Explaining Human Activity Recognition with SHAP: Validating insights with perturbation and quantitative measures

Felix Tempel [a] [iD],*, Espen Alexander F. Ihlen [b] [iD], Lars Adde [c] [iD], Inga Strümke [a] [iD]

[a] *Faculty of Informatics, Norwegian University of Science and Technology, Trondheim, Norway*
[b] *Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway*
[c] *Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Clinic of Rehabilitation, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway*

ARTICLE INFO

ABSTRACT

In Human Activity Recognition (HAR), understanding the intricacy of body movements within high-risk applications is essential. This study uses SHapley Additive exPlanations (SHAP) to explain the decision-making process of Graph Convolution Networks (GCNs) when classifying activities with skeleton data. We employ SHAP to explain two real-world datasets: one for cerebral palsy (CP) classification and the widely used NTU RGB+D 60 action recognition dataset. To test the explanation, we introduce a novel perturbation approach that modifies the model's edge importance matrix, allowing us to evaluate the impact of specific body key points on prediction outcomes. To assess the fidelity of our explanations, we employ informed perturbation, targeting body key points identified as important by SHAP and comparing them against random perturbation as a control condition. This perturbation enables a judgment on whether the body key points are truly influential or non-influential based on the SHAP values. Results on both datasets show that body key points identified as important through SHAP have the largest influence on the accuracy, specificity, and sensitivity metrics. Our findings highlight that SHAP can provide granular insights into the input feature contribution to the prediction outcome of GCNs in HAR tasks. This demonstrates the potential for more interpretable and trustworthy models in high-stakes applications like healthcare or rehabilitation.

1. Introduction

In recent years, there has been a notable expansion in the field of developing Graph Convolution Networks (GCNs) for Human Activity Recognition (HAR). GCNs have established themselves as the default method for classifying human activities based on skeleton data [1]. Nevertheless, the principal objective of researchers has been to enhance the utilized GCN architectures, particularly on benchmark datasets such as NTU RGB+D 60/120 [2,3], and Kinetics [4]. Because of this, exploring the underlying mechanisms of the developed GCN models has been largely neglected. At the same time, these models' increasing complexity and size have led to a lack of interpretability, effectively turning them into black boxes. Consequently, there is a conspicuous research gap between examining the accuracy and interpretability of GCNs in the HAR domain, limiting the applicability of these methods in high-risk fields such as medical and clinical decision-making [5,6].

To address this issue, methods within the field of Explainable Artificial Intelligence (XAI) could be employed to shed light on the internal mechanisms of GCNs, thereby fostering trust and providing informed decision-making foundations for the users of such systems. This is particularly crucial within the medical domain and other high-risk areas, where understanding the reasoning behind AI-generated outcomes is vital for users to make informed decisions [5,7–9]. Therefore, meaningful explanations provided by – or alongside the predictions of – GCN models must be transparent and comprehensible for the users to accept and safely use such systems within their respective environment [6,10].

Most XAI methods for GCNs rely on gradient-based explanation concepts to interpret their models [11–13]. These methods are currently preferred due to their direct correlation with the skeleton body key points, allowing the identification of the specific body parts important for the model's prediction. However, while such approaches effectively highlight body key points, they fail to capture the nuanced contributions of other essential input features – such as position, velocity, acceleration, and segment orientation – vital for a deeper understanding.

SHapley Additive exPlanations (SHAP) [14] can overcome this fundamental shortcoming of gradient-based methods as it provides a direct relationship between the different input features and model predictions, offering a more comprehensive understanding of how each input feature contributes to the final prediction of the GCNs. This level of detailed insight is not possible with gradient-based methods, which typically provide only a prevailing sense of the general importance of the features by probing the overall gradient at the last network layer. Conversely, SHAP assigns an exact contribution value to each feature for a given prediction, facilitating a more granular and interpretable understanding of how each feature influences the model's output. Although SHAP exhibits substantial potential compared to gradient-based methods, it has yet to be implemented in the context of GCNs within HAR. This gap presents an opportunity to enhance the accuracy and granularity of model interpretation. The comprehensive explanations provided by SHAP have the potential to make it an effective method for interpreting HAR models.

While XAI plays a pivotal role in making machine learning (ML) more understandable, trustworthy, and accessible, commonly used methods often rely heavily on visual, qualitative validation methods, which is also the case for the few endeavors within the HAR domain. In many cases, researchers tend to cherry-pick results that support their stated claim, emphasizing the critical need for comprehensive and standardized evaluation frameworks for XAI methodologies [15,16]. This reliance highlights the need for quantitative metrics to objectively evaluate explanation methods beyond what visual and qualitative assessments alone can deliver [17]. Since single performance metrics such as a high accuracy score do not provide insights into how and why an ML model makes its predictions, the interpretability of a complex ML system cannot be reduced to one simple metric, necessitating a multifaceted perspective instead [17–19].

A prominent method for evaluating XAI methods is the perturbation-based approach [20,21]. In essence, by observing how sensitive a model's predictions are to changes in particular inputs, one can gain insight into which features are most important or influential for decision-making. This method allows for an investigation of how modifications to input features impact the model's predictions, thus enabling an examination of the model's explanation. This technique is powerful due to its intuitive nature and broad applicability across various ML models, rendering it a straightforward and effective technique for XAI testing. Despite its potential, this approach has not yet been explored within the HAR domain, and there is a conspicuous absence of perturbation methods that provide relevant insights. Within HAR, where understanding the model's reasoning process can be crucial for interpreting predictions related to physical motion, the lack of tailored perturbation methods suggests an opportunity for further research and development. Extending perturbation techniques to HAR at the model level, where the model is treated as a dynamic entity rather than a static, unchangeable component, could yield more meaningful insights from explanation methods such as SHAP. Such an approach can enhance the transparency and reliability of the model's decision-making processes.

1.1. Contributions

In this paper, we present several novel contributions addressing the gaps mentioned earlier within current HAR and XAI research:

- First work applying Shapley value-based explanation to a GCN model for human skeleton data on the primary input features, encapsulated within a new algorithmic framework named ShapGCN.
- Introduction of a novel perturbation approach tailored to GCNs within the domain of HAR to test the explanation.
- Apply quantitative metrics to validate the explanations with the perturbation on two real-world HAR datasets.

These contributions collectively advance the understanding of HAR by bridging the gap between model performance and interpretability, offering valuable insights for researchers and practitioners in the respective HAR domain.

2. Related work

2.1. XAI methods

Several XAI approaches have been proposed to explain model predictions within the healthcare and HAR domain since modern ML systems require explainability tools to ensure transparency and trust [5, 6,9,22,23]. Among the most common are SHAP [14] used in our study, saliency maps [24], Local Interpretable Model-agnostic Explanations (LIME) [25], Gradient-weighted Class Activation Mapping (GradCAM) [26], and Concept Activation Vectors (CAVs) [27].

Saliency maps are used to visualize the region of interest of an input attribution based on the model's decision [24]. Saliency maps can be applied to visualize the importance of the individual input channels, which can then be used for dimensionality reduction and improved computational efficiency as done in [28]. However, saliency maps are limited in that they typically only provide a global and sometimes unstable representation of the model's decision process [29].

LIME is a popular model-agnostic approach for interpretability, providing model explanations by use of interpretable surrogate models. This is achieved by perturbing points from the input dataset in a neighborhood in order to generate training data points used to fit an interpretable model in this region. This interpretable model is used as a directly interpretable surrogate of the full model around the explained data point [25]. While LIME is effective across a variety of models, it has not been applied within the HAR domain to the best of our knowledge. This may be due to the challenge of generating meaningful perturbations in skeleton data, as perturbations to this kind of data may not translate to realistic kinematic sound human movement, making the generated explanations less reliable.

GradCAM is a gradient-based method providing visual explanations by highlighting regions of the input based on the gradient activation from a chosen layer from the model [26]. However, GradCAM focuses on the global relevance of the gradients taken from the model's chosen layer without precisely quantifying the individual input feature contributions to the output.

Another XAI approach is the use of CAVs, which explain model predictions in terms of high-level human-understandable concepts [27]. This method can be applied to domains where well-defined concepts exist, such as in medical imaging, where features like the presence of a tumor can be interpreted as a meaningful concept for a model's decision [30,31]. However, applying CAVs to skeleton HAR poses significant challenges. Constructing well-defined and human-intuitive concept functions for human activities in skeleton data, which often involve complex and subtle movements across multiple body key points, would likely be excessively involved and time-consuming. As such, concept-based explanations have only been applied to accelerometer sensor data within the HAR domain [32].

2.2. XAI within HAR

Within the domain of skeleton-based HAR, most XAI approaches are mostly limited to gradient-based explanation techniques like class activation maps (CAM) [33] or GradCAM [26]. In contrast, the widely used SHAP [14], a method based on cooperative game theory that provides consistent and locally accurate feature attributions, has, to the best of our knowledge, only been employed in one work within the HAR domain although only on the secondary input features.

One relevant work that adapts the gradient-based explanation concept is from [11]. The authors use GradCAM to localize and visually

highlight the important body key points within the skeleton data. Additionally, the validity is tested by masking and perturbing parts of the skeleton to calculate metrics, including faithfulness and contrastivity [34]. However, their gradient approach misses the potential to reveal the individual contribution of the input features to the model output, which other model-agnostic methods like SHAP could achieve. Moreover, their perturbation approach reduces the overall accuracy, regardless of the importance of the specific joint being masked, which is the key parameter used in their metrics. An alternative approach to the perturbation, which would not entirely occlude the skeleton body key points from the model, could potentially provide more realistic and kinematic sound explanation results.

Another work using a gradient-based approach is [12], where the authors explore the utilization of CAM mapped to the skeleton sequences. Their results only offer visual representations of the derived explanations without incorporating contextual analysis or evaluative measures. Consequently, in the absence of objective metrics to assess the explanation produced, the overall contribution of this approach remains challenging to compare to other concepts. As with the work from [11], this gradient approach misses the potential to reveal the individual contribution of the features to the final output. This is particularly important since [12] uses multiple branches to train their network on individual input representations derived from the skeleton data fused later in the model. Again, approaches like SHAP could investigate the importance of those individual input features more deeply.

In [13], the authors also use a gradient-based approach to examine how well XAI techniques work for HAR by assessing the metrics Stability and Faithfulness, first proposed in [35,36], on GradCAM and CAM values. Their work aims to simulate realistic changes in input data by applying perturbations to the NTU RGB+D 60 dataset. The EfficientGCN [12] architecture is used as the model architecture for testing the generated GradCAM and CAM values through the used metrics. However, the metrics for the Prediction Gap of important (PGI) and the Prediction Gap of unimportant (PGU) features show the same trend as randomly perturbing body key points. This indicates that the perturbation method applied does not demonstrate any significant advantage over random perturbation, thereby questioning its effectiveness in distinguishing between important and unimportant body key points. Since higher values of the stability metrics from [36], indicate higher instability of the explanation, a logical conclusion would be that no perturbation would result in the lowest value. However, the trend observed in [13] reveals a distinct pattern, showing that the highest values occur with minor perturbation. This contrasts the metric discussed in [36], again causing us to question the perturbation procedure. A different approach for the perturbation, which also encodes the spatial information of the skeleton, could lead to more realistic explanation results. Since this work uses the architecture from [12], the importance of the individual feature contribution also holds up here.

In [37], the authors use SHAP to enhance the model's explanations for early cerebral palsy (CP) screening. Their work employs three branches, which combine infant videos and essential characteristical inputs to classify CP. However, the explainability aspect is restricted to the secondary input features (sex, gestational age, birth weight, and the corrected age), limiting its overall impact on the model's interpretability. This narrow focus on secondary input features restricts the generalizability of the findings and overlooks significant insights that could be derived from the primary features in the video. Expanding the focus to include primary input features could fully leverage SHAP's capabilities to improve the model's transparency and interpretability. Overall, it can be stated that SHAP has some advantages compared to gradient-based explanations in depicting the particular contribution of individual features.

3. ShapGCN

This section provides the essential technical background to follow our contributions. We begin with an overview of SHAP and its application to GCNs in HAR. Next, we present the novel perturbation approach developed to evaluate and validate the explanations generated by our framework ShapGCN. We conclude by describing the metrics used to evaluate the obtained explanations. In Fig. 1, the general workflow of the proposed ShapGCN algorithm is shown.

3.1. SHAP

SHAP is a unified framework for explaining ML model predictions by attributing an importance score to each model input feature [14]. This is done by adapting a solution concept from game theory, the Shapley value [38], to the context of a predictive ML model by formulating the model input features and prediction in the game theoretic setting, thus creating the SHAP value. Representing the importance of a model input feature i, the SHAP value ϕ is computed as follows:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]. \tag{1}$$

To estimate the Shapley value, the model's input features must be systematically removed from the model during prediction. This process requires evaluating the model on all input feature subsets $S \subseteq F$, where F represents the set of all features and results in importance scores for each feature, denoting their impact on the model's prediction. This impact is computed by comparing predictions when the feature is included ($f_{S \cup \{i\}}$) versus excluded (f_S). These differences are computed across all the subsets of $S \subseteq F \setminus \{i\}$, and the final SHAP values are derived as a weighted average of these differences. Since a trained ML model cannot be evaluated with input features missing, SHAP simulates the absence of input features by replacing their values with a mean or uninformative value. The SHAP algorithm generates an attribution score for each model input feature, denoted as $\phi_i(x)$, quantifying its contribution to the model's output. These SHAP values break down the model's prediction into contributions from each feature. The sum of these contributions for all features approximates the difference between the reference value ϕ_0 and the model's prediction $f(x)$ for a given input x, which can be written as:

$$f(x) \approx \sum_{i=1}^{N} \phi_i(x) + \phi_0, \tag{2}$$

where each $\phi_i(x)$ quantifies how much the specific feature drives the model prediction away from the reference value ϕ_0 for the data instance x. Here, the reference value, ϕ_0, is the expected output of the model, calculated as the average over the entire input distribution:

$$\phi_0 = \mathbb{E}_x[f(x)]. \tag{3}$$

In our approach, the input features are computed from the skeleton data, divided into four distinct groups, namely: **(1) Joint** (J), **(2) Velocity** (V), **(3) Bone** (B), and **(4) Acceleration** (A) features, as defined in [39,40]. **(1)** J: These are the 3D or 2D coordinates of each body key point, capturing the spatial layout of the skeleton. **(2)** V: Velocity measures how fast each body key point moves between frames, calculated as the difference in position over time, to capture the dynamics of the motion. **(3)** B: These features represent the relative positions between connected body key points, encoding the length and angular orientation to capture the spatial relationships during motion. **(4)** A: Acceleration captures the rate of change of velocity, highlighting sudden changes in motion like starts, stops, or rapid shifts in direction. Aggregating the SHAP values of each of these feature groups, we calculate the total SHAP value as $\phi = \phi_J + \phi_V + \phi_B + \phi_A$. The data used to calculate ϕ_0 in our approach is randomly sampled from $\mathcal{D}_{train}$ with $n = 100$.
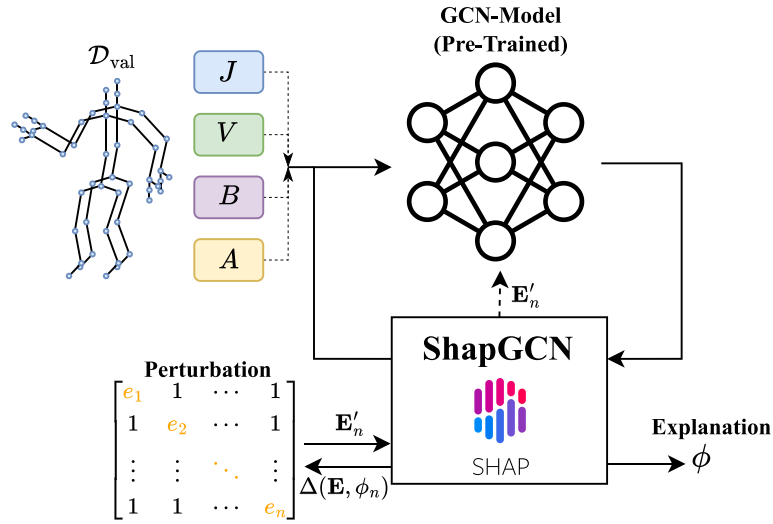
Fig. 1. Overview of the proposed XAI pipeline for HAR-ShapGCN. The features computed from the skeleton data are fed into a pre-trained model, followed by the explanation through ShapGCN to determine the feature attributions on the primary input features. The explanation ϕ is validated with the proposed quantitative metrics $\mathcal{M}$ and interpreted visually. Based on ϕ, the perturbed edge matrix $\mathbf{E}'_n$ is constructed, and the model architecture is updated. Afterward, $\mathcal{D}_{val}$ is inferred on the updated model architecture.

The computation of exact SHAP values has a time complexity of $\mathcal{O}(2^n)$, where n is the number of input features, making it computationally expensive for high-dimensional datasets. However, several approximation methods have been developed to make SHAP more applicable for real-world usage. These include algorithms such as "Kernel SHAP" for model-agnostic interpretation, "Tree SHAP" for tree-based models, and "Deep SHAP", which utilizes the "Deep Lift" algorithm for deep learning models [14]. These approximations significantly reduce the complexity while maintaining the interpretability and consistency of SHAP values. Beyond explaining individual predictions or local explanations, SHAP values can be aggregated across multiple data points to get global explanations [6]. By summing the SHAP values for each feature across the dataset, we obtain a global understanding of the most influential features overall. This provides insight into how the model makes decisions on a broader scale, allowing us to identify the most critical features driving predictions across the dataset.

3.2. Perturbation for GCNs

In XAI, perturbation techniques are commonly employed to perform sensitivity analyses of a model on its input features by perturbing portions of the input data and observing the resulting impact on the model's predictions [41]. In our novel perturbation approach, we extend this technique by perturbing the parts of the model architecture directly correlating with the skeleton body key points instead of perturbing the input features. In the following, we give a concise overview of the GCN architecture utilized in skeleton HAR before elaborating further on our methodology.

GCNs extend the concept of convolutional neural networks to graph data, making them suitable for skeleton data. In a GCN, each node aggregates features from its neighboring nodes, effectively capturing local and spatial relationships inherent to human movements. The core operation involves multiplying the node feature matrix by a normalized adjacency matrix $\mathbf{A}$, identity matrix $\mathbf{I}$, and a learnable edge importance weight matrix $\mathbf{E}$ to capture the spatial dependencies of the skeleton. Depending on the chosen partition strategy $s = \{\text{spatial}, \text{distance}, \text{uniform}\}$, $\mathbf{A}$ is parsed into $\mathbf{A}_j$, where $\mathbf{A} + \mathbf{I} = \sum_j \mathbf{A}_j$. Based on s, the maximum graph distance j defines the output as follows [42]:

$$\mathbf{f}_{out} = \sum_j \mathbf{W}_j \mathbf{f}_{in} (\mathbf{\Lambda}_j^{-\frac{1}{2}} \mathbf{A}_j \odot \mathbf{E}_j \mathbf{\Lambda}_j^{-\frac{1}{2}}). \tag{4}$$

Here, $\mathbf{f}_{out}$ denotes the output features and $\mathbf{f}_{in}$ the input features respectively. $\mathbf{W}_j$, $\mathbf{\Lambda}_j$, and $\mathbf{A}_j$ contribute to the convolutional operation,

normalization of $\mathbf{A}$, and the adjacency matrix itself with graph distance j. Matrix $\mathbf{E}$ has the same shape as $\mathbf{A}$, and its elements e are the learnable parameters along the diagonal, defined as

$$\mathbf{E} = \begin{bmatrix} e_1 & 1 & \cdots & 1 \\ 1 & e_2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & e_n \end{bmatrix}.$$

The diagonal elements e_n are emphasized in orange, indicating their learnable nature during training, while the off-diagonal elements are fixed to 1.

Given that this matrix encodes spatial information, it represents the significance of edges in different architecture regions to the model. This opens the possibility of applying an informed perturbation to the model by masking the respective edges corresponding to a skeletal body key point. The individual importance of those edges at a specific time is obtained through the SHAP values; see Fig. 1 for an overview of the overall procedure. In our experimental setting, we assigned a value of $e_n = 1 \times 10^{-5}$ to the respective edges, simulating a masking effect. Alternatively, we applied a percentage-based perturbation to introduce variation, modifying the original edge values as $e_n = e_n \times \epsilon$, where ϵ represents the percentage adjustment.

3.3. XAI metrics

The validity of the explanations obtained from SHAP is tested through the presented perturbation procedure. To assess it, the PGI and PGU metrics are utilized [35]. Additionally, we provide the specificity, sensitivity, and class accuracy metrics to further examine the influences of the perturbation.

3.3.1. PGI & PGU

First introduced in [43] and later formalized in [35], these metrics measure the difference between the original prediction and the prediction after an applied perturbation. Based on the results from the explainability method, we determine the important and unimportant values, which are, in our case, the body key points with the highest and lowest SHAP values for the respective class. By perturbing the top-K important body key points for PGI and ¬ top-K body key points for PGU, the prediction gap between the unperturbed prediction $\hat{y} = f(x)$ and the perturbed prediction $f(x')$ can then measure the prediction difference.

$$\text{PGI}(x, f, \mathbf{e}_x, k) = \mathbb{E}_{x' \sim \text{perturb}(x, \mathbf{e}_x, \text{top-}K)} \left[|\hat{y} - f(x')| \right], \tag{5}$$

(a) NTU RGB+D skeleton (b) CP skeleton (c) CP recording environment
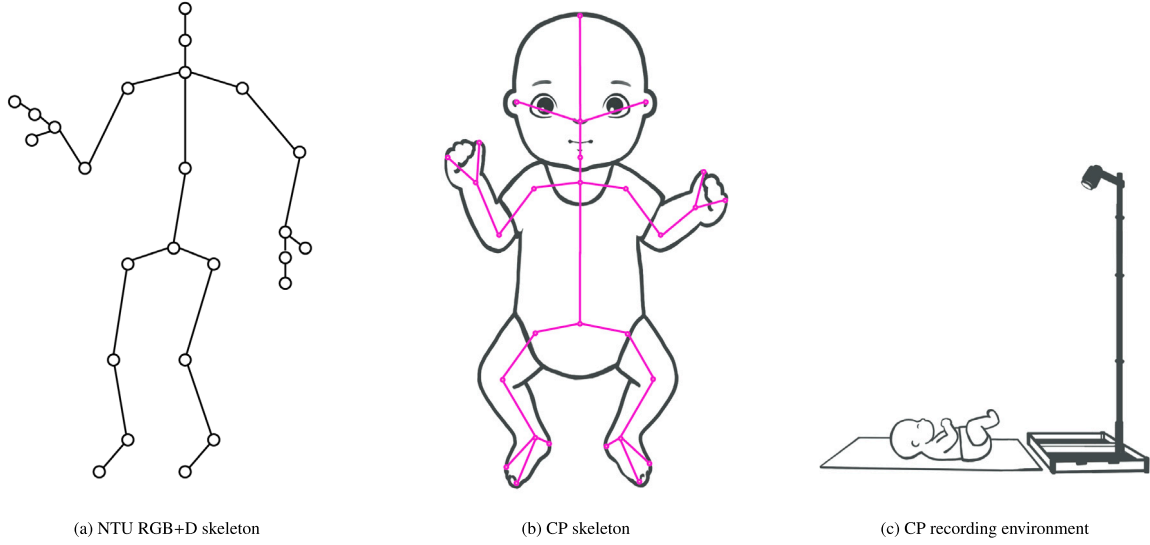
Fig. 2. Skeletons of the NTU RGB+D and CP dataset and the recording environment. (a) The NTU RGB+D skeleton with $n = 19$ body key points. (b) The CP skeleton with $n = 29$ body key points adds finer detail in the toes and the head region. (c) The recording environment for the CP dataset.

$$\text{PGU}(x, f, \mathbf{e}_x, k) = \mathbb{E}_{x' \sim \text{perturb}(x, \mathbf{e}_x, \neg\text{top-}K)} \left[|\hat{y} - f(x')| \right]. \qquad (6)$$

Higher PGI values imply that perturbing important body key points results in performance degradation, implying that these are important for the model prediction. PGU, on the contrary, captures the loss in performance when unimportant body key points are perturbed. Lower values indicate that the perturbed body key points are less influential in determining the outcome, thereby verifying the unimportance of these body key points. In summary, faithful importance attribution scores result in high values of PGI and low values of PGU.

4. Experiments

In this section, we conduct experiments on two real-world datasets to show the effectiveness of our proposed perturbation procedure and provide the SHAP explanations. The arrangement of the experiment section is as follows. First, we concisely present the used datasets and the implementation details. Afterward, we will explain the two datasets and the importance of the features obtained with SHAP. Furthermore, we describe the behavior of the metrics when informed and random perturbation is applied.

4.1. Datasets

The NTU RGB+D 60 is a 3D human activity dataset for action recognition [2]. It consists of 56,880 videos, holding 60 action classes. Our experiments use the defined subset cross-view (X-View) from [2]. This subset partitions the dataset according to camera perspectives. Here, camera views two and three are used as the training data $\mathcal{D}_{train}$, while camera view one is dedicated to the validation data $\mathcal{D}_{val}$. The data is preprocessed by parallelizing the shoulder and hip joints, shifting the skeleton so that the spine joint is centered at the origin, scaling the lengths of the limbs, and normalizing the reference viewpoint. Missing skeletons in the dataset are listed and provided by the authors from [2]. For our experiments, only actions displaying one person are enclosed. The reader is referred to [39] for more details and the preprocessing steps for this dataset.

The second dataset comprises 557 videos of infants with medical risk factors for CP, recorded between 2001 and 2018 across four countries: the United States ($n = 248$), Norway ($n = 190$), Belgium ($n = 37$), and India ($n = 82$). These recordings, including infant spontaneous

movements in a supine position at 9 to 18 weeks, are standardized according to [44,45] and assessed by two medical experts. For each frame within a video, the positions of 29 body key points of the infant in motion are captured as X and Y coordinates, resulting in the infant skeleton. The video sequences are processed to improve their quality and consistency. This involves resampling to 30 Hz, removing noise and inconsistencies in the data using a median filter, and adjusting the data to a common scale based on the trunk length as the reference point. The normalization is done by doubling the infant's trunk length and centering on the median mid-pelvis position. Finally, the processed data is divided into short windows, each lasting 5 s with a 2.5-s overlap. The reader is referred to [46] for more details on the preprocessing and the dataset itself.

Each dataset requires a different dimensionality for $\mathbf{E}$ based on the number of body key points, ($n = 29$) for the CP dataset and ($n = 25$) for the NTU RGB+D dataset. Two example skeletons illustrating the edge connections for both datasets are presented in Fig. 2, alongside the depiction of the recording environment used for the CP dataset.

4.2. Implementation

We use the best-performing model architecture for the NTU RGB+D 60 X-View dataset experiments from [39]. Furthermore, we use the best-performing architecture from [40] for the CP dataset. For the explanation, the DeepExplainer implementation from the SHAP library [14] is adopted to calculate the SHAP values. We define a random reference dataset with 100 instances possibly chosen from every class for the NTU RGB+D 60 dataset. For the CP dataset, we generate a reference dataset consisting of 100 randomly chosen window samples from both classes. As the GPU can only be used with a limited number of background samples, the experiments are conducted using subsamples ($n = 20$) of the background, with the resulting SHAP values later aggregated. The pseudocode for the implementation is presented in Algorithm 1, where the input ϕ_0 is calculated by SHAP and represents the average model prediction across the provided reference dataset, see Eq. (3). The experiments are performed on a single NVIDIA-V100 with 32 GB GPU RAM on the PyTorch framework (version 2.3.1) [47] with the global seed set to 1234. In our setup, computing SHAP values for the CP dataset (two classes) takes 16 h and 15 min. For the NTU RGB+D dataset (three classes), the SHAP value calculation requires 23 h and 41 min.

Algorithm 1: ShapGCN

Input: Data: $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}$; Model; ϕ_0
Output: $\phi, \mathcal{M}$
$\mathcal{M} \leftarrow \{\text{PGI, PGU, Specificity, Sensitivity, Accuracy}\}$

Initialize trained model
$\phi_0 = \mathbb{E}\left[\text{Model}(\mathcal{D}_{\text{train}})\right]$
if D is NTU RGB+D **then**
   $p \leftarrow \epsilon$
else
   $p \leftarrow 1 \times 10^{-5}$
$\phi \leftarrow SHAP(\mathcal{D}_{\text{val}}, \phi_0)$
for $k \leftarrow 1$ **to** *10* **do**
   // Generate **E** for top-k most/least influential
     body key points
   $\mathbf{E}_{\max \phi, k} \leftarrow p$
   $\mathbf{E}_{\min \phi, k} \leftarrow p$
   for d in $\mathcal{D}_{val}$ **do**
     $\mathcal{M}_{\max} \leftarrow d, \mathbf{E}_{\max \phi, k}$
     $\mathcal{M}_{\min} \leftarrow d, \mathbf{E}_{\min \phi, k}$

4.3. Averaging technique

Given the multidimensional SHAP output, which matches the model input's dimensions, e.g., NTU RGB+D: $\mathbb{R}^{4 \times 6 \times 300 \times 25 \times 2 \times 60}$, CP: $\mathbb{R}^{4 \times 4 \times 225 \times 29 \times 1 \times 2}$, the SHAP values are averaged to generate interpretable explanations. The indices represent the following values:

$$[\text{Input feature}(f), \text{Feature dimension}(k), \text{Time frames}(t),$$
$$\text{body key points}(n), \#\text{ persons}(m), \text{SHAP values}(s)]$$

For the NTU RGB+D subset X-View, we compute the SHAP values for three randomly chosen classes and take the average over each. To address padding, where classes repeat the same scene after the initial action sequence to maintain consistent input sizes, we reduce the temporal dimension (t) by half. The input features (f) and feature dimensions (k) are merged into a single index. Additionally, we remove the second skeleton (m), focusing solely on the first skeleton by squeezing the dimensionality. Thus, the new dimensionality for one skeleton sequence's SHAP values is $\mathbb{R}^{24 \times 150 \times 25 \times 60}$.

In the CP dataset, unlike NTU RGB+D, there is no second skeleton, so the index (m) is directly squeezed. As for the NTU RGB+D dataset, the input features (f) and feature dimensions (k) are combined into a single index. The resulting dimension for one window is $\mathbb{R}^{16 \times 225 \times 29 \times 2}$.

4.4. CP dataset

4.4.1. Global feature importance

To illustrate the global feature importance associated with CP predictions, we use the SHAP *beeswarm* visualization [14]. With this representation, the contributions of the specific input feature joint, velocity, bone, and acceleration (J, V, B, A) for the validation dataset $\mathcal{D}_{val}$ are highlighted. For this analysis, SHAP values are aggregated and averaged over the duration of each individual window. The resulting plot is presented in Fig. 3.

The y-axis of each subplot lists the nine most important body key points for the respective feature in descending order of importance. The importance is measured based on the average magnitude of the SHAP values across all windows. The 10th element represents a cumulative score that combines the SHAP values of all remaining body key points, capturing their collective contribution to the model output. The x-axis corresponds to the SHAP value, which reflects the influence of each feature on the model's output. Positive values indicate a push towards

the CP class (class 1), and negative values indicate a push towards the No-CP class (class 0). The colored dots represent the actual values of the input features for each joint, where the color gradient provides a visual cue about the magnitude of the feature values, ranging from low to high across the body key points.

For features J and B, a strong and consistent correlation is observed between the actual input feature values and their corresponding SHAP values across all windows. This implies that these features have a stable relationship with the model's prediction, where higher or lower values of J and B align with the likelihood of a particular class (CP or No-CP). This stability suggests that these features have a relatively straightforward and interpretable influence on the model's output across different windows from infants in the same class.

In contrast, the features A and V exhibit a more complex relationship. The SHAP values for these features do not consistently align with the input feature values across all windows. This lack of a clear, direct correlation is due to how these features are computed. Acceleration and velocity have both positive and negative values, which introduces more variability in their contribution to the predictions. The distribution of the input feature values across the SHAP values for A and V suggests their influence on the model's decision is more context-dependent. Each individual window varies depending on the specific kinematics at this point. Because of this variability, the features A and V do not lend themselves to a simple, global interpretation like J and B. The contribution of A and V to the model's predictions may shift depending on the local patterns in the data, and a detailed, window-by-window exploration is necessary to fully understand how they impact the final predictions. Thus, while J and B provide consistent and interpretable information, A and V demand a more granular investigation to discern their roles in the model's decision-making process.

4.4.2. Local feature importance

The SHAP values for one randomly chosen infant with CP are aggregated and directly displayed on the respective body key points in the skeleton data to visualize the local explanation.

Fig. 4 illustrates the skeletal representation of the infant in five windows. In this visualization, the aggregated SHAP values from all features are mapped to the body key points at each time window. The size of each dot corresponds to the magnitude of the SHAP values, while the color indicates the direction of influence exerted by each joint. Specifically, red signifies a pull towards a prediction of CP, whereas blue indicates a pull towards No-CP. Body key points with SHAP values near zero are represented as smaller dots, highlighting their minimal influence on the prediction at the respective window. The SHAP values indicate that the model output is most influenced by the infant's four limbs freely moving against gravity in the supine position. The observation of smaller SHAP values in the head, trunk, and pelvis region is plausible because the infant is lying on its back during the video recording, resulting in minimal spatial movement from these body regions.

The local explanation of the individual features (J, V, B, and A) for the same infant are provided in Appendix A.

4.4.3. Perturbation

To assess the validity of the SHAP explanation, we perform a series of experiments with our proposed edge matrix perturbation. In these experiments, the features derived from the dimension (k), namely x and y orientations, are combined since these features are directly coupled. The mean of each body joint is calculated over the (k) time frames of each window. The obtained mean values are then aggregated on a joint-by-joint basis. Finally, the resulting values are sorted in descending order according to their corresponding SHAP values for each class. Body key points identified as influential and non-influential for CP prediction are selected for modification, and these body key points are perturbed within the edge matrix **E** at their respective position e_n.
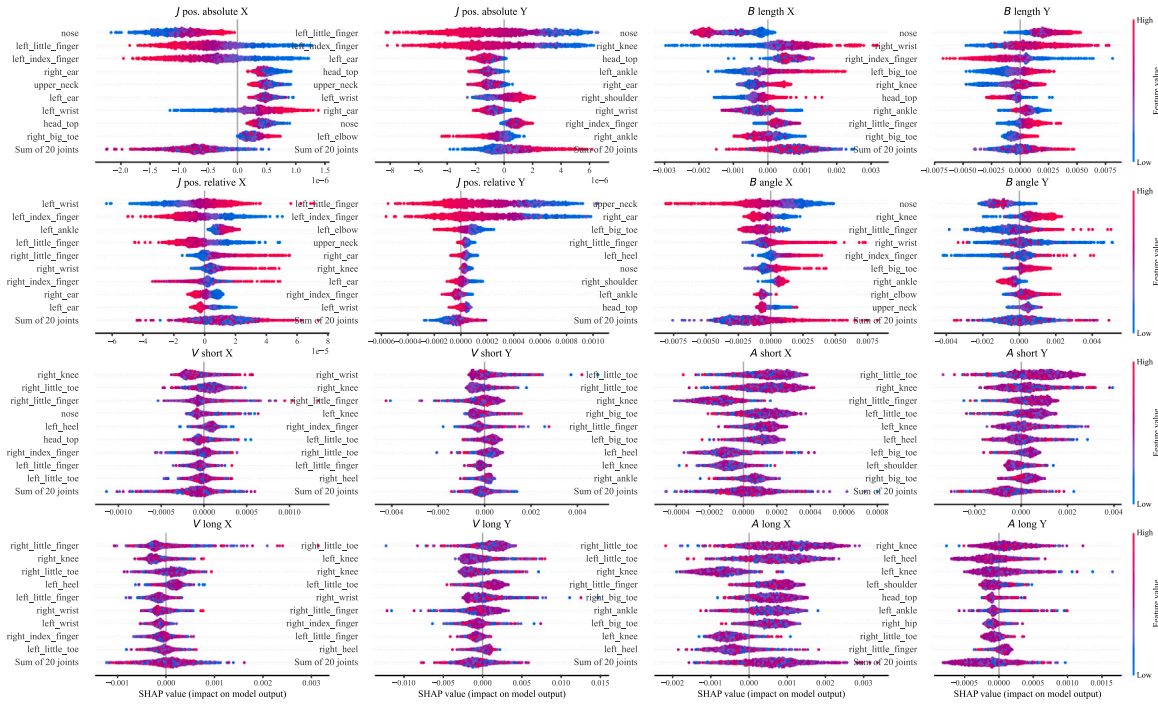
Fig. 3. SHAP values of the primary input features. The figures show the SHAP value on the *x*-axis and the corresponding body key point on the *y*-axis, sorted in descending order of importance. The color of the dots indicates the feature value, with red corresponding to a high value, while blue is linked to a low feature value. The values are mean aggregated over the individual window time frames (*t*) and belong to class 1 (CP).
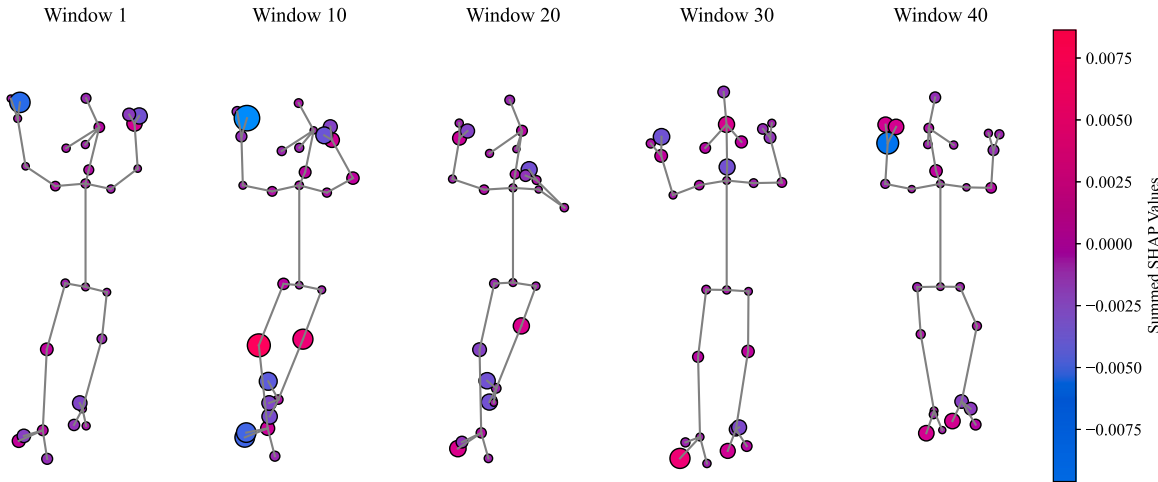
Fig. 4. Local SHAP values of the combined input features. The features (*J*, *V*, *B*, and *A*) are added together and shown for five different windows of one infant with CP. The size and color of the dots indicate the value of the summed SHAP values. SHAP assigns the importance to the knees in window 10 and to the toes within windows 20, 30, and 40. Whereas a negative importance to CP is assigned to the hand region in windows 1 and 10.

Table 1 shows the perturbation results for the $\mathcal{D}_{val}$ CP dataset. The table summarizes the performance metrics, comparing the effects of perturbing the top-K body key points identified as important to the two possible classification outcomes (CP vs. No-CP). This table provides insights into how an informed perturbation – particularly focusing on perturbing the top-K most influential body key points versus the least influential body key points – affects model performance regarding sensitivity, specificity, PGI, and PGU. The first row is the baseline performance in which none of the body key points are perturbed, where the model achieves a sensitivity of 0.944 and a specificity of 0.873. The two major columns describe the perturbation on the top-K body key points, either important or unimportant, for classifying CP or No-CP. The arrows indicate the importance of those body key points towards one or the other class.

When focusing on the values in the first major column, in which the SHAP values have a low impact on CP and a high impact on No-CP, the $PGI_{spec.}$ remains relatively stable. Even when more influential body key points are perturbed, the model maintains its ability to correctly identify No-CP cases, reflecting resilience in how the No-CP class is handled. This implies that the features critical for No-CP classification are either more evenly distributed across multiple body key points or less sensitive to the perturbations of the top-ranked body key points, allowing the model to recover from modifications to important body key points.

In the second major column, it can be observed that $PGI_{sens.}$ generally increases with the number of body key points perturbed, peaking at ten perturbed body key points ($PGI_{sens.}$ = 0.572). This indicates that perturbing the ten most important body key points leads to the

Table 1

Performance metrics on the CP dataset under various perturbation scenarios. The table compares the effects of perturbing the most important body key points versus the least important ones, as determined by the respective SHAP value for this specific window. Different numbers of top-K influential body key points are perturbed, with the perturbation value for the edge matrix set to $1e-5$, simulating a masking effect. Due to the class imbalance in the dataset – where healthy infants outnumber those with CP – both specificity and sensitivity metrics are reported for each perturbation scenario. The PGI and PGU values are reported based on important and unimportant body key points for the respective setting gains/losses under perturbation—higher values for PGI imply greater explanation faithfulness while it is the opposite for PGU.

| # top-K | SHAP value CP ($\downarrow$) No-CP ($\uparrow$) | | | | SHAP value CP ($\uparrow$) No-CP ($\downarrow$) | | | |
body key points	$PGI_{spec.}(\uparrow)$	$PGU_{sens.}(\downarrow)$	Specificity	Sensitivity	$PGI_{sens.}(\uparrow)$	$PGU_{spec.}(\downarrow)$	Specificity	Sensitivity
No perturb.	–	–	0.873	0.944	–	–	0.873	0.944
1	0.108	0.234	0.765	0.710	0.306	0.080	0.793	0.638
2	0.137	**0.215**	0.736	0.729	0.326	0.094	0.779	0.618
3	0.152	0.223	0.721	0.721	0.362	0.104	0.769	0.582
4	**0.157**	0.239	0.716	0.705	0.395	0.113	0.760	0.549
5	0.154	0.255	0.719	0.689	0.429	0.108	0.765	0.515
6	0.143	0.257	0.730	0.687	0.440	0.100	0.773	0.504
7	0.137	0.263	0.736	0.681	0.477	0.099	0.774	0.467
8	0.131	0.276	0.742	0.668	0.505	0.084	0.789	0.439
9	0.126	0.298	0.747	0.646	0.532	0.069	0.804	0.412
10	0.120	0.309	0.753	0.635	**0.572**	**0.060**	0.813	0.372

most significant drop in performance, supporting the claim that these are crucial for correctly identifying CP. Overall, the specificity for the most important body key points is fairly stable, ranging from 0.96 to 0.84, but the sensitivity decreases with each added joint, from 0.638 to 0.372. This confirms that the SHAP values correctly identify the body key points essential for classifying CP, as the model's performance on sensitivity is directly impacted when these body key points are perturbed. Furthermore, the stability in specificity, despite the sensitivity drop, indicates that the model's ability to classify non-CP cases correctly relies less on the same set of body key points.

4.4.4. Random perturbation

Random perturbation is a control to evaluate the perturbations in the informed setting. This method randomly perturbs body key points without relying on SHAP values. The goal is to determine whether targeting specific body key points based on their importance (as in the SHAP-guided, informed perturbation) leads to significantly different outcomes than randomly perturbing body key points.

From Table 2, we observe that random perturbations have a milder effect on both specificity and sensitivity when compared to informed perturbations from Table 1. For instance, the sensitivity under random perturbation decreases to 0.612 at most, whereas in Table 1, informed perturbation for the CP class reduces the sensitivity to 0.372. This indicates that perturbing the most important body key points, identified through the SHAP values, has a far more detrimental impact on the model's performance. Specifically, informed perturbation leads to more significant drops in sensitivity, particularly when perturbing the top 10 most important body key points, indicating that these body key points hold critical information for an accurate prediction of CP.

This supports the validity of using SHAP-based explanations, as perturbing the most important body key points meaningfully degrades the model's predictive performance. Hence, informed perturbation provides a more faithful assessment of the model's reliance on the specific features at a specific body key point.

4.5. NTU RGB+D

4.5.1. Global feature importances

We show the global feature importance for class 6 in the NTU RGB+D X-View subset, corresponding to the action *pick up*. As for the CP dataset, we utilize the SHAP *beeswarm* plot to show the feature pulls across the body key points for the individual input features. In Fig. 5, the SHAP values on the individual input features are shown, where we have the additional coordinate orientation Z.

For the input features J and B, the features show a separation and noticeable trend across the body key points. This trend, however, is not as distinctive as for the CP dataset, where the windows show a clear correlation between the SHAP and feature values.

Table 2

Performance metrics on the CP dataset with random perturbation. The table displays the effects of randomly perturbing a specified number of body key points on specificity and sensitivity.

| # body key points | Random perturbation | |
	Specificity	Sensitivity
No perturb.	0.873	0.944
1	0.792	0.728
2	0.789	0.711
3	0.790	0.708
4	0.789	0.691
5	0.789	0.696
6	0.790	0.680
7	0.793	0.664
8	0.790	0.641
9	0.800	0.648
10	0.804	0.612

Furthermore, it can be observed that the feature and SHAP values are not as distinctly correlated for the input features V and A. For example, the SHAP values for the *A short* along the X, Y, and Z axes show a distribution where the contribution of individual key points is spread evenly across the shown body key points, and no dominant body key point clearly stands out. Similarly, for the input features V, different axes exhibit variability in the SHAP and feature values. The SHAP values do not correlate strongly and consistently with the actual input values across the individual time points. This highlights the more complex nature of how A and V contribute to the model's predictions for this dataset, suggesting that these features likely capture more nuanced kinematics that are harder to interpret globally without considering the specific context and time point where the action occurs.

This is a notable difference from the CP dataset, where specific key points exhibited a more pronounced relationship with the SHAP values of these features, and is due to the long time frame (t) on which the SHAP values are computed for the NTU RGB+D dataset. However, it can be observed that the body joints of the right and left hand appear to be the most important feature in almost all feature dimensions. This is plausible since the action *pick up* is mainly executed by those limbs.

4.5.2. Local feature importance

The SHAP values for a randomly selected individual from class 6 (*pick up*) are aggregated and visually represented on the corresponding body key points within the skeleton data. This allows for a localized interpretation similar to that of the CP dataset. Fig. 6 shows this skeletal representation of the person performing the action across six Windows.

This visualization maps the aggregated SHAP values derived from all features onto the body key points for each time window. The size of each dot reflects the magnitude of the SHAP values, while the color indicates the direction of influence of each joint. Specifically, red dots
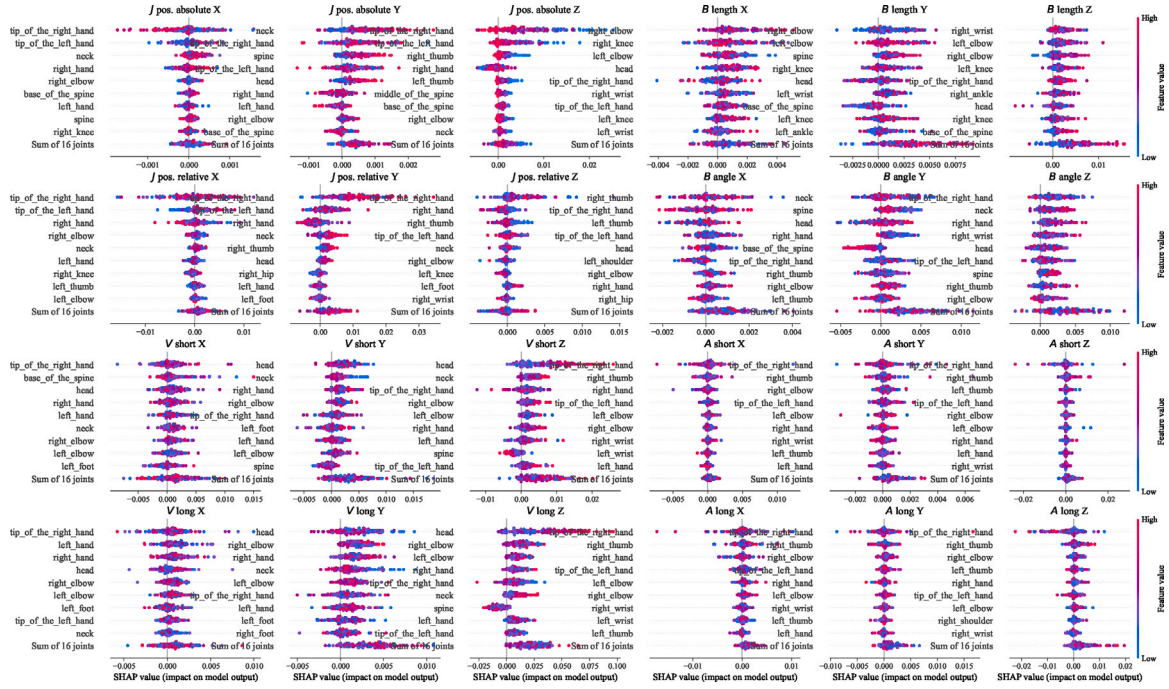
Fig. 5. SHAP values of the primary input features. The figures show the SHAP value on the *x*-axis and the corresponding body key point on the *y*-axis, sorted in descending order of importance. The color of the dots indicates the feature value, with red corresponding to a high value, while blue is linked to a low feature value. The values are mean aggregated over the individuals performing the action and belong to class 6 (*pick up*).
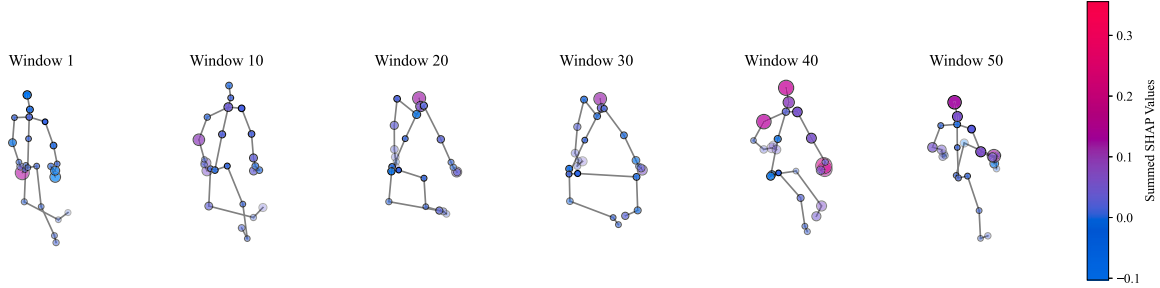
Fig. 6. Local SHAP values of the combined input features for class 6 (*pick up*). The features ($J, V, B,$ and A) are added together and shown for five different windows. The size and color of the dots indicate the value of the summed SHAP values. SHAP assigns the highest importance for the action to the hand and head region in windows 40 and 50, while other body key points remain neutral (with SHAP values near zero) throughout the remaining windows.

signify a pull towards class 6, whereas blue dots indicate a pull away from this class. Body key points associated with SHAP values close to zero are represented as smaller dots, thereby emphasizing their minimal impact on the prediction for the respective window. Notably, the SHAP values suggest that the model's output for this individual is predominantly influenced by the hands engaged in the action during window 40. The observed smaller SHAP values in Windows 1 and 10 are reasonable, as the action has not yet commenced during these intervals.

4.5.3. Perturbation

As with the CP dataset, we assess the validity of the obtained SHAP explanations by perturbing the influential and non-influential body key points. Unlike the CP dataset, where action sequences are segmented into 5-s windows, the NTU RGB+D dataset stores each action sequence as a single continuous time frame. Given the dataset's sensitivity to edge masking, where fixing edges to near-zero values greatly impacts the model prediction, we do the perturbations by adjusting edge weights with a percentage-based threshold.

Table 3 reports the results for perturbing the edges with a $\epsilon = 35\%$ change. In Appendix B, we show the perturbation experiment when

the edges are perturbed with threshold changes from $\epsilon \in [0.05, 0.45]$. This systematic manipulation of body key points allows us to investigate how perturbations influence the model prediction based on the importance of features as determined by SHAP.

Notably, when body key points associated with positive SHAP values are perturbed, a sharp decline in classification accuracy is observed across all the shown actions. Conversely, although perturbing joints classified as non-influential also result in a decline in accuracy, this decline is less severe than that observed with influential joints. This trend is plausible and suggests that while the accuracy decreases for unimportant joints, the impact is moderated relative to the notable effect of perturbing important joints. Furthermore, across all three classes, the PGI value is constantly higher than the PGU value, which implies that the body key points determined by SHAP are truly important or unimportant. These findings highlight the efficacy of SHAP in identifying critical body key points in this dataset.

4.5.4. Random perturbation

We run the same experiment by randomly perturbing the edges to further validate our proposed perturbation approach. Table 4 shows the results of perturbing between one to ten body key points for classes 6,

Table 3

Performance metrics on the NTU RGB+D dataset under informed perturbation scenarios. The accuracy progressively decreases for each class as the top-K body points are perturbed. We show the impact of perturbing important (imp.) vs. unimportant (¬ imp.) body key points with the threshold set to $\epsilon = 35\%$ on classification accuracy for three action classes and report the PGI and PGU metrics. Accuracy drops significantly in all three classes when perturbing influential points, confirming SHAP's ability to correctly identify crucial body key points.

# top-K body key points	Class 6 (*pick up*)				Class 11 (*reading*)				Class 16 (*put on a shoe*)			
	Acc. imp.	Acc. ¬ imp.	PGI (↑)	PGU (↓)	Acc. imp.	Acc. ¬ imp.	PGI (↑)	PGU (↓)	Acc. imp.	Acc. ¬ imp.	PGI (↑)	PGU (↓)
No perturb.	0.971	0.971	–	–	0.797	0.797	–	–	0.845	0.845	–	–
1	0.883	0.940	0.088	**0.031**	0.597	0.733	0.200	**0.064**	0.533	0.784	0.312	**0.061**
2	0.851	0.918	0.120	0.053	0.568	0.667	0.229	0.130	0.444	0.727	0.401	0.118
3	0.759	0.883	0.212	0.088	0.505	0.632	0.292	0.165	0.406	0.686	0.439	0.159
4	0.617	0.839	0.354	0.132	0.470	0.568	0.327	0.229	0.397	0.629	**0.448**	0.216
5	0.560	0.807	0.411	0.164	0.371	0.486	0.426	0.311	0.400	0.575	0.445	0.270
6	0.541	0.782	**0.430**	0.189	0.305	0.397	0.492	0.400	0.429	0.540	0.416	0.305
7	0.566	0.772	0.405	0.199	0.232	0.368	0.565	0.429	0.406	0.521	0.439	0.324
8	0.573	0.763	0.398	0.208	0.187	0.337	0.610	0.460	0.422	0.511	0.423	0.334
9	0.557	0.753	0.414	0.218	0.117	0.321	0.680	0.476	0.432	0.505	0.413	0.340
10	0.544	0.734	0.427	0.237	0.086	0.314	**0.711**	0.483	0.416	0.502	0.429	0.343

Table 4

Performance metrics on the NTU RGB+D dataset with random perturbation. The effects of randomly perturbing a specified number of body key points on the class accuracy of classes 6 (*pick up*), 11 (*reading*), and 16 (*put on a shoe*).

# body key points	Class accuracy		
	Class 6	Class 11	Class 16
No perturb.	0.971	0.797	0.845
1	0.956	0.778	0.790
2	0.937	0.702	0.775
3	0.886	0.616	0.724
4	0.769	0.524	0.679
5	0.744	0.483	0.568
6	0.639	0.321	0.508
7	0.541	0.283	0.489
8	0.459	0.244	0.406
9	0.361	0.159	0.356
10	0.294	0.089	0.283

11, and 16 randomly.

From Table 4, it is observable that random perturbations induce a more gradual decline in accuracy compared to informed perturbations from Table 3. This can be attributed to the fact that, in the random case, both influential and non-influential body key points have equal chances of being perturbed, leading to a diluted effect. For instance, with the SHAP-guided perturbations, accuracy decreases sharply when critical body key points are disrupted, as seen in Class 6 (*pick up*), where perturbing five influential points drops accuracy from 0.971 to 0.560. In contrast, random perturbations reduce accuracy more moderately, with only a 0.227 drop in accuracy after perturbing five body key points.

However, random perturbations eventually also lead to substantial accuracy degradation, as seen when perturbing a larger number of body key points (e.g., 10), where the model's accuracy diminishes almost to the same extent as with the informed perturbation scenario. This behavior illustrates that while random perturbations are initially less disruptive, the model becomes increasingly unstable as more body key points are altered.

5. Discussion

The results of this study constitute significant progress towards understanding how SHAP can be used to explain GCNs in HAR. With the novel perturbation strategy, we can selectively alter the most important body key points identified through SHAP values. We do this by probing the model's reliance on the features associated with the highest SHAP value in the two datasets, i.e., perturbing the most important body key points for the model. This leads to a sharp decline in sensitivity for the CP dataset, which measures the ability of the model to correctly identify positive cases among all actual positives. For the NTU RGB+D

dataset, we see a decrease in class accuracy for the individual action. This serves as a strong verification of SHAP's role in offering faithful explanations, as the features it deemed most important are indeed those that, when perturbed, most affect the model's prediction outcome. As a baseline, we show that random perturbation has a significantly lesser effect on the model's performance: When body key points are perturbed randomly rather than based on SHAP values, the model's sensitivity and specificity decline more moderately. This contrast between random and SHAP-guided perturbations indicates that SHAP indeed effectively identifies important features.

Trust from the users in ML models is important, especially in medical settings [5,6,9]. Our findings show that HAR models in clinical applications need not be considered "black boxes". With ShapGCN, we lay the foundation for clinicians to understand why a model makes its specific decisions. This way, clinicians can gain trust in ML models by understanding how important different input features are for the model, retaining the ability to choose *not* to trust the model, e.g., if it makes use of input features the clinician knows to be unimportant. However, to ensure that the provided explanations are genuinely useful, it is essential to research what specific information clinicians need and how they prefer it to be visualized or communicated. Tailoring these insights to their requirements can significantly enhance their trust and make the tool more practical within the clinical routine. For instance, if certain body key points or limb areas are consistently flagged as important for diagnosing CP, this could lead to more targeted therapies or insights into the diagnostic processes.

While our perturbation strategy works well within our GCN framework, it is not guaranteed to generalize across other architectures or datasets. Other perturbation techniques on the input data, such as a naive perturbation approach of zeroing the coordinate orientation of the body key points, would collapse the body key points at the origin. This renders the data representations meaningless and distorts the spatial relationships within the skeleton too much for an informed perturbation. Further investigation is needed to explore whether matching perturbation strategies can be successfully applied to other data types, such as RGB video or time-series sensor data, as often used in HAR. Finally, additional datasets should be tested to validate our proposed method further.

The scope of perturbation techniques within SHAP is constrained to post-hoc methods. This hinders its deployment in dynamic or real-time systems, where the immediate interpretation of prediction results is required. A notable challenge associated with the use of SHAP is its high computational cost, particularly when processing large datasets. This computational burden originates from the computational complexity of calculating Shapley values, which presents a challenge to the scalability of the approach for real-time or large-scale applications. This issue can be resolved by parallelizing the SHAP calculation on different nodes, but it requires more computational resources, which may not be available in every setting. Additionally, approximation techniques like

(a) SHAP values of the acceleration feature.

(b) SHAP values of the velocity feature.

(c) SHAP values of the bones feature.

(d) SHAP values of the joint feature.
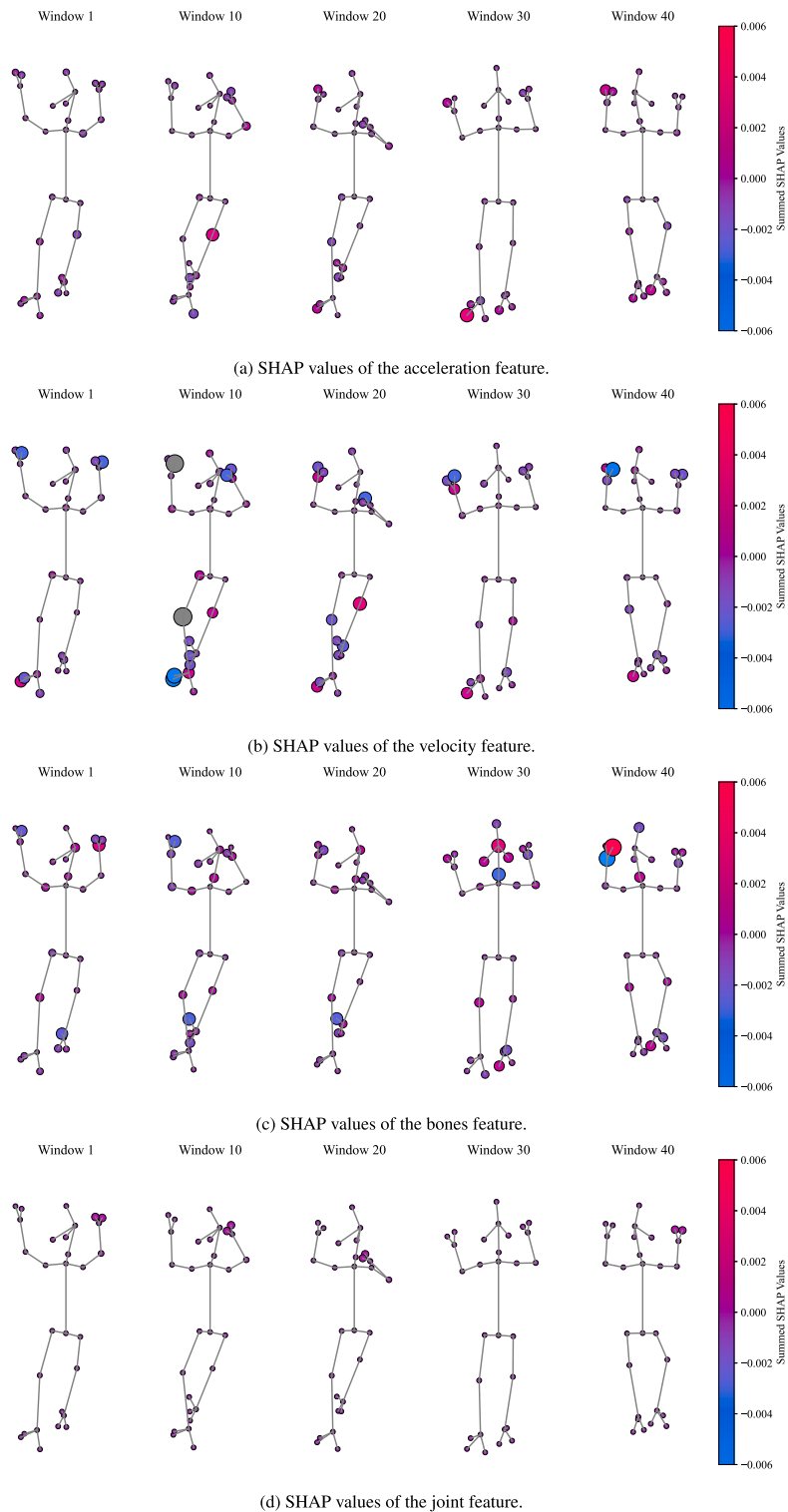
Fig. A.7. Comparison of the local SHAP values for the different input features.

"Kernel-", "Tree-", and "Deep SHAP" can be used to speed the process of obtaining the explanation [14].

SHAP explanations offer spatial–temporal interpretations of the impact of different skeleton body key points. However, the size and composition of the reference dataset have an impact on these explanations. Selecting an appropriate reference dataset that reflects the dataset distribution is crucial to ensure reliable and meaningful explanations. Although a larger reference dataset offers a more detailed explanation, it also requires more computational effort, increasing exponentially. In this context, loading the entire dataset combined with a large model on the GPU may be currently not possible when resources are restricted, necessitating several iterations to obtain the explanations. GradCAM may be the better alternative in an environment with resource restrictions due to its lower computational requirements, as shown in [48].

Biomarkers depict unique features or patterns tied to specific movements and hold critical importance in HAR. This is especially important

Table B.5
Performance metrics of class 6 (*pick up*) with informed perturbation scenarios with varying percentual threshold settings. The table presents the accuracy of the important body key points (Acc. imp.), the accuracy of the unimportant body key points (Acc. ¬ imp.), PGI, and PGU for different perturbation percentages.

Thresh. (%)	Perturb. key points (#)	Acc. imp.	Acc. ¬ imp.	PGI (↑)	PGU (↓)	Thresh. (%)	Perturb. key points (#)	Acc. imp.	Acc. ¬ imp.	PGI (↑)	PGU (↓)
5	1	0.975	0.972	0.004	0.001	30	1	0.924	0.965	0.047	0.006
5	2	0.972	0.972	0.001	0.001	30	2	0.908	0.930	0.063	0.041
5	3	0.975	0.972	0.004	0.001	30	3	0.902	0.911	0.069	0.060
5	4	0.972	0.972	0.001	0.001	30	4	0.886	0.886	0.085	0.085
5	5	0.972	0.972	0.001	0.001	30	5	0.858	0.883	0.113	0.088
5	6	0.972	0.972	0.001	0.001	30	6	0.848	0.873	0.123	0.098
5	7	0.972	0.975	0.001	0.004	30	7	0.848	0.867	0.123	0.104
5	8	0.972	0.972	0.001	0.001	30	8	0.851	0.864	0.120	0.107
5	9	0.972	0.972	0.001	0.001	30	9	0.848	0.858	0.123	0.113
5	10	0.972	0.972	0.001	0.001	30	10	0.835	0.854	0.136	0.117
10	1	0.968	0.972	0.003	0.001	35	1	0.883	0.940	0.088	0.031
10	2	0.972	0.978	0.001	0.007	35	2	0.851	0.918	0.120	0.053
10	3	0.972	0.978	0.001	0.007	35	3	0.759	0.883	0.212	0.088
10	4	0.972	0.975	0.001	0.004	35	4	0.617	0.839	0.354	0.132
10	5	0.972	0.975	0.001	0.004	35	5	0.560	0.807	0.411	0.164
10	6	0.972	0.972	0.001	0.001	35	6	0.541	0.782	0.430	0.189
10	7	0.981	0.968	0.010	0.003	35	7	0.566	0.772	0.405	0.199
10	8	0.975	0.972	0.004	0.001	35	8	0.573	0.763	0.398	0.208
10	9	0.975	0.975	0.004	0.004	35	9	0.557	0.753	0.414	0.218
10	10	0.972	0.975	0.001	0.004	35	10	0.544	0.734	0.427	0.237
15	1	0.965	0.975	0.006	0.004	40	1	0.775	0.930	0.196	0.041
15	2	0.962	0.978	0.009	0.007	40	2	0.766	0.886	0.205	0.085
15	3	0.965	0.975	0.006	0.004	40	3	0.646	0.826	0.325	0.145
15	4	0.968	0.975	0.003	0.004	40	4	0.424	0.763	0.547	0.208
15	5	0.965	0.972	0.006	0.001	40	5	0.244	0.722	0.727	0.249
15	6	0.972	0.968	0.001	0.003	40	6	0.177	0.699	0.794	0.272
15	7	0.972	0.968	0.001	0.003	40	7	0.212	0.696	0.759	0.275
15	8	0.975	0.972	0.004	0.001	40	8	0.266	0.671	0.705	0.300
15	9	0.975	0.972	0.004	0.001	40	9	0.228	0.674	0.743	0.297
15	10	0.968	0.972	0.003	0.001	40	10	0.203	0.668	0.768	0.303
20	1	0.956	0.975	0.015	0.004	45	1	0.655	0.899	0.316	0.072
20	2	0.959	0.965	0.012	0.006	45	2	0.611	0.813	0.360	0.158
20	3	0.959	0.968	0.012	0.003	45	3	0.516	0.753	0.455	0.218
20	4	0.968	0.968	0.003	0.003	45	4	0.361	0.677	0.610	0.294
20	5	0.959	0.965	0.012	0.006	45	5	0.165	0.630	0.806	0.341
20	6	0.972	0.959	0.001	0.012	45	6	0.063	0.608	0.908	0.363
20	7	0.959	0.959	0.012	0.012	45	7	0.051	0.598	0.920	0.373
20	8	0.965	0.956	0.006	0.015	45	8	0.089	0.592	0.882	0.379
20	9	0.968	0.956	0.003	0.015	45	9	0.082	0.595	0.889	0.376
20	10	0.962	0.959	0.009	0.012	45	10	0.079	0.589	0.892	0.382
25	1	0.949	0.972	0.022	0.001						
25	2	0.943	0.956	0.028	0.015						
25	3	0.927	0.956	0.044	0.015						
25	4	0.959	0.940	0.012	0.031						
25	5	0.946	0.930	0.025	0.041						
25	6	0.956	0.934	0.015	0.037						
25	7	0.956	0.934	0.015	0.037						
25	8	0.953	0.934	0.018	0.037						
25	9	0.953	0.934	0.018	0.037						
25	10	0.934	0.930	0.037	0.041						

within clinical applications such as the CP dataset is used. SHAP offers insights into the contribution of input features and has the potential to identify biomarkers. Due to the cooperative game theory concept, SHAP's close relationship with the input data makes it a promising tool for biomarker identification. Identifying these biomarkers with SHAP could help diagnose and monitor infant conditions. SHAP can also serve as a valuable method for discovering how certain input features contribute to recognizing such movement-related biomarkers. However, this concept remains unexplored in this work and requires further validation within a clinical study.

Finally, the averaging technique for the SHAP values may obscure nuanced yet important variations in data. This might be especially important within the windows of the CP dataset, where absent fidgety movements can indicate CP [45]. For instance, subtle short temporal dynamics in the windows may be lost through averaging and potentially impact the interpretability of these essential patterns. Future work could refine this technique to preserve such details and offer more exhaustive explanations. However, which information is finally communicated to the clinicians has to be decided.

6. Conclusion

This study highlights the efficacy of SHAP in explaining the decision-making processes of GCNs used in HAR tasks based on skeleton data. Our novel perturbation strategy enables examining how specific body key points influence model predictions, and we identify the importance of body key points to the model on two real-world datasets, classifying CP and human actions. Perturbing the most important body key points significantly compromises the model's sensitivity on the CP dataset and class accuracy on the model for the NTU RGB+D dataset, validating the importance of the specific body key points as obtained using SHAP. This reinforces the reliability of SHAP values as an attribution method for feature importance. We can also gain granular insight into the specific feature contributions to the model output through SHAP.

Table B.6

Performance metrics of class 11 (*reading*) with informed perturbation scenarios with varying percentual threshold settings. The table presents the accuracy of the important body key points (Acc. imp.), the accuracy of the unimportant body key points (Acc. ¬ imp.), PGI, and PGU for different perturbation percentages.

Thresh. (%)	Perturb. key points (#)	Acc. imp.	Acc. ¬ imp.	PGI (↑)	PGU (↓)	Thresh. (%)	Perturb. key points (#)	Acc. imp.	Acc. ¬ imp.	PGI (↑)	PGU (↓)
5	1	0.806	0.800	0.009	0.003	30	1	0.727	0.771	0.07	0.026
5	2	0.816	0.800	0.019	0.003	30	2	0.746	0.724	0.051	0.073
5	3	0.819	0.797	0.022	0.000	30	3	0.714	0.702	0.083	0.095
5	4	0.816	0.797	0.019	0.000	30	4	0.683	0.663	0.114	0.134
5	5	0.813	0.797	0.016	0.000	30	5	0.629	0.625	0.168	0.172
5	6	0.825	0.784	0.028	0.013	30	6	0.594	0.606	0.203	0.191
5	7	0.822	0.787	0.025	0.010	30	7	0.549	0.578	0.248	0.219
5	8	0.819	0.784	0.022	0.013	30	8	0.489	0.565	0.308	0.232
5	9	0.813	0.784	0.016	0.013	30	9	0.435	0.556	0.362	0.241
5	10	0.822	0.784	0.025	0.013	30	10	0.362	0.546	0.435	0.251
10	1	0.819	0.797	0.022	0.000	35	1	0.597	0.733	0.200	0.064
10	2	0.825	0.803	0.028	0.006	35	2	0.568	0.667	0.229	0.130
10	3	0.838	0.800	0.041	0.003	35	3	0.505	0.632	0.292	0.165
10	4	0.851	0.797	0.054	0.000	35	4	0.470	0.568	0.327	0.229
10	5	0.851	0.790	0.054	0.007	35	5	0.371	0.486	0.426	0.311
10	6	0.854	0.781	0.057	0.016	35	6	0.305	0.397	0.492	0.400
10	7	0.854	0.781	0.057	0.016	35	7	0.232	0.368	0.565	0.429
10	8	0.854	0.784	0.057	0.013	35	8	0.187	0.337	0.610	0.460
10	9	0.860	0.784	0.063	0.013	35	9	0.117	0.321	0.680	0.476
10	10	0.863	0.784	0.066	0.013	35	10	0.086	0.314	0.711	0.483
15	1	0.825	0.797	0.028	0.000	40	1	0.432	0.695	0.365	0.102
15	2	0.848	0.787	0.051	0.010	40	2	0.349	0.600	0.448	0.197
15	3	0.867	0.787	0.070	0.010	40	3	0.235	0.495	0.562	0.302
15	4	0.879	0.790	0.082	0.007	40	4	0.178	0.365	0.619	0.432
15	5	0.876	0.781	0.079	0.016	40	5	0.124	0.302	0.673	0.495
15	6	0.876	0.778	0.079	0.019	40	6	0.095	0.251	0.702	0.546
15	7	0.873	0.778	0.076	0.019	40	7	0.076	0.206	0.721	0.591
15	8	0.883	0.784	0.086	0.013	40	8	0.041	0.190	0.756	0.607
15	9	0.879	0.781	0.082	0.016	40	9	0.016	0.184	0.781	0.613
15	10	0.876	0.781	0.079	0.016	40	10	0.013	0.181	0.784	0.616
20	1	0.829	0.787	0.032	0.010	45	1	0.333	0.673	0.464	0.124
20	2	0.848	0.778	0.051	0.019	45	2	0.184	0.505	0.613	0.292
20	3	0.870	0.765	0.073	0.032	45	3	0.117	0.346	0.680	0.451
20	4	0.883	0.775	0.086	0.022	45	4	0.089	0.232	0.708	0.565
20	5	0.886	0.759	0.089	0.038	45	5	0.035	0.175	0.762	0.622
20	6	0.870	0.762	0.073	0.035	45	6	0.029	0.140	0.768	0.657
20	7	0.854	0.752	0.057	0.045	45	7	0.022	0.127	0.775	0.670
20	8	0.844	0.756	0.047	0.041	45	8	0.006	0.111	0.791	0.686
20	9	0.854	0.756	0.057	0.041	45	9	0.003	0.111	0.794	0.686
20	10	0.854	0.752	0.057	0.045	45	10	0.000	0.111	0.797	0.686
25	1	0.790	0.778	0.007	0.019						
25	2	0.810	0.756	0.013	0.041						
25	3	0.813	0.746	0.016	0.051						
25	4	0.819	0.752	0.022	0.045						
25	5	0.816	0.740	0.019	0.057						
25	6	0.794	0.727	0.003	0.070						
25	7	0.781	0.714	0.016	0.083						
25	8	0.759	0.708	0.038	0.089						
25	9	0.721	0.692	0.076	0.105						
25	10	0.670	0.689	0.127	0.108						

Table B.7

Performance metrics of class 16 (*put on a shoe*) with informed perturbation scenarios with varying percentual threshold settings. The table presents the accuracy of the important body key points (Acc. imp.), the accuracy of the unimportant body key points (Acc. ¬ imp.), PGI, and PGU for different perturbation percentages.

Thresh. (%)	Perturb. key points (#)	Acc. imp.	Acc. ¬ imp.	PGI (↑)	PGU (↓)	Thresh. (%)	Perturb. key points (#)	Acc. imp.	Acc. ¬ imp.	PGI (↑)	PGU (↓)
5	1	0.838	0.841	0.007	0.004	30	1	0.635	0.810	0.210	0.035
5	2	0.838	0.841	0.007	0.004	30	2	0.638	0.775	0.207	0.070
5	3	0.848	0.841	0.003	0.004	30	3	0.629	0.737	0.216	0.108
5	4	0.844	0.841	0.001	0.004	30	4	0.644	0.692	0.201	0.153
5	5	0.841	0.841	0.004	0.004	30	5	0.606	0.692	0.239	0.153
5	6	0.844	0.841	0.001	0.004	30	6	0.610	0.667	0.235	0.178
5	7	0.838	0.841	0.007	0.004	30	7	0.603	0.641	0.242	0.204
5	8	0.841	0.841	0.004	0.004	30	8	0.600	0.641	0.245	0.204
5	9	0.854	0.841	0.009	0.004	30	9	0.622	0.641	0.223	0.204
5	10	0.848	0.841	0.003	0.004	30	10	0.587	0.629	0.258	0.216
10	1	0.829	0.838	0.016	0.007	35	1	0.533	0.784	0.312	0.061
10	2	0.838	0.829	0.007	0.016	35	2	0.444	0.727	0.401	0.118
10	3	0.841	0.841	0.004	0.004	35	3	0.406	0.686	0.439	0.159
10	4	0.841	0.844	0.004	0.001	35	4	0.397	0.629	0.448	0.216
10	5	0.832	0.844	0.013	0.001	35	5	0.400	0.575	0.445	0.270
10	6	0.841	0.841	0.004	0.004	35	6	0.429	0.540	0.416	0.305
10	7	0.838	0.835	0.007	0.010	35	7	0.406	0.521	0.439	0.324
10	8	0.838	0.835	0.007	0.010	35	8	0.422	0.511	0.423	0.334
10	9	0.844	0.835	0.001	0.010	35	9	0.432	0.505	0.413	0.340
10	10	0.841	0.835	0.004	0.010	35	10	0.416	0.502	0.429	0.343
15	1	0.813	0.835	0.032	0.010	40	1	0.422	0.756	0.423	0.089
15	2	0.810	0.816	0.035	0.029	40	2	0.324	0.670	0.521	0.175
15	3	0.825	0.819	0.020	0.026	40	3	0.260	0.600	0.585	0.245
15	4	0.819	0.819	0.026	0.026	40	4	0.260	0.521	0.585	0.324
15	5	0.822	0.822	0.023	0.023	40	5	0.248	0.473	0.597	0.372
15	6	0.822	0.819	0.023	0.026	40	6	0.273	0.438	0.572	0.407
15	7	0.819	0.819	0.026	0.026	40	7	0.267	0.416	0.578	0.429
15	8	0.829	0.819	0.016	0.026	40	8	0.254	0.397	0.591	0.448
15	9	0.825	0.822	0.020	0.023	40	9	0.302	0.397	0.543	0.448
15	10	0.813	0.822	0.032	0.023	40	10	0.270	0.394	0.575	0.451
20	1	0.771	0.835	0.074	0.010	45	1	0.333	0.714	0.512	0.131
20	2	0.775	0.803	0.070	0.042	45	2	0.283	0.597	0.562	0.248
20	3	0.756	0.806	0.089	0.039	45	3	0.225	0.492	0.620	0.353
20	4	0.759	0.813	0.086	0.032	45	4	0.175	0.387	0.670	0.458
20	5	0.775	0.806	0.070	0.039	45	5	0.165	0.362	0.680	0.483
20	6	0.787	0.797	0.058	0.048	45	6	0.143	0.324	0.702	0.521
20	7	0.778	0.790	0.067	0.055	45	7	0.152	0.308	0.693	0.537
20	8	0.759	0.781	0.086	0.064	45	8	0.162	0.286	0.683	0.559
20	9	0.762	0.787	0.083	0.058	45	9	0.162	0.283	0.683	0.562
20	10	0.746	0.787	0.099	0.058	45	10	0.156	0.283	0.689	0.562
25	1	0.698	0.816	0.147	0.029						
25	2	0.721	0.800	0.124	0.045						
25	3	0.717	0.787	0.128	0.058						
25	4	0.721	0.768	0.124	0.077						
25	5	0.733	0.756	0.112	0.089						
25	6	0.714	0.743	0.131	0.102						
25	7	0.711	0.749	0.134	0.096						
25	8	0.721	0.743	0.124	0.102						
25	9	0.708	0.740	0.137	0.105						
25	10	0.686	0.743	0.159	0.102						

While our findings are promising, they also have some limitations, including the computational burden associated with SHAP calculations and the potential lack of generalizability of the perturbation technique across diverse datasets and tasks. Adapting our perturbation approach to different datasets and architectures and exploring the use of SHAP in biomarker identification are promising areas for future research. Our study lays a foundation for further advances in the explainability of models used in HAR, especially in the medical domain.

CRediT authorship contribution statement

Felix Tempel: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Espen Alexander F. Ihlen:** Writing – review & editing, Supervision, Data curation. **Lars Adde:** Writing – review & editing, Data curation. **Inga Strümke:** Writing – review & editing, Supervision, Project administration.

Code availability

The code and the experimental results are publicly available at https://github.com/DeepInMotion/ShapGCN.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Lars Adde reports a relationship with In-Motion Technologies AS that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Individual feature importance CP dataset

In Fig. A.7, the individual SHAP values for the input features (J, V, B, and A) are shown. These plots illustrate the contribution of each feature to the model's predictions on a specific body key point. The variation in SHAP values across body key points, input features, and windows highlights their differing levels of importance, allowing for a detailed comparison of their individual impact on the prediction.

Appendix B. NTU RGB+D threshold experiments

This section presents the experimental results to show the performance behavior for the NTU RGB+D dataset under different edge matrix threshold perturbations. The results of our experiments on classes 6, 11, and 16 are presented in Tables B.5, B.6, and B.7, respectively. Unlike the CP dataset, the NTU RGB+D dataset is especially sensitive to the masking perturbation, in which the corresponding edge e_n in $\mathbf{E}$ is set to 1×10^{-5}, which results in a great performance degradation already after one perturbed joint. Therefore, we conduct experiments in which the corresponding edge is perturbed with a percental threshold away from its original value.

Data availability

The NTU RGB+D 60 dataset that supports the findings of this study is available in https://rose1.ntu.edu.sg/dataset/actionRecognition/ [2]. Due to ethical considerations, the CP dataset is not publicly available.

References

[1] Z. Shuchang, A survey on human action recognition, 2022, arXiv:2301.06082.

[2] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, NTU RGB+D: A large scale dataset for 3D human activity analysis, 2016, arXiv:1604.02808.

[3] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, A.C. Kot, NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding, IEEE Trans. Pattern Anal. Mach. Intell. 42 (10) (2020) 2684–2701, http://dx.doi.org/10.1109/TPAMI.2019.2916873.

[4] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The kinetics human action video dataset, 2017, arXiv:1705.06950.

[5] S. Bharati, M.R.H. Mondal, P. Podder, A review on explainable artificial intelligence for healthcare: Why, how, and when? IEEE Trans. Artif. Intell. 5 (4) (2024) 1429–1442, http://dx.doi.org/10.1109/TAI.2023.3266418.

[6] A. Chaddad, J. Peng, J. Xu, A. Bouridane, Survey of explainable AI techniques in healthcare, Sensors 23 (2) (2023) 634, http://dx.doi.org/10.3390/s23020634.

[7] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, Inf. Fusion 99 (2023) 101805, http://dx.doi.org/10.1016/j.inffus.2023.101805.

[8] S. Nazir, D.M. Dickson, M.U. Akram, Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks, Comput. Biol. Med. 156 (2023) 106668, http://dx.doi.org/10.1016/j.compbiomed.2023.106668.

[9] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): toward medical XAI, IEEE Trans. Neural Networks Learn. Syst. 32 (11) (2021) 4793–4813, http://dx.doi.org/10.1109/TNNLS.2020.3027314.

[10] W. Saeed, C. Omlin, Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities, Knowl.-Based Syst. 263 (2023) 110273, http://dx.doi.org/10.1016/j.knosys.2023.110273.

[11] P. Das, A. Ortega, Gradient-weighted class activation mapping for spatio temporal graph convolutional network, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, Singapore, Singapore, 2022, pp. 4043–4047, http://dx.doi.org/10.1109/ICASSP43922.2022.9746621.

[12] Y. Song, Z. Zhang, C. Shan, L. Wang, Constructing stronger and faster baselines for skeleton-based action recognition, IEEE Trans. Pattern Anal. Mach. Intell. PP (2022) http://dx.doi.org/10.1109/TPAMI.2022.3157033.

[13] K.N. Pellano, I. Strümke, E.A.F. Ihlen, From movements to metrics: Evaluating explainable AI methods in skeleton-based human activity recognition, Sensors 24 (6) (2024) 1940, http://dx.doi.org/10.3390/s24061940.

[14] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, 2017, arXiv:1705.07874.

[15] M.L. Leavitt, A. Morcos, Towards falsifiable interpretability research, 2020, arXiv:2010.12016.

[16] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38, http://dx.doi.org/10.1016/j.artint.2018.07.007.

[17] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI, ACM Comput. Surv. 55 (13s) (2023) 1–42, http://dx.doi.org/10.1145/3583558, arXiv:2201.08164.

[18] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017, arXiv:1702.08608.

[19] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, second ed., Christoph Molnar, Munich, Germany, 2022.

[20] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, J. Mach. Learn. Res. 20 (177) (2019) 1–81, http://dx.doi.org/10.48550/arXiv.1801.01489.

[21] M. Ivanovs, R. Kadikis, K. Ozols, Perturbation-based methods for explaining deep neural networks: A survey, Pattern Recognit. Lett. 150 (2021) 228–234, http://dx.doi.org/10.1016/j.patrec.2021.06.030.

[22] J. Allgaier, L. Mulansky, R.L. Draelos, R. Pryss, How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare, Artif. Intell. Med. 143 (2023) 102616, http://dx.doi.org/10.1016/j.artmed.2023.102616.

[23] Z. Sadeghi, R. Alizadehsani, M.A. Cifci, S. Kausar, R. Rehman, P. Mahanta, P.K. Bora, A. Almasri, R.S. Alkhawaldeh, S. Hussain, B. Alatas, A. Shoeibi, H. Moosaei, M. Hladík, S. Nahavandi, P.M. Pardalos, A review of explainable artificial intelligence in healthcare, Comput. Electr. Eng. 118 (2024) 109370, http://dx.doi.org/10.1016/j.compeleceng.2024.109370.

[24] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014, http://dx.doi.org/10.48550/arXiv.1312.6034, arXiv:1312.6034.

[25] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, 2016, arXiv:1602.04938.

[26] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2) (2020) 336–359, http://dx.doi.org/10.1007/s11263-019-01228-7.

[27] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (Tcav), in: International Conference on Machine Learning, PMLR, 2018, pp. 2668–2677.

[28] Z. Yan, R. Younes, J. Forsyth, ResNet-like CNN architecture and saliency map for human activity recognition, in: S. Deng, A. Zomaya, N. Li (Eds.), Mobile Computing, Applications, and Services, vol. 434, Springer International Publishing, Cham, 2022, pp. 129–143, http://dx.doi.org/10.1007/978-3-030-99203-3_9.

[29] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc., 2018.

[30] K. Borys, Y.A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C.M. Friedrich, F. Nensa, Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches, Eur. J. Radiol. 162 (2023) 110786, http://dx.doi.org/10.1016/j.ejrad.2023.110786.

[31] A. Lucieri, M.N. Bajwa, S.A. Braun, M.I. Malik, A. Dengel, S. Ahmed, On interpretability of deep learning based skin lesion classifiers using concept activation vectors, in: 2020 International Joint Conference on Neural Networks, IJCNN, 2020, pp. 1–10, http://dx.doi.org/10.1109/IJCNN48605.2020.9206946, arXiv:2005.02000.

[32] J.V. Jeyakumar, A. Sarker, L.A. Garcia, M. Srivastava, X-CHAR: A concept-based explainable complex human activity recognition model, Proc. the ACM Interact. Mob. Wearable Ubiquitous Technol. 7 (1) (2023) 1–28, http://dx.doi.org/10.1145/3580804.

[33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, Las Vegas, NV, USA, 2016, pp. 2921–2929, http://dx.doi.org/10.1109/CVPR.2016.319.

[34] P.E. Pope, S. Kolouri, M. Rostami, C.E. Martin, H. Hoffmann, Explainability methods for graph convolutional neural networks, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, Long Beach, CA, USA, 2019, pp. 10764–10773, http://dx.doi.org/10.1109/CVPR.2019.01103.

[35] C. Agarwal, D. Ley, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, H. Lakkaraju, OpenXAI: Towards a transparent evaluation of model explanations, 2024, arXiv:2206.11104.

[36] C. Agarwal, N. Johnson, M. Pawelczyk, S. Krishna, E. Saxena, M. Zitnik, H. Lakkaraju, Rethinking stability for attribution-based explanations, 2022, arXiv:2203.06877.

[37] Q. Gao, S. Yao, Y. Tian, C. Zhang, T. Zhao, D. Wu, G. Yu, H. Lu, Automating general movements assessment with quantitative deep learning to facilitate early screening of cerebral palsy, Nat. Commun. 14 (1) (2023) 8294, http://dx.doi.org/10.1038/s41467-023-44141-x.

[38] L. Shapley, 7. a value for n-person games. contributions to the theory of games II (1953) 307-317., in: H.W. Kuhn (Ed.), Classics in Game Theory, Princeton University Press, 1953, pp. 69–79, http://dx.doi.org/10.1515/9781400829156-012.

[39] F. Tempel, E.A.F. Ihlen, I. Strümke, AutoGCN-toward generic human activity recognition with neural architecture search, IEEE Access : Pr. Innov. Open Solutions 12 (2024) 39505–39516, http://dx.doi.org/10.1109/ACCESS.2024.3377103.

[40] F. Tempel, E.A.F. Ihlen, I. Strümke, Lightweight neural architecture search for cerebral palsy detection, 2024, arXiv:2409.20060.

[41] H. Xiong, X. Li, X. Zhang, J. Chen, X. Sun, Y. Li, Z. Sun, M. Du, Towards explainable artificial intelligence (XAI): A data mining perspective, 2024, arXiv:2401.04374.

[42] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, Thirty- Second. AAAI Conf. Artif. Intell. (2018) arXiv:1801.07455.

[43] V. Petsiuk, A. Das, K. Saenko, RISE: Randomized input sampling for explanation of black-box models, 2018, arXiv:1806.07421.

[44] H.F. Prechtl, C. Einspieler, G. Cioni, A.F. Bos, F. Ferrari, D. Sontheimer, An early marker for neurological deficits after perinatal brain lesions, Lancet 349 (9062) (1997) 1361–1363, http://dx.doi.org/10.1016/S0140-6736(96)10182-3.

[45] C. Einspieler, H.F.R. Prechtl, A.F. Bos, Prechtl's Method on the Qualitative Assessment of General Movements in Preterm, Term and Young Infants, Mac Keith Press, London, 2004.

[46] D. Groos, L. Adde, S. Aubert, L. Boswell, R.-A. de Regnier, T. Fjørtoft, D. Gaebler-Spira, A. Haukeland, M. Loennecken, M. Msall, U.I. Möinichen, A. Pascal, C. Peyton, H. Ramampiaro, M.D. Schreiber, I.E. Silberg, N.T. Songstad, N. Thomas, C. Van den Broeck, G.K. Øberg, E.A. Ihlen, R. Støen, Development and validation of a deep learning method to predict cerebral palsy from spontaneous movements in infants at high risk, JAMA Netw. Open 5 (7) (2022) e2221325, http://dx.doi.org/10.1001/jamanetworkopen.2022.21325.

[47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, 2019, arXiv:1912.01703.

[48] F. Tempel, D. Groos, E.A.F. Ihlen, L. Adde, I. Strümke, Choose your explanation: A comparison of SHAP and gradCAM in human activity recognition, 2024, http://dx.doi.org/10.48550/arXiv.2412.16003, arXiv:2412.16003.