

Emotion Recognition in Body and Dance

Introduction

Human emotion is expressed not only through the face and voice but also through the body—through posture, gesture, rhythm, weight shifts, and movement dynamics. However, most existing emotion-recognition systems focus primarily on facial expressions or speech, overlooking the rich emotional information conveyed through full-body motion. This limitation is particularly significant in fields such as performance arts analysis, movement-based mental health assessment, immersive human–computer interaction, and accessibility scenarios where facial cues may be unreliable or unavailable. Addressing this gap requires models that can interpret emotion directly from body movement, motivating our project, *Emotion Recognition in Body and Dance*.

To explore this challenge, we first experimented with the EMOKINE dataset, a controlled collection of dance-based emotional expressions. Although EMOKINE offered clean recordings and consistent labels, our experiments with LSTMs, attention-based LSTMs, Transformers, and classical machine-learning models revealed a fundamental constraint: despite training well, all models struggled to generalize to new sequences. The dataset’s small size, limited stylistic variation, and subtle emotional differences made it difficult for deep-learning approaches to extract reliable patterns. These observations—discussed in detail later—motivated us to transition to the larger and more diverse Kinematic Actors Dataset, which contains 1,402 motion-capture trials of actors performing seven distinct emotions. Its BVH-format recordings include detailed skeletal hierarchies and rich joint-motion data, providing the temporal and structural information needed for sequence models like LSTMs and graph-based models such as ST-GCN. Working with this dataset enabled us to develop a robust motion-based emotion-recognition pipeline and achieve far more reliable model performance.

Related Work

Research on emotion recognition from full body movement has grown steadily, particularly within affective computing, movement sciences, and embodied interaction. Several studies using the EMOKINE dataset have demonstrated that emotional intention can indeed be reflected in body movement, but they also highlight the challenges of building reliable machine learning models from such limited and highly controlled data. The creators of EMOKINE primarily used the dataset for human perception experiments, showing that observers can often distinguish emotions from movement alone. However, computational models trained directly on EMOKINE tend to struggle with generalization due to the small number of samples, subtle emotional distinctions, and the fact that movements are performed by a single dancer. Prior work typically focuses on analyzing global kinematic features such as speed, expansion, and symmetry rather than using deep learning, because traditional sequence models often overfit on EMOKINE’s small dataset size.

Beyond EMOKINE, related datasets such as AffectKinetics, CMU Motion Capture, and actor-based kinematic datasets have been used to explore motion-driven emotion recognition. Many of these studies show that richer datasets with multiple performers and varied emotional expressions enable robust results using LSTMs, Transformers, Graph Convolutional Networks and ST-GCN.

Datasets such as Human3.6M, CMU Mocap, and AMASS remain some of the most widely used resources for learning general motion representations. To address the lack of affective labels in traditional mocap collections, several works focus specifically on emotion-oriented kinematic datasets, which provide detailed recordings of bodily expressions produced under controlled conditions. In the PhysioNet Kinematic Dataset of Actors Expressing Emotions, performers enact prototypical emotional states such as happiness, anger, sadness, fear, and neutral captured using marker based motion capture systems. The dataset offered clean, high-quality joint trajectories and clear emotion annotations, making them well suited for computational modeling of expressive body movement. However, the controlled environment and acted expressions introduce concerns about ecological validity. Recently developed resources attempt to address this limitation. For instance, the Bodily Emotion Recognition Dataset (BERD) investigates how acting expertise, recording devices, and stimulus conditions influence emotion recognizability, reflecting a broader shift toward understanding how production factors shape affective body motion. Complementary resources such as BoLD (Body Language Dataset), while not mocap-based, provide large scale in the wild annotations of bodily emotion from video, offering more natural variability that can be integrated with kinematic datasets for multimodal training.

System overview

Application Context and Impact

Emotion recognition from the body focuses on extracting affective state signals from posture, gestures, and motion dynamics rather than facial expression or voice. This is valuable where face/voice are absent, occluded, privacy-sensitive, or intentionally masked for dancers, VR avatars, low light, CCTV. Body based emotion recognition enhances accuracy in applications from human-robot interaction to healthcare by adding depth beyond ambiguous facial expressions, revealing true feelings, and improving human-computer interaction. Body language, posture, gestures especially hands, and even internal bodily sensations provide vital emotional information. This results in boosting system reliability and natural interaction.

The impact is visible across various domains where body emotion recognition can be implemented for creating context aware systems. Robots use body cues for natural, empathetic interactions, adapting to human emotional states for better collaboration. Customer emotions can be analyzed through body language to tailor experiences and advertisements in the retail industry. Monitoring patient emotional states, detecting distress, or assessing treatment response through subtle body shifts in healthcare.

Target users & stakeholders

- Researchers can study emotions through the body in affective computing, computer vision, HCI, and computational dance analysis.
- Product teams and engineers building emotion aware features for AR/VR, social robots, games, or streaming platforms.
- Healthcare practitioners and wellbeing apps can use it as passive indicators of mood
- Animators / VFX artists for emotion-driven motion retargeting and automatic labeling.

Use cases and application scenarios

- Adaptive User Interfaces: VR/AR systems can be used to capture user frustration if there is change in difficulty or content.
- Social robotics: Robots adapt dialogue and proxemics to perceived human emotion.
- Mental health monitoring : Longitudinal movement patterns can be used as adjunct signals for depression and anxiety monitoring.
- Dance and performance analytics: Classify emotional intent of choreography; assist choreography search.
- Assistive tech: Detect distress in elderly or mobility-impaired users at home

Kinematic Actors Dataset

The experiments use the Kinematic Actors Dataset, consisting of 1,402 motion trials recorded using 17 inertial motion-capture sensors. Each trial corresponds to a performance of one of seven emotions such as Angry, Disgust, Fearful, Happy, Neutral, Sad, Surprise. Recordings in this dataset are provided as BVH i.e Biovision Hierarchy files. Every BVH file contains two sections:

HIERARCHY Section

This section was useful for reconstruction of the human skeleton and mapping of joint trajectories.

- 72 nodes: 1 root (Hips), 58 joints, 13 end sites
- Joint tree structure
- Bone lengths via OFFSET
- Motion degrees of freedom via CHANNELS

MOTION Section

This section contains the actual movement data having continuous time series and enabled to form the raw basis for feature extraction.

- Number of frames
- Frame time
- For each frame:
 - Global position of the root (Hips)
 - Rotation parameters for all joints

Preprocessing Steps

BVH motion data was in the raw format so a tabular dataset was constructed by performing feature extraction on 17 joints such as LeftArm, RightFoot, Head, Hips etc using steps described above

- Custom BVH parser was constructed by reading hierarchical joint tree, extracting joint channels, iterating through motion frames and converting BVH rotations into numeric trajectory arrays of 17 skeletal joints.
- Centering and calibration was performed by using the mass center of the first frame and all subsequent joint trajectories were translated relative to this origin. This ensured that motions are captured and to remove actor dependent offsets.
- Motion based features such as speed, acceleration, joint movement range, overall movement intensity were engineered per trial.
- All frame level features are aggregated into statistical features as all sequences have variable length.
- This resulted in one feature vector representing a BVH file and the final dataset had 1402 samples with almost 400 to 500 engineered features.
- Metadata columns such as 'filename', 'actor_ID', 'emotion', 'gender', 'scenario_ID', 'version', 'num_frames', 'duration' were excluded and rest of the numerical features were kept.
- Emotion labels were encoded using Label Encoder.
- 80% of data was used for training the models and 20% was kept for evaluation.

Methods

Baseline Models

1. Stratified K Fold Cross validation was used to ensure every fold preserves class proportions and reduces performance variance over random splits
2. Random Forest with parameters such as, SVM and Gradient Boosting using stratified cross validation were applied on this dataset.

Multilayer Perceptron

1. 150 most predictive kinematic features were selected using mutual information.
2. Applied augmentation to minority classes for handling class imbalance.
3. Stratified K-Fold cross-validation was applied which ensured that each fold preserves class distribution.
4. Best hyperparameters such as learning rate, dropout, hidden_dim, weight_decay and label smoothing were selected by applying optuna hyperparameter optimization.
5. The classifier was implemented as a fully connected Multi-Layer Perceptron (MLP) whose architecture was determined by Optuna-based hyperparameter optimization. The network consisted of an input layer matching the dimension of the MI-selected features, followed by 1–4 hidden layers composed of linear transformations, ReLU or LeakyReLU nonlinearities, and dropout regularization (0.0–0.5). The number of neurons per layer ranged from 64 to 512 depending on the selected trial. A final fully connected output layer projected the last hidden representation to the number of emotion classes. The network was trained using cross-entropy loss and the Adam optimizer, with learning rate, batch size, and dropout probability also tuned through Optuna

ST-GCN

The BVH Parser was adapted for both ST-GCN and LSTM to capture time based movements. The BVH Parser class parses the joint hierarchy such as joint names, parent child relationship, offsets, and channel definitions. Also, it loads the raw motion data such as position and rotation over time using functions such as `get_joint_channels` and `get_joint_position_indices` which allow you to extract the translation and rotation channel ranges for each joint. This class enables one to construct the ST-GCN skeleton graph and to convert motion channels into (T, V, 3) joint trajectories used for training.

The ST-GCN Dataset Builder is responsible for converting raw BVH motion-capture files into a standardized graph-structured format suitable for spatio-temporal graph convolutional networks. It loads each BVH file resamples all sequences to a fixed number of frames, ensuring temporal consistency across samples. The builder establishes a global joint order and skeleton graph (edge list) so that every sample shares the same node indexing required by GCNs. It then centers the motion on a root joint, applies scale normalization, and computes motion features such as velocities, producing data shaped as (C, T, V) for each sequence. This preprocessing pipeline ensures spatial alignment, temporal uniformity, and consistent graph topology which are used for ST-GCN training on human motion or emotion-from-movement tasks.

The proposed ST-GCN architecture extends the classical spatio-temporal graph convolutional framework by integrating hierarchical graph reasoning with multi-dimensional attention. Each ST-GCN block performs spatial graph convolution to aggregate information across the skeletal adjacency graph, followed by temporal convolution to capture dynamic evolution of motion. To enhance discriminative power, every block incorporates temporal, spatial, and channel-wise attention modules that adaptively weight important frames, joints, and feature channels. Residual connections ensure stable gradient flow while deeper layers progressively downsample temporal resolution to learn high-level motion semantics. After global attention aggregation and adaptive pooling, a fully connected classifier predicts the target emotion class. This design allows the model to jointly learn pose structure, motion patterns, and expressive movement characteristics essential for body- and dance-based emotion recognition.

LSTM

For the LSTM model, we began by preparing the kinematic dataset so that it could be understood as a time-based sequence of body movements. Each sample in the dataset contains 3D joint positions recorded across time, and we reorganized this information so that the model receives the movement frame by frame for every joint. To help the model better understand not just where the joints are but how they move, we also computed additional motion features such as velocity and acceleration. These describe how fast each joint is moving and how its speed changes, which are important cues for recognizing emotional expression in dance. To make the model more robust, we applied data augmentation during training. This included adding small amounts of random noise to joint positions so the model does not become overly sensitive to tiny variations, and using a technique called Mixup, where two samples are blended together to encourage the model to learn general patterns rather than memorizing examples. We then split the dataset into training and validation sets while keeping the emotion class distribution balanced.

The model we designed, called MotionLSTM, processes sequences of motion over time. Before the sequence enters the LSTM, we included a lightweight feature-attention module that helps the model focus on the most important aspects of the movement, such as certain joints or specific motion features. The core of the model is a bidirectional LSTM, which learns patterns in the movement by looking both forward and backward in time. Because not every frame in a dance sequence is equally expressive, we added a temporal attention mechanism that learns to highlight the key moments that contribute most to the emotion being expressed. The final layers convert the learned representation into an emotion prediction. We trained the model over multiple epochs using the AdamW optimizer, a gradually adjusting learning rate schedule, and gradient clipping for stability. During training, Mixup was incorporated at the loss level so that blended samples were handled correctly. Throughout this process, we monitored validation accuracy and saved the best-performing version of the model. After training, we evaluated the model using standard tools such as confusion matrices, class-wise accuracy, and classification reports, which will be discussed in the results section.

Evaluation Results

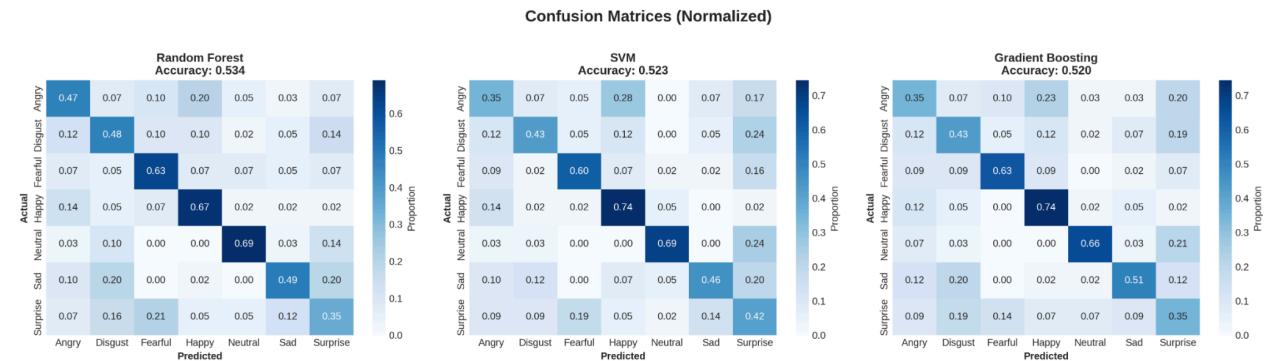
Models were evaluated using Accuracy, Precision, Recall, F1-score, Confusion matrices, Per-class breakdowns

Machine Learning Model Results

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.534	0.534	0.534	0.532
SVM	0.523	0.540	0.523	0.525
Gradient Boosting	0.520	0.527	0.520	0.520

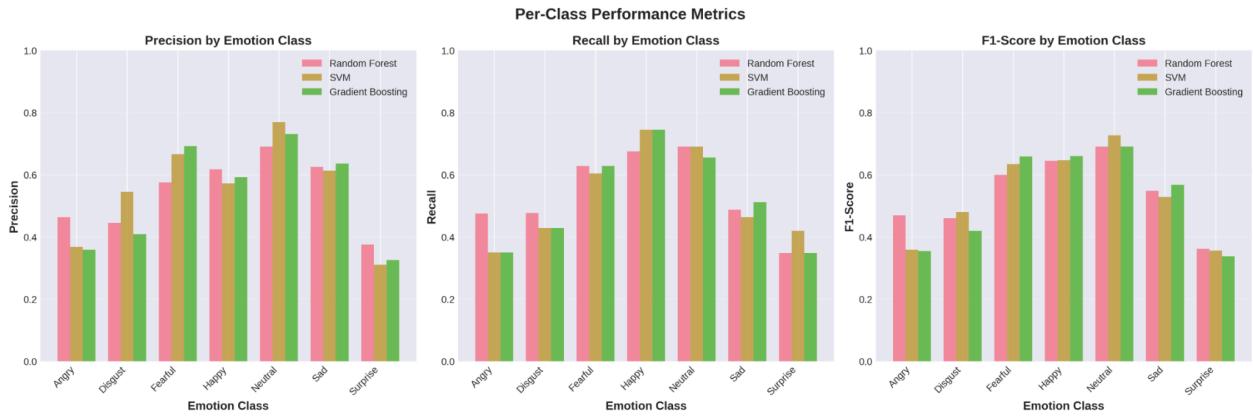
All three models perform in a similar manner with scores ranging from 0.52 to 0.55 across accuracy, precision, recall and F1 score. Model choice is not the primary limiting factor. SVM yields highest precision better at avoiding false positives. Random Forest gives highest recall better at covering difficult classes. Gradient Boosting shows balanced results.

Confusion Matrix



Random Forest performs slightly better on Sad (0.49), Disgust (0.48) and Fearful (0.63). SVM shows best performance for Happy (0.74) and Surprise(0.42). Gradient Boosting performs well across classes except surprise and angry on comparison.

Per-Class Precision, Recall, F1



Precision

- Neutral shows the highest precision across all models, with SVM performing best (0.78), followed by Gradient Boosting (0.74) and Random Forest (0.70).
- Fearful also exhibits strong precision where Gradient Boosting performs similar to that of SVM.
- Angry, Disgust, and Surprise show the lowest precision across all models.
- SVM consistently gives the highest or near-highest precision, indicating it produces fewer false positives.

Recall

- Happy shows the highest recall across SVM and Gradient Boosting being 0.74 and Random Forest showing 0.65.
- Neutral has the second highest recall across of all the models (0.65-0.70).
- Fearful, Happy, Neutral, Sad all show relatively high recall across models (0.60–0.75).
- SVM and Gradient Boosting show high recall across Fearful, Happy and Sad whereas Random Forest performs best on recall for Angry and Disgust

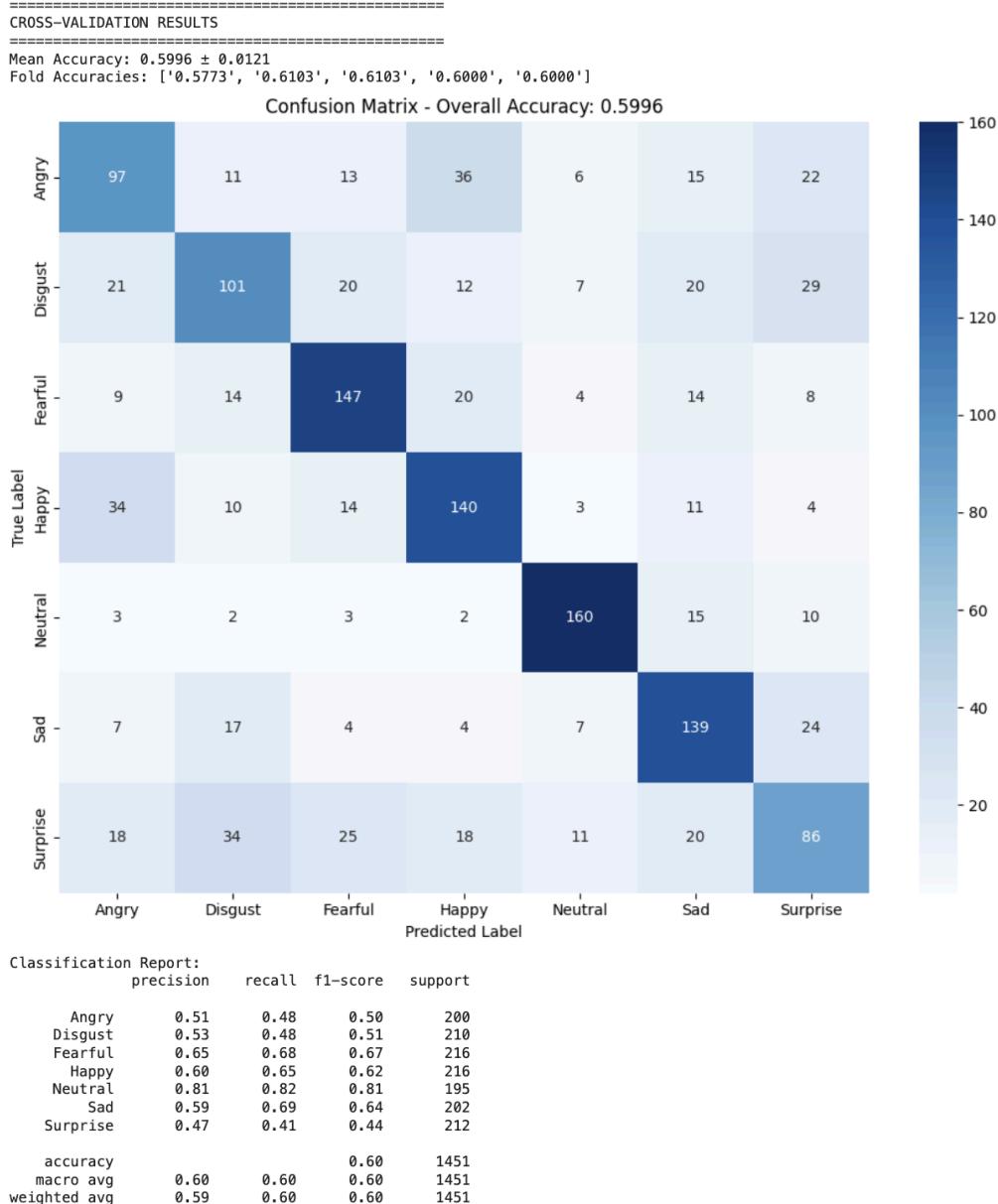
F1 Score

- Neutral, Happy, Fearful, and Sad achieve the strongest F1-scores across all models (0.60–0.75).
- Surprise, Angry, and Disgust have substantially lower F1-scores across all models (0.30-0.45)
- Gradient Boosting often shows slightly higher F1 on Fearful, Sad, while SVM performs best on Neutral.

SVM demonstrates the highest overall precision for well-defined emotions such as Neutral, Fearful, and Happy, indicating its stronger ability to reduce false positives. In contrast, Random Forest achieves the highest recall for several challenging classes, including Angry, Disgust, and Surprise, suggesting it captures a larger proportion of true samples even when class boundaries

are ambiguous. Gradient Boosting provides the most balanced behavior, often matching SVM or slightly exceeding the other models in F1-score for classes such as Fearful, Happy and Sad without dominating any single metric

MLP Performance Results



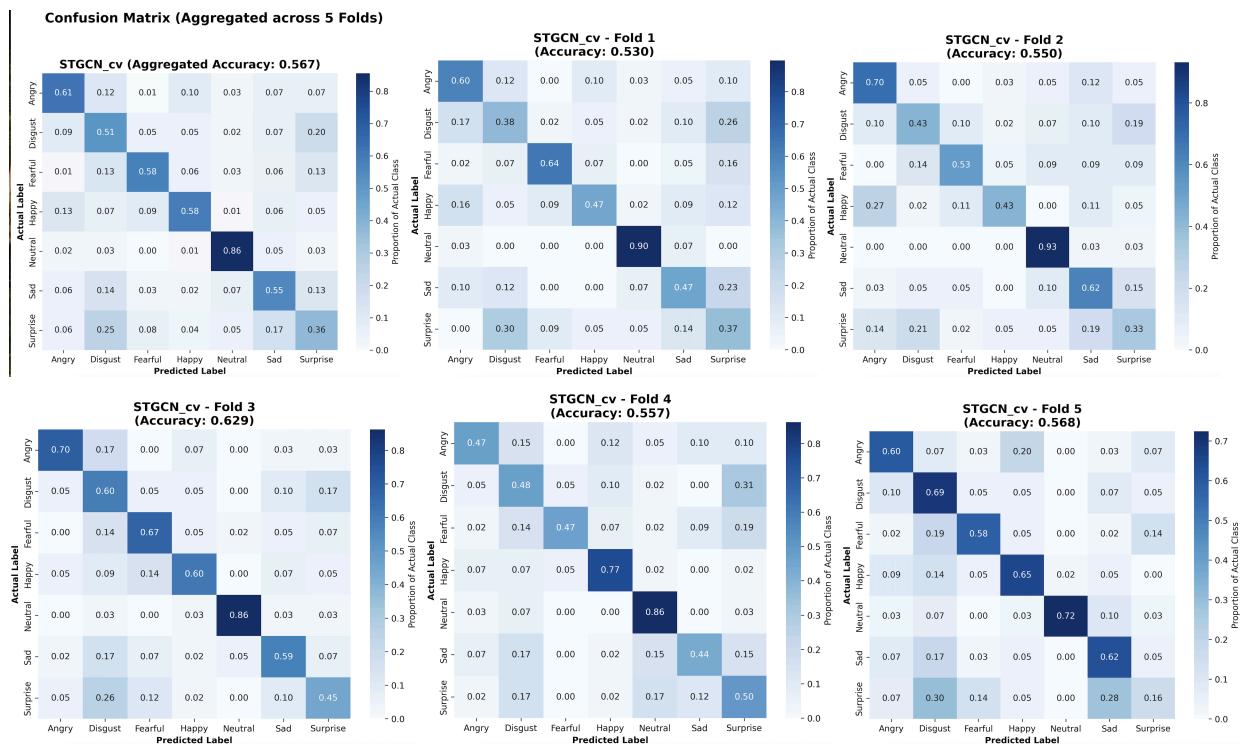
Mean Accuracy: 0.5996 ± 0.0121 Fold Accuracies: 0.5773, 0.6103, 0.6103, 0.6000, 0.6000

The MLP shows stable and consistent performance across folds, with very small variance showing that the model generalizes reasonably well. The classes that show strong performance in prediction are Neutral (160 correct), Fearful (147 correct), Happy (140 correct), Sad (139 correct). Angry and Surprise still show lowest performance.

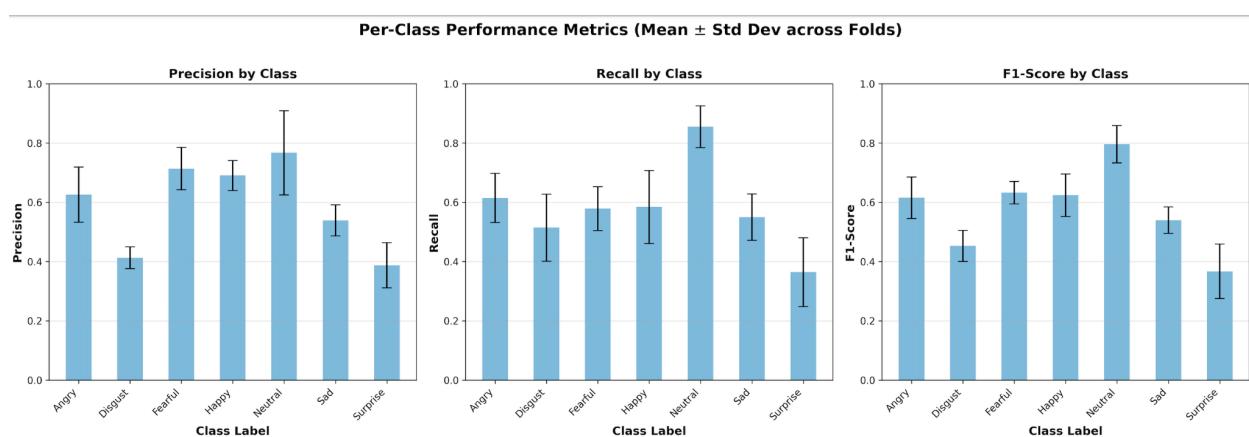
MLP shows highest performance for Neutral emotion (0.80-0.82) across F1, Precision and Recall. Apart from Neutral emotion, MLP shows decent results (0.60-0.62) for Happy, Fearful and Sad classes. Lowest performance is visible among the classes such as Angry, Disgust and Surprise. The MLP represents better performance (60%) compared to traditional ML baselines (52–53%).

ST-GCN Evaluation Results

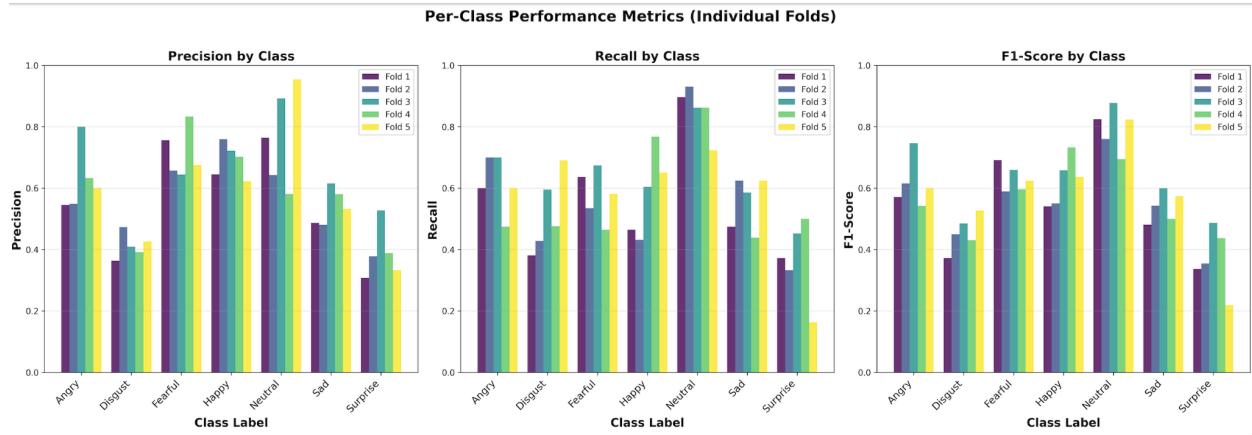
Overall and Per Fold Confusion Matrix



Aggregated accuracy across folds is 0.567 which confirms that the model generalizes reasonably. Accuracy across folds ranges 0.53 to 0.629 with the fold 3 performing the best.



- Neutral has the highest recall, precision and F1 score. This means that neutral movements are most stable and consistent across participants.
- Fearful, Happy, Angry have decent precision (0.60 to 0.70) and F1 score (0.62 approx)
- Disgust and Surprise have shown least performance across recall, precision and F1 score.

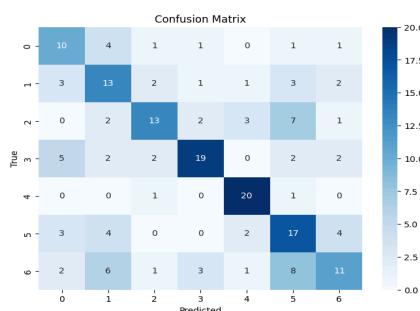


- Neutral, Fearful, Happy show high and stable scores across folds.
- Fold 3 (green) peaks in precision, recall, and F1 suggesting that it contains representative or balanced emotion sequences.
- Precision varies widely (0.54–0.80) and Recall remains steady (0.60–0.70) for Angry emotion showing that the model detects Angry better than it predicts.
- Happy has balanced precision, recall, and F1 across folds (0.55–0.75). This shows that happy gestures have expansive posture, arm movements which are stable patterns across subjects.
- Fearful has strong scores such as precision (0.66–0.84), recall (0.48–0.68) and F1 (0.58–0.68) showing that model is able to capture movements such as shrinking, recoiling, raised arms.

Results for LSTM / Evaluation of LSTM

To understand how well our LSTM model recognizes emotions from body movement, we used several evaluation metrics. Each of these metrics looks at the model's performance from a slightly different angle, helping us get a complete view of how well the model is doing and where it struggles.

1. Confusion Matrix



What this confusion matrix shows:

- The diagonal values (top-left to bottom-right) represent correct predictions.
- Neutral has the strongest diagonal value (20 correct), showing it is recognized very reliably.
- Happy also has strong correct predictions (19 correct).
- Angry, Disgust and Fearful and Sad have moderate correctness.
- Surprise shows more confusion, meaning the model struggles with this emotion.
- Misclassifications appear as off-diagonal values and highlight which emotions look similar in movement patterns.

2. Classification Report

	precision	recall	F1-score	Support
Angry	0.43	0.56	0.49	18
Disgust	0.42	0.52	0.46	25
Fearful	0.65	0.46	0.54	28
Happy	0.73	0.59	0.66	32
Neutral	0.74	0.91	0.82	22
Sad	0.44	0.57	0.49	30
Surprise	0.52	0.34	0.42	32
Accuracy			0.55	187
Macro Avg	0.56	0.56	0.55	187
Weighted Avg	0.57	0.55	0.55	187

Accuracy: 0.55

This report shows initial performance. Some classes did well while others struggled.

Key observations:

- Neutral was the strongest performer, with an F1-score of 0.82.
- Happy also performed well, with an F1-score of 0.66.
- Angry, Disgust, Sad and Surprise had moderate to low F1 scores, indicating difficulty differentiating them.
- Macro avg and weighted avg around 0.55 suggest that performance is fairly consistent across classes and not dominated by any single class.

Final Classification Report (after tuning and best weights)

	precision	recall	F1-score	Support
Angry	0.5200	0.7222	0.6047	18
Disgust	0.6000	0.6000	0.6000	25
Fearful	0.5417	0.4643	0.5000	28
Happy	0.6429	0.5625	0.6000	32
Neutral	0.7407	0.9091	0.8163	22
Sad	0.5000	0.6667	0.5714	30
Surprise	0.6667	0.3750	0.4800	32
Accuracy			0.5936	187
Macro Avg	0.6017	0.6143	0.5961	187
Weighted Avg	0.6028	0.5936	0.5858	187

Final accuracy: 0.5936 ($\approx 59\%$)

- The model's accuracy improved from 55% \rightarrow 59%.
- Neutral remains the strongest class with Precision (0.74), Recall (0.91), F1: 0.82
- Angry improved significantly with an F1 of 0.60.
- Disgust, Happy, Sad achieved balanced and moderate performance (around 0.57–0.60).
- Surprise continues to be the most challenging class (F1 = 0.48).
- The macro and weighted averages (~0.60) show that the model treats most classes fairly evenly.
- The model can correctly identify the emotional category about 6 out of 10 times, which is good for body-movement emotion recognition, a task known to be ambiguous and difficult.

3. Per-Class Accuracy

Class	Accuracy	Meaning
Angry	0.722	The model correctly recognizes this emotion most of the time.
Disgust	0.6	Moderately well-recognized.
Fearful	0.464	Harder class—almost half are misclassified.
Happy	0.562	Reasonable recognition.
Neutral	0.909	Extremely well-recognized; very distinctive movement pattern.
Sad	0.667	Fairly solid performance.
Surprise	0.375	Most difficult emotion; gestures or motion patterns likely overlap heavily with others.

4. Predicted Probability Inspection

Why we use it:

Probability inspection shows how confident the model is in its predictions. Instead of just the final label, we can see how it distributes belief across all emotion classes.

```
Probabilities for first batch sample: [0.00218036 0.01224436 0.1614127  
0.01997477 0.00134924 0.73668104 0.06615743]
```

- The model assigned the highest probability (0.736) to Class 5.
- All other classes received very low probabilities.
- This indicates the model was highly confident that the sample belongs to Class 5.
- This is a healthy sign because it shows the model is not guessing randomly—it forms strong opinions when patterns are clear.

6. Cross validation

The 5-fold cross-validation results show that our LSTM model performs consistently across different subsets of the kinematic dataset, with accuracies between 0.55 and 0.61 and a mean of approximately 0.58. This stability indicates that the model is learning meaningful motion patterns rather than relying on any particular train-validation split. Cross-validation is important because it provides a more reliable estimate of real-world performance: instead of depending on a single split, it tests the model on multiple configurations of the data, reducing the risk of overfitting and increasing confidence that the results generalize to unseen samples. By ensuring that every example is used for validation at least once, cross-validation strengthens the credibility of our evaluation and confirms that the model's behavior is robust across the full dataset.

K-Fold Accuracy Summary

Fold	Accuracy
Fold 1	0.6096
Fold 2	0.5775
Fold 3	0.5484
Fold 4	0.5591
Fold 5	0.6129
Mean Accuracy	0.5815
Standard Deviation	0.026

Average Classification Report

Class	Precision	Recall	F1-Score
Angry	0.5678	0.5209	0.5402
Disgust	0.6701	0.463	0.5313
Fearful	0.5664	0.5527	0.5568
Happy	0.6397	0.6605	0.6484
Neutral	0.7106	0.902	0.7899
Sad	0.4993	0.5819	0.5313
Surprise	0.5126	0.4236	0.4575

Limitations

1. Some emotions are inherently difficult for the model to distinguish.

Classes like 6 and 2 have significantly lower accuracy and F1-scores. This suggests that the movements related to these emotions are either subtle, overlap heavily with other emotions, or are inconsistently expressed by different performers. As a result, the model struggles to form strong, reliable patterns for these classes.

2. The dataset size and distribution limit model performance.

Although some classes like 4 perform very well (over 90% accuracy), others have fewer or more variable samples, which makes it difficult for the model to generalize. Emotion-through-movement datasets are naturally small and imbalanced, and our results reflect that challenge.

3. Movement-only features may not capture the full emotional expression.

Our model relies entirely on skeletal joint positions, velocity, and acceleration. While useful, these features do not include facial expressions, hand gestures, speed changes over longer intervals, or other contextual cues that humans naturally use to interpret emotion. This restricts the model's ability to fully capture emotional nuance.

4. LSTM architecture has limitations in capturing long-term dependencies.

Although LSTMs are good at sequence modeling, they can struggle with very long sequences or highly complex movement patterns. Some emotional performances may extend beyond the window length or require stronger temporal reasoning than LSTMs offer.

5. Augmentation strategies may not reflect real emotional variability.

Noise and Mixup augmentation help with regularization but do not generate *realistic* variations of emotional movement. Emotional expression is complex, and simple augmentations may not be enough to improve low-performing classes.

6. Model interpretability is limited.

While we included channel and temporal attention modules, it is still challenging to explain *exactly* why the model misclassifies certain emotions. This limits transparency when presenting to non-technical audiences or clients.

Future Work to Address These Limitations

1. Expanding the dataset (more samples, more emotions, more performers).

Collecting additional motion sequences would give the model a richer understanding of how emotions are expressed. More balanced class distributions would also help improve performance for weaker classes like 6 and 2.

2. Incorporating multimodal information (face, audio, posture style).

Adding facial expression analysis, audio cues (such as breathing, vocal sounds during dance), or rhythm patterns can significantly boost accuracy. Emotions are rarely expressed through body movement alone, and multimodal fusion would make the system more realistic.

3. Personalization models (adaptation to performer styles).

Individuals express emotions differently. A future system could learn small personalized adjustments so the model adapts to different dancers' movement styles.

4. Improving interpretability tools.

Adding visual explanations—such as highlighting which joints the model paid attention to—would help users understand how the system makes its decisions and increase trust in predictions.

5. Evaluating on continuous emotional scales instead of fixed classes.

Emotions are not always discrete. In future work, predicting emotional intensity or valence/arousal could give a more nuanced and realistic output.

Challenges faced with EMOKINE Dataset:

Our initial goal was to determine whether emotional intention could be recognized from full-body movement alone, and the EMOKINE pilot dataset appeared to be a strong starting point. EMOKINE consists of 63 choreographed sequences performed by a single professional dancer, with each take expressing one of six emotions. While the dataset is visually clean and tightly controlled—featuring consistent lighting, fixed camera setup, and precisely structured movements—these same strengths quickly revealed themselves as limitations for machine-learning models. Only 54 sequences were suitable for modeling, and the lack of performer variability alongside the subtle differences between emotional categories made it difficult for algorithms to learn robust and generalizable motion patterns.

To explore the dataset fully, we implemented both 2D and 3D pose-extraction pipelines. MediaPipe Pose provided 33 visible landmarks in 2D, while XSENS sensor data supplied accurate 3D joint positions. After applying preprocessing steps such as centering, scaling, and temporal alignment, we generated clean pose sequences for modeling. Despite building these rich representations, the dataset's limited size became a persistent obstacle. Our first modeling attempts used LSTM networks, which are well-suited for temporal data. These models trained without difficulty, but their predictive accuracy remained extremely low—typically around **15–20%**. Even when we added attention mechanisms designed to highlight emotionally salient frames, performance barely improved. The models frequently collapsed to predicting a single emotion for most samples, suggesting that the sequences did not contain enough distinct temporal cues for the networks to learn meaningful differences.

Recognizing the constraints of raw temporal modeling, we shifted toward feature engineering inspired by movement psychology. By computing motion descriptors such as velocity, acceleration, jerk, symmetry, and body-expansion measures, we reduced each movement sequence to a compact representation capturing global motion qualities. Classical machine-learning models like SVMs performed noticeably better than deep networks, reaching approximately **37% accuracy**. This result aligned with prior research showing that handcrafted kinematic features often outperform raw sequence models on very small motion datasets. However, even with more expressive architectures like Transformers, we encountered a similar pattern: models quickly memorized the training data—achieving nearly perfect accuracy—but failed to generalize.

To test generalization more rigorously, we adopted Leave-One-Out Cross-Validation, the most reliable evaluation method for a dataset of this size. The results were decisive. Across all model types—including LSTMs, attention models, SVMs, Transformers, and even multi-model fusion systems—the accuracy dropped to roughly **3.7%**, indicating almost random predictions. These findings made it clear that EMOKINE's small sample count, single-performer structure, and subtle emotional distinctions fundamentally limit its usefulness for data-driven emotion-recognition models. This realization ultimately motivated our transition to a larger and more diverse dataset for the remainder of the project.

Bonus

Ethical and Social Concerns

Emotion recognition from body movement raises several ethical and social concerns, especially when using models such as LSTMs that attempt to infer emotional states from subtle motion cues. Body language varies widely across individuals, cultures, and physical abilities, which means the model may misinterpret or oversimplify emotional expressions. There is also the risk that users might overtrust the system, assuming its predictions represent objective truth rather than probabilistic estimates. Additionally, although the data is represented as skeleton coordinates rather than video, motion patterns can still reveal sensitive information, potentially exposing health conditions or personal states. If deployed without proper safeguards, such systems could be misused in surveillance, workplace monitoring, or other contexts where individuals have not explicitly consented to emotional analysis.

To mitigate these risks, the system should only be used in environments where informed consent is clearly obtained and the purpose of emotion recognition is transparent to all users. Model outputs should include confidence scores and disclaimers stating that predictions are approximate and should not be used for high-stakes decision-making. Data collected for training or evaluation must be anonymized, securely stored, and handled according to ethical research standards. Future iterations of the model can incorporate more diverse datasets, fairness-aware training techniques, and explainability tools to reduce bias and improve transparency. Restricting the system to supportive or creative applications such as performance analysis, movement therapy, or interactive art can further ensure that the technology is used responsibly and avoids contexts where emotional inference could cause harm.

Working with the Kinematic Actors Dataset raises several ethical and social considerations that must be acknowledged when designing emotion-recognition systems. Although motion-capture data does not reveal facial identity, it contains unique behavioral signatures—such as gait, posture, and joint movement patterns—that can function as biometric identifiers. This means individuals could potentially be re-identified or profiled through their motion alone, presenting privacy risks if such data were misused in surveillance or workplace monitoring.

Mitigating ethical and social risks in emotion recognition from kinematic data requires a combination of technical, procedural, and governance-focused strategies. First, robust privacy protections should be implemented by ensuring data minimization, anonymizing identifiers, restricting access to motion signatures, and adhering to regulatory frameworks such as GDPR. Motion-derived embeddings should be encrypted and stored separately from metadata to reduce re-identification risks.

Cross-Modal Integration

While our primary system focuses on recognizing emotion from full-body movement using LSTM models, there is strong evidence that combining multiple modalities can significantly enhance emotion-recognition performance. Body movement alone often provides incomplete or ambiguous cues, especially for subtle emotional states or classes with overlapping motion patterns. Cross-modal integration introduces an additional modality—such as audio, facial expressions, textual descriptions, or sensor-derived features—to complement and enrich the information captured by skeletal trajectories. In the context of our project, one promising avenue involves leveraging audio or other sensor channels already present in datasets like EMOKINE or extending the Kinematic Actors Dataset with synchronized sound or inertial sensor metadata. An LSTM can then process the temporal structure of motion while a second model (e.g., a CNN for audio spectrograms or a Transformer for text cues) processes the supplementary modality. The outputs can be fused through concatenation, late fusion, or attention-based weighting, allowing the system to consider both motion dynamics and auxiliary emotional signals.

To assess the value of cross-modal integration, the enriched system would be compared against the baseline LSTM that relies solely on skeletal data. Based on our experience with EMOKINE and the Kinematic Actors Dataset, we anticipate that adding another modality would increase both performance and interpretability. For example, motion-only LSTMs struggled with ambiguous classes such as Fearful and Disgust or with classes showing low per-joint distinctiveness. An audio or contextual cue could help disambiguate these cases by providing rhythmic, energetic, or affective information that is not always reflected in joint movement alone. Additionally, cross-modal fusion can reveal which modality contributes more strongly to certain emotions—movement may dominate for expressive emotions like Anger or Surprise, while audio or timing cues may improve recognition of more subtle states like Neutral or Sad. Even if accuracy improvements are modest, integrating another modality provides richer interpretability, helping us understand how emotion emerges from the interplay of movement and other expressive channels. This analysis would guide future development toward more robust, multimodal emotion-recognition systems.

Cross-modality within the Kinematic Actors Dataset and ST GCN framework refers to the integration of skeletal motion data with additional complementary feature streams to enrich emotional understanding beyond joint trajectories alone. While ST GCN naturally captures spatial temporal structure from body joints, kinematic motion is often influenced by latent factors such as movement intensity, rhythm, speed variations, or joint-level dynamics—that may not be fully represented through raw positions and rotations. By combining skeletal coordinates with engineered time based features such as velocities, accelerations, joint angle changes, or global movement energy, the model gains access to a richer description of expressive behavior. This cross-modal fusion allows ST-GCN to reason simultaneously about structure and dynamics , improving its ability to distinguish subtle emotional cues such as tension, fluidity, or abruptness in movement. Integrating these modalities using attention spatial and temporal based helped mitigate limitations of relying solely on skeletal geometry and results in a more robust representation of emotional expression.

Individual Reflection — Shobha Gupta

Working on *Emotion Recognition in Body and Dance* has been one of the most challenging and rewarding experiences of this course, pushing me to grow both technically and conceptually. When we began, I anticipated that modeling emotion from body movement would be similar to working with other temporal datasets. However, I quickly realized that human movement is far more nuanced and complex than most time-series data. My primary responsibilities included dataset preprocessing, feature engineering, building the LSTM-based models, conducting extensive evaluations, and contributing to the documentation and overall structure of the project. Each of these components exposed me to different layers of the problem and taught me important lessons about the relationship between data quality, model design, and real-world performance.

Our initial work with the EMOKINE dataset was eye-opening. Although the recordings were clean, consistent, and well-labeled, the dataset's small size and limited variation made it unsuitable for generalizable deep-learning models. I spent considerable time experimenting with LSTMs, attention mechanisms, Transformers, and classical baselines, only to find that although the models trained well, they consistently failed to generalize. This was frustrating at first, but it ultimately helped me understand a fundamental principle of machine learning: even sophisticated architectures cannot compensate for insufficient or non-representative data. This experience shaped my thinking for the rest of the project and taught me how important it is to evaluate datasets critically before committing to a modeling strategy.

Transitioning to the Kinematic Actors Dataset shifted the trajectory of my work. With over 1,400 trials and rich BVH motion data, this dataset allowed me to fully implement and refine the LSTM architecture. I worked on converting joint coordinates into LSTM-ready sequences, generating motion features such as velocity and acceleration, applying data augmentation, and integrating channel and temporal attention mechanisms. These steps significantly deepened my understanding of sequential deep learning. I also implemented extensive evaluation methods—including classification reports, confusion matrices, per-class accuracy, probability inspection, and 5-fold cross-validation—to gain a more holistic understanding of the model's behavior. Through this process, I learned that evaluating a model requires more than a single accuracy number; it involves understanding *which* classes succeed, *which* ones fail, and *why* those patterns occur.

A major takeaway for me was the importance of clear communication, both within the group and in the written documentation. Ensuring that our methodological decisions were well explained—especially for an audience with varying technical backgrounds—helped me improve my ability to translate complex processes into accessible language. Additionally, coordinating dataset preparation, model experiments, and report writing taught me how essential organization and consistency are in collaborative research.

.

Individual Reflection — Kriti Shahi

Working on the Emotion Recognition from Body and Dance project has been challenging but a great learning experience. I was really looking forward to work on this project which helps to classify human emotions from movement. This project quickly evolved into a deep exploration of motion-capture formats, sequence modeling, graph neural architectures. The project pushed me because I had to design fully custom BVH parser. This helped me to curate clean dataset with effective engineered features used for classical machine learning models.

One of the turning points of the project was my decision to construct the BVH motion-capture parser from scratch. At first, I underestimated the difficulty of working directly with raw BVH files. Unlike typical CSV or pre-cleaned datasets, BVH files contain hierarchical skeletal structures, variable channel definitions, and frame-level motion encoded through rotation and translation parameters. Understanding the HIERARCHY section alone required me to think about mapping parent child joints, preserving offsets, and correctly handling the degrees of freedom associated with each joint. Writing parser functions such as `get_joint_channels()` and `get_joint_position_indices()` forced me to internalize how motion is represented computationally, and not merely as sequences of numbers. When the first end-to-end extraction finally succeeded producing valid joint trajectories, it felt like unlocking a new understanding of human movement through the lens of computation. This parser later became the backbone of both LSTM and ST-GCN pipelines.

Another major learning curve involved understanding ST-GCN (Spatio-Temporal Graph Convolutional Networks). Initially, the architecture felt intimidating because it combines graph reasoning for spatial relationships with temporal convolution for motion dynamics. To truly understand the design, I immersed myself in research papers, tutorials, and open-source implementations. What initially appeared abstract, graph kernels, adjacency matrices, temporal receptive fields gradually became intuitive once I connected them to the structure of the human body. The idea that each joint is a node and each bone is an edge created a natural bridge between human anatomy and deep learning. Extending the classical ST-GCN with attention mechanisms, normalization, and improved temporal handling taught me how to adapt state of the art architectures for specialized tasks rather than treating existing code as a black box. This experience strengthened my confidence in reading, interpreting, and modifying complex research models.

Simultaneously, I explored MLP-based classification using engineered features, which turned out to be surprisingly powerful. Instead of relying solely on deep sequence models, I used mutual information to select the most predictive features from hundreds of kinematic descriptors. Running Optuna for hyperparameter tuning further helped me appreciate the sensitivity of neural networks to learning rates, dropout, and architecture depth.

Perhaps the most personally rewarding part of this project was the transformation in how I approach research problems. I learned that building a high-performing model is rarely linear; it is iterative, diagnostic, and deeply experimental.

Key References and Links

- Christensen, J. F., Fernández, A., Smith, R. A., Michalareas, G., Yazdi, S. H. N., Farahi, F., Schmidt, E.-M., Bahmanian, N., & Roig, G. "EMOKINE: A Software Package and Computational Framework for Scaling Up the Creation of Highly Controlled Emotional Full-Body Movement Datasets." *Behavior Research Methods* (2024). DOI: 10.3758/s13428-024-02433-0. [SpringerLink+2ResearchGate+2](#). Dataset & code: EMOKINE GitHub repository. [GitHub+1](#)
- Sapiński, T., Kamińska, D., Pelikant, A., & Anbarjafari, G. "Emotion Recognition from Skeletal Movements." *Entropy* (2019). [ResearchGate](#)
- Paiva, P. V. V., "Attention Model for Pose-Based Emotion Recognition." *SCITEPRESS Proceedings* (2023). [SciTePress](#)
- Ghaleb, E., Mertens, A., Asteriadis, S., & Weiss, G. "Skeleton-Based Explainable Bodily Expressed Emotion Recognition Through Graph Convolutional Networks." (2021). [esamghaleb.github.io](#)
- Wang, T., et al. "Emotion Recognition From Full-Body Motion Using a Multiscale Spatio-Temporal Network." [publisher] (2024). [computer.org](#)
- Bhatia, Y., Bari, A. H., Hsu, G.-S. J., & Gavrilova, M. "Motion Capture Sensor-Based Gait Emotion Recognition Using a Bi-Modular Sequential Neural Network." *Sensors* 22(1): 403 (2022).
- Zhang, M., Yu, L., Zhang, K., Du, B., Zhan, B., Chen, S., Jiang, X., Guo, S., Zhao, J., Wang, Y., Wang, B., Liu, S., & Luo, W. (2020). Kinematic dataset of actors expressing emotions (version 2.0.0). *PhysioNet*. RRID:SCR_007345. <https://doi.org/10.13026/cdb3-8925>
- BVH Parser Class - <https://github.com/UPC-ViRVIG/pymotion> and <https://github.com/20tab/bvh-python>
- Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition(Sijie Yan, Yuanjun Xiong, Dahua Lin) - <https://arxiv.org/abs/1801.07455>
- ST-GCN Implementation - <https://github.com/yysijie/st-gcn>
- Skeleton-Based Emotion Recognition Based on Two-Stream Self-Attention Enhanced Spatial-Temporal Graph Convolutional Network by Jiaqi Shi, Chaoran Liu, Carlos Toshinori Ishi and Hiroshi Ishiguro. <https://doi.org/10.3390/s21010205>
- Spatial Temporal Variation Graph Convolutional Networks (STV-GCN) for Skeleton-Based Emotional Action Recognition - <https://ieeexplore.ieee.org/document/9328124>