotion labels, multi-view pose data, and actor consistency across scenarios.

Beyond MEED, several works have explored automatic interpretation of body language for emotion and behavior understanding. Research in this area often emphasizes the importance of kinematic and postural features—including velocity, acceleration, and symmetry—as critical cues in affect recognition. For instance, studies that model body joint dynamics demonstrate that such features can provide meaningful discriminative information when compared to static pose snapshots.

Compared to these prior works, our project builds on the MEED dataset by not only employing raw 2D pose coordinates but also deriving biomechanically informed features such as joint angles, motion dynamics, and symmetry measures. This enables more expressive representations of emotional behavior beyond simple coordinate statistics, and our comparative evaluation of baseline and engineered feature approaches illustrates the value of richer feature design in emotion classification.

3. Models

Model 1 – Baseline Model

Model 1 is adapted from a publicly available GitHub repository implementing a pose-based emotion classification baseline. The repository was selected due to its clean structure, reproducibility, and clear separation between data loading, feature extraction, and classification stages.

Feature Representation

For each clip, Model 1 summarizes the pose sequence by computing:

- Mean of X coordinates
- Mean of Y coordinates
- Standard deviation of X and Y coordinates
- Mean and standard deviation of confidence scores

With 25 body joints, this results in a 150-dimensional feature vector. While computationally efficient, this representation captures only coarse spatial statistics and ignores explicit motion, posture, or symmetry information.

Classifiers

The following classical machine learning models are evaluated:

- Logistic Regression
- Ridge Classifier
- Random Forest
- Gradient Boosting

All models are trained using the same feature representation.

Training Setup

A subject-based train–test split is applied using group-aware splitting to avoid identity leakage. However, model selection is performed directly on test-set performance, which introduces a potential evaluation bias.

Model 2

Model 2 is a significantly enhanced pipeline designed to address the limitations of Model 1.

Feature Engineering

Instead of raw coordinate statistics, Model 2 extracts biomechanically and behaviorally meaningful features, including:

- Posture features (torso tilt, shoulder and hip alignment)
- Joint angle statistics (elbows and knees)
- Motion features (velocity of key joints such as wrists, shoulders, and torso)
- Symmetry measures (left-right differences)
- Confidence coverage metrics

All features are normalized, producing a compact 45-dimensional feature vector that captures posture, movement, and reliability more explicitly than Model 1.

In addition to classical classifiers, Model 2 introduces: Multi-layer Perceptron (MLP) for non-linear decision boundaries.

Training Scheme

Data is split into train, validation, and test sets in a subject-independent manner. Hyperparameters are selected based on validation performance, and final results are reported on the held-out test set, ensuring methodological rigor.

Github repository: <https://github.com/manthan89-py/Body-Language-Decoder-Using-Mediapipe>

4. Experiments

4.1 Baseline Model (Model 1)

We first evaluate the baseline approach using simple statistical summaries of pose coordinates combined with classical machine learning classifiers. Performance is measured using Macro-F1 to account for class imbalance.

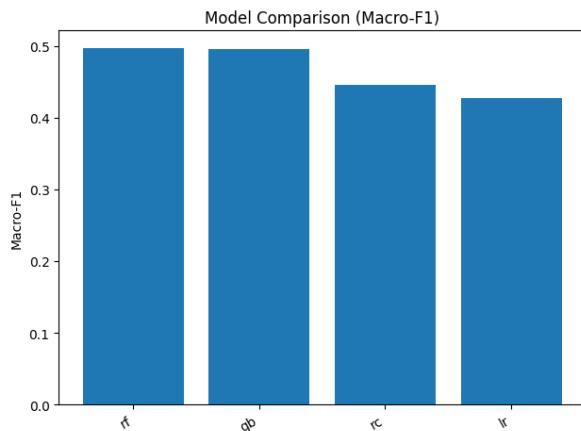
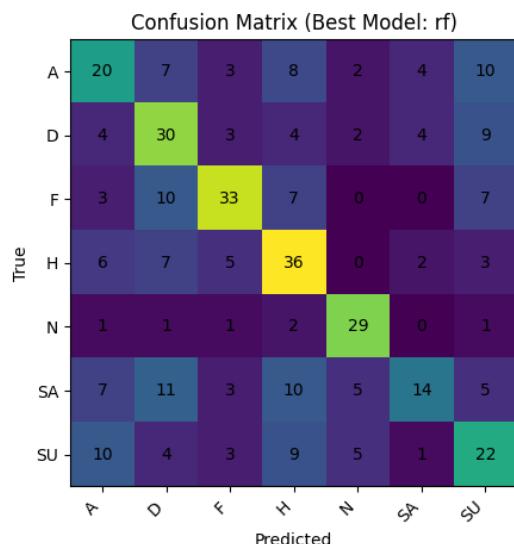
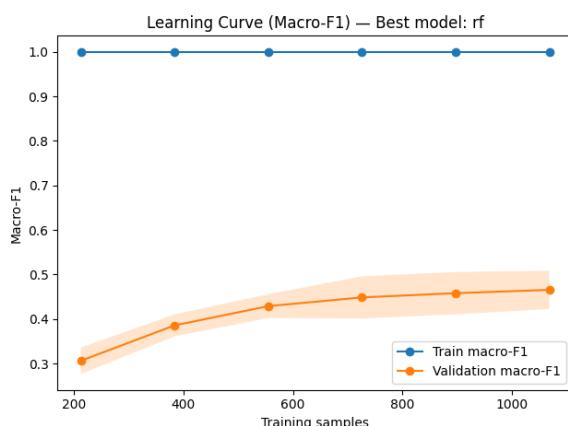


Figure 6 compares the Macro-F1 scores of different baseline classifiers. Among the evaluated models, Random Forest achieves the highest performance, indicating that non-linear models are more suitable even with simple pose statistics.



To analyze class-level behavior, Figure 7 represents the confusion matrix of the best baseline model (Random Forest). Several emotion classes are frequently confused, suggesting that mean and variance-based pose features are insufficient to fully capture discriminative emotional patterns.



The learning behavior of the baseline model is illustrated in Figure 8. While training performance quickly saturates, validation performance improves slowly, indicating overfitting and limited generalization capability.

4.2 Enhanced Model (Model 2)

Model 2 introduces biomechanically meaningful features and a validation-based model selection strategy. Multiple classifiers are evaluated using the same subject-based splits to ensure fair comparison.

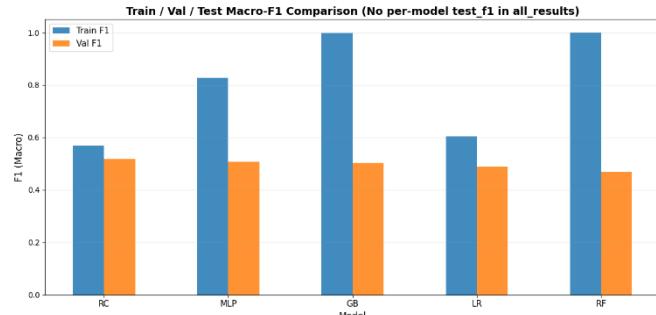
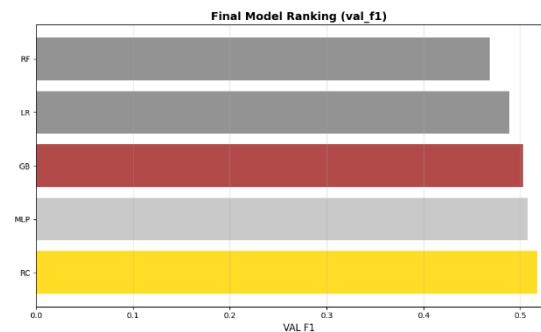
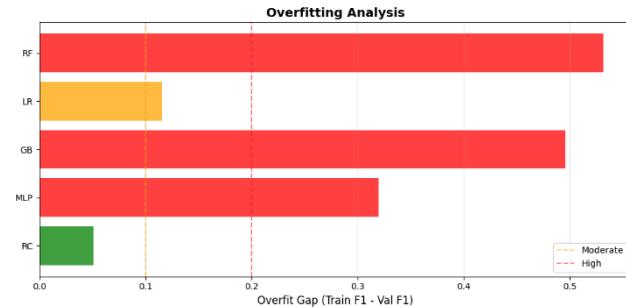


Figure 9 compares training and validation Macro-F1 scores across all Model 2 classifiers. Unlike Model 1, validation performance is explicitly used for model selection, preventing test-set leakage.



Based on validation performance, models are ranked as shown in Figure 10. This ranking identifies the best-performing model without using the test set.



To further quantify generalization behavior, Figure 11 represents the overfitting gap between training and validation Macro-F1. Tree-based models exhibit larger gaps, whereas neural models show a more balanced bias-variance trade-off.

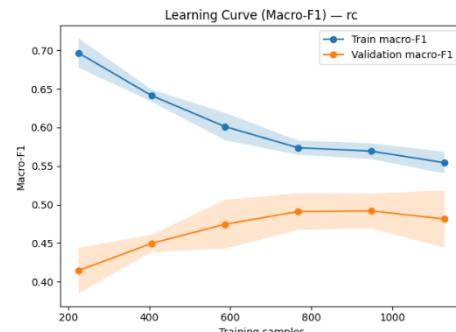
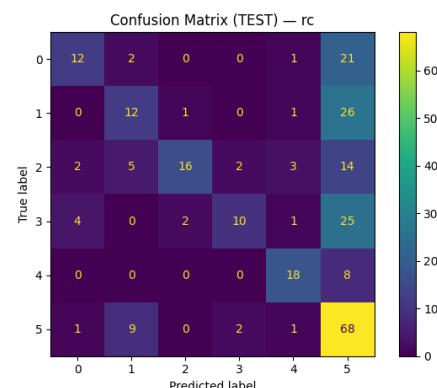


Figure 12 shows the learning curve of the selected best model. Compared to the baseline, Model 2 demonstrates improved validation performance and more stable learning behavior.



Finally, the confusion matrix in Figure 13 highlights improved class-level discrimination compared to the baseline model.

5. Comparison

This section compares the baseline approach (Model 1) and the enhanced pipeline (Model 2) in terms of performance, generalization, and experimental rigor.

Model 2 achieves higher Macro-F1 scores than Model 1, indicating improved class-balanced emotion recognition. As shown in Figures 6 and 9, Model 1 provides a reasonable baseline using simple statistical pose features, while Model 2 benefits from biomechanically meaningful feature engineering and validation-based model selection.

Learning curve analysis reveals a clear difference in generalization behavior. Model 1 rapidly overfits, with training performance saturating early and limited improvement in validation Macro-F1 (Figure 8). In contrast, Model 2 explicitly evaluates train-validation gaps across multiple classifiers (Figure 11), enabling more transparent assessment of overfitting. Neural models, such as MLP, exhibit a more balanced bias-variance trade-off compared to tree-based methods.

Class-level analysis further highlights the advantages of Model 2. The confusion matrix of Model 1 (Figure 7) shows frequent misclassifications among visually similar emotions, reflecting the limitations of simple coordinate statistics. Model 2 demonstrates improved class discrimination (Figure 13), particularly for emotions that depend on posture symmetry and motion dynamics.

Beyond performance improvements, Model 2 represents a methodological advancement. The use of subject-based train/validation/test splits and validation-driven model selection ensures fair comparison and prevents test-set leakage. These design choices directly address the limitations observed in Model 1.

Overall, the results show that combining meaningful feature engineering with rigorous evaluation leads to more reliable emotion recognition from body language. While Model 2 still faces challenges related to data size and class imbalance, it provides a stronger and more principled solution than the baseline approach.

6. Conclusion

In this project, we investigated emotion recognition from body language using pose-based representations and compared two modeling approaches with increasing levels of complexity and rigor. The baseline model (Model 1) relied on simple statistical summaries of pose coordinates and served as a reference point for evaluating more advanced methods.

The enhanced pipeline (Model 2) demonstrated clear improvements by incorporating biomechanically meaningful features such as posture, joint angles, motion dynamics, and symmetry measures. Combined with subject-based train/validation/test splits and validation-driven model selection, Model 2 achieved higher and more reliable Macro-F1 performance while providing better insight into generalization behavior.

Experimental results showed that simple pose statistics are insufficient to capture subtle emotional cues expressed through body posture and movement. In contrast, explicitly modeling these characteristics led to improved class-level discrimination and reduced overfitting. Although some limitations remain—particularly related to dataset size and class imbalance—the proposed approach offers a more principled and interpretable solution than the baseline.

Overall, our study highlights the importance of feature design and experimental rigor in pose-based emotion recognition. Future work may explore temporal deep learning architectures, data augmentation strategies, or multimodal fusion to further improve robustness and performance.

Yağız Alp Zeybek – 2201633

Mehmet Atahan Külekçi - 2201364