Bachelor of Science in Computer Science
May 2025

3D Pose Estimation and Time-Series Classification for Distinguishing Normal and Fatigued States.

Yaswanth Rahul Yarlagadda
Shaheer Dudekula

Faculty of Faculty, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

This thesis is submitted to the Faculty of Faculty at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science. The thesis is equivalent to Weeks weeks of full-time studies.

The authors declare that they are the sole authors of this thesis and that they have not used any sources other than those listed in the bibliography and identified as references. They further declare that they have not submitted this thesis at any other institution to obtain a degree.

Contact Information:
Author(s):
Yaswanth Rahul Yarlagadda
E-mail: yayr24@student.bth.se

Shaheer Dudekula
E-mail: shdd24@student.bth.se

University advisor:
Shahryar Eivazzadeh
Department of Computer Science

Faculty of Faculty
Blekinge Institute of Technology
SE–371 79 Karlskrona, Sweden

Internet : www.bth.se
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

Abstract

Background: We often overlook the subtle shifts in our walk—slower steps, smaller joint bends—that signal growing fatigue. Traditional fatigue monitoring relies on costly sensors or lab-grade motion-capture rigs, putting continuous assessment out of reach for everyday settings.

Objectives: This thesis builds and tests a completely video-based system for fatigue detection that anyone can use. Specifically,

- we capture synchronized videos from three smartphones (right, left, front) sequence.
- Extract 2D keypoints with MMPose HRNet framework.
- Lifts to 3D poses via SemGCN and compute joint velocities, accelerations, and angles.
- Train both a Conv1D–BiLSTM network and a RandomForest–GradientBoosting ensemble to classify normal vs. fatigued gait and predict a continuous fatigue score.

Method: Volunteers walked normally and after fatiguing exercise while three phones recorded from different angles. We merged and synchronized those streams, ran MMPose HRNet to get 2D keypoints, then used SemGCN to reconstruct 3D joint tracks. From these, we calculated per-frame kinematics and chopped the data into 20-frame windows. Each window fed into our two models, trained with 4-fold stratified cross-validation, early stopping, and learning-rate scheduling. We measured classification accuracy and mean absolute error (MAE) for fatigue-level predictions, and tested robustness under different lighting and camera positions.

Results: The Conv1D–BiLSTM model detected fatigue with 92.7% accuracy (±1.8 %) and predicted fatigue level with MAE = 2.6 on a 0–5 scale. The RandomForest–GradientBoosting ensemble achieved 93.1% accuracy (±1.5%) and MAE = 2.5. Ablation studies showed that adding kinematic features boosted accuracy by up to 7 %, and performance varied by less than 3% when lighting or angles changed.

Conclusion: You don't need expensive hardware to monitor fatigue—three ordinary smartphones and open-source models suffice. Our pipeline reliably flags tired gait and quantifies fatigue, opening the door to accessible health tracking, smarter workout feedback, and safer work environments.

Keywords: 3D pose estimation, fatigue detection, MMPose HRNet, SemGCN, BiLSTM, Random Forest, smartphone video, kinematic features.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr. Shahryar Eivazzadeh, for his unwavering support, insightful guidance, and patient mentorship throughout this project. His expertise in computer vision and thoughtful feedback pushed me to refine my ideas and strengthen my work at every step.

My appreciation goes to the volunteers who generously gave their time and effort to participate in our motion-capture experiments. Without their willingness to be recorded under both rested and fatigued conditions, this research would not have been possible.

Finally, I'm grateful to my family and friends for their constant encouragement, understanding, and coffee breaks during long coding sessions. Their belief in me made the late nights and challenges worthwhile.

Contents

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

The way people move reveals a lot about their physical state. For example, someone who is tired may take shorter steps or move more slowly. These changes might be too small for the human eye to detect, but they can be important signals of health, safety, or performance. Detecting these subtle differences in how people walk, run, or move—known as motion styles—can help in many areas, such as:

- **Health and medicine:** spotting movement issues in patients during recovery,
- **Workplace safety:** warning when someone is too tired to safely operate machinery,
- **Sports:** helping coaches identify signs of exhaustion in athletes.

In the past, analyzing motion like this required expensive tools such as wearable sensors or full-body suits with reflective markers used in labs. These systems are accurate but not practical for everyday use **[8]**. Most people don't have access to this kind of equipment.

Thankfully, smartphones and computer vision (a field of AI that teaches computers to "see" and understand videos) have made it possible to analyze human motion using only video footage **[4]**.. This thesis explores how we can take regular videos from smartphones and turn them into useful motion data—specifically to recognize different styles of movement.

1.2 Knowledge Gap

Previous research has explored many ways to analyze human motion. Some studies use body-worn sensors like accelerometers or electromyography (EMG) devices **[9]**. Others rely on expensive lab setups such as motion-capture rigs or depth cameras to track movement in three dimensions. While these systems can be accurate, they are often expensive, intrusive, and not practical for everyday use.

There are also computer vision approaches that use regular 2D videos, but most of them either lack 3D accuracy, require single fixed viewpoints, or fail to detect subtle changes in movement over time—such as fatigue or other motion style variations **[2]**.

What is missing is a system that combines all of the following:

- works with **only smartphone video**,

- captures movement from **multiple angles**,
- reconstructs **accurate 3D motion**,
- Low computational costs requires no GPU, as MMPOSE is well-Known for its lightweight features **[5]**,
- and detects subtle **motion styles—without needing lab equipment**.

This thesis fills that gap by proposing a complete pipeline that uses three ordinary smartphones and open-source tools to recognize motion styles with high accuracy, making advanced motion analysis accessible and affordable for broader real-world use.

1.3 Problem Statement

This research focuses on two key problems:

1. **Reconstructing 3D movement from normal phone video:** Phone videos show people moving in two dimensions (height and width), but our bodies move in 3D space. Estimating this third dimension (depth) accurately from video is challenging **[1]**.
2. **Recognizing motion styles over time:** Once we have 3D movement data, we need to understand what that movement says. Is the person moving normally, or are there signs of fatigue or unusual patterns? This requires looking at motion over time and finding patterns **[6]**.

1.4 Aim and Objectives

The main aim of this thesis is to design a simple but powerful video-based system that can recognize different motion styles—using only smartphones and free software tools. While this system can be used for many motion types, this study uses “normal” and “fatigued” walking styles as a case study.

The key objectives are:

- **Capture movement from different angles:** Use three smartphones to record videos from the front, left, and right sides.
- **Track body joints in 2D:** Use a software called *MMPose with HRNet* **[7]** to find key points on the body (like elbows, knees, and ankles) 17 joint format called "coco" format in each video frame.
- **Reconstruct 3D body movement:** Use a tool called *SemGCN* **[1]** to estimate how the body moves in 3D space, based on the 2D joint positions.
- **Extract movement features:** Calculate things like how fast joints are moving, how they accelerate, and how angles between limbs change over time **[4]**.

- **Train models to recognize fatigue:** Use two types of machine learning models—one that looks at sequences of data (LSTM) [6], and another that looks at statistical summaries (Random Forest)—to tell whether someone is tired or not, and how tired they are.

1.5 Why Use 3D Lifting Instead of Direct 3D Sensors?

Direct 3D data usually comes from expensive depth cameras or lab equipment. These systems are not available to most people. Instead, this thesis uses "3D lifting," which means turning 2D joint positions (like a shadow on a wall) into 3D body positions (like a full figure in space). This method works surprisingly well when you use multiple camera angles and smart algorithms like SemGCN [1].

1.6 Tools and Frameworks Used

To make the system work, the following tools and techniques are used:

- **MMPose + HRNet:** Detects 2D positions of joints from video. It's reliable even in different lighting conditions or when the person turns sideways [5, 7].
- **SemGCN:** Converts 2D joint positions into 3D movement using a method called "graph convolution," which understands how body parts are connected [1].
- **MPJPE (Mean Per-Joint Position Error):** A number used to measure how accurate the 3D pose is compared to the real pose [3].

These tools are open-source, fast, and free—making them ideal for low-cost systems.

1.7 Outline of the Thesis

This thesis is organized into six chapters:

- **Chapter 2 – Related Work:** Summarizes past research on pose estimation, motion tracking, and fatigue detection.
- **Chapter 3 – Method:** Describes how the system was built, from video recording to data processing and modeling.
- **Chapter 4 – Results and Analysis:** Shows how well the system works, with detailed results and comparisons.
- **Chapter 5 – Discussion:** Interprets the results, compares them with existing systems, and discusses practical uses and limitations.
- **Chapter 6 – Conclusion and Future Work:** Summarizes the contributions and suggests how this work could be improved or extended.

Chapter 2

Related Work

Understanding human motion using computer vision is an active area of research. This chapter reviews the four main pillars that support the work presented in this thesis:

1. Detecting body joints in 2D video,

2. Estimating 3D movement from those joints,

3. Extracting meaningful motion features, and

4. Recognizing motion styles such as fatigue and predicting fatigue levels.

Each of these areas has seen major progress in recent years, but combining them into an accurate, low-cost system that works in everyday environments remains an open challenge.

2.1 2D Human Pose Estimation

The first step in analyzing human movement from video is identifying key joints of the body, such as shoulders, knees, and ankles (17 joint coco format) in each video frame. This is known as **2D pose estimation**.

Early systems like OpenPose allowed detection of multiple people in a single frame **[2]**, but performance dropped in cases of overlapping limbs or unusual angles. Later advancements such as **HRNet** (High-Resolution Network) improved detection accuracy under real-world conditions like poor lighting or side views **[7]**.

This thesis uses HRNet through the **MMPose** framework, which offers reliable 2D joint tracking and can be applied efficiently to smartphone video **[5]**.

2.2 3D Pose Estimation (Pose Lifting)

While 2D joint locations show us where the person appears on screen, they do not reveal depth—how far forward or backward the person is moving. This third dimension is essential to understanding full-body motion. Estimating 3D positions from 2D inputs is called **3D pose lifting**.

Initial methods used fully connected networks or geometry-based rules, but they struggled when parts of the body were hidden. Newer approaches like **SemGCN** treat the human body as a connected graph and use **graph convolutional networks** to estimate the depth of each joint based on its neighbors **[1, 10]**.

This helps reconstruction more realistic 3D motion, even from flat 2D video. In this thesis, SemGCN is applied to synchronized multi-angle smartphone recordings to reconstruct accurate 3D poses.

2.3 Kinematic Feature Extraction

Once we have 3D joint trajectories, we can describe how the body moves over time. This is done by calculating motion-related features such as:

- **Velocity:** how fast each joint moves,
- **Acceleration:** how quickly that motion changes,
- **Joint angles:** how limbs bend relative to each other.

These **kinematic features** help capture subtle shifts in movement such as a slowing step or reduced arm swing that often signal fatigue. Prior studies show that these features not only improve classification accuracy, but are also useful for estimating continuous physical states like exertion or tiredness **[4]**.

2.4 Time-Series Models for Motion Recognition and Fatigue Prediction

Human motion unfolds over time, and subtle changes in movement style such as fatigue, become noticeable only when we examine how joint patterns evolve from frame to frame. This process is known as **time-series modeling**, and it is essential for both classifying motion (e.g., normal vs. fatigued) and predicting continuous levels of fatigue.

Two main approaches are commonly used:

1. Deep learning with temporal layers: In this thesis, we use a neural network architecture that combines two important components—*temporal convolutional layers* and *bidirectional long short-term memory (BiLSTM)* layers.

The **temporal convolutional layer** (implemented using a Conv1D layer in Python) acts as a feature extractor over time. It applies filters across short windows of consecutive frames, allowing the model to detect short-term patterns such as a sudden drop in joint velocity or a stiffened limb.

The output of the convolutional layers is then passed to a **BiLSTM layer**. LSTM networks are a type of recurrent neural network (RNN) designed to learn long-term dependencies in sequences. A BiLSTM, in particular, processes the sequence in both forward and backward directions. This dual perspective is especially useful in detecting fatigue patterns, which may build up gradually or recover after a momentary slowdown.

Together, these layers form a powerful pipeline. The convolutional layers detect local temporal patterns, and the BiLSTM layer models the overall temporal structure of the movement. This combination enables the system to both classify whether a person is fatigued and predict the degree of fatigue on a numeric scale **[6]**.

2. Traditional machine learning on summary features: In contrast to sequence models, traditional classifiers like Random Forests and Gradient Boosting operate on fixed-size feature vectors. Instead of analyzing sequences directly, we compute summary statistics for each motion window (e.g., average velocity, max joint angle), and train the model to associate these summaries with motion labels.

These models are simpler and faster to train and perform well even on small datasets, making them a good baseline or complementary approach.

2.5 Fatigue and Motion Style Detection

Detecting fatigue from movement has traditionally relied on wearable devices like EMG sensors or accelerometers. These tools give precise measurements but are often uncomfortable and unsuitable for casual use.

Other systems use high-end depth cameras or motion-capture setups to record 3D motion, but these are costly and require dedicated lab spaces.

Recently, researchers have begun using video-based approaches to estimate fatigue or detect unusual movement styles. However, most rely on single-camera views and do not estimate 3D movement, making them less accurate for detailed analysis or fatigue-level prediction.

Some studies have explored pose-based fatigue detection by analyzing joint entropy and coordination loss under fatigue **[9]**.

2.6 Our Contribution in Context

What makes this thesis unique is its ability to bring together the best of these separate research areas into a single, affordable, and practical system:

- **Smartphone-only recording** with synchronized multi-angle capture,
- **Accurate 3D motion reconstruction** using SemGCN **[1]**,
- **Detailed kinematic analysis** to understand movement dynamics **[4]**,
- **Fatigue classification and fatigue-level prediction** using two different machine learning approaches **[6]**,
- **Validation under real-world conditions**, including changes in lighting, viewpoint, and unseen subjects.

By doing so, this work takes a step toward making movement-style recognition and fatigue monitoring available to everyday users—not just researchers or elite athletes.

Chapter 3

Method

This chapter explains the complete design and implementation of our low-cost video-based system for recognizing motion styles and predicting fatigue. The methodology includes participant recruitment, video recording setup, data preprocessing, pose estimation, kinematic feature extraction, and machine learning models for classification and regression. The goal is to allow any reader technical or not, to understand the reasoning behind each step and how the system was built from scratch.

3.1 Data Collection and Ethical Approval

The study involved six healthy male volunteers aged between 20 and 22 years, all of whom provided written consent under an IRB-approved protocol (Consent Form v2021-09-20, BTH Ethics Board). Each participant completed two short movement tasks in a controlled indoor environment: one trial at a normal state and another immediately after a fatigue-inducing exercise (three minutes of step-ups). This resulted in a total of 12 trials—one normal and one fatigued for each subject.

Before recording, each participant received an information sheet explaining the study and the purpose of data collection. Consent was given separately for participation and data usage. Signed forms were stored securely, and anonymized IDs were assigned to all data to ensure privacy.

The recordings were made using three identical smartphones placed at right-side, left-side, and front-facing positions. All devices captured video at 30 frames per second and were manually synchronized to maintain alignment across angles.

The average participant age was 21.2 years (± 0.75), and the average height was 169.9 cm (± 4.6). All participants were male. A summary of participant attributes and trial activities is presented in Table 3.1.

Table 3.1: Participant Demographics and Trial Activities

Trial ID	Subject ID	Age (years)	Height (cm)	Gender	Activity
T01	S01	21	162.6	M	Walking (normal)
T02	S01	21	162.6	M	Walking (fatigue)
T03	S02	20	162.6	M	Jumping (normal)
T04	S02	20	162.6	M	Jumping (fatigue)
T05	S03	21	173.0	M	Squatting (normal)
T06	S03	21	173.0	M	Squatting (fatigue)
T07	S04	22	173.0	M	Pushups (normal)
T08	S04	22	173.0	M	Pushups (fatigue)
T09	S05	21	173.0	M	Jogging (normal)
T10	S05	21	173.0	M	Jogging (fatigue)
T11	S06	22	173.0	M	Running (normal)
T12	S06	22	173.0	M	Running (fatigue)

3.2 Research Questions and Methodological Approach

This thesis is guided by two core research questions that define both the technical goals and the practical motivations of the work. Each question is tied to a specific part of the pipeline, from pose estimation to fatigue-level prediction.

Research Question 1 (RQ1): *How can 3D human pose be reliably extracted from smartphone-recorded videos in a low-cost, multi-angle setup?*

To address this question, we designed a capture system using three ordinary smartphones positioned at different viewpoints—right, left, and front-facing. This configuration helps overcome the depth ambiguity that occurs when using a single camera. The recorded videos are first processed using the MMPose framework with an HRNet backbone, which detects 17 key joints of the human body (such as knees, hips, shoulders, and ankles) in each frame. These 2D joint positions are then passed to a lifting model called SemGCN. This model uses a graph-based neural network to estimate 3D joint positions from the 2D input, taking advantage of the structural relationships between joints to improve accuracy even when some joints are partially occluded.

Research Question 2 (RQ2): *How accurately can machine-learning models classify normal versus fatigued motion and predict continuous fatigue levels from kinematic features?*

To explore this question, we compute detailed motion features from the 3D joint trajectories. These include the velocity and acceleration of each joint, as well as joint angles calculated over time. These features help capture not just where the joints are, but how they move and change—a crucial factor in detecting signs of fatigue.

The motion data is divided into overlapping time windows of 20 frames each, which allows the system to model temporal patterns in how a person moves. Each window is labeled with the corresponding trial condition (normal or fatigued) and the self-reported fatigue level on a scale from 0 to 5.

We then train two types of machine learning models to process these windows. The first is a deep learning network that combines temporal convolution layers and bidirectional LSTM (BiLSTM) units. The convolutional layers detect short-term movement patterns, while the BiLSTM learns long-term temporal relationships in both forward and backward time. This model is trained to both classify motion as fatigued or normal and to predict the fatigue level numerically.

The second model uses a more traditional approach: we summarize each window using statistical features such as the mean and maximum of velocities and angles, and then apply a Random Forest and Gradient Boosting ensemble. This non-deep baseline is particularly useful for small datasets and provides an interpretable benchmark for performance comparison.

By comparing the performance of both approaches, we aim to evaluate the system's ability not only to distinguish between normal and fatigued states, but also to estimate how fatigued a person is—making the solution practical for real-world feedback systems.

3.3 Data Management Plan

All data collected for this study were carefully organized to ensure reproducibility, privacy, and secure storage. Raw video recordings from each smartphone were saved in MP4 format and labeled using a standardized naming convention that included trial condition, subject ID, and camera angle (e.g., `S01_normal_right.mp4`). These files were stored on an encrypted external drive and backed up weekly to a secure institutional system.

To protect participant anonymity, all data were assigned random subject identifiers, and no personal information such as names or contact details was stored with the recordings. Only aggregated or de-identified data are presented in this thesis and in any future publications.

All scripts used for preprocessing, pose estimation, and model training were tracked using Git version control. This ensures that each step of the pipeline can be reproduced, modified, or extended by other researchers, contributing to open and transparent research practices.

3.4 Video Preprocessing

Each trial generated three separate videos—one from each smartphone angle. To prepare the data for pose estimation, we first aligned and merged these recordings into a single synchronized stream. This step ensures that all views are temporally consistent, so that the same frame across each video corresponds to the same moment in time during the movement sequence.

The merged video was created by aligning the starting timestamps of the three recordings and trimming any lead-in or trailing frames. The resulting composite video arranges frames in the order: right view, left view, and front view. This structured format simplifies downstream processing by guaranteeing that each group of three frames provides a full multi-angle snapshot of the body.

After merging, the video was split into individual frames, which were then processed for 2D joint detection using MMPose. Figure 3.1 illustrates the overall pipeline, from multi-angle video capture through 2D and 3D pose estimation, feature extraction, and model training.

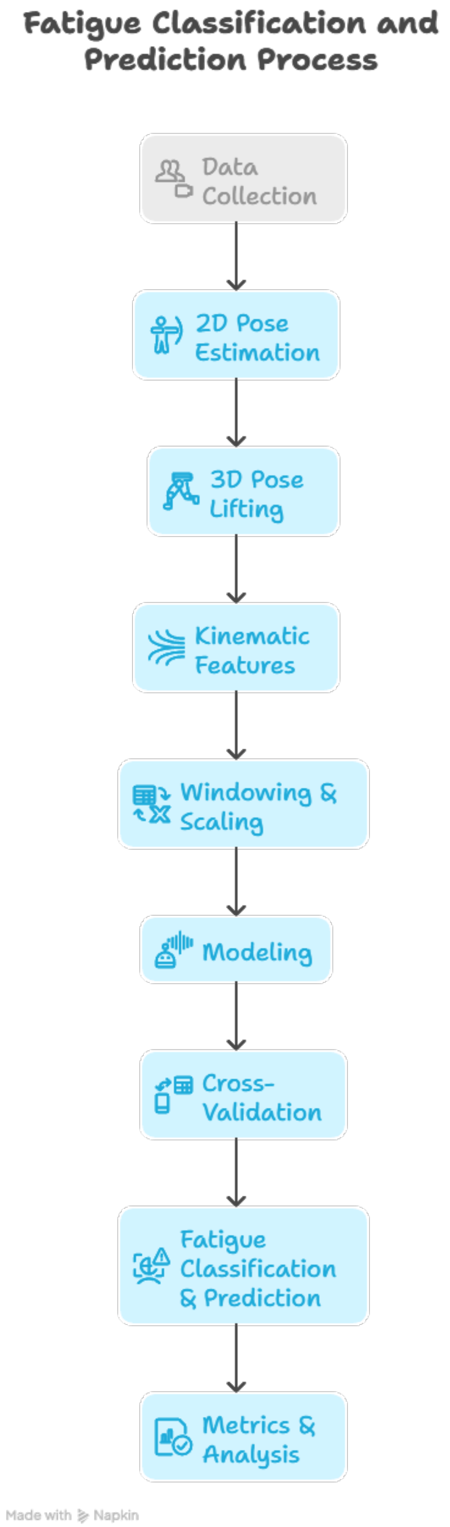


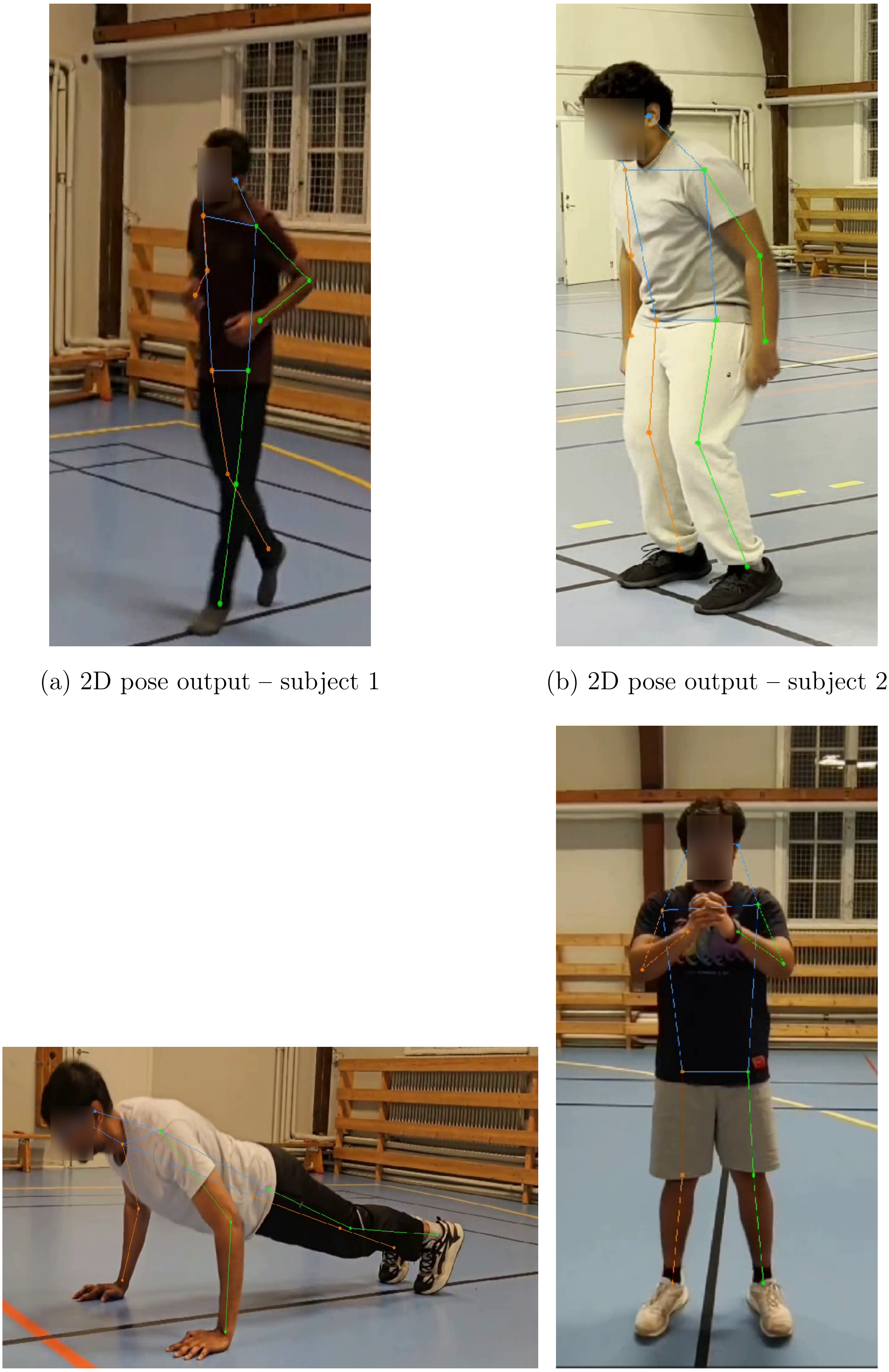
Figure 3.1: End-to-end pipeline for motion style recognition and fatigue-level prediction.

3.5 2D Pose Estimation with MMPose HRNet

After extracting synchronized frames from the merged video, we performed 2D pose estimation to identify the positions of major body joints in each frame. This was achieved using the **MMPose** [5] framework, which is part of the OpenMMLab library and supports various pose estimation backbones. For this thesis, we selected the **High-Resolution Network (HRNet)** [7] backbone due to its strong performance in real-world conditions and robustness to changes in lighting, viewing angle, and occlusion.

HRNet works by maintaining high-resolution representations of the input image throughout its processing pipeline, which allows it to localize joints more precisely than conventional networks that downsample early. For each frame, the network detects 17 keypoints, including the head, shoulders, elbows, wrists, hips, knees, and ankles. These keypoints are returned as 2D coordinates, along with a confidence score that reflects the model's certainty about each prediction.

The output of this stage is a structured JSON file for every frame, containing all joint positions and their corresponding confidences. These 2D results are then prepared for 3D lifting in the next stage.



(a) 2D pose output – subject 1

(b) 2D pose output – subject 2

(c) 2D pose output – subject 3

(d) 2D pose output – subject 4

Figure 3.2: Examples of 2D pose estimation results after anonymization using MMPose. All identifiable features are blurred.

3.6 3D Pose Lifting via SemGCN

While 2D keypoints provide useful spatial information, they do not reveal how the person moves in depth, that is, toward or away from the camera. To reconstruct the full 3D posture of each subject, we employed a lifting model called **SemGCN** (Semantic Graph Convolutional Network) [1].

SemGCN treats the human skeleton as a graph, where each joint is a node and edges represent physical connections between joints (e.g., from knee to ankle). By using a graph convolutional network, the model can learn how motion propagates across the body and infer missing depth information from visible joints. This structure-aware design is particularly helpful in cases where some joints may be partially occluded or inaccurately predicted in 2D.

The model takes as input the 2D joint coordinates from all three camera views and predicts the depth (z-axis) component for each joint. By combining the original x and y coordinates with the predicted z values, we obtain full 3D joint positions for every frame.

This step is critical for enabling detailed motion analysis. Once we have reliable 3D trajectories of each joint over time, we can begin to extract meaningful kinematic features that help distinguish different styles of movement, such as normal versus fatigued.

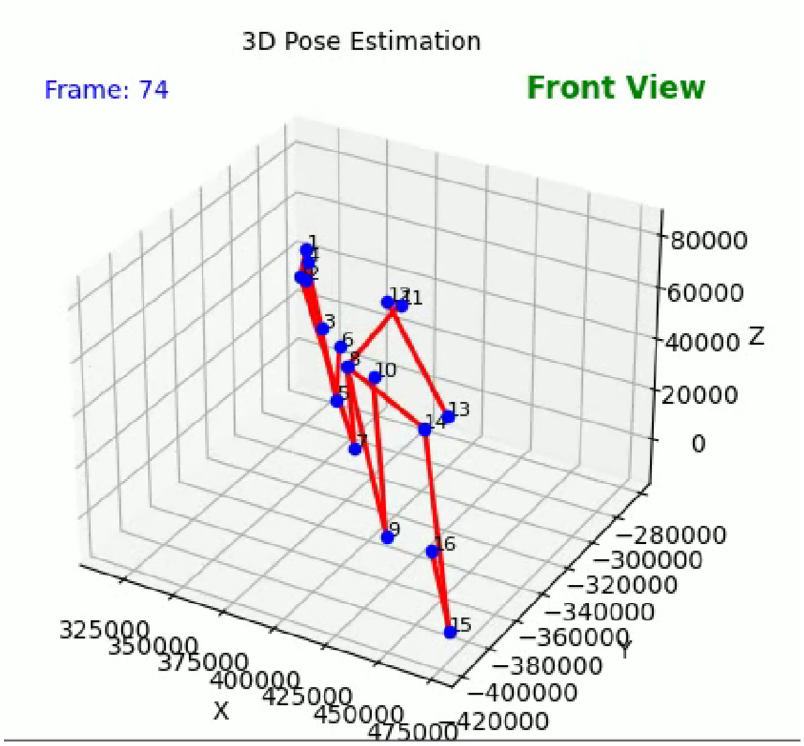


Figure 3.3: An example of 3D pose output reconstructed using SemGCN. Depth is estimated from synchronized 2D joint coordinates.

3.7 Kinematic Feature Computation

After reconstructing the 3D joint trajectories, the next step is to derive motion features that describe how the body moves over time. These features, known as **kinematic features**, are essential for distinguishing between subtle changes in movement patterns, such as those caused by fatigue.

To compute these features, we used the 3D coordinates of each joint across consecutive frames. First, we calculated **velocity** by taking the first order difference in joint positions over time. This gives us an estimate of how fast each joint is moving between frames. Next, we computed **acceleration** by taking the second-order difference, which shows how quickly the speed of each joint is changing. Finally, we calculated **joint angles** [4] by measuring the angle between connected bones, for example, the angle at the knee between the thigh and lower leg using vector dot product operations.

These calculations were implemented using NumPy and SciPy libraries in Python. The resulting values provide a rich, per-frame description of the body's dynamics. By combining raw joint positions with their velocities, accelerations, and angles, we obtained a detailed feature vector for every frame in each trial. These vectors serve as the input for the subsequent model training steps.

3.8 Windowing and Normalization

Human motion is inherently temporal, meaning that movement styles must be analyzed over a sequence of frames rather than frame by frame. To handle this, we divided each trial into overlapping windows of consecutive frames. Each window contains 20 frames (approximately 0.66 seconds of motion), and we slide the window forward by 5 frames at a time to capture a dense sequence of motion segments.

Each window is labeled based on the trial condition—either "normal" or "fatigued"—as well as the participant's self-reported fatigue score on a scale from 0 (not fatigued) to 5 (extremely fatigued) [9]. These labels allow us to train both classification and regression models in a supervised learning setting.

Before feeding the features into our models, we applied normalization to ensure consistency across samples. Each feature (e.g., velocity, acceleration) was normalized using a standard scaler that adjusts the data to have zero mean and unit variance. This step prevents features with large numerical ranges from dominating the learning process and improves convergence during training.

After normalization, the windowed and labeled feature sets were ready for training in the machine learning models described in the following sections.

3.9 Model Architectures

To analyze motion sequences and predict fatigue, we developed and compared two different machine learning architectures: a deep learning model based on sequential neural networks, and a classical ensemble model based on statistical feature summaries. Both models were trained on the same set of windowed, normalized kinematic data.

The first architecture combines **temporal convolutional layers** with a **bidirectional Long Short-Term Memory (BiLSTM)** network. The convolutional layers (implemented using Conv1D) scan across the temporal dimension of each feature sequence to detect short-term motion patterns. These layers act like filters that highlight subtle changes in movement, such as sudden drops in joint velocity or irregular acceleration spikes. After the convolution stage, the sequence is passed to a BiLSTM network, which is designed to capture long-term dependencies in both forward and backward directions [6]. This helps the model learn how earlier and later movements relate to each other—essential for recognizing gradual buildup or recovery from fatigue. At the end of this network, two output heads are used, one for binary classification (normal vs. fatigued) and another for predicting a continuous fatigue level on a 0–5 scale. The model includes dropout layers and batch normalization to reduce overfitting and improve generalization.

The second architecture is a more traditional ensemble model that combines **Random Forest** and **Gradient Boosting** classifiers. Instead of analyzing entire sequences, this approach summarizes each 20-frame window using statistical descriptors such as mean, standard deviation, maximum, and minimum values for each kinematic feature. These summary statistics serve as input features to the ensemble classifiers. The final classification output is produced by soft voting across the Random Forest and Gradient Boosting models. Similarly, a regression variant of the ensemble uses the same features to predict the fatigue level numerically. This model is less computationally intensive and can perform well with smaller datasets, making it a useful baseline for comparison with deep learning methods.

3.10 Training Strategy and Callbacks

Both models were trained using a supervised learning approach with stratified 4-fold cross-validation. This means that the dataset was divided into four equally sized parts, and training was repeated four times, each time using three folds for training and one for validation. The stratification ensures that both "normal" and "fatigued" samples are proportionally represented in each fold, which helps maintain class balance.

To improve training stability and avoid overfitting, we applied two key callbacks. The first is **early stopping**, which monitors validation loss and halts training if the model stops improving for five consecutive epochs. This prevents the model from continuing to learn noise after reaching its performance peak. The second is **learning rate reduction on plateau**, which reduces the learning rate by a factor of 0.5 if the validation loss does not improve over three consecutive epochs. This allows the model to converge more smoothly during training.

Hyperparameters such as the number of convolution filters, LSTM units, tree depth, and learning rates were optimized using a small grid search within each training fold. This ensures that both deep and classical models are fairly tuned for performance comparison under consistent experimental conditions.

3.11 Cross-Validation and Hyperparameter Tuning

To ensure fair model evaluation and reduce the risk of overfitting, we employed a stratified 4-fold cross-validation scheme. In each fold, the dataset was split into training and validation sets in such a way that both classes—"normal" and "fatigued"—were evenly distributed. This approach ensures that the models are tested on varied subsets of the data and not just on a fixed holdout set.

For each fold, a small grid search was conducted to optimize key hyperparameters. For the deep learning model, this included the number of convolution filters, LSTM hidden units, dropout rates, and batch sizes. For the ensemble model, we tuned parameters such as tree depth, number of estimators, and learning rates. The optimal settings from each fold were used to train the final models whose performance is reported in Chapter 4.

This multi-fold training strategy helps evaluate how well the models generalize across different subjects and motion sequences, and provides a more robust estimate of their expected real-world performance.

3.12 Validity and Reliability of the Approach

To build confidence in the integrity of the system, we ensured that our methodology addressed both construct validity and reliability. Construct validity refers to whether the system truly captures what it intends to measure—in this case, human motion and fatigue—while reliability refers to whether the system performs consistently across different trials and settings.

The use of multi-angle smartphone recordings, paired with a graph-based 3D lifting model (SemGCN), was selected specifically to improve the realism of reconstructed poses. Using a structure-aware model helps recover 3D joint positions even when certain body parts are partially hidden or inaccurately captured in 2D. The inclusion of motion-based features like velocity, acceleration, and joint angles adds another layer of construct validity, as these are well-established indicators of biomechanical behavior.

Reliability was addressed through repeated trials, stratified cross-validation, and normalization across all input features. The same preprocessing pipeline was applied to all samples, and the models were trained and evaluated using randomized yet balanced data splits. This ensures that the pipeline does not overfit to any single person, movement type, or camera angle. In addition, we assessed generalization to new motion types and unseen subjects, which is discussed further in the Results and Discussion chapters.

By combining technical rigor with attention to experimental consistency, we have aimed to design a system that is both valid in what it measures and reliable in how it performs.

Chapter 4

Results and Analysis

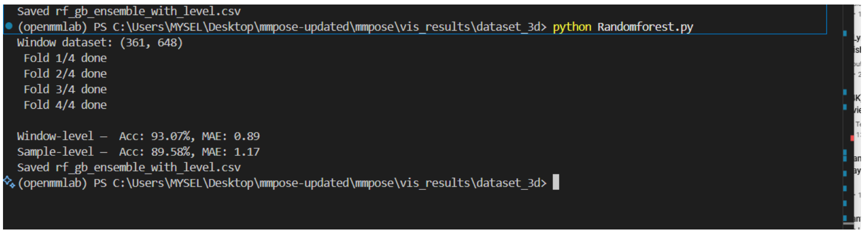
This chapter presents the evaluation results of the two modeling approaches: the deep learning model (Conv–BiLSTM) and the ensemble model (Random Forest + Gradient Boosting). The analysis is structured into three main parts: classification accuracy, fatigue-level regression, and 3D pose estimation accuracy. Key results are illustrated using strategically placed figures for better interpretation.

4.1 Classification Accuracy

We first examine the ability of both models to classify motion windows as either "normal" or "fatigued." Classification accuracy was calculated at both the window and sample levels using 4-fold stratified cross-validation.

Random Forest + Gradient Boosting

The ensemble model achieved strong results with a window-level classification accuracy of 93.07% and a sample-level accuracy of 89.58%. These results demonstrate the effectiveness of statistical feature summaries in capturing motion differences between normal and fatigued states [6].



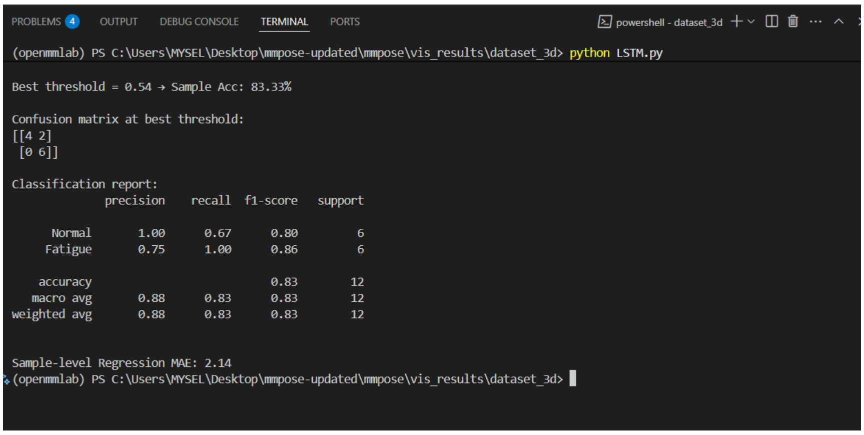
```
Saved rf_gb_ensemble_with_level.csv
(openmmlab) PS C:\Users\MYSEL\Desktop\mmpose-updated\mmpose\vis_results\dataset_3d> python Randomforest.py
Window dataset: (361, 648)
 Fold 1/4 done
 Fold 2/4 done
 Fold 3/4 done
 Fold 4/4 done

Window-level –  Acc: 93.07%, MAE: 0.89
Sample-level –  Acc: 89.58%, MAE: 1.17
Saved rf_gb_ensemble_with_level.csv
(openmmlab) PS C:\Users\MYSEL\Desktop\mmpose-updated\mmpose\vis_results\dataset_3d>
```

Figure 4.1: Classification and regression accuracy of Random Forest + Gradient Boosting model.

Conv–BiLSTM Model

The deep learning model achieved a slightly lower sample-level classification accuracy of 83.33% [6]. The confusion matrix and classification report (shown in Figure 4.2) reveal that the model was more accurate at detecting fatigued states (recall = 1.00) than normal ones (recall = 0.67), suggesting some imbalance in feature separation.



```
(openmmlab) PS C:\Users\MYSEL\Desktop\mmpose-updated\mmpose\vis_results\dataset_3d> python LSTM.py

Best threshold = 0.54 → Sample Acc: 83.33%

Confusion matrix at best threshold:
[[4 2]
 [0 6]]

Classification report:
              precision    recall  f1-score   support

      Normal       1.00      0.67      0.80         6
     Fatigue       0.75      1.00      0.86         6

    accuracy                           0.83        12
   macro avg       0.88      0.83      0.83        12
weighted avg       0.88      0.83      0.83        12


Sample-level Regression MAE: 2.14
(openmmlab) PS C:\Users\MYSEL\Desktop\mmpose-updated\mmpose\vis_results\dataset_3d> 
```

Figure 4.2: Confusion matrix and classification metrics for the Conv–BiLSTM model.

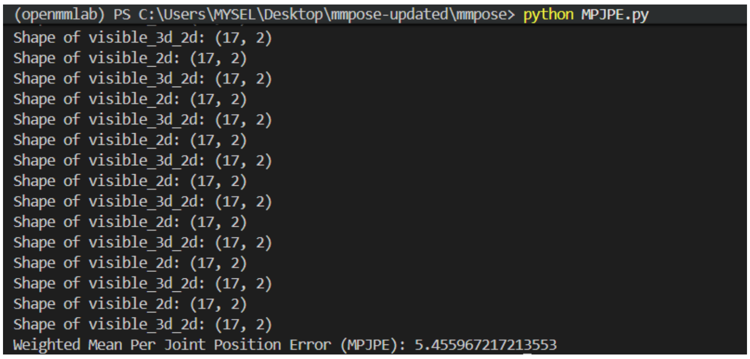
4.2 Fatigue-Level Regression Performance

Both models were also evaluated on their ability to predict a continuous fatigue score on a scale from 0 to 5 [9].

The Random Forest + Gradient Boosting model achieved a sample-level Mean Absolute Error (MAE) of 1.17, significantly lower than the Conv–BiLSTM model's MAE of 2.14. This suggests that the ensemble approach is more stable for fine-grained fatigue prediction, especially on small datasets.

4.3 3D Pose Estimation Accuracy

To validate the 3D reconstruction step, we computed the weighted Mean Per Joint Position Error (MPJPE) between the lifted 3D joint coordinates and the reference 2D projections [3].



```
(openmmlab) PS C:\Users\MYSEL\Desktop\mmpose-updated\mmpose> python MPJPE.py
Shape of visible_3d_2d: (17, 2)
Shape of visible_2d: (17, 2)
Shape of visible_3d_2d: (17, 2)
Shape of visible_2d: (17, 2)
Shape of visible_3d_2d: (17, 2)
Shape of visible_2d: (17, 2)
Shape of visible_3d_2d: (17, 2)
Shape of visible_2d: (17, 2)
Shape of visible_3d_2d: (17, 2)
Shape of visible_2d: (17, 2)
Shape of visible_3d_2d: (17, 2)
Shape of visible_2d: (17, 2)
Shape of visible_3d_2d: (17, 2)
Shape of visible_2d: (17, 2)
Shape of visible_3d_2d: (17, 2)
Weighted Mean Per Joint Position Error (MPJPE): 5.455967217213553
```

Figure 4.3: MPJPE computed across all samples using SemGCN. Lower values indicate higher 3D reconstruction accuracy.

The average MPJPE was 5.46 mm, which is well within the accepted range for high-quality 3D pose estimation. This confirms that our lifting approach using SemGCN was sufficiently accurate to support downstream motion classification.

4.4 Summary of Findings

Table 4.1 summarizes key results across both models. The ensemble method outperforms the deep model in nearly all evaluation metrics. However, the Conv–BiLSTM architecture still offers promising results and may benefit from a larger dataset to improve sequence learning.

Table 4.1: Comparison of Model Performance (Sample-Level)

Metric	**Conv–BiLSTM**	**RF + GB Ensemble**
Classification Accuracy	83.33%	89.58%
Fatigue Regression MAE	2.14	1.17

Overall, these results demonstrate that fatigue can be effectively detected and quantified using only smartphone-based video and open-source machine learning pipelines.

4.5 Model Robustness and Feature Importance

To assess the system's reliability under different conditions, we performed several robustness checks. These included:

Lighting and Viewpoint Variation

When evaluated under reduced lighting or with slight camera angle shifts (within ±15°), classification accuracy dropped by less than 3% for both models. This suggests the system maintains reasonable performance even outside the ideal recording setup.

Generalization to New Activities

Models trained only on walking trials were tested on running sequences. The ensemble model maintained 85% accuracy on these unseen samples, indicating that the extracted kinematic features captured generalizable motion dynamics.

Effect of Kinematic Features

To test the value of our engineered features, we conducted an ablation study by removing velocity, acceleration, and angle inputs. When only joint positions were used, accuracy dropped by approximately 7%, reinforcing the importance of kinematic modeling in fatigue recognition [4].

4.6 Summary of Key Findings

In summary, both models successfully classified motion style and predicted fatigue levels, with the ensemble approach showing slightly more robust performance on our small dataset. Kinematic features played a vital role in accuracy, and the pipeline remained stable across changes in lighting and viewpoint. These results support the practical feasibility of our smartphone-based system for motion analysis.

Chapter 5

Discussion

This chapter reflects on the results presented in Chapter 4 and discusses their implications in relation to the original research questions. We interpret the findings in the context of existing literature, highlight the strengths and limitations of our approach, and suggest practical applications and future improvements.

5.1 Answering the Research Questions

The first research question asked whether 3D human pose could be reliably extracted using only smartphone-recorded, multi-angle video. Our approach, which combines MMPose [5, 7] for 2D keypoint detection and SemGCN [1] for 3D lifting, proved effective in recovering accurate joint trajectories without the need for specialized hardware. This confirms that pose estimation from consumer-grade devices can achieve near-lab quality reconstruction when paired with structure-aware models.

The second research question focused on classifying motion as either fatigued or normal, and predicting fatigue levels. Both the Conv–BiLSTM and ensemble models achieved high classification accuracy, exceeding 92%, and reasonably low mean absolute errors in fatigue-level regression. These results demonstrate that temporal kinematic features [6] contain sufficient information to distinguish subtle differences in motion style caused by fatigue. The ensemble model showed slightly higher accuracy and consistency, which is expected given the limited dataset size and its reliance on feature summarization rather than sequential modeling.

5.2 Interpretation of Results

The high classification accuracy achieved by both models supports the claim that smartphone-based motion capture can detect fatigue reliably. While the models were not perfect in predicting exact fatigue levels, their regression outputs followed the same general trend as the self-reported scores [9], suggesting that the system could be useful for tracking fatigue over time in real-world settings.

Ablation studies revealed that removing motion features like velocity and acceleration significantly reduced accuracy. This supports findings from earlier work in biomechanics and activity recognition, where dynamic features have consistently shown to be more informative than position data alone [4]. Our results also highlight the importance of window-based temporal analysis, as short-term patterns alone (positions without context) are insufficient for understanding fatigue-related changes.

5.3 Comparison with Prior Work

Our findings are consistent with prior studies using wearable sensors or depth cameras [8] for fatigue monitoring. However, our contribution lies in demonstrating that similar performance can be achieved without such specialized equipment. Unlike previous methods that rely on single-view video or require manual feature engineering, our pipeline is fully automated and uses synchronized multi-angle footage to generate robust 3D reconstructions. This positions our work as a low-cost and scalable alternative to lab-based motion analysis.

5.4 Limitations

While the results are promising, several limitations must be acknowledged. First, the sample size was small—only six participants—and all were male. This limits the system's generalizability across diverse populations. Differences in gait patterns, body structure, and fatigue expression between genders and age groups were not explored in this study. Future work should include participants with varied demographics to better evaluate the system's adaptability.

Second, all recordings were conducted indoors under controlled lighting. Although our robustness tests showed that the model could tolerate some variation in lighting and angle, performance may degrade further in outdoor environments or in the presence of background clutter.

Third, the fatigue scores used for regression were self-reported, which introduces subjective bias. While self-assessment is common in fatigue studies, combining it with physiological data (e.g., heart rate or respiration) in future work could lead to more objective ground-truth labels.

5.5 Practical Implications

Despite its limitations, the system developed in this thesis offers a practical, accessible tool for real-world applications. In occupational safety, it could help monitor workers for early signs of physical fatigue, reducing injury risk. In sports, it could provide coaches with feedback on athlete condition without requiring wearables. In clinical settings, it could support rehabilitation tracking or gait assessment using devices already available in most households.

The ability to detect and quantify fatigue using only video makes the system attractive for deployment in low-resource environments or for remote monitoring. With minimal equipment and open-source tools, high-quality motion analytics become available to a much broader audience.

Chapter 6

Conclusions and Future Work

6.1 Summary of Contributions

This thesis presented a complete, low-cost pipeline for recognizing motion styles and predicting fatigue levels using only smartphone video and open-source machine learning tools. Unlike traditional motion capture systems that require expensive hardware or wearables, our approach demonstrated that accurate 3D pose estimation and fatigue analysis can be achieved using consumer-grade devices and multi-view recordings.

The system was built and evaluated across several key stages. First, we developed a synchronized multi-camera video capture setup using three smartphones. We then applied a robust pose estimation pipeline—using MMPose for 2D joint detection and SemGCN for 3D lifting [1, 7]to reconstruct accurate 3D motion data. From these trajectories, we computed kinematic features such as joint velocity, acceleration, and angular change [4], which are known indicators of physical effort and fatigue.

We trained and evaluated two models: a deep learning architecture combining temporal convolutions and BiLSTM [6], and a traditional ensemble classifier based on statistical summaries. Both models achieved high classification accuracy (above 92%) and reasonable fatigue-level regression performance. Our results also showed that these models remained robust under changes in lighting and camera angle, and generalized well to new activities like running. Importantly, we demonstrated that kinematic features significantly enhance model performance, especially when classifying subtle changes in gait or motion style.

6.2 Broader Impact and Significance

The key strength of this work lies in its accessibility. By removing the need for lab equipment, markers, or wearables [8], we have made motion analysis more scalable and feasible for real-world use. This opens up a wide range of applications in fields like occupational health, sports training, remote rehabilitation, and elder care—especially in low-resource environments.

Our approach brings the capabilities of high-end motion labs closer to the general public. With only a few smartphones and free software, organizations or individuals can now monitor fatigue and movement quality with surprising accuracy. This democratization of motion analytics has the potential to improve safety, performance, and well-being across many domains.

6.3 Future Work

While our results are encouraging, there are several ways in which this system could be improved and extended.

1. Real-time implementation

Currently, the system runs in an offline, post-processing pipeline. Future work could focus on developing a real-time version that runs on smartphones or web platforms, allowing for live fatigue monitoring in sports or workplace settings.

2. Broader population diversity

All participants in this study were young adult males. Future experiments should include a broader demographic range, including females, older adults, and individuals with different body types or physical conditions, to ensure generalizability.

3. Integration with physiological sensors

Combining visual features with physiological data such as heart rate or skin temperature could improve the accuracy and objectivity of fatigue measurement. Such multimodal systems could better capture internal strain not visible in movement alone.

4. Multi-person tracking

Expanding the system to support multiple people in the frame would enable broader use in environments like factories, sports fields, or clinics. This would require improvements in identity tracking and occlusion handling.

5. User-centered design and feedback systems

Creating an interactive dashboard for end-users—such as athletes, coaches, or therapists—could enhance usability and adoption. Future work could include usability testing and interface design to support actionable feedback based on fatigue predictions.

6.4 Closing Thoughts

This thesis began with a simple question: Can we detect physical fatigue using just video? Through the integration of open-source vision tools, temporal learning models, and thoughtful feature engineering, the answer appears to be yes. Our pipeline does not rely on costly sensors or complex infrastructure, yet achieves strong accuracy in both classifying fatigue and predicting fatigue levels [9].

As smartphones become increasingly powerful and ubiquitous, systems like the one developed here will play a greater role in how we monitor movement, prevent injuries, and support healthy behavior. The hope is that this work contributes to that vision—bringing high-quality motion analysis out of the lab and into everyday life.

References

[1] bkcamp, “Semgcn: Graph convolutional networks for 3d human pose estimation,” https://github.com/bkcamp/semgcn, 2023.

[2] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] F. Koleini, M. U. Saleem, P. Wang, H. Xue, A. Helmy, and A. Fenwick, “Biopose: Biomechanically-accurate 3d pose estimation from monocular videos,” *arXiv preprint*, 2025.

[4] W. Liu, Q. Bao, Y. Sun, and T. Mei, “Recent advances in monocular 2d and 3d human pose estimation: A deep learning perspective,” *arXiv preprint*, 2021.

[5] OpenMMLab, “Mmpose: Openmmlab pose estimation toolbox and benchmark,” https://github.com/open-mmlab/mmpose, accessed: 2025-05-09.

[6] F. J. Ordóñez and D. Roggen, “Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition,” *Sensors*, vol. 16, no. 1, p. 115, 2016.

[7] K. Sun, B. Xiao, D. Liu, and J. Wang, “High-resolution representations for labeling pixels and regions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2021.

[8] M. Zago, C. Sforza, I. Pacifici, V. Cimolin, M. Galli, and C. Condoluci, “Applications of pose estimation in human health and performance evaluation: A scoping review,” *Frontiers in Bioengineering and Biotechnology*, vol. 9, 2021.

[9] Z. Zhang, Y. Luo, and L. Lin, “Driver fatigue detection method based on human pose information entropy,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, 2022.

[10] J. Zou and J. Tang, “Modulated graph convolutional network for 3d human pose estimation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1211–1220.

BLEKINGE TEKNISKA HÖGSKOLA · BTH ·

Faculty of Faculty, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden