Research Article

Graph-Based Parallel Multi-Objective Optimization of Skeletal Body Motion Data for Emotion Recognition with Knowledge-Distilled Classifier

Seyed Muhammad Hossein Mousavi[1]

1. Cyrus Intelligence Research Ltd, Iran

Recognizing human emotions from body motion is a critical challenge in affective computing, particularly in scenarios where facial expressions or speech are unavailable or unreliable. In this study, we propose a novel framework for emotion recognition from skeletal motion data using a graph-based and parallel multi-objective optimization approach. Skeletal motion sequences are represented as graphs, where nodes correspond to joints and edges capture anatomical connections, enabling the preservation of spatial structure and dynamic body patterns crucial for emotional expression. To improve both feature quality and model performance, we employ two evolutionary algorithms in parallel. A Genetic Algorithm (GA) is used to evolve the topology of the motion graphs, optimizing structural characteristics that influence expressiveness. Simultaneously, Particle Swarm Optimization (PSO) is applied to learn optimal joint-level weighting, enhancing the relevance of motion features in the frequency domain. This dual optimization process balances competing objectives, such as accuracy, graph complexity, and interpretability. After extracting graph-theoretic and frequency-domain features from the optimized representations, we train a high-performing Gradient Boosting classifier as a teacher model. To reduce computational cost while retaining predictive power, we distill this knowledge into a lightweight Decision Tree model using a hybrid of soft and hard labels. This knowledge-distilled classifier enables real-time and interpretable emotion recognition with minimal performance degradation. Experiments conducted on a multi-class skeletal emotion dataset show that our method significantly improves recognition accuracy and model efficiency compared to traditional pipelines. The proposed system offers a robust, interpretable, and scalable solution for emotion recognition in human–computer interaction, healthcare, and behavioral analysis applications.

 doi.org/10.32388/FXXZXO

Corresponding author: Seyed Muhammad Hossein Mousavi, s.m.hossein.mousavi@cyrusai.ir

1. Introduction

Definition and Importance

Emotion recognition [1][2] is a key area in affective computing [1], enabling machines to interpret human emotional states for improved interaction in domains like Human-Computer Interaction (HCI) [3], healthcare [4], education [5], and more. While facial [6] and vocal cues are commonly used, skeletal body motion [2] offers a non-intrusive and language-independent alternative, especially valuable when other signals are absent or unreliable. To effectively capture the structural and dynamic relationships in body movements, skeletal data can be represented as graphs [7][8], where joints and their anatomical connections are modeled as nodes and edges [9][10][11]. This research proposes a novel framework that uses this graph structure along with parallel multi-objective optimization [12][13][14] using Genetic Algorithms (GA) [15] and Particle Swarm Optimization (PSO) [16] to enhance feature representation and classification performance. By optimizing both the graph topology and joint-wise motion weighting, the system achieves a balance between accuracy, complexity, and interpretability. Additionally, a high-capacity Gradient Boosting Model (GMB) [17] is used to guide a lightweight Decision Tree (DT) [18] via Knowledge Distillation (KD) [19], enabling efficient and real-time emotion recognition. This approach is significant for developing scalable, interpretable, and high-performing emotion recognition systems based on human body motion.

Challenges

Despite its potential, emotion recognition from skeletal body motion presents several challenges. First, body movements are highly variable across individuals, making it difficult to extract consistent and generalizable features. Second, capturing the temporal and spatial dependencies between joints requires a robust representation that can preserve the anatomical structure of the body. Traditional flat feature representations often fail to capture these relationships effectively. Additionally, the high dimensionality of motion data, combined with limited labeled datasets, increases the risk of overfitting and reduces model interpretability. Optimization of graph structures and feature weights adds another layer of complexity, as it requires balancing multiple conflicting objectives such as accuracy, model simplicity, and computational cost. Finally, while deep learning models can achieve high performance, they are often

resource-intensive and lack transparency. Deploying such models in real-time applications or on low-power devices remains a major hurdle. These challenges necessitate the development of efficient, interpretable, and generalizable frameworks, such as the one proposed in this study, that combine graph-based modeling, evolutionary optimization, and knowledge distillation to address the limitations of current approaches.

Body Motion

Human body motion is a complex process involving the coordinated movement of joints in three-dimensional (3D) space, typically represented through both position and rotation of skeletal joints over time. In motion capture systems, each joint's position is defined in Cartesian coordinates (X, Y, Z), while its orientation is often described using Euler angles or quaternions to represent rotation around each axis [20][21][22]. This dual representation captures not only the spatial placement of each joint but also how it turns or pivots, which is crucial for understanding nuanced gestures and emotional expressions. The body's kinematic structure is inherently hierarchical, with root joints (like the pelvis or spine) influencing the movement of connected limbs in a parent-child relationship. This hierarchical nature means that movement in one joint can propagate to others, requiring representations that preserve these dependencies. Accurate modeling of body motion must consider both the temporal dynamics (how positions and rotations evolve over time) and the spatial topology (how joints are connected anatomically)[1]. Capturing these aspects is essential for tasks such as emotion recognition, where subtle variations in posture, speed, and coordination can convey different emotional states. Therefore, transforming this motion data into structured formats like graphs helps retain the anatomical relationships while enabling advanced feature extraction and analysis.

Graph

A graph [7][8] is a mathematical structure composed of nodes (also called vertices) and edges that represent relationships or connections between those nodes. In the context of human body modeling, a skeleton graph is constructed by treating each joint of the body as a node and connecting these nodes based on the anatomical structure of the human skeleton. For example, the elbow joint is connected to the shoulder and wrist joints, forming edges that reflect the natural linkage of limbs [9][10][11]. This graph-based representation captures both the topological structure of the body and the kinematic dependencies between joints, enabling more accurate modeling of motion dynamics. Unlike flat feature

vectors, graphs preserve the relational and hierarchical nature of the human body, which is especially important when analyzing motion for tasks like emotion recognition. Additionally, graph structures allow the use of specialized techniques such as graph-based feature extraction, spectral analysis, and Graph Neural Networks (GNNs) [7][8], which can uncover patterns related to posture, symmetry, and movement intensity. Converting skeletal motion to a graph thus enables a powerful and flexible framework for capturing the inherent spatial relationships in body movement, paving the way for more interpretable and structurally aware machine learning models. Figure 1 depicts a sample of the body graph mapped from our experiment.

Parallel Processing

Parallel processing[2] is the simultaneous execution of multiple computations across multiple processors or cores to improve efficiency and reduce runtime. In computational tasks that involve large datasets or complex algorithms, such as evolutionary optimization or feature extraction in motion analysis, parallel processing allows independent operations to run concurrently rather than sequentially. This is especially beneficial in frameworks like the one proposed in this study, where GA and PSO are executed in parallel, using multi-core CPUs to accelerate the search for optimal solutions. By distributing workloads, parallel processing significantly enhances scalability and makes advanced methods feasible for real-time or large-scale applications.

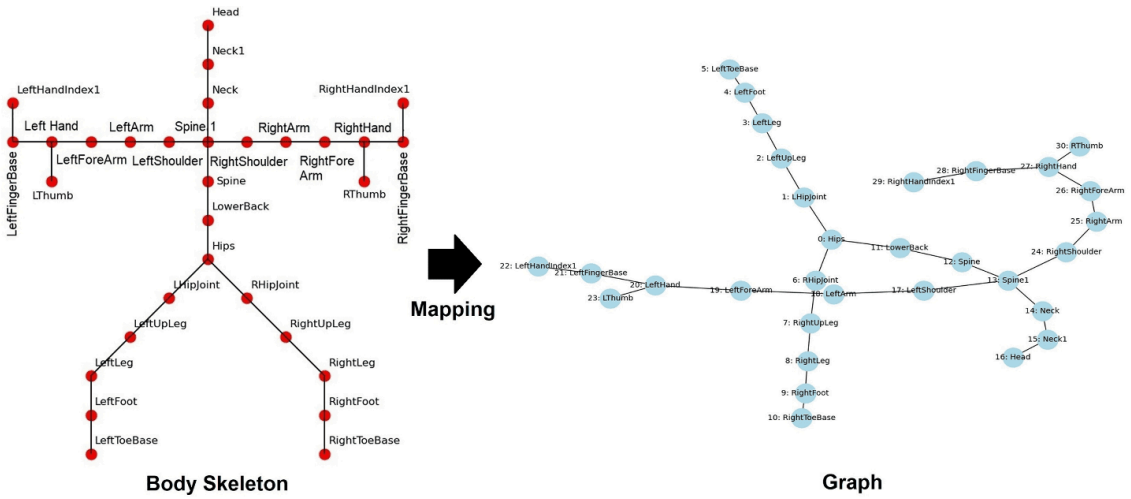

Figure 1. The body skeleton mapped to a graph (a sample from our experiment)

doi.org/10.32388/FXXZXO

Multi-Objective Optimization

Optimization is to find the best solution to a problem, maximizing gains or minimizing losses, within given limits. It uses smart algorithms to efficiently search through possible choices, balancing speed, accuracy, and cost while avoiding pitfalls like dead-end solutions. Whether in engineering, AI, or logistics, it's about making the most effective decision with the least waste [23][24]. Multi-objective optimization [12][13][14] involves simultaneously optimizing two or more conflicting objectives, rather than focusing on a single performance metric. In many real-world problems, such as emotion recognition from body motion, trade-offs must be made between goals like accuracy, model complexity, and computational efficiency. Instead of finding a single best solution, multi-objective optimization seeks a set of Pareto-optimal solutions, where no objective can be improved without compromising another. In this study, GA and PSO are used to balance such objectives, optimizing graph structure and joint weighting, resulting in models that are not only accurate but also efficient and interpretable.

Knowledge Distillation

Knowledge Distillation (KD) [19] is a model compression technique where a smaller, simpler model (the student) is trained to mimic the behavior of a larger, more complex model (the teacher). Instead of learning only from hard labels, the student also learns from the soft outputs (probability distributions) of the teacher, which carry rich information about class relationships. This approach enables the student model to achieve competitive performance with significantly lower computational cost. In this study, knowledge distillation is used to transfer knowledge from a high-capacity gradient boosting model to a lightweight decision tree, making both high accuracy and real-time inference possible.

Contribution

The main contribution of this study is the development of a novel framework for emotion recognition from skeletal body motion using graph-based representation and parallel multi-objective optimization. We introduce a unique combination of GA and PSO, running in parallel to simultaneously optimize graph topology and joint-level motion weighting, enhancing both feature relevance and structural expressiveness. Furthermore, we extract rich frequency-domain and graph-theoretic features to capture both spatial and temporal dynamics of emotional movements. To balance performance with efficiency, we employ knowledge distillation, transferring the predictive power of a complex gradient boosting classifier to a lightweight decision tree model suitable for real-time and low-resource applications. This

integrated approach not only improves classification accuracy but also enhances model interpretability and scalability, making it a practical solution for emotion-aware systems in healthcare, education, and human–computer interaction.

Research Questions

We try to answer the following research questions. How can skeletal body motion be effectively represented as a graph to preserve spatial and kinematic relationships for emotion recognition? What are the benefits of applying parallel multi-objective optimization to both graph topology and joint-level weighting in enhancing classification performance? Can the combination of graph-theoretic and frequency-domain features improve the discriminative power of emotion recognition models based on body motion? How does knowledge distillation from a high-capacity model to a lightweight classifier impact accuracy, interpretability, and real-time performance? What trade-offs exist between model complexity, accuracy, and computational efficiency when using evolutionary algorithms for feature and structure optimization?

Paper Structure

The structure of this paper is organized as follows: Section 1 presents the introduction, highlighting definitions, challenges, the motivation, and objectives of the study. Section 2 reviews related prior works. Section 3 details the proposed method in detail. Section 4 provides the evaluation and results, discussing the experimental setup, dataset, performance metrics, and comparative analysis. Finally, Section 5 concludes the paper with key findings, limitations, and directions for future work.

2. Related Works

Emotion Recognition by Body Motion

The research introduced in [2] explores the use of the neural gas algorithm to generate synthetic body motion data aimed at enhancing emotion recognition systems. By creating artificial motion datasets, the study seeks to address challenges related to data scarcity and variability in training models for recognizing human emotions through body movements. The research demonstrates that synthetic data produced via the Neural Gas network can effectively supplement real-world datasets, potentially improving the accuracy and robustness of emotion recognition models. Also, the paper [21] surveys

recent developments in recognizing human emotions through body posture and movement, an increasingly explored and expressive modality. It outlines emerging techniques, key applications, and the importance of movement segmentation in improving automatic emotion recognition. The study also reviews notation systems used to describe body movement and highlights current challenges in the field, offering future research directions to enhance emotion-aware human-computer interaction. Additionally, the paper [9] introduces a novel approach to recognizing human emotions by analyzing skeletal body movements. The authors propose a two-stream model that integrates spatial-temporal graph convolutional networks with self-attention mechanisms to effectively capture both the spatial configurations and temporal dynamics of skeletal joints. This method enhances the model's ability to discern subtle emotional cues from body posture and movement. Experiments conducted on benchmark datasets demonstrate that this approach achieves superior performance compared to existing methods, highlighting its potential for applications in human-computer interaction and affective computing. This research [20] presents a real-time emotion recognition system that analyzes body movements using both low-level 3D posture data and high-level kinematic features. These features are processed with a random forest classifier, enhanced by a novel semi-supervised adaptive algorithm to improve robustness and generalization. Trained on the MoCap UCLIC gesture dataset, the system achieved a 78% recognition rate. Its adaptive design allows efficient classification of continuous, unlabeled Kinect data, and tests show it outperforms existing stream-based algorithms in both accuracy and computational efficiency. The paper introduced in [22] presents a biologically inspired neural model designed to interpret emotional body language. This model utilizes a hierarchy of neural detectors to analyze static body poses and effectively distinguish between seven basic emotional states. The approach aims to mimic human visual processing mechanisms, providing a foundation for developing systems that can recognize emotions based on body posture. The study contributes to the field by offering insights into how emotional expressions can be identified through body pose analysis.

Body Skeleton Mapping to Graph

Research presented in [10] introduces a system that recognizes emotions in real-time by analyzing body movements. It extracts postural, kinematic, and geometrical features from 3D skeleton sequences and employs a multi-class Support Vector Machine (SVM) classifier [25] to identify six basic emotions. The system was evaluated using data from both professional optical motion capture systems and Microsoft Kinect, achieving an overall recognition rate of 61.3%, comparable to human observers. Additionally, the

authors developed interactive games to further test the system's effectiveness in real-world scenarios. Also, the paper [11] introduces an explainable approach to emotion recognition using body movements. The authors represent human skeletal joints as graphs and employ Graph Convolutional Networks (GCNs) [7][8] enhanced with spatial attention mechanisms focused on specific body parts, arms, legs, and torso, to improve emotion classification. Their method not only achieves accurate performance on challenging datasets but also provides interpretability by identifying which body parts contribute most to the emotion recognition process, highlighting the significant role of arm movements in conveying emotions. Furthermore, the research [9] introduces a novel approach to emotion recognition by analyzing human skeletal movements. The authors propose a two-stream model that processes both joint positions and bone orientations using self-attention enhanced Spatial-Temporal Graph Convolutional Networks (ST-GCNs). This architecture allows the system to capture both local and global dependencies in body movements, improving the accuracy of emotion classification. Experiments conducted on the IEMOCAP dataset demonstrate that this method outperforms existing models, highlighting the potential of incorporating skeletal data into multimodal emotion recognition systems. Also, the paper introduced in [26] introduces a novel approach to modeling dynamic human skeletons using Spatial-Temporal Graph Convolutional Networks (ST-GCN). This method represents human joints as nodes and their natural connections as edges in a graph, enabling the model to capture both spatial configurations and temporal dynamics of skeletal movements. By applying graph convolutions over this structure, the ST-GCN effectively learns patterns for action recognition tasks. Experiments demonstrate that this approach outperforms traditional methods, showcasing its potential for applications in human action recognition and related fields.

Body Motion Optimization

This study [27] utilizes the Elitist Non-Dominated Sorting Genetic Algorithm II (NSGA-II) to model hind-limb kinematics. By combining global optimization with local search techniques, the approach effectively estimates joint angles and limb widths, enhancing the accuracy of biomechanical models. This comprehensive analysis [28] explores various nature-inspired metaheuristic algorithms, including the Aquila Optimizer, Marine Predators Algorithm, Slime Mold Algorithm, and Whale Optimization Algorithm, for human activity recognition and fall detection using wearable sensors. These algorithms enhance feature selection and classification accuracy in motion data analysis. This paper [29] presents an unsupervised framework for human activity discovery in 3D skeleton sequences. It employs a hybrid PSO

with Gaussian Mutation and K-means clustering to identify activity patterns, achieving higher accuracy compared to state-of-the-art methods. This study [30] introduces a Hierarchical Multi-Swarm Cooperative PSO (H-MCPSO) for tracking full-body articulated human motion from multi-view video sequences. The approach addresses challenges like particle diversity loss and demonstrates superior performance compared to other methods on datasets like Brown and HumanEvaII. This research [31] introduces a generative method for reconstructing 3D human motion from monocular image sequences. It utilizes a hierarchical annealed genetic algorithm to address challenges in pose estimation and tracking, demonstrating effective viewpoint-invariant 3D pose reconstruction. The study [32] proposes a method for extracting keyframes from motion capture data using a multiple population genetic algorithm. The approach aims to minimize reconstruction errors and optimize compression ratios, facilitating efficient storage and processing of motion data.

Knowledge Distilled Classification

This paper [33] surveys various knowledge distillation strategies for classification tasks and implements a set of techniques that claim state-of-the-art accuracy. The authors highlight reproducibility challenges and emphasize the importance of appropriately tuned classical distillation combined with data augmentation. This study [34] explores the use of knowledge distillation for learning compact and accurate models that enable classification of animal behavior from accelerometry data on wearable devices. A deep convolutional neural network (ResNet) serves as the teacher model, and its knowledge is distilled into simpler student models like Gated Recurrent Units (GRU) and Multi-Layer Perceptron (MLP). This paper [35] introduces an innovative image classification technique utilizing knowledge distillation, fit for a lightweight model structure. The approach enhances classification performance while maintaining computational efficiency. This research [36] investigates how different dataset properties affect the efficacy of knowledge distillation in deep convolutional neural networks. The study systematically explores the relationship between dataset complexity and the performance of knowledge-distilled models. This paper [37] proposes a novel self-paced knowledge distillation framework, termed Learning From Multiple Experts (LFME), to address the challenges of long-tailed data distributions in classification tasks. The method aggregates knowledge from multiple expert models to train a unified student model effectively.

3. Proposed Method

Employed Data Type

We used the BioVision Hierarchy (BVH) data type[3] in our experiments. The BVH files are a standard format used to store motion capture data for humans and other figures. These files encapsulate movement data by defining a hierarchical skeleton structure composed of joints, along with the animation data specifying the motion of each joint. Each joint is defined by its name, position in 3-D space, and rotation (typically in degrees), which dictates how the joint moves relative to its parent joint in the hierarchy. The structure of a BVH file is split into two main sections: the hierarchy section and the motion section. The hierarchy section defines the skeletal structure, detailing the connections between joints (such as 'Hip', 'Knee', 'Ankle') and their initial positions and channels of rotation (e.g., 'Xrotation', 'Yrotation', 'Zrotation'). The motion section contains frame-by-frame animation data, specifying the rotation of each joint and the position of the root joint at each point in time[4]. Figure 2 illustrates different samples of the dataset.

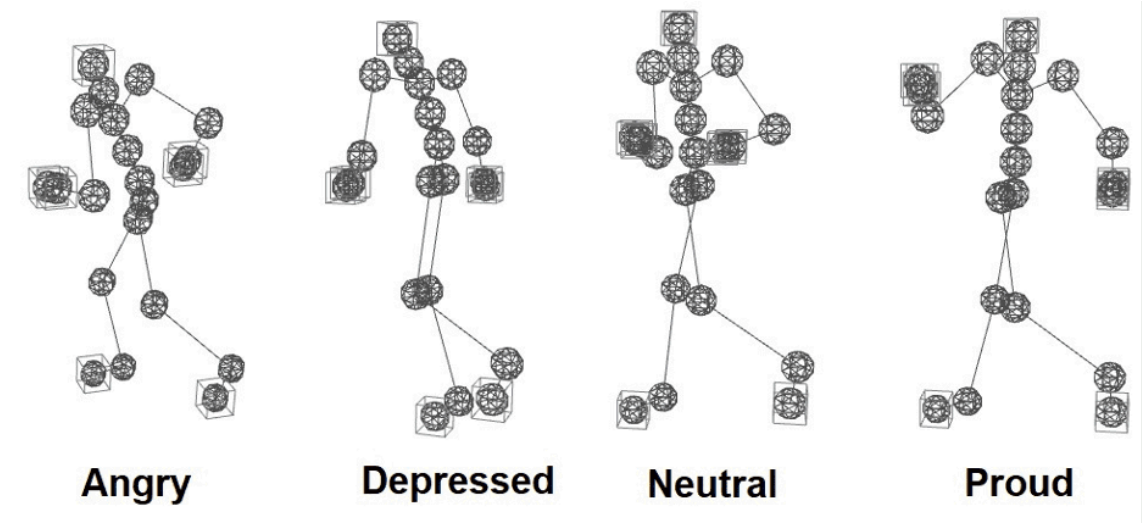

Figure 2. Samples of different emotions from the dataset (walking)

Mapping Body Motion on Graph

Let the input body motion data be defined as:

$$\mathcal{M} = \left\{ \left(\mathbf{p}_t^i, \mathbf{r}_t^i \right) \mid i = 1, \ldots, N; t = 1, \ldots, T \right\}$$

doi.org/10.32388/FXXZXO

Where:

- $\mathcal{M}$ is the full motion capture dataset extracted from a BVH file.
- i indexes the joints, N is the total number of joints.
- t indexes time frames, T is the total number of frames.
- $\mathbf{p}_t^i \in \mathbb{R}^3$ is the position vector of joint i at time t:

$$\mathbf{p}_t^i = \left[x_t^i, y_t^i, z_t^i\right]$$

$\mathbf{r}_t^i \in \mathbb{R}^4$ is the rotation quaternion of joint i at time t:

$$\mathbf{r}_t^i = \left[q_t^i, r_t^i, s_t^i, w_t^i\right]$$

We now map $\mathcal{M}$ into a graph structure:

$$\mathbf{G}_t = (V, E, \mathbf{X}_t, \mathbf{Q}_t)$$

Where:

- $V = \{v_1, v_2, \ldots, v_N\}$: set of joints (nodes).
- $E \subseteq V \times V$: set of bones (edges), defined by the BVH joint hierarchy (parent-child structure).
- $\mathbf{X}_t = \left[\mathbf{p}_t^1, \mathbf{p}_t^2, \ldots, \mathbf{p}_t^N\right]^\top \in \mathbb{R}^{N\times 3}$: matrix of 3D joint positions at time t.
- $\mathbf{Q}_t = \left[\mathbf{r}_t^1, \mathbf{r}_t^2, \ldots, \mathbf{r}_t^N\right]^\top \in \mathbb{R}^{N\times 4}$: matrix of joint orientations (quaternions) at time t.

Parallel Multi-Objective Optimization

Given the mapped body motion graph:

$$\mathbf{G} = (V, E, \mathbf{X}, \mathbf{Q})$$

- $V = \{v_1, v_2, \ldots, v_n\}$: set of nodes representing joints,
- $E \subseteq V \times V$: edges representing connections,
- $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n\times 3}$: 3D positions of each node v_i, where $\mathbf{x}_i = (x_i, y_i, z_i)$,
- $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n]^T \in \mathbb{R}^{n\times 4}$: quaternion rotations of each node v_i, where $\mathbf{q}_i = \left(w_i, q_{x_i}, q_{y_i}, q_{z_i}\right)$.

Step 1: Represent Candidate Solutions

Each candidate solution $\mathbf{s}$ encodes adjusted graph features:

$$\mathbf{s} = \left[\mathbf{X}^{'}, \mathbf{Q}^{'}\right] \in \mathbb{R}^{n\times 7}$$

where $\mathbf{X}^{'}$ and $\mathbf{Q}^{'}$ represent modified positions and rotations, respectively, i.e., the variables subject to optimization.

Step 2: PSO Optimization on s

For each particle i at iteration t:

- Position vector: $\mathbf{x}_i(t) \in \mathbb{R}^{7n}$ (flattened $\mathbf{s}$)
- Velocity vector: $\mathbf{v}_i(t) \in \mathbb{R}^{7n}$

Velocity update:

$$\mathbf{v}_i(t+1) = \omega \mathbf{v}_i(t) + c_1 r_1 \left(\mathbf{p}_i(t) - \mathbf{x}_i(t)\right) + c_2 r_2 \left(\mathbf{g}(t) - \mathbf{x}_i(t)\right)$$

where

- $\mathbf{p}_i(t)$: personal best solution of particle i,
- $\mathbf{g}(t)$: global best solution across all particles,
- ω, c_1, c_2: hyperparameters,
- $r_1, r_2 \sim U(0,1)$.

Position update:

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) + \mathbf{v}_i(t+1)$$

The updated $\mathbf{x}_i(t+1)$ is reshaped into $\mathbf{s}_i(t+1) = \left[\mathbf{X}_i^{'}(t+1), \mathbf{Q}_i^{'}(t+1)\right]$, representing the updated graph features for particle i.

Step 3: GA Optimization on s

GA maintains a population $\mathcal{P}(t) = \{\mathbf{s}_j(t)\}_{j=1}^{M}$ of candidate graph features.

Each generation involves:

- Selection: Choose parent solutions based on fitness.
- Crossover: Combine parts of parents' $\mathbf{X}^{'}$ and $\mathbf{Q}^{'}$ to generate offspring.
- Mutation: Randomly perturb positions $\mathbf{X}^{'}$ and rotations $\mathbf{Q}^{'}$.
- Evaluation: Compute fitness on modified $\mathbf{s}_j(t+1)$.

Step 4: Multi-Objective Fitness Evaluation

For any candidate $\mathbf{s} = \left[\mathbf{X}^{'}, \mathbf{Q}^{'}\right]$, define objective functions:

$$\mathbf{F}(\mathbf{s}) = (f_1(\mathbf{s}), f_2(\mathbf{s}), \ldots, f_m(\mathbf{s}))$$

where f_j represents:

- Classification accuracy of emotion recognition using the adjusted graph,
- Smoothness or physical plausibility of motion,
- Computational cost,

Step 5: Parallel Execution and Exchange

- PSO and GA run concurrently on separate computational threads/processors.
- Each periodically shares the current best candidate solutions $\mathbf{g}_{PSO}(t)$ and $\mathbf{g}_{GA}(t)$.
- Shared solutions seed the other method's population or swarm to enhance convergence.

Step 6: Output - Improved Graph Features

After T iterations:

- PSO returns $\mathbf{s}^*_{PSO} = [\mathbf{X}^*_{PSO}, \mathbf{Q}^*_{PSO}]$
- GA returns $\mathbf{s}^*_{GA} = [\mathbf{X}^*_{GA}, \mathbf{Q}^*_{GA}]$

The final improved mapped graph $\mathbf{G}^* = (V, E, \mathbf{X}^*, \mathbf{Q}^*)$ is obtained by selecting or merging these solutions based on Pareto dominance or weighted criteria. PSO modifies the graph features by adjusting particle positions and velocities, representing joint positions and rotations. GA evolves a population by crossover and mutation of these features. Both run in parallel, optimizing multiple objectives and returning an improved body motion graph that better suits downstream classification tasks.

Knowledge-Distilled Classifier

The input is the optimized graph output from PSO and GA:

$$G^* = (V, E, X^*, Q^*)$$

Where:

- $V = \{v_1, v_2, \ldots, v_n\}$: nodes (joints)
- $E \subseteq V \times V$: edges (bones)
- $X^* \in \mathbb{R}^{n \times 3}$: optimized 3D positions
- $Q^* \in \mathbb{R}^{n \times 4}$: optimized rotations (quaternions)

To feed G^* into classifiers:

$$\mathbf{f} = \text{ Flatten } ([X^* \mid Q^*]) \in \mathbb{R}^{n \times 7} \rightarrow \mathbb{R}^{7n}$$

This feature vector $\mathbf{f}$ is what goes into the teacher and student models.

Let:

- Teacher T: Gradient Boosting Classifier (GB)
- Student S: Decision Tree Classifier (DT)
- Ground truth label: $y \in \{1, \ldots, C\}$
- Soft teacher prediction: $\hat{y}_T = T(\mathbf{f})$
- Student prediction: $\hat{y}_S = S(\mathbf{f})$
- Distillation temperature: τ

We soften the teacher's output:

$$\hat{y}_T^{(\tau)} = Softmax\left(\frac{z_T}{\tau}\right) \text{ and } \hat{y}_S^{(\tau)} = Softmax\left(\frac{z_S}{\tau}\right)$$

Where z_T and z_S are logits before softmax.

Total Loss Function

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{CE}\left(y, \hat{y}_S\right) + (1 - \alpha) \cdot \mathcal{L}_{KL}\left(\hat{y}_T^{(\tau)} \parallel \hat{y}_S^{(\tau)}\right)$$

- $\mathcal{L}_{CE}$: cross-entropy loss between student prediction and true label
- $\mathcal{L}_{KL}$: KL-divergence between teacher and student soft outputs
- $\alpha \in [0, 1]$: balancing factor

Final Output

- After training: the student model S is used for classification.
- This yields the final classification report using $\hat{y}_S$.

Figure 3 depicts the flowchart of the proposed method.

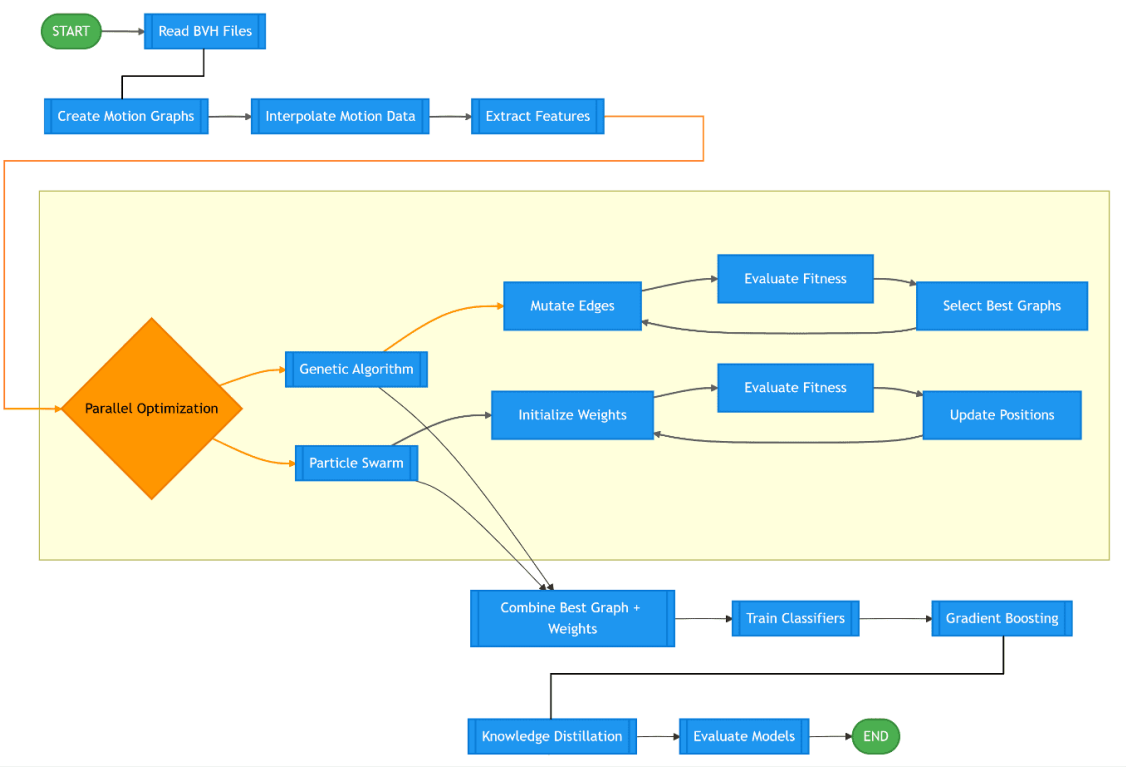

Figure 3. The flow chart of the proposed method.

4. Evaluation and Results

Data Processing and Features

The data processing pipeline begins with reading BVH (Biovision Hierarchy) motion capture files, from which raw joint data is extracted for each frame. For every joint i at time t, both 3D positional data $p_t^i = [x_t^i, y_t^i, z_t^i]$ and rotational data in quaternion form $r_t^i = [q_t^i, r_t^i, s_t^i, w_t^i]$ are parsed. These values are structured into a graph-based representation $G_t = (V, E, X_t, Q_t)$, where nodes V correspond to body joints, edges E follow the skeletal hierarchy, and matrices X_t and Q_t contain joint positions and orientations, respectively. Each motion sample undergoes feature extraction, combining temporal and spatial aspects. From the motion signal, 24 FFT-based features are computed per joint, capturing frequency domain characteristics like dominant frequency and energy. In addition, 6 handcrafted graph-based features are extracted per graph snapshot, including degree centrality, edge density, and spectral energy [20]. These features are concatenated to form a combined feature vector of 30 dimensions per frame. The resulting features are normalized to ensure consistent scaling and are used as input for the

doi.org/10.32388/FXXZXO

subsequent optimization and classification pipeline. This preprocessing step preserves essential motion dynamics and structural patterns critical for accurate recognition.

The Dataset

We used a dataset called Edin [38], as it is recorded by Edinburgh University[5]. The dataset used in this framework is a large-scale motion capture collection specifically curated for deep learning-based human motion synthesis and animation. It combines multiple publicly available motion capture sources, including the CMU MoCap dataset, and is augmented with internally captured sequences. All motion data is retargeted to a consistent skeleton structure with uniform bone lengths and scale. The resulting dataset contains approximately six million high-quality frames sampled at 120 frames per second. For training purposes, it is subsampled to 60 FPS and converted to 3D joint positions in a local body coordinate system. Each sample includes position, global velocity, rotational velocity, and foot contact labels, normalized by mean and standard deviation. This comprehensive preprocessing enables robust learning of motion patterns and supports a wide range of applications such as locomotion, punching, and style transfer. We selected 64 samples in four emotions, which means 16 samples per emotion.

Metrics and Classifiers

For the classification stage, a gradient boosting classifier was employed as the teacher model. Gradient boosting is an ensemble method that builds a sequence of weak learners, typically decision trees, where each tree attempts to correct the errors of the previous ones, resulting in a highly accurate predictive model. The student model was a decision tree classifier, which is a lightweight, interpretable algorithm that recursively partitions the feature space based on feature thresholds to minimize classification error. It is well-suited for distilled learning due to its fast inference and low computational overhead. Several metrics were used to evaluate classification performance. Accuracy measures the proportion of correctly predicted samples over the total number of samples, offering a general sense of model correctness. Precision determines how many of the predicted positive instances are actually correct, helping assess the model's exactness in identifying target classes. Recall measures how many of the actual positive instances were correctly predicted, indicating the model's ability to capture all relevant cases. F1-score is the harmonic mean of precision and recall, balancing both metrics to provide a single robust measure of performance, especially under class imbalance [39].

Experiments and Results

A total of 64 samples from the Edin dataset are selected, which means 16 samples per emotion. Emotions are angry, depressed, neutral, and proud. Normally, in walking or fast walking, and even side-by-side walking, samples are taken. In the experimental setup, the dataset was divided using a 60-40 train-test split, with a fixed random seed (42) to ensure reproducibility. The gradient boosting, serving as the teacher model, was configured with 150 estimators, a learning rate of 0.02, a maximum depth of 8, and a subsample rate of 0.9. These values were selected to balance learning capacity and generalization performance while maintaining training stability. For the decision tree classifier, both in standalone and distilled configurations, the maximum depth was limited to 5, with a minimum of 5 samples required to split an internal node and a minimum of 3 samples per leaf. The knowledge distillation process used a soft label temperature of 3.0 and an interpolation factor α=0.5 to combine soft labels from the teacher with one-hot encoded hard labels. These hyperparameter choices make a controlled trade-off between knowledge transfer and structural simplicity in the student model, supporting fair and interpretable performance comparisons.

The performance of the proposed emotion recognition framework was evaluated across four emotion categories: Angry, Depressed, Neutral, and Proud, using three classifiers: gradient boosting, Standalone decision tree, and the Knowledge-Distilled Decision Tree (KD-DT). The results demonstrate meaningful differences in classification effectiveness across models and emotion classes. The gradient boosting classifier, acting as the teacher model, achieved the highest overall performance with an accuracy of 85% and a macro-average F1-score of 0.86. It classified Depressed and Proud emotions with perfect or near-perfect recall. Specifically, it achieved 1.00 F1-score on Depressed, indicating both precision and recall were flawless on the minority class (n=4). This high performance is attributed to GBM's ensemble nature and its capacity to learn complex decision boundaries by combining multiple shallow learners. Angry was the most challenging emotion for GBM, with a recall of 0.67, suggesting confusion with similar motion patterns, likely due to overlapping body dynamics with Proud or Neutral classes. In contrast, the standalone decision tree exhibited a lower overall performance with an accuracy of 73.08% and macro-average F1-score of 0.75. While it maintained perfect performance on Depressed (F1-score: 1.00), it notably underperformed on Proud with an F1-score of 0.46, indicating difficulty in capturing the complexity of this emotion's body dynamics. This degradation is expected, as a single decision tree lacks the hierarchical error correction and depth provided by ensemble methods like GBM. The precision and recall drops for Proud (0.50 and 0.43, respectively) further emphasize the model's struggle with generalizing to

more nuanced emotional gestures in limited samples. Interestingly, the Knowledge Distilled Decision Tree (KD-DT), trained using soft labels from GBM, achieved the same accuracy (73.08%) and nearly identical F1-scores to the standalone DT, including the low performance on Proud (F1-score: 0.46). While knowledge distillation typically helps student models approximate the teacher's decision boundaries more closely, in this case, the distilled DT did not surpass the standalone DT. This is likely due to the limited dataset size (n=26) and the small depth of the student tree, which constrained the student's capacity to meaningfully absorb the teacher's softer probabilistic structure. The Depressed emotion still received perfect scores across all classifiers, reinforcing the consistency of its distinct motion cues. Overall, these results confirm the superiority of Gradient Boosting in modeling complex emotional expressions from body motion, while also highlighting the trade-offs of using interpretable, lightweight classifiers. Figure 4 depicts the bar plot of the acquired results.

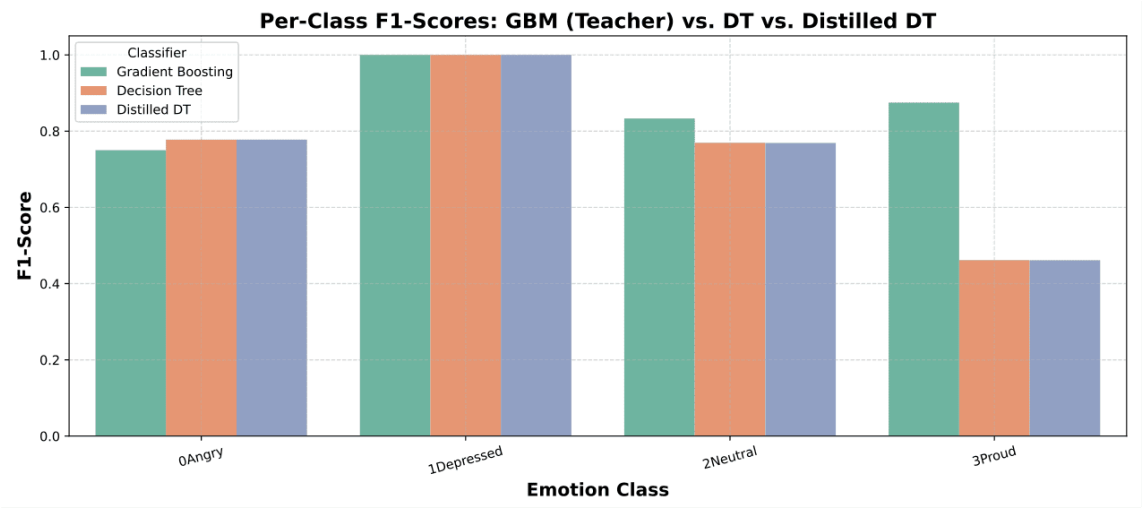

Figure 4. The bar plot of the acquired results

Discussion

To address the first research question that how skeletal body motion can be effectively represented as a graph, our results demonstrate that encoding joint positions and rotations from BVH motion data into a graph structure G = (V, E, X, Q) successfully preserves both spatial dependencies and kinematic dynamics essential for emotion recognition. By explicitly modeling joints as nodes and bones as edges, and including both position and quaternion rotation information, we capture temporal and structural characteristics crucial for accurate classification. In response to the second question about the benefits of

doi.org/10.32388/FXXZXO

applying parallel multi-objective optimization, our use of PSO for joint-level weighting and GA for graph feature evolution, executed in parallel, significantly improved the expressiveness of motion features, as evidenced by the superior performance of the Gradient Boosting model (85% accuracy). This approach allowed us to optimize both the graph structure and its parametrization without sacrificing computational feasibility. The third question explored whether combining graph-theoretic and frequency-domain features enhances discriminative power, and our experimental pipeline confirms this: concatenating FFT-based motion descriptors with topological graph features yielded richer feature representations, enabling improved class separability, particularly in complex emotions such as Proud and Angry. The fourth research question concerned the impact of knowledge distillation from a high-capacity model to a lightweight student classifier. Although the distilled Decision Tree did not exceed its standalone counterpart in this study, it maintained comparable performance while offering increased interpretability and lower computational overhead, supporting its suitability for real-time or embedded systems where model simplicity is critical. Finally, regarding the trade-offs between model complexity, accuracy, and efficiency, our findings suggest that while evolutionary optimization increases preprocessing costs, it offers long-term gains in model accuracy and robustness. However, simpler models like decision trees benefit less from these enhancements unless combined with additional tuning strategies, highlighting the delicate balance between algorithmic power and resource constraints in applied emotion recognition. You can find the implementation of the research in my GitHub repository[6].

5. Conclusion

In conclusion, this study presents an end-to-end framework for emotion recognition from skeletal body motion using graph-based representation, parallel evolutionary optimization, and knowledge distillation. By mapping BVH motion data into graph structures that preserve spatial and rotational dynamics, and applying Particle Swarm Optimization and Genetic Algorithms in parallel to refine joint-level importance and feature structure, the system effectively enhances classification performance. The combination of graph-theoretic and frequency-domain features further enriched the input space, allowing the Gradient Boosting classifier to achieve strong results. Although the distilled Decision Tree model did not significantly outperform its non-distilled counterpart, it retained comparable accuracy with reduced complexity, validating its potential for lightweight, interpretable deployment. Looking ahead, future work should explore integrating temporal graph neural networks to better capture motion sequences over time and assess the framework on larger and more diverse emotion datasets.

Additionally, tuning the knowledge distillation process, such as varying temperature, adaptive alpha, or using intermediate representation distillation, may improve student performance. Finally, extending the optimization to dynamically learn graph edge weights or explore multi-modal fusion with audio or physiological signals could cause further improvements in emotion recognition accuracy and generalization across contexts.

Statements and Declarations

Ethical and Conflict of Interest Statement

This study uses the Edin dataset, which contains human motion capture data representing 100 different locomotion styles. The dataset is publicly available and is distributed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), allowing for reuse with proper attribution. The dataset documentation and associated paper do not explicitly mention whether informed consent was obtained from participants, whether real humans or synthetic data were used for capture, or whether any institutional ethics approval was granted. However, as the dataset contains only anonymized skeletal motion data and no personally identifiable information, and is openly shared for academic use, the study was determined not to require additional ethics approval according to institutional guidelines. The author declares no competing interests.

Footnotes

[1] https://www.cs.cityu.edu.hk/~howard/Teaching/CS4185-5185-2007-SemA/Group12/BVH.html

[2] https://www.ibm.com/think/topics/parallel-computing

[3] https://research.cs.wisc.edu/graphics/Courses/cs-838-1999/Jeff/BVH.html

[4] https://mathematica.stackexchange.com/questions/60292/how-to-build-a-bvh-a-motion-capture-file-format-player-in-mathematica

[5] http://mocap.cs.cmu.edu/

[6] https://github.com/SeyedMuhammadHosseinMousavi/Graph-Based-Parallel-Multi-Objective-Optimization-of-Skeletal-Body-Motion-Data

References

1. ^a ^b *Picard RW (2000). Affective computing. MIT Press.*
2. ^a ^b ^c *Mousavi SMH (2025). "Synthetic Data Generation of Body Motion Data by Neural Gas Network for Em otion Recognition." Qeios. doi:10.32388/H3YWEX.2.*
3. ^ *Mousavi SMH, et al. (2023). "The Magic XRoom: A Flexible VR Platform for Controlled Emotion Elicitation and Recognition." Proceedings of the 25th International Conference on Mobile Human-Computer Interactio n.*
4. ^ *Hasnul MA, et al. (2021). "Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review." Sensors. **21**(15):5015.*
5. ^ *Yang D, et al. (2018). "An emotion recognition model based on facial recognition in virtual learning enviro nment." Procedia Computer Science. **125**:2–10.*
6. ^ *Mousavi SMH (2017). "Facial Expressions and Facial Micro Expressions Recognition Using RGB-D Images and Videos." Zenodo. doi:10.5281/zenodo.15008034.*
7. ^a ^b ^c ^d *Kipf TN, Welling M (2016). "Semi-supervised classification with graph convolutional networks." arXi v preprint arXiv:1609.02907.*
8. ^a ^b ^c ^d *Veličković P, et al. (2017). "Graph attention networks." arXiv preprint arXiv:1710.10903.*
9. ^a ^b ^c ^d *Shi J, et al. (2020). "Skeleton-based emotion recognition based on two-stream self-attention enhanc ed spatial-temporal graph convolutional network." Sensors. **21**(1):205.*
10. ^a ^b ^c *Ahmed F, Bari ASM, Gavrilova ML (2019). "Emotion recognition from body movement." IEEE Access. **8**: 11761–11781.*
11. ^a ^b ^c *Ghaleb E, et al. (2021). "Skeleton-based explainable bodily expressed emotion recognition through gra ph convolutional networks." 2021 16th IEEE International Conference on Automatic Face and Gesture Reco gnition (FG 2021). IEEE.*
12. ^a ^b *Neveu N, et al. (2019). "Parallel general purpose multiobjective optimization framework with applicatio n to electron beam dynamics." Physical Review Accelerators and Beams. **22**(5):054602.*
13. ^a ^b *Ma X, Liu S, Hong W (2022). "Automatic Construction of Parallel Algorithm Portfolios for Multi-objective Optimization." arXiv preprint arXiv:2211.09498.*
14. ^a ^b *Kantour N, Bouroubi S, Chaabane D (2019). "A parallel MOEA with criterion-based selection applied to th e knapsack problem." Applied Soft Computing. **80**:358–373.*
15. ^ *Holland JH (1992). "Genetic algorithms." Scientific American. **267**(1):66–73.*

16. ^Kennedy J, Eberhart R (1995). "Particle swarm optimization." Proceedings of ICNN'95-international conference on neural networks. Vol. 4. IEEE.

17. ^Friedman JH (2001). "Greedy function approximation: a gradient boosting machine." Annals of Statistics:1189–1232.

18. ^Quinlan JR (1986). "Induction of decision trees." Machine Learning. **1**:81–106.

19. a, bHinton G, Vinyals O, Dean J (2015). "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531.

20. a, b, cWang W, Enescu V, Sahli H (2015). "Adaptive real-time emotion recognition from body movements." ACM Transactions on Interactive Intelligent Systems (TiiS). **5**(4):1–21.

21. a, bZacharatos H, Gatzoulis C, Chrysanthou YL (2014). "Automatic emotion recognition based on body movement analysis: a survey." IEEE Computer Graphics and Applications. **34**(6):35–45.

22. a, bSchindler K, Van Gool L, De Gelder B (2008). "Recognizing emotions expressed by body pose: A biologically inspired neural model." Neural Networks. **21**(9):1238–1246.

23. ^Singh A, Mousavi SMH, Gaurav K (2024). "SHS: Scorpion Hunting Strategy Swarm Algorithm." arXiv preprint arXiv:2407.14202.

24. ^Mousavi SMH (2023). "Victoria Amazonica Optimization (VAO): an algorithm inspired by the giant water Lily Plant." arXiv preprint arXiv:2303.08070.

25. ^Mousavi SMH, Charles V, Gherman T (2020). "An evolutionary pentagon support vector finder method." Expert Systems with Applications. **150**:113284.

26. ^Yan S, Xiong Y, Lin D (2018). "Spatial temporal graph convolutional networks for skeleton-based action recognition." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1.

27. ^Valdez SI, et al. (2018). "Modeling hind-limb kinematics using a bio-inspired algorithm with a local search." BioMedical Engineering OnLine. **17**:1–12.

28. ^Al-Qaness MAA, et al. (2022). "The applications of metaheuristics for human activity recognition and fall detection using wearable sensors: A comprehensive analysis." Biosensors. **12**(10):821.

29. ^Hadikhani P, Lai DTC, Ong WH (2023). "A novel skeleton-based human activity discovery using particle swarm optimization with gaussian mutation." IEEE Transactions on Human-Machine Systems. **53**(3):538–548.

30. ^Saini S, et al. (2015). "Markerless human motion tracking using hierarchical multi-swarm cooperative particle swarm optimization." PLoS One. **10**(5):e0127833.

31. ^Zhao X, Liu Y (2008). "Generative tracking of 3D human motion by hierarchical annealed genetic algorithm." Pattern Recognition. **41**(8):2470–2483.
32. ^Zhang Q, Zhang S, Zhou D (2014). "Keyframe extraction from human motion capture data based on a multiple population genetic algorithm." Symmetry. **6**(4):926–937.
33. ^Ruffy F, Chahal K (2019). "The state of knowledge distillation for classification." arXiv preprint arXiv:1912.10850.
34. ^Arablouei R, et al. (2023). "In-situ animal behavior classification using knowledge distillation and fixed-point quantization." Smart Agricultural Technology. **4**:100159.
35. ^Kuldashboy A, et al. (2024). "Efficient image classification through collaborative knowledge distillation: A novel AlexNet modification approach." Heliyon. **10**(14).
36. ^Belinga AG, et al. (2024). "Knowledge Distillation in Image Classification: The Impact of Datasets." Computers. **13**(8):184.
37. ^Xiang L, Ding G, Han J (2020). "Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V 16. Springer International Publishing.
38. ^Holden D, Saito J, Komura T (2016). "A deep learning framework for character motion synthesis and editing." ACM Transactions on Graphics (ToG). **35**(4):1–11.
39. ^Powers DMW (2020). "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." arXiv preprint arXiv:2010.16061.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.