ISSN 1213-6972 (print)
ISSN 1213-6964 (online)
Journal of WSCG
http://www.wscg.eu
Vol.33, No-1-2, 2025

Object Detection and Human Activity Recognition for Improved Patient Mobility and Caregiver Ergonomics

Mahesh Madhavan
Sustainable Digitalization Research Center (SDRC)
Department of Computer Science and Media Technology
Malmö University
mahesh87madhav@gmail.com

Patford Nkhoma
Sustainable Digitalization Research Center (SDRC)
Department of Computer Science and Media Technology
Malmö University
psnkhoma@outlook.com

Reza Khoshkangini
Sustainable Digitalization Research Center (SDRC)
Department of Computer Science and Media Technology
Malmö University
reza.khoshkangini@mau.se

Mahtab Jamali
Sustainable Digitalization Research Center (SDRC)
Department of Computer Science and Media Technology
Malmö University
mahtab.jamali@mau.se

Paul Davidsson
Sustainable Digitalization Research Center (SDRC)
Department of Computer Science and Media Technology
Malmö University
paul.davidsson@mau.se

Jan Åberg
Arjo Sverige AB
Malmö, Sweden 20120
jan.aberg@arjo.com

Martin Ljungqvist
Axis Communications AB
Lund, Sweden 22369
martin.ljungqvist@axis.com

ABSTRACT

This study explores the use of machine learning to enhance patient mobility and caregiver ergonomics by optimizing the use of mobility aids. Traditional manual assessments can be subjective and inaccurate, so this research develops a data-driven model for object detection and human activity recognition. A computer vision dataset was created using video recordings of controlled caregiving scenarios. The study leverages advanced machine learning models, including YOLO for object detection, pose estimation, ResNet-18 for frame classification, Inception-v4 for feature extraction, and LSTM for sequence modeling. The findings provide valuable insights into integrating machine learning into mobility aids, improving both patient outcomes and caregiver well-being.

Keywords

Mobility aid, Ergonomics, Caregiver, Machine Learning, Musculoskeletal disorders.

1 INTRODUCTION

Humans are designed for movement, and limited mobility affects daily tasks, social interactions, and overall well-being. In healthcare environments, understanding visual scenes is essential for supporting patients and assisting caregivers in monitoring and decision-making [Khoshkangini et al., 2024]. Caregivers in healthcare settings, from hospitals to nursing homes, face high rates of Musculoskeletal Disorders (MSDs)[Jamali et al., 2024b], making it crucial to improve their quality of life and work conditions. Mechanical lift equipment reduces the risk of MSDs associated with patient handling. Observational postural assessment methods, such as Rapid Entire Body Assessment (REBA), Rapid Upper Limb Assessment (RULA), and Ovako Working Posture Analysis System (OWAS), are commonly used to mitigate MSD risks, but these methods are manual, time-consuming, and prone to variability due to subjective expert evaluations [Yuan and Zhou, 2023]. Recently, advanced computer vision technologies have been explored for automatic ergonomic assessments, offering a more efficient and objective alternative [Jamali et al., 2025, Kim et al., 2021, Zhang et al., 2018, Jamali et al., 2024a].

MSDs pose significant challenges in healthcare, affecting nurses, therapists, and employers through absenteeism, medical expenses, and workers' compensation claims [Yuan and Zhou, 2023]. Despite the critical nature of this issue, ensuring proper product usage, maintaining patient comfort, and optimizing caregiver ergonomics remains a persistent challenge during caregiving. While existing studies have explored healthcare technology and posture recognition [Kim et al., 2021], [Zhang et al., 2018], they often fail to address the specific application of mobility aids and their correct usage in caregiving scenarios. Additionally, limited research has focused on improving patient mobility outcomes and caregiver ergonomics through machine learning. This study aims to bridge these gaps by developing and validating machine learning models for object detection, pose estimation, and activity recognition in simulated clinical environments, focusing on mobility aids like lifts, wheelchairs, and beds. Using a novel dataset created from videos recorded at a controlled clinical setting, where subjects simulated caregivers and patients, this research represents an innovative approach to classifying caregiving activities in relation to mobility aid equipment. While the study does not encompass the full real-world variability or explore all poten-

tial algorithms, it targets correct product usage and enhances patient mobility and caregiver posture, addressing a critical and underexplored area in healthcare technology.

In order to better understand the applicability of machine learning techniques in the identification of correct product usage of mobility aids equipment and posture assessment, the following Research Questions (RQ) have been formulated.

RQ 1. To what extent can machine learning be applied for object and activity recognition to effectively utilize and identify correct product usage and assess posture in caregiving using mobility equipment?

RQ 2. Can transfer learning techniques be effectively applied to improve the accuracy and efficiency of object detection and human activity recognition systems in real-world scenarios?

2 BACKGROUND

2.1 Caregiver Ergonomics

Ergonomics in healthcare focuses on optimizing caregivers' tasks and environments to enhance well-being, productivity, and care quality [Waters, 2010]. Derived from the Greek words "ergo" (work) and "nomos" (natural laws), it is defined as the science of improving safety and productivity by aligning jobs, equipment, and human interaction [Mansoor et al., 2022]. The physically demanding tasks of caregiving, such as lifting, transferring, and repetitive movements, often strain the musculoskeletal system, increasing the risk of MSDs like back pain, strains, and sprains [Waters, 2010].

2.2 Mobility Aid equipment

Mobility aid equipment includes devices like wheelchairs, crutches, scooters, canes, and walkers to support individuals with mobility impairments. Advances in technology have introduced innovative aids such as ceiling lifts, bed lifts, and transfer aids, designed to improve accessibility and facilitate patient transfers. These assistive devices aim to enhance safety for both patients and caregivers [Schoenfisch et al., 2019].

2.3 MaxiMove Lift

The MaxiMove is a mobile patient lifting device designed to aid caregivers in transferring and repositioning patients with limited mobility. It typically consists of a wheeled base for maneuverability, a standing frame with a lifting mechanism and adjustable height as shown in Figure 1(a) [Owens and Tapley, 2018], and supportive sling attachments that secure the patient during transfers.

2.4 SaraFlex Lift

The SaraFlex lift Figure 1(b) is designed to assist caregivers in transferring patients who require support during activities such as standing, walking, or repositioning, enabling a single caregiver to aid a patient from a seated to a standing position in one natural movement [Arjo, 2024].

(a) MaxiMove (b) SaraFlex

Figure 1: Mobility Aids

3 RELATED WORKS

Few studies classify correct mobility aid usage, making it largely unexplored. However, there are similarities in dataset creation efforts highlighting the need for customized data to support machine learning models. [Gauen et al., 2017] analyze major machine learning datasets and propose a novel dataset creation method using network cameras, focusing on object detection with "people" labels. Our study uses RGB cameras to capture specific events. [Tian et al., 2022] present the Construction Motion Data Library (CML) for activity recognition in construction, using ergonomic labels to identify unsafe postures. Our project leverages similar methods for posture recognition in specialized activity datasets. [Agrawal and Ertel, 2018] introduce the ERTRAG project, a virtual trainer for nursing staff to reduce MSDs by identifying incorrect postures. Our project involves MaxiMove and SaraFlex lifts to facilitate patient transfers, reducing direct physical handling by caregivers. [Yuan and Zhou, 2023] propose an ergonomic posture risk assessment (EPRA) method using the ROMP algorithm with a single RGB camera. Our work calculates angles between 2D body segment vectors. [Oh et al., 2011] introduced the VIRAT Video Dataset to support continuous visual event recognition (CVER) in outdoor environments using footage from both stationary and aerial cameras. While our work focuses on indoor settings typical of caregiving contexts, it similarly emphasizes continuous monitoring of human-object interactions. Recent work by [Azadvatan and Kurt, 2024] demonstrates the effectiveness of training YOLO-based architectures from scratch for vehicular detection, while our approach leverages transfer learning to adapt pre-trained YOLO weights for healthcare-specific objects, achieving faster convergence and higher accuracy on domain-specific tasks.

4 METHODOLOGY

4.1 Data Collection and Creation

We created a dataset capturing caregiver-patient interactions with mobility aids, inspired by safety and productivity monitoring methodologies [Gauen et al., 2017]. This dataset simulates realistic caregiving scenarios, addressing the lack of specificity in existing datasets like MOCAP and NTU + RGBD 120 [Tian et al., 2022]. Data were collected through video recordings in a simulated healthcare environment using two Sony alpha-6 RGB cameras, following ethical guidelines with informed consent from participants. Efforts were made to ensure diversity in height, weight, gender, and ethnicity among participants, all of whom were 18 years or older. Although professional nursing staff were not involved, the data collection was guided and validated by domain experts to ensure procedural accuracy. Our dataset focuses on two primary activities: transferring from bed to wheelchair and vice versa using the Maxi-MOVE lift, and transferring from bed to wheelchair and vice versa using the Sara-Flex lift. Eight sub-activities were identified from each of the main activities, and only the relevant frames for each sub-activity were included for processing. Irrelevant portions, such as repeated or incorrect actions, were excluded, ensuring a focused and accurate dataset for model training.

4.2 Experimental Setup

Experiments were conducted in a simulated healthcare setting with hospital beds, wheelchairs, and lift systems to mirror typical caregiving scenarios. Two Sony alpha 6 RGB cameras were positioned 2.5 meters from the scene and 3 meters apart, angled at 45 degrees to capture views. The setup included a wheeled hospital bed, MaxiMove and SaraFlex lifts, and necessary slings and straps. Activities involving the MaxiMove lift lasted 3 to 5 minutes, while SaraFlex activities averaged around 2 minutes, depending on the complexity of the activity.

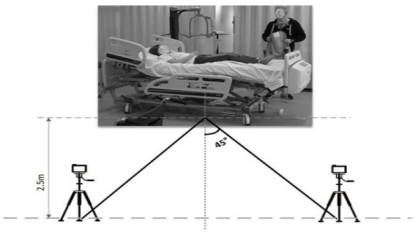


Figure 2: Experimental setup of the shooting scene

5 PROPOSED APPROACH

The proposed approach is shown in Figure 3. The process starts by converting raw videos into frames. YOLOv8 Object Detection identifies and localizes objects, while YOLOv8 Pose tracks movements by mapping key skeletal points. A Face Detection module was optionally included to blur faces for privacy considerations. The ResNet18 frame classifier, which distinguishes relevant frames from irrelevant ones, categorizes frames before Inception-v4 extracts high-level feature vectors. While both ResNet and Inception architectures are effective for image processing and classification, Inception-v4 was chosen for its superior feature extraction capabilities, whereas ResNet18 was optimized for frame classification in this study. The ergonomic analysis module uses pose keypoints to compute an ergonomic score. Feature vectors are processed by an LSTM network to analyze sequential activities, with fully connected layers supporting high-level reasoning and a sigmoid activation for binary classification of ergonomic correctness.

5.1 Data preprocessing

Data preprocessing involved annotating the dataset for object detection and pose estimation using Roboflow, as well as preparing the dataset for training the ResNet and LSTM models for activity detection and scoring. Videos were uploaded to the Roboflow platform and manually annotated.

For object detection, bounding boxes were drawn, and class labels (caregiver, patient, wheelchair, bed, Maxi-Move, and SaraFlex) were assigned at a rate of 1 frame per second. The dataset comprised 4,932 labeled images, preprocessed with auto-orientation and resizing. Data augmentation (flipping, brightness adjustments, blurring) expanded it to 11,834 images. The dataset was then split into 10,354 samples for training, 986 for validation, and 494 for testing.

In order to annotate a dataset for pose estimation, we identified keypoints representing human joints such as the head, shoulders, elbows, wrists, hips, knees, and ankles. A 13-keypoint skeletal structure was used, as illustrated in Figure 4, connecting joints through lines (e.g., shoulder to elbow, elbow to wrist) to form a mapped human pose. Unlike more complex 17- or 21-keypoint systems, the 13-point structure minimized annotation time while maintaining the necessary detail to evaluate ergonomic postures. Our pose estimation dataset preparation included labeling 1,042 images, followed by preprocessing (auto-orientation, resizing) and augmentations (saturation, brightness, blurring, noise) to enhance robustness, expanding it to 3,229 images. The dataset was split into 2,916 for training, 211 for validation, and 102 for testing.

The dataset for training the frame classifier and LSTM models was prepared through systematic steps. Frames were extracted from raw videos at 25 fps, categorized into distinct sub-activities and saved in separate folders with sequential filenames. This organization enables focused training and evaluation on specific activities. The preprocessing (cropping) retained only objects of interest detected by the object detection model,

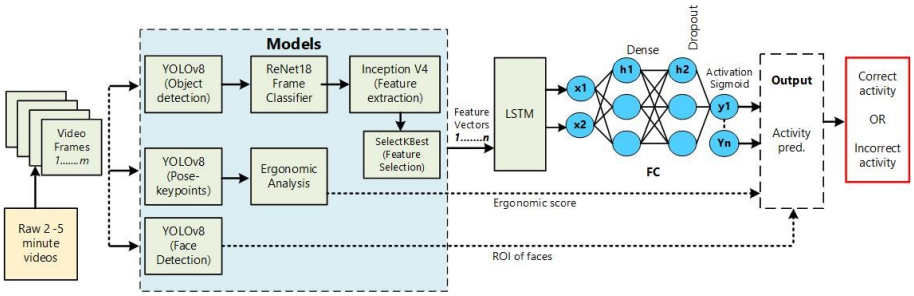
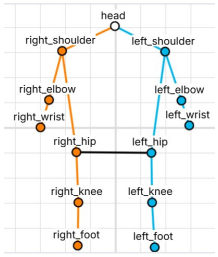


Figure 3: The conceptual diagram of the proposed approach for classifying activities as correct or incorrect.



1 Head
2 Left_shoulder
3 Left_elbow
4 Left_wrist
5 Right_shoulder
6 Right_elbow
7 Right_wrist
8 Left_hip
9 Left_knee
10 Left_foot
11 Right_hip
12 Right_knee
13 Right_foot

Figure 4: 13-point keypoint system

minimizing background noise. To enhance generalization, five data augmentation techniques, including rotation, flipping, color adjustments, brightness and contrast variations, were applied.

5.2 Object Detection Module

The object detection component (illustrated in figure 3) of this study utilizes the YOLOv8 model as the primary framework for identifying and locating various objects involved in caregiving activities. The YOLOv8n variant was chosen for its computational efficiency and capability to operate on a CPU, yielding satisfactory results for object detection.

During the training phase, key metrics such as box loss and class loss were computed. Box loss quantifies the error in bounding box predictions, while class loss assesses the accuracy of object class predictions. Following the training phase, the model's performance was assessed using the validation dataset.

5.3 Pose Estimation Module

Pose estimation was performed using the YOLOv8 Pose model, adapted for tracking and analyzing caregiver movements to assess ergonomic postures. While this study focuses on caregiver posture assessment, evaluating patient posture is proposed for future research. The YOLOv8n-Pose variant was trained with custom hyperparameters, setting 'pretrained' parameter to *false* to accommodate our 13-keypoint annotation system instead of the default 17-keypoint setup. Model accuracy was assessed using precision metrics, including the mean absolute error (MAE) of joint angles.

5.4 Angle Calculation and Posture Assessment

Accurate calculation of joint angles and posture assessment are crucial for evaluating caregiving activities and identifying risks of MSDs [Yuan and Zhou, 2023]. The model predicts keypoints on video frames of caregivers and patients, corresponding to body parts like the head, shoulders, elbows, hips, knees, and ankles. These coordinates are used to compute joint angles, with a focus on assessing the caregiver's back bending posture to determine MSD risks.

The study assesses the degree of hip bending from a lateral trunk view, classifying posture as normal, moderate, or severe based on RULA and OWAS guidelines [Yuan and Zhou, 2023], [Zhang et al., 2018], [Paudel et al., 2022]. These tools correlate bending angles with risk categories: normal posture (low risk), moderate bending (medium risk), and severe bending (high risk). The following section outlines techniques for analyzing caregiver ergonomics, focusing on angle calculation and posture evaluation.

Keypoint Detection and Pose Analysis: Keypoints correspond to critical body landmarks, such as the hips and shoulders. Midpoints between keypoints are calculated to simplify angle computations. For instance, the midpoints between the hips and shoulders are represented as (x_i, y_i) coordinates for subsequent calculations.

The first step involves extracting the coordinates of relevant keypoints detected by the pose estimation model. The detection process is performed with a confidence threshold, ensuring accurate and reliable identification of keypoints. The coordinates of keypoints are represented as:

$$P_i = (x_i, y_i) \tag{1}$$

Where P_i denotes the index of the keypoint. For instance, keypoints for shoulders, hips, and other significant joints are indexed accordingly.

Angle Calculation: Angle calculation involves determining the angles formed by vectors between keypoints that represent caregiver and patient postures. Specifically, to evaluate body segment alignment, we calculate the angle at the hip by using lines connecting the shoulder midpoint and the hip midpoint. Another line is drawn from the shoulder midpoint to intersect an imaginary line that extends vertically from the hip midpoint, forming a right angle at the intersection point, referred to as the vertical reference.

1. Midpoints Calculation

Shoulder midpoint:

$$P_{\text{shoulder\_mid}} = \left(\frac{x_L + x_R}{2}, \frac{y_L + y_R}{2} \right) \tag{2}$$

where x_L and y_L represent the coordinates of the left shoulder, and x_R and y_R represent the coordinates of the right shoulder.

Hip midpoint:

$$P_{\text{hip\_mid}} = \left(\frac{x_L + x_R}{2}, \frac{y_L + y_R}{2} \right) \tag{3}$$

where x_L and y_L represent the coordinates of the left hip, and x_R and y_R represent the coordinates of the right hip.

2. Vector Definitions

Vectors u and v are defined between these keypoints

$$u = (P_{\text{shoulder\_mid}}, P_{\text{hip\_mid}}) \tag{4}$$

$$v = (P_{\text{vertical\_reference}}, P_{\text{hip\_mid}}) \tag{5}$$

where $P_{\text{vertical\_reference}}$ is a point directly above $P_{\text{hip\_mid}}$ chosen to form a right angle with the horizontal line through $P_{\text{shoulder\_mid}}$

3. Angle computation

This is achieved by using trigonometric functions. The cosine of the angle θ between the vectors u and v is computed as:

$$\cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} \tag{6}$$

Thus, the angle θ can be calculated as:

$$\theta = \arccos\left(\frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} \right) \tag{7}$$

where arcos() finds the inverse of the cosine angle. The angle θ is then converted from radians to degrees by the formula:

$$\theta_{\text{degrees}} = \theta \times \frac{180}{\pi} \tag{8}$$

Angle Adjustment: To account for the 2D effect and body alignment, the calculated angle is adjusted based on the slope of the hip line. This adjustment ensures that the angle measurement accurately reflects true ergonomic conditions, compensating for any tilt in the hip line. The adjustment is formulated as follows:

1. Slope Calculation

$$\text{Slope} = \frac{y_{\text{right\_hip}} - y_{\text{left\_hip}}}{x_{\text{right\_hip}} - x_{\text{left\_hip}}} \tag{9}$$

2. Adjusted Factor

The adjustment factor incorporates the slope to modify the calculated angle:

$$\text{adjustment\_factor} = \frac{\text{scaling\_factor}}{1 + |\text{slope}|} \tag{10}$$

Where scaling_factor is a manually fixed parameter determined through empirical analysis to ensure accurate adjustments.

3. Adjusted Angle

$$\theta_{\text{adjusted}} = \theta \times (1 + \text{adjustment\_factor}) \tag{11}$$

Ergonomic Scoring: To quantify ergonomic quality, an ergonomic score is calculated based on the measured angles of the caregiver's posture. The RULA/OWAS scores for the trunk flexion angle (TFA) are categorized into angles between 0° and 20° are considered neutral and acceptable, while angles between 20° and 60° indicate moderate lateral inclinations. Upper body angles exceeding 60° for sagittal inclinations are deemed unacceptable [Yan et al., 2017], [Abobakr et al., 2019], [Hellmers et al., 2022], [Zhang et al., 2018]. Different weightings are assigned to various strain levels to classify the angles into Normal, Moderate, or Severe categories. where:

- **Normal** - with a score of 1 and angles between 0° to 20° at the hip indicating optimal posture.
- **Moderate** - a score of 0.5 and angles between 20° to 60° indicating some deviation from the optimal posture.
- **Severe** - a score of 0 and angles greater than 60° reflecting significant deviations that could lead to musculoskeletal disorders (MSDs).

$$\text{strain\_score} = \begin{cases} 1 & \text{if } \theta_{\text{adjusted}} < 20 \\ 0.5 & \text{if } 20 \leq \theta_{\text{adjusted}} \leq 60 \\ 0 & \text{if } \theta_{\text{adjusted}} > 60 \end{cases} \tag{12}$$

The overall ergonomic score is derived as the average strain score across all analyzed angles. This score provides an assessment of the caregiver's posture, facilitating the identification of potential ergonomic risks and implementation of corrective measures.

$$\text{Overall Ergonomic Score} = \frac{1}{N} \sum_{i=1}^{N} \text{strain\_score}_i \tag{13}$$

where N is the total number of frames analyzed.

5.5 Frame Classification Module

The frame classification module is essential for classifying individual video frames to aid in detecting and analyzing human activities and movements. We employed the ResNet-18 architecture, pre-trained on ImageNet [He et al., 2015] to classify frames into distinct activity categories. The dataset included 36,072 frames, with 4,509 frames allocated to each of the eight classes. This transfer learning approach utilizes the pre-trained weights of ResNet-18 enhancing performance and accelerating convergence.

5.5.1 Model Training and Validation

To narrow down the research, we focused on a single activity: "transferring from bed to wheelchair using the Sara-Flex lift." The dataset for frame classification consisted of a varied number of frames per activity, ranging from 59,740 to 37,575 frames in the training set. To standardize, one frame was selected for every 10 frames, resulting in 3,757 frames per activity, based on the class with the fewest frames. The test set comprised 20% of the training set, yielding 751 frames per activity.

Two CSV files were created for training and testing, each containing file paths and corresponding labels based on the following label mapping:

$$\text{label\_mapping} = \begin{cases} 0 : \text{'sara\_sling\_placement\_bed'} \\ 1 : \text{'sara\_sling\_attach\_bed'} \\ 2 : \text{'sara\_lifting\_from\_bed'} \\ 3 : \text{'sara\_transfer\_to\_wheelchair'} \\ 4 : \text{'sara\_lowering\_to\_wheelchair'} \\ 5 : \text{'sara\_detach\_sling\_wheelchair'} \\ 6 : \text{'sara\_remove\_sling\_wheelchair'} \\ 7 : \text{'sara\_lifting\_from\_bed\_wrong'} \end{cases}$$

During preprocessing, the frames were resized to 224×224 pixels and normalized to ensure consistency. To manage the large volume of frames, every 10th frame was sampled for inclusion in the training dataset.

In training the frame classifier, we utilized a CNN based on the pre-trained ResNet-18 model, replacing the final layer with eight output neurons corresponding to the sub-activity classes. The performance was evaluated using a confusion matrix on the test set to ensure generalization to new data.

5.5.2 Sequence Creation

Frames extracted from the input video are classified into one of eight predefined classes or labeled as unknown if the maximum probability is below 0.40. Predictions of frames in the original sequence are smoothed using a 500-frame window to reduce noise. Frames are grouped into sequences based on transitions in smoothed predictions. A change in predicted class ends the current sequence and starts a new one. Sequences shorter than a specified proportion of total frames are labeled as unknown. If an unknown label appears between two sequences of the same activity, it is merged with them; otherwise, it remains labeled as unknown.

5.6 Feature Extraction

We implemented the Inception-v4 model. The model was set to evaluation mode to process each frame and extract features from the video frames in sequences of length 50. As each frame passes through the pretrained model, a 1536-dimensional feature vector is derived. These features are aggregated across the sequence, ensuring that each sequence of frames is consistently represented by 50 sets of feature vectors. If a sequence had fewer than 50 frames, frames were duplicated; if more, the sequence was truncated to maintain consistency. The extracted features are then used as inputs for the LSTM model.

5.7 Feature Selection

Before inputting video frame features into the LSTM for confidence scoring, we used the SelectKBest [Aguilera et al., 2022] method with the chi-squared (χ^2) scoring function. This is a univariate feature selection method that ranks features based on their importance in predicting the target variable, selecting the top k features most relevant for classification. The highest values indicate features with a strong association with the target variable. The chi-squared statistic for each feature is calculated as:

$$\chi^2 = \sum_{c=1}^{C} \frac{(O_{fc} - E_{fc})^2}{E_{fc}} \tag{14}$$

Where:

O_{fc} is the observed frequency of the feature in class c,
E_{fc} is the expected frequency of the feature in class c,
$E_{fc} = \frac{\text{total count of feature in the dataset} \times \text{total count of class } c}{\text{total number of examples}}$.

Prior to feature selection, the LSTM model struggled with convergence, often leading to oscillation or divergence of the loss function. By applying SelectKBest, we reduced the dimensionality of input features from 1,536 to 1,024, retaining only the most informative features and eliminating irrelevant data. Each of the eight corresponding LSTM models was paired with its SelectKBest model, resulting in a significant reduction in average test loss and improved generalization during inference.

5.8 LSTM Model

The LSTM model is designed to evaluate whether activities are performed correctly, where each activity includes eight specific sub-activities that must follow a set sequence. The model computes a confidence score for each detected sub-activity, determining if it exceeds a predefined threshold, thereby verifying proper execution.

5.8.1 *Training and Validation*

The training process involved developing eight distinct LSTM models with the same architecture, each dedicated to a specific sub-activity to capture its unique characteristics. The dataset contained 150 training examples (80%) and 37 test examples (20%) per sub-activity. To address potential bias from intra-video similarity, the dataset was split strictly at the video level, ensuring no frames from the same video appeared in multiple sets. While the background remained constant, participants alternated roles (patient, caregiver), and recordings were captured from two different camera angles. Data augmentation was also applied to increase training diversity. This setup ensured robust training while retaining sufficient data for reliable evaluation. Each sequence was downsampled to a maximum length of 50 frames by selecting an optimal downsampling rate. Additional sequences were generated from the same videos by selecting subsequent frames not included in previous sequences while maintaining the same step size, resulting in 7225 training and 1776 testing examples. Feature selection reduced the feature set from 1536 to 1024 dimensions, leading to final dataset dimensions of (7225, 50, 1024) for training and (1776, 50, 1024) for testing.

To prepare the data for the LSTMs, binary labels were generated, and eight separate data loaders were created. Each data loader contained all examples of a specific class labeled as positive (label 1) and a random selection of examples from other classes labeled as negative (label 0). The ratio of positive to negative examples was maintained at 1:0.5. For example, if class 0 (`activity_0`) has 1000 examples, the corresponding model for class 0 will be trained with 1000 positive and 500 negative examples.

The LSTM architecture comprised:

- **LSTM Layer:** Parameters include `input_size = 1024`, `hidden_size = 256`, `num_layers = 1`, learning rate = `0.00001`, and `epochs = 10`.
- **Fully Connected (Linear) Layer:** Reducing LSTM output to a single confidence score.
- **Activation Function:** A sigmoid activation function providing a confidence score between 0 and 1.

Adam optimizer [Kingma and Ba, 2015] was used to adjust learning rates based on recent gradients.

5.8.2 *Testing*

Each LSTM model was evaluated with its respective test dataset, yielding confidence scores for each sub-activity. Performance was assessed using Binary Cross Entropy Loss, with the average test loss for each sub-activity recorded.

6 RESULTS

6.1 Object Detection Results

We evaluated object detection performance using Mean Average Precision (mAP), Recall, and F1 Score, with mAP50-95 scores indicating the model's precision across Intersection over Union (IoU) thresholds from 0.50 to 0.95. Table 1 presents F1, Precision, and Recall metrics.

Epochs	mAP50-95 (Box)	Precision	Recall	F1-score
50	0.8453	0.96638	0.9689	0.9677

Table 1: Validation set performance results for object detection.

IoU measures the overlap between predicted and ground truth bounding boxes, with higher thresholds requiring more precise overlap [Everingham et al., 2010]. The YOLOv8 model achieved an mAP50-95 (Box) score of 0.8453, reflecting an 84.53% average precision, demonstrating reliable object localization accuracy. Figure 5 below shows examples of detected objects with their respective confidence levels.

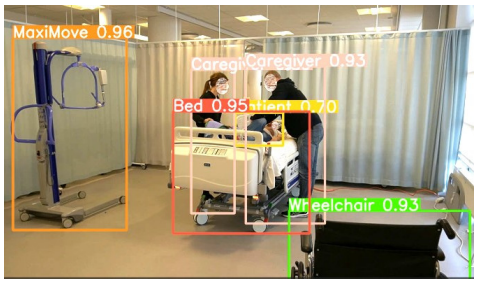


Figure 5: Object detection output and confidence levels.

The model accurately identifies "Bed," "Patient," "SaraFlex," "Wheelchair," and "Background" classes, with minimal misclassification. Errors mainly occur between "Caregiver" and "Patient" due to their close interactions with equipment. A background class is included to reduce false positives (FP) [Ultralytics, 2023].

6.2 Pose Estimation Results

Pose estimation was assessed using mAP and accuracy metrics, which are crucial for posture analysis. The model achieved a mAP50-95 of 91.20% with over 99% precision and recall, indicating better model performance in detecting and localizing keypoints accurately. Table 2 summarizes the model's performance, and Figure 6 shows detected keypoints in an inference image.

Epochs	mAP50-95		Precision		Recall	
	Box	Pose	Box	Pose	Box	Pose
50	0.938	0.912	0.995	0.995	0.995	0.995

Table 2: Validation set performance results for pose estimation.

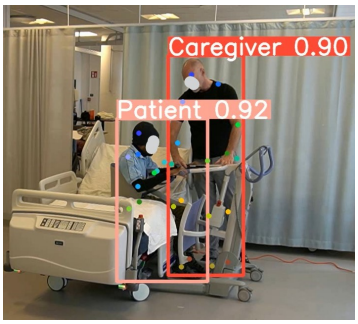


Figure 6: Keypoint detection.

6.2.1 Ergonomic Assessment

The ergonomic assessment used a scoring system based on key body joint angles to determine the level of ergonomic strain experienced by caregivers during patient handling tasks. Table 3 lists frames with various angles and their ergonomic scores, categorizing strain severity. Frames with low angles, such as 5 and 3 degrees in Frame1 and Frame2, received a score of 1, indicating 'normal' severity, while higher angles like 22 and 21 degrees (Frame828 and Frame829) scored 0.5, signaling 'low strain.' The majority of frames fell within the 'normal' range, with a mean score indicating acceptable ergonomic strain levels. However, instances of 'low strain' scores suggest specific postures where caregivers could experience musculoskeletal discomfort. Figure 7 illustrates the pose estimation and keypoints with angle calculations used for ergonomic scoring.

Frames	Angle	Severity	Score	Mean score
Frame1	5.0	normal	1	1
Frame2	3.0	normal	1	1
...	...	...	...	...
Frame828	22.0	low_strain	0.5	0.69
Frame829	21.0	low_strain	0.5	0.69

Table 3: Ergonomic posture scoring from different video frames.

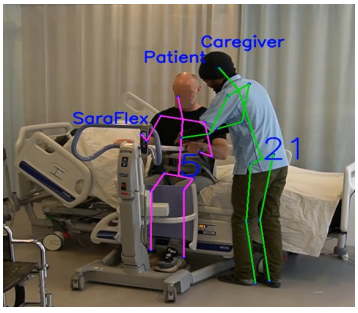


Figure 7: Pose estimation and Keypoints with Angle.

6.3 Frame Classifier Results

The ResNet-18 model was employed to classify frames into different activity categories, focusing on the SaraFlex. The model achieved 99.14% accuracy in training and 98.85% in testing, demonstrating strong classification performance. The high accuracy suggests that it effectively learned to classify the different activities from the dataset.

6.4 LSTM Model Results

LSTM training was conducted in two phases. Initially, using all extracted features without feature selection, the test loss showed insufficient convergence, often due to overfitting. Table 4 presents the training, validation, and test losses over 10 epochs, showing that while training and validation losses appeared to converge, the test loss for most sub-activity classes did not achieve the desired improvement. Applying feature selection significantly enhanced performance, with test losses dropping as low as 0.10 across sub-activities, as shown in Table 5. This indicates better generalization compared to the initial approach.

Sub-activity	Epoch	Training Loss	Validation Loss	Test Loss
0	10	0.3503	0.3255	0.6283
1	10	0.3296	0.3598	0.4213
2	10	0.1867	0.2060	0.4061
3	10	0.0987	0.0852	0.3566
4	10	0.2007	0.2065	0.2184
5	10	0.2806	0.2696	0.7026
6	10	0.1569	0.1472	0.1957
7	10	0.0452	0.0388	0.0443

Table 4: Training results for LSTM models without feature selection.

Sub-activity	Epoch	Training Loss	Validation Loss	Test Loss
0	10	0.1133	0.0924	0.1006
1	10	0.2113	0.1952	0.2024
2	10	0.1113	0.0980	0.0982
3	10	0.0630	0.0534	0.0950
4	10	0.0618	0.0532	0.0591
5	10	0.1401	0.1238	0.1418
6	10	0.1304	0.1102	0.1096
7	10	0.0718	0.0584	0.0768

Table 5: Training results for LSTM models with feature selection.

6.5 Analysis of Results

Overall, the combined model's performance across object detection, pose estimation, frame classification, and sequence modeling indicates capability in identifying correct product usage and assessing posture for caregivers. The results from each module of the system were analyzed to assess their overall effectiveness and key observations include:

Object Detection model's mAP score of 0.8453 indicates good performance.

Pose Estimation model achieved 91.20% mAP for pose accuracy. Despite the high mAP values, the model's performance could further be improved by incorporating a more comprehensive range of pose scenarios in the training dataset.

The Frame Classifier model demonstrated good performance across many classes, though with some

misclassifications likely due to frame overlap; addressing these overlaps could enhance precision.

LSTM model's implementation without feature selection has test losses ranging from approximately 0.2 to 0.7 for various sub-activities. In particular, sub activity 5 exhibited high test losses, indicating that the model struggled to generalize well. feature selection, losses stabilized between 0.025 and 0.2, indicating better generalization.

6.6 Inference Workflow and Evaluation

A snapshot of Inference results from each workflow stage is shown in Figure 8.

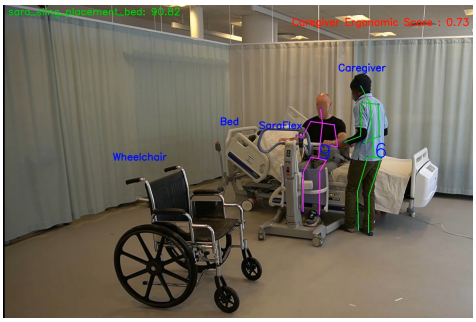


Figure 8: A snapshot of combined results of all models.

The frame classifier displays the current activity and confidence score in the top left, updating as the caregiver transitions between tasks. The top right displays the ergonomic score for the caregiver during each activity. The Object Detection and Pose Detection models identify relevant objects and key skeletal points, drawing the skeleton and calculating ergonomic scores based on bending angles (6° for the caregiver, 9° for the patient). Detected objects are labeled within the scene. The model is designed to analyze entire activities, even those involving repeated sub-activities (e.g., transferring a patient from bed to wheelchair). This allows comprehensive assessment of prolonged activities. While the LSTM model's confidence score may lower with prolonged tasks, it remains effective for evaluating caregiver performance across complete activities.

7 DISCUSSION

The primary aim of this project was to explore how machine learning could facilitate the classification of correct mobility aid usage and ergonomic assessment in caregiving. Additionally, this research provided insights into dataset creation and human activity recognition to monitor interactions between caregivers and mobility aids. The findings indicate that integrating YOLO, ResNet, Inception-v4 and LSTM, effectively identifies and classifies objects and activities within caregiving settings. The object detection model showed high accuracy in recognizing essential items like mobility aids, while the pose estimation model offered valuable insights into caregiver and patient interactions. These high-accuracy results highlight machine learning's potential in improving product usage and ergonomics in healthcare.

Given this, we could answer RQ1 by stating that "*The ML models demonstrated effectiveness in identifying objects and activities relevant to caregiving. Object detection successfully recognized key items like mobility aids, while pose estimation accurately assessed caregiver and patient postures, showing that ML can assist in identifying correct product usage and ergonomic assessment.*"

Regarding RQ2, and upon the results obtained, we could state that" *The use of Models like YOLO and ResNet18 confirmed that using pre-trained models and transfer learning improves the accuracy of object detection and activity recognition tasks"*

This study has several limitations. The dataset may not fully represent real-world healthcare variability, as subjects imitating patients may not capture the full range of actual behaviors, limiting generalizability. Testing in controlled settings may not directly translate to real-world scenarios, impacting the model's external validity. The manual annotation process can introduce human error and bias, potentially affecting data quality and model performance.

8 CONCLUSION

This study demonstrated the application of machine learning techniques in healthcare, particularly for improving patient care and caregiver safety. Key objectives included creating a robust dataset for object detection, pose estimation, and activity recognition; leveraging machine learning models for accurate recognition and assessment; and applying transfer learning to enhance model performance. The study integrated YOLO for object detection, YOLO Pose for pose estimation, ResNet for frame classification, Inception-v4 for feature extraction, and LSTM for sequence modeling. This cohesive system effectively recognizes human activities and assesses caregiver ergonomics, thereby contributing to improved healthcare practices.Future research could focus on collecting more comprehensive data, exploring advanced ML models, and developing refined ergonomic scoring techniques to better measure caregiver posture.

9 ACKNOWLEDGMENT

This study is supported by the 'Synergy' project at Malmo University which was funded by the Knowledge Foundation in Sweden.

10 REFERENCES

[Abobakr et al., 2019] Abobakr, A., Nahavandi, D., Hossny, M., Iskander, J., Attia, M., Nahavandi, S., and Smets, M. (2019). RGB-D ergonomic assessment system of adopted working postures. *Appl. Ergon.*, 80:75–88.

[Agrawal and Ertel, 2018] Agrawal, A. and Ertel, W. (2018). Automatic nursing care trainer based on machine learning. pages 53–59.

[Aguilera et al., 2022] Aguilera, A., Pezoa, R., and RodrÃguez-Delherbe, A. (2022). A novel ensemble feature selection method for pixel-level segmentation of her2 overexpression. *Complex Intelligent Systems*, 8(6):5489–5510.

[Arjo, 2024] Arjo (2024). Patient Handling. `https://www.arjo.com/en-us/products/patient-handling/`. Accessed: 2024-04-19.

[Azadvatan and Kurt, 2024] Azadvatan, Y. and Kurt, M. (2024). Melnet: A real-time deep learning algorithm for object detection. arXiv preprint arXiv:2401.17972.

[Everingham et al., 2010] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338.

[Gauen et al., 2017] Gauen, K., Dailey, R., Laiman, J., Zi, Y., Asokan, N., Lu, Y.-H., Thiruvathukal, G. K., Shyu, M.-L., and Chen, S.-C. (2017). Comparison of visual datasets for machine learning. In *2017 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 770–778.

[Hellmers et al., 2022] Hellmers, S., Brinkmann, A., Böhlen, C. F.-v., Lau, S., Diekmann, R., and Hein, A. (2022). Posture and mechanical load assessment during patient transfers. *SN Comput. Sci.*, 3(5).

[Jamali et al., 2024a] Jamali, M., Davidsson, P., Khoshkangini, R., Ljungqvist, M. G., and Mihailescu, R.-C. (2024a). Specialized indoor and outdoor scene-specific object detection models. In *Sixteenth International Conference on Machine Vision (ICMV 2023)*, volume 13072, pages 201–210. SPIE.

[Jamali et al., 2025] Jamali, M., Davidsson, P., Khoshkangini, R., Ljungqvist, M. G., and Mihailescu, R.-C. (2025). Context in object detection: a systematic literature review. *Artificial Intelligence Review*, 58(6):1–89.

[Jamali et al., 2024b] Jamali, M., Davidsson, P., Khoshkangini, R., Mihailescu, R.-C., Sexton, E., Johannesson, V., and Tillström, J. (2024b). Video-audio multimodal fall detection method. In *Pacific Rim International Conference on Artificial Intelligence*, pages 62–75. Springer.

[Khoshkangini et al., 2024] Khoshkangini, R., Tajgardan, M., Jamali, M., Ljungqvist, M. G., Mihailescu, R.-C., and Davidsson, P. (2024). Hierarchical transfer multi-task learning approach for scene classification. In *International Conference on Pattern Recognition*, pages 231–248. Springer.

[Kim et al., 2021] Kim, W., Sung, J., Saakes, D., Huang, C., and Xiong, S. (2021). Ergonomic postural assessment using a new open-source human pose estimation technology (OpenPose). *Int. J. Ind. Ergon.*, 84:103164.

[Kingma and Ba, 2015] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*. arXiv:1412.6980.

[Mansoor et al., 2022] Mansoor, S. N., Al Arabia, D. H., and Rathore, F. A. (2022). Ergonomics and musculoskeletal disorders among health care professionals: Prevention is better than cure. *J. Pak. Med. Assoc.*, 72(6):1243–1245.

[Oh et al., 2011] Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J. T., Mukherjee, S., Aggarwal, J. K., Lee, H., Davis, L., Swears, E., Wang, X., Ji, Q., Reddy, K., Shah, M., Vondrick, C., Pirsiavash, H., Ramanan, D., Yuen, J., Torralba, A., Song, B., Fong, A., Roy-Chowdhury, A., and Desai, M. (2011). A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*. IEEE.

[Owens and Tapley, 2018] Owens, T. and Tapley, C. (2018). Pediatric mobility: The development of standard assessments and interventions for pediatric patients for safe patient handling and mobility. *Crit. Care Nurs. Q.*, 41(3):314–322.

[Paudel et al., 2022] Paudel, P., Kwon, Y.-J., Kim, D.-H., and Choi, K.-H. (2022). Industrial ergonomics risk analysis based on 3d-human pose estimation. *Electronics (Basel)*, 11(20):3403.

[Schoenfisch et al., 2019] Schoenfisch, A. L., Kucera, K. L., Lipscomb, H. J., McIlvaine, J., Becherer, L., James, T., and Avent, S. (2019). Use of assistive devices to lift, transfer, and reposition hospital patients. *Nurs. Res.*, 68(1):3–12.

[Tian et al., 2022] Tian, Y., Li, H., Cui, H., and Chen, J. (2022). Construction motion data library: an integrated motion dataset for on-site activity recognition. *Sci Data*, 9(1):726.

[Ultralytics, 2023] Ultralytics (2023). Tips for best training results. `https://docs.ultralytics.com/yolov5/tutorials/tips_for_best_training_results`. Accessed: 2024-8-5.

[Waters, 2010] Waters, T. R. (2010). Introduction to ergonomics for healthcare workers. *Rehabil. Nurs.*, 35(5):185–191.

[Yan et al., 2017] Yan, X., Li, H., Wang, C., Seo, J., Zhang, H., and Wang, H. (2017). Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion. *Adv. Eng. Inform.*, 34:152–163.

[Yuan and Zhou, 2023] Yuan, H. and Zhou, Y. (2023). Ergonomic assessment based on monocular RGB camera in elderly care by a new multi-person 3D pose estimation technique (ROMP). *Int. J. Ind. Ergon.*, 95:103440.

[Zhang et al., 2018] Zhang, H., Yan, X., and Li, H. (2018). Ergonomic posture recognition using 3D view-invariant features from single ordinary camera. *Autom. Constr.*, 94:1–10.