EJUST-GYM-3: A Multi-Modal Dataset and Dual-Task Framework for Exercise Recognition and Performance Evaluation Using RGB, Thermal, and IMU Sensors

Ahmed Attia, Mazen Seif, Mohamed Mohy, Ali Ismaeel, Yahia Ali
Egypt-Japan University of Science and Technology (E-JUST), Alexandria, Egypt

***Abstract*—Human Activity Recognition (HAR) is a widely explored field with applications spanning healthcare, sports analytics, and fitness monitoring [1], [2]. While prior studies have investigated HAR in sport-related activities [3] and technique analysis [4], there remains a need for datasets that combine multi-modal data with real-time performance evaluation. To address this, we introduce EJUST-GYM-3, a large-scale, multi-modal dataset that supports both exercise classification and performance assessment.**

The dataset contains synchronized recordings of four common gym exercises—steam engine, jumping jacks, sit-ups, and high knees—performed by 128 participants. RGB cameras (frontal and lateral), a thermal camera, and four IMU sensors positioned on the limbs were used to capture diverse perspectives of motion. Additionally, participant metadata such as height, weight, sleep duration, and lifestyle factors was collected to enable personalized modeling [6].

A dual-task deep learning pipeline was developed. The first model uses bidirectional LSTM networks for exercise classification based on 3D pose sequences. The second model estimates a percentage-based correctness score for each repetition. The system was deployed on a mobile application using TensorFlow Lite for on-device inference. EJUST-GYM-3 serves as a valuable resource for developing AI-driven fitness coaching systems and benchmarking multi-modal HAR models.

***Index Terms*—Index Terms — Human Activity Recognition, Exercise Classification, Pose Estimation, Deep Learning, Mobile Inference, TensorFlow Lite, Multi-modal Dataset**

I. Introduction

Human Activity Recognition (HAR) has emerged as a pivotal area of research within artificial intelligence, playing a critical role in a diverse array of applications spanning healthcare monitoring, physical rehabilitation, elderly care, and fitness technology platforms [1], [2], [6]. With the growing adoption of smart wearables, home-based exercise regimes, and virtual coaching systems, the need for accurate, real-time recognition of human physical activities in unconstrained, real-world environments has intensified significantly [7].

Despite advancements in computer vision and sensor technologies, current HAR systems often face limitations that constrain their utility and generalizability. These include restricted dataset diversity, reliance on unimodal data sources (e.g., solely vision or inertial sensors), and the lack of end-to-end systems capable of deploying HAR models efficiently on resource-limited edge devices [3], [4], [14]. To bridge these gaps and push the boundaries of HAR research, we present EJUST-GYM-3, a comprehensive, large-scale, and multi-modal dataset explicitly designed for exercise recognition in naturalistic gym-like settings.

Collected at the Egypt-Japan University of Science and Technology (E-JUST), this dataset captures the activities of 128 diverse participants as they perform four foundational calisthenic exercises: steam engine, jumping jacks, sit-ups, and high knees. What distinguishes EJUST-GYM-3 from existing benchmarks is its multi-sensor setup, which synchronously records activity data using two RGB cameras (frontal and lateral views), a thermal imaging camera, and four Inertial Measurement Units (IMUs) strategically positioned on the participant's limbs. This multi-angle, multi-modal configuration enriches the dataset by offering complementary data streams for robust activity understanding under varying conditions.

Furthermore, to facilitate personalized HAR research, detailed metadata was collected, including anthropometric measurements (e.g., height, weight), demographic information (e.g., age), and lifestyle attributes. This auxiliary information is crucial for developing adaptive and personalized HAR models that can account for individual variability in movement patterns and performance [6].

Beyond the dataset itself, we introduce a novel deep learning pipeline specifically tailored for exercise classification using pose-based temporal features. The proposed framework leverages 3D keypoint trajectories extracted from RGB video recordings. Each activity repetition is represented as a fixed-length sequence of 30 frames, where each frame consists of 33 anatomical keypoints annotated with (x, y, z) coordinates. These sequences are processed using Long Short-Term Memory (LSTM) network, enabling the model to capture temporal dependencies and spatial dynamics inherent in human movement. The model, implemented in TensorFlow, achieved robust classification performance and exhibited strong generalization to unseen subjects, a critical requirement for real-world applicability.

To ensure practical usability and real-time deployment, the

This work was supported by the Information Technology Industry Development Agency (ITIDA), Egypt.

trained network was optimized and converted to TensorFlow Lite, facilitating on-device inference on mobile platforms. This allows end-users to record exercise routines with standard mobile devices and receive instantaneous feedback through offline classification. Unlike previous approaches that primarily focus on visual recognition or segment-level classification [8], [9], our system operates on structured pose sequences stored as CSV files and supports single-clip predictions using efficient majority voting strategies. The framework also accommodates inference on unseen samples through a compact .h5 model and an accompanying label encoder, enabling lightweight integration into various HAR-enabled applications.

Through the integration of multi-modal sensing, efficient preprocessing pipelines, deep temporal modeling, and mobile-friendly deployment, the EJUST-GYM-3 initiative not only provides a rich benchmark for HAR but also serves as a robust foundation for building the next generation of context-aware, scalable, and personalized activity recognition systems.

II. Related Work

A. Thematic Organization

The field of Human Activity Recognition (HAR) has progressed rapidly in recent years, especially in domains like fitness monitoring, rehabilitation, and automated coaching. Emerging trends reflect a shift from unimodal systems and controlled laboratory studies to *multi-modal*, *context-rich*, and *real-world deployable* systems. This section systematically reviews the progress and gaps across four interrelated themes:

- Multimodal Sensing and Synchronization
- Temporal Deep Learning for Exercise Understanding
- Exercise-Specific and Personalized HAR Datasets
- Real-Time and Edge Deployment of HAR Systems

The **EJUST-GYM-3** dataset—capturing synchronized RGB, thermal, and IMU signals across four common exercises performed by 128 individuals—addresses several critical limitations in prior studies. Below, we position it in the context of state-of-the-art contributions.

B. Recent Developments (Last 5 Years)

1) Multimodal Sensing and Synchronization: Recent HAR research has increasingly favored multimodal sensing, leveraging the complementary strengths of vision-based and inertial sensing technologies. The MoVi dataset by Ghorbani et al. [12] stands out for combining motion capture, video, and IMUs to record everyday activities and sports motions. Despite its diversity, MoVi lacks synchronized thermal imaging and participant lifestyle metadata—both of which are critical for real-world modeling in exercise contexts.

Similarly, the ValS dataset [13] focuses on squats, combining IMUs and video to support applications in gender recognition and gym training. However, its narrow focus on a single exercise limits generalizability to other movements.

In contrast, **EJUST-GYM-3** includes four varied calisthenic exercises—steam engine, sit-ups, jumping jacks, and high knees—captured via synchronized RGB cameras (frontal and lateral views), a thermal camera, and four limb-mounted IMUs. Unlike previous datasets, thermal data adds an infrared modality to handle low-light conditions or thermographic fatigue analysis, as explored in recent thermal-based HAR systems [7].

The dataset structure—organized per participant, per exercise, and per modality—ensures high synchronization fidelity, allowing models to learn from temporally aligned sensor streams. This enables researchers to study sensor fusion more effectively than in loosely structured multimodal datasets like MM-Fit [14].

2) Temporal Deep Learning for Exercise Understanding: The complexity of physical exercises requires models that can capture both spatial posture and temporal dynamics. Traditional convolutional models have proven limited for these tasks. Instead, recurrent architectures, especially LSTMs and Temporal Convolutional Networks (TCNs), have become dominant in capturing the temporal evolution of human motion.

In the context of squats, Ogata et al. [15] proposed temporal distance matrices to distinguish between correct and incorrect forms. While innovative, their approach is limited to squats and binary classification. Similarly, GymCam [15] uses CNNs on RGB frames for repetition counting but lacks interpretability regarding movement correctness.

EJUST-GYM-3 improves upon these works by:

- Encoding exercise repetitions as pose sequences of 30 frames, each frame containing 33 3D keypoints, enabling fine-grained motion tracking.
- Training a LSTM classifier on these sequences to recognize exercise type and assess correctness with a percentage score.
- Leveraging multiple modalities to enrich sequence modeling: video for spatial features, IMUs for accelerative force, and thermal data for physiological patterns.

This dual-task learning architecture—classification plus correctness estimation—is an advancement over prior unimodal or single-objective HAR models.

3) Exercise-Specific and Personalized HAR Datasets: Most HAR datasets focus on daily living activities such as sitting, standing, walking, and lying [16]. However, the specificity of sports-related exercises, and their impact on muscle groups and posture, necessitates targeted datasets.

Datasets like ValS [13] and Squat-Var [15] focus exclusively on squats. The MM-Fit dataset [14] includes 21 exercises but is limited to 10 participants and lacks participant diversity or lifestyle metadata.

EJUST-GYM-3 is distinct in several ways:

- It includes 128 participants, capturing significant variability in body dimensions and movement execution.
- It integrates participant metadata such as age, height, weight, sleep duration, smoking status, and dietary habits. This enables personalized HAR, a growing subfield where models adapt to user-specific biomechanics and lifestyle factors [6].
- Exercises were recorded naturally, in semi-controlled environments (gym and dorms), without externally enforced

synchronization or exaggerated postures. This supports the training of models intended for deployment in home and gym environments, where movement variability is high.

By combining multi-sensor input with detailed participant profiles, **EJUST-GYM-3** offers a foundation for studies in adaptive HAR, error-aware coaching, and behavioral movement analysis—an area currently underserved in public datasets.

4) Real-Time and Edge Deployment of HAR Systems: A significant challenge in HAR research is the translation from lab models to deployable systems. While Kinect-based systems [16] enable near real-time recognition using SDK-acquired skeletons, they rely on expensive, immobile hardware and struggle in varied lighting conditions.

Systems like GymCam [15] attempt real-time recognition from RGB videos, achieving a 93.6% accuracy in repetition counting. However, GymCam cannot evaluate correctness of repetitions and lacks support for on-device inference.

EJUST-GYM-3 is explicitly designed with deployment in mind:

- The pipeline includes model conversion to TensorFlow Lite, enabling offline inference on mobile devices.
- The model consumes CSV-based pose sequences, making it compatible with standard video + OpenPose workflows or IMU-only setups.
- A lightweight `.h5` file and label encoder facilitate easy deployment across platforms.

Incorporating thermal data also enables infrared-guided feedback systems in poorly lit environments or for users who prefer privacy during training.

Together, these attributes position **EJUST-GYM-3** not only as a dataset but as a reference platform for real-world HAR systems that combine robustness, efficiency, and personalization.

III. Dataset: EJUST-GYM-3

The EJUST-GYM-3 dataset was developed at the Egypt-Japan University of Science and Technology to support multi-modal human activity recognition (HAR) and personalized exercise evaluation. It includes synchronized data from 128 participants performing four exercises: steam engine, sit-ups, jumping jacks, and high knees. Data was collected using RGB cameras, thermal imaging, and wearable inertial sensors.

A. Sensor Modalities and Setup

Each participant was recorded using three synchronized sensing modalities:

- **RGB Video:** Two smartphones (Honor 20 and Mi A3) recorded frontal and side views at 1280×720 resolution and 30 FPS.
- **Thermal Video:** FLIR One Pro thermal camera (160×120, 8.7 FPS).
- **IMUs:** Four MetaMotion units (wrists and ankles) recording 9-axis motion and environmental data at 50 Hz.

Sensors were synchronized via global timestamps, and alignment was aided by a synchronization beep sound. RGB recordings included on-screen timestamps; thermal videos were manually trimmed.

IV. Visual Dataset Overview

The EJUST-GYM-3 dataset includes synchronized multi-view recordings of physical exercises using three visual modalities: a frontal RGB view, a lateral (side) RGB view, and a thermal infrared view. Each modality provides a unique perspective on movement quality and body dynamics, allowing comprehensive human activity recognition (HAR) and performance evaluation. The RGB cameras were positioned to capture anatomical symmetry and motion patterns, while the thermal camera was used to extract physiological signals such as body temperature variation during exercises.

A. Participant Demographics and Metadata

To support personalized modeling, metadata such as age, height, weight, sleep duration, and lifestyle indicators were collected for each participant. This enables researchers to adapt HAR models to subject-specific attributes and physical variability.

TABLE I
Participant Metadata Summary

Attribute	**Minimum**	**Average**	**Maximum**
Height (cm)	160	178.68	197
Weight (kg)	50	76	120
Daily Sleep (hrs)	4	7.25	12
Age (years)	18	20.4	34

Smoking Status	**Number**	**Percentage**
Smokers	16	14.3%
Non-smokers	108	85.7%

Dietary Status	**Number**	**Percentage**
Healthy Eaters	56	45.5%
Non-healthy Eaters	70	55.5%

B. Frontal and Side RGB Views

The RGB recordings were captured using two smartphones placed at the front and left side of the subject. The frontal view captures full-body visibility for tracking joint alignments, particularly useful in exercises involving symmetry (e.g., jumping jacks). The side view is essential for observing motion trajectories, angles, and body posture during dynamic movements such as sit-ups and steam engines. These two views are synchronized frame-by-frame, enabling 3D posture reconstruction when combined.

C. Thermal Camera View

Thermal imaging was integrated using the FLIR One Pro thermal sensor to provide additional information related to heat distribution across the body during exercise. Unlike RGB views, thermal imaging can be utilized in low-light conditions and offers privacy-preserving monitoring. It is particularly beneficial for evaluating exertion and fatigue, as

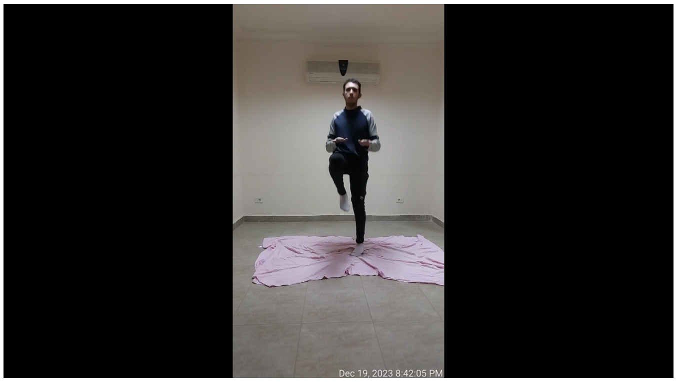

Fig. 1. Frontal RGB View

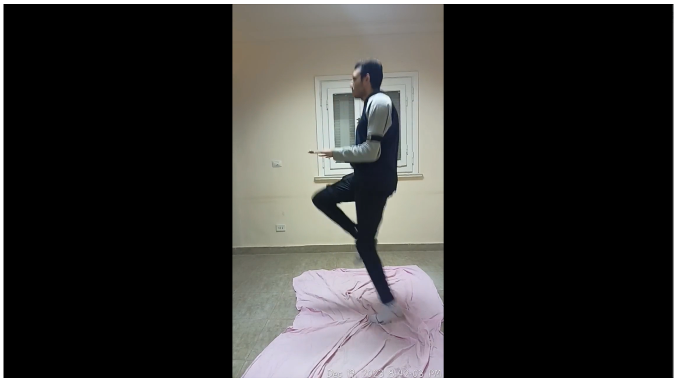

Fig. 2. Side RGB View

body temperature patterns often vary with physical intensity and physiological load.

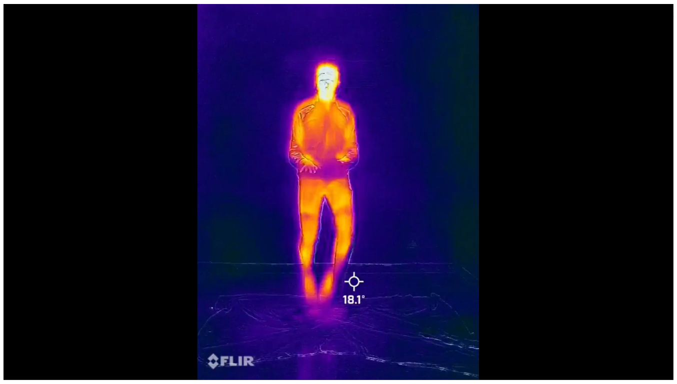

Fig. 3. Thermal View

D. Data Files and Formats

Each recorded data instance in the EJUST-GYM-3 dataset is stored using a structured naming convention to facilitate traceability and automation. The file naming format is as follows:

```
<firstname-lastname>-
<exercise-abbreviation>-<modality>.csv
```

The dataset includes the following components:

- Over 5000 synchronized video clips across frontal, side, and thermal modalities.
- CSV files containing raw IMU sensor readings (accelerometer, gyroscope, magnetometer) for each participant and exercise.
- Annotated metadata files with demographic and lifestyle details for every subject.

Kinect V2 was used for a subset to capture 25 3D skeletal joint coordinates at 30 FPS, synchronized with video and IMU streams.

E. Data Processing

Following the collection of synchronized thermal, RGB, and IMU data for each exercise session, we implemented a series of processing steps to extract meaningful features and enable advanced modeling.

The key stages of the data processing pipeline are as follows:

1) **Thermal Frame Conversion to Gait Energy Images (GEIs):** Each thermal video was converted into a Gait Energy Image, a static 2D representation that encodes the average energy (intensity) of motion over time. This technique is especially effective in highlighting consistent movement patterns and spatial heat distributions during repetitive activities such as jumping jacks or high knees.
2) **IMU-Based Pose Tracking:** Simultaneously recorded IMU signals from the four wearable sensors (wrists and ankles) were used to extract temporal motion dynamics. These readings include accelerometer, gyroscope, and magnetometer data, enabling precise localization of joint-level activity throughout the exercise.
3) **Pose Estimation and Alignment:** 2D and 3D poses were estimated using pose-tracking models on RGB frames. These pose sequences were aligned with corresponding thermal frames and IMU signals based on global timestamps.
   Figure 4 shows an example of a detected pose using RGB-based keypoint extraction.
4) **Segmentation of Thermal Frames:** The resulting GEIs and thermal videos were segmented by repetition using both IMU peaks and pose sequence changes. This segmentation isolates each exercise repetition into a clean sequence, allowing downstream tasks like correctness scoring, per-rep heat pattern comparison, and fatigue estimation.
   A sample segmented thermal image sequence is shown in Figure 5.

These multimodal steps enable a rich, temporally aligned feature space where visual, inertial, and thermal cues can be jointly modeled. This allows us to move beyond basic activity recognition into domains like personalized feedback, health monitoring, and low-light exercise coaching.

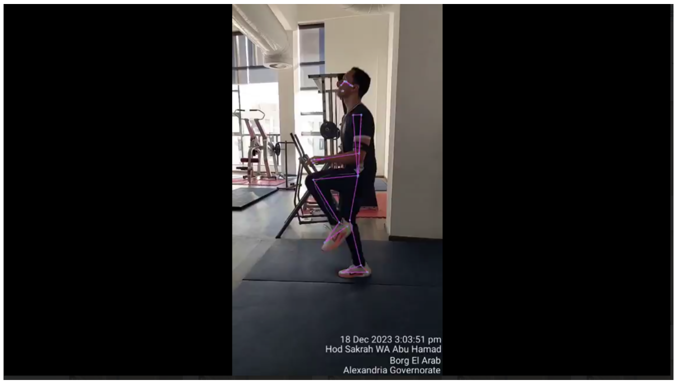

Fig. 4. Example of pose estimation output showing detected anatomical keypoints overlaid on the RGB frame.

Fig. 5. Segmented thermal frames from an exercise clip showing individual repetition boundaries.

V. Methodology

This section outlines the end-to-end methodology adopted for building our smart fitness coaching system based on the EJUST-GYM-3 dataset. Our approach was designed to support two primary tasks: (1) exercise classification and (2) performance evaluation. We employ a multi-modal architecture that leverages RGB video (frontal and lateral views), thermal imaging, and inertial signals from four wearable IMUs. The entire pipeline—from data acquisition to on-device deployment—was constructed with efficiency, robustness, and real-time responsiveness in mind.

A. System Overview

The overall system pipeline consists of four main stages:

1) **Pose Extraction:** Raw RGB videos are processed to extract 3D human pose keypoints using the MediaPipe Pose model.
2) **Thermal and IMU Feature Extraction:** Thermal videos are converted into Gait Energy Images (GEIs), and IMU readings are parsed into structured time-series signals.
3) **Deep Sequence Modeling:** Pose sequences are used to train Bidirectional LSTM (BiLSTM) networks for exercise recognition and correctness estimation.
4) **Mobile Deployment:** The trained models are optimized using TensorFlow Lite and deployed for real-time inference on mobile devices.

B. Pose-Based Feature Engineering

Pose data serves as the primary modality for both classification and assessment tasks. Using MediaPipe Pose, we extract 33 anatomical keypoints per frame, each represented by (x, y, z) coordinates, resulting in 99-dimensional vectors per frame. Each exercise repetition is encoded as a fixed-length sequence of 30 frames, leading to an input tensor of shape $(30, 99)$ per sample. Pose sequences are globally normalized and cleaned to eliminate outliers and maintain spatial consistency across frames. This enables the learning model to capture invariant motion dynamics across different participants.

C. Thermal Image Processing

Thermal recordings were captured using a FLIR One Pro sensor at 8.7 FPS. Instead of relying on frame-by-frame modeling, we transform each thermal video segment into a Gait Energy Image (GEI), which encodes the average pixel intensity across frames. GEIs provide a compact yet informative representation of motion and physiological exertion.

For segmentation, thermal videos were aligned with IMU and RGB pose sequences using timestamp synchronization and segmentation beeps. Repetition-level slicing was achieved by detecting motion peaks in IMU signals and discontinuities in pose trajectories. Each segmented GEI corresponds to a single repetition and is used to complement pose-based models during correctness evaluation.

D. IMU Signal Processing

Four MetaMotion IMUs were placed on the wrists and ankles of each participant. Each device recorded 9-axis data (accelerometer, gyroscope, magnetometer) at 50 Hz. IMU signals were segmented into windows of 2 seconds (100 time steps) with 50% overlap, resulting in input tensors of shape $(100, 12)$.

These signals provide fine-grained information about joint-level motion intensity, orientation, and frequency, especially useful for exercises like high knees or jumping jacks where limb acceleration plays a key role. Although the main classification model relies on pose sequences, IMU features were used to support segmentation and correctness modeling in fusion experiments.

E. Model Architecture and Implementation

To classify exercise types based on temporal pose information, we employ a deep neural network architecture designed to learn from sequential 3D keypoint data. The primary goal of the model is to capture the temporal evolution of human body posture during physical activity and predict the corresponding exercise class. For this purpose, we utilize a **Bidirectional Long Short-Term Memory (BiLSTM)** network, followed by fully connected layers and a softmax classifier. This architecture is well-suited for recognizing structured and repetitive human movements such as calisthenic exercises.

1) Architectural Rationale: Traditional Recurrent Neural Networks (RNNs) often struggle with long-term dependencies and unidirectional sequence processing, which limits their ability to capture full contextual patterns in time-series data. In contrast, our choice of a BiLSTM enables the model to process the sequence in both forward and backward directions, effectively utilizing the entire temporal context of the exercise repetition. This is critical for tasks like exercise classification, where movements can have overlapping characteristics at the beginning but differ substantially by the end.

The input to the model consists of sequences of 3D pose keypoints extracted using OpenPose from synchronized RGB video. Each frame in the sequence is represented by 33 joints, each with (x, y, z) coordinates, yielding a total of 99 features per frame. Sequences are either padded or truncated to a fixed length of $T = 30$ frames to ensure uniform input dimensions. The complete input for a sample is thus:

$$\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_T\}, \quad \mathbf{p}_t \in \mathbb{R}^{99}, \quad \mathbf{P} \in \mathbb{R}^{T\times 99} \tag{1}$$

2) Network Layers: The implemented model architecture is composed of the following components:

- **BiLSTM Layer (256 units)**: Processes the input sequence in both forward and backward directions and outputs the concatenated hidden state:

$$\mathbf{h}_t = \left[\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t\right] \tag{2}$$

- **Dense Layers with ReLU Activation and Dropout**:

$$\mathbf{z}_1 = \text{ReLU}(\mathbf{W}_1\mathbf{h} + \mathbf{b}_1), \quad \text{Dropout}(0.4) \tag{3}$$

$$\mathbf{z}_2 = \text{ReLU}(\mathbf{W}_2\mathbf{z}_1 + \mathbf{b}_2), \quad \text{Dropout}(0.4) \tag{4}$$

- **Output Layer (Softmax Classifier)**:

$$\hat{\mathbf{y}} = \text{Softmax}(\mathbf{W}_3\mathbf{z}_2 + \mathbf{b}_3), \quad \hat{\mathbf{y}} \in \mathbb{R}^C \tag{5}$$

3) Loss Function and Optimization: The model is trained using the *sparse categorical cross-entropy loss*, given as:

$$\mathcal{L}_{\text{CE}} = -\log(\hat{y}_c) \tag{6}$$

where $\hat{y}_c$ is the predicted probability for the correct class c.

Training is performed using the Adam optimizer, with a batch size of 32 and up to 100 epochs. Early stopping (patience = 15), learning rate scheduling (ReduceLROnPlateau), and model checkpointing are employed to improve generalization.

4) Data Handling: Pose sequences are normalized and validated before being passed to the model. Minor augmentation is applied to simulate temporal noise. The dataset is stratified and split into 70% training, 15% validation, and 15% testing to ensure balanced evaluation across different participants and exercise types.

F. Data Augmentation and Generalization

To improve model generalization and robustness, several data augmentation techniques were applied during training:

- **Random Frame Dropping and Shuffling:** Applied to simulate temporal jitter or frame loss common in mobile-captured video.
- **Gaussian Noise Injection:** Small Gaussian noise was added to pose coordinates to mimic sensor inaccuracies and enhance tolerance to noise.
- **Horizontal Flipping:** Pose sequences were flipped laterally in cases where exercise symmetry (e.g., jumping jacks) allows such transformations.

A cross-subject evaluation protocol was employed to assess the model's generalization. In this setting, the model was trained on a subset of participants and tested on entirely unseen subjects. This ensures that the learned representations are not overfitted to specific individuals and are capable of capturing the underlying motion patterns across different users.

G. Mobile Deployment

Following training, both the classification and correctness estimation models were converted to TensorFlow Lite (TFLite) format for deployment on mobile platforms. This allows the system to perform efficient on-device inference with minimal computational overhead.

The real-time deployment pipeline proceeds as follows:

1) Pose keypoints are extracted from video frames directly on the mobile device.
2) A single exercise repetition is segmented and formatted into a fixed-length input of shape $(30, 99)$.
3) The pose sequence is passed to the TFLite classification model to determine the exercise type.
4) Based on the predicted class, a corresponding TFLite correctness model is loaded and executed.
5) The model returns a correctness score in the range $[0, 1]$ as feedback to the user.

This architecture enables real-time feedback in fully offline settings, eliminating the need for cloud-based inference. The lightweight nature of the models and preprocessing steps ensures suitability for deployment in resource-constrained environments such as smartphones and wearables, supporting real-world fitness coaching applications.

VI. Results and Evaluation

This section presents the evaluation outcomes of the proposed BiLSTM-based classification model trained on the EJUST-GYM-3 dataset. The analysis includes both quantitative metrics and visual diagnostics, illustrating the model's classification performance, generalization capability, and training behavior.

A. Classification Performance

The model was evaluated on a stratified test set using cross-subject partitioning. Table II summarizes the performance metrics per exercise class, including precision, recall, F1-score, and accuracy. The model achieved a near-perfect accuracy of 99.9%, with only one misclassified sample out of the total 832 test instances.

B. Confusion Matrix Analysis

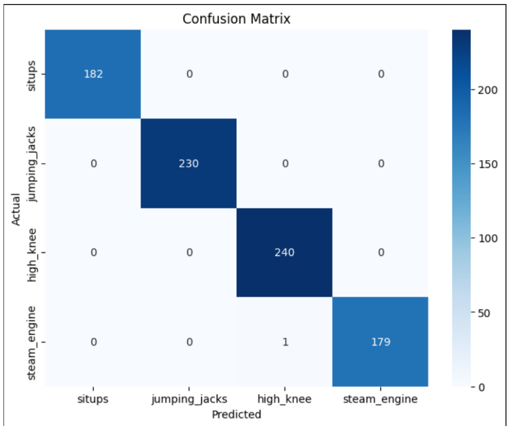

Fig. 6. Confusion matrix showing predicted vs. actual classes across the four exercises.

The confusion matrix in Fig. 6 illustrates the model's ability to distinguish between the four exercise classes with extremely high confidence. Only a single sample from the "High Knees" class was misclassified, suggesting high discriminative power in the learned temporal pose features.

C. Training and Validation Behavior

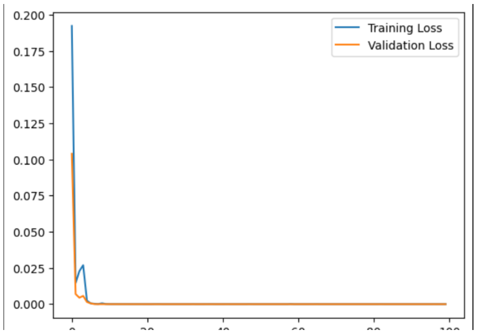

Fig. 7. Training vs. validation loss over 100 epochs.

As depicted in Fig. 7, both training and validation losses converged smoothly within the first 10 epochs and maintained a stable plateau. The minimal gap between them indicates strong generalization and no significant overfitting. This convergence is attributed to:

- Effective data augmentation (frame jittering, noise injection, mirroring),
- Dropout regularization after dense layers,
- Use of early stopping and adaptive learning rate scheduling.

TABLE II
PER-CLASS EVALUATION METRICS ON THE TEST SET

Exercise Class	Precision	Recall	F1-Score	Accuracy
Sit-Ups	1.00	1.00	1.00	100%
Jumping Jacks	1.00	1.00	1.00	100%
High Knees	1.00	0.996	0.998	99.6%
Steam Engine	1.00	1.00	1.00	100%
Average	**1.00**	**0.999**	**0.9995**	**99.9%**

D. Evaluation Summary

- **Model:** BiLSTM + Dense layers, Softmax output.
- **Input:** 30-frame sequences of 99-dimensional 3D pose vectors.
- **Metrics:** 99.9% accuracy; F1-score $\geq$ 0.998 for all classes.
- **Generalization:** Cross-subject stratified evaluation confirmed robustness.
- **Deployment:** Model exported to TensorFlow Lite for mobile inference.

These results demonstrate the system's reliability for fine-grained exercise recognition and validate its readiness for real-world deployment on mobile platforms in offline fitness coaching scenarios.

VII. Future Work

In future iterations, we aim to significantly expand the functionality, usability, and scalability of the proposed system by incorporating new features in both the machine learning pipeline and the mobile application deployment.

A primary direction will be the implementation of *real-time feedback and form correction* capabilities. Building upon the current correctness scoring model, we plan to introduce a frame-wise posture evaluation mechanism that can identify deviations from optimal form in specific joints or limb trajectories. By leveraging keypoint-level temporal dynamics and pose embeddings, the system could deliver *live correction cues* through the mobile interface using audio prompts, visual highlights, or haptic feedback. This would transform the current post-exercise evaluation into a *real-time digital coaching assistant*.

In addition, we intend to integrate *automatic exercise repetition counting* using a combination of IMU peak detection and temporal pattern recognition from pose sequences. This functionality will allow users to monitor the quantity of repetitions without manual input and support structured workout session logging within the mobile app.

To increase the diversity and generalizability of the system, we plan to *expand the dataset and classification models to support additional exercises*. Beyond the four calisthenic movements currently supported (steam engine, sit-ups, jumping jacks, and high knees), we will incorporate other compound and bodyweight movements such as squats, lunges, push-ups, and planks. Each new exercise will be labeled and scored using domain expert supervision to preserve annotation quality.

From a deployment perspective, we will continue optimizing the system for *mobile platforms*. This includes improving *on-device inference latency* using model quantization and pruning, and extending the current app interface to accommodate new features like per-exercise history tracking, user progress analytics, and voice-assisted session guidance. The mobile app will also be designed with modularity in mind to support *custom workout plans*, user profiles, and privacy-preserving settings such as offline operation and local storage of motion data.

Finally, we foresee the extension of this platform to *healthcare and rehabilitation* contexts. By adapting the feedback models to clinical correctness standards and integrating additional sensors (e.g., heart rate monitors, EMG), the system could serve as a digital assistant for physiotherapy, injury recovery, or elder care—delivering personalized, remote guidance while reducing the burden on clinical resources.

VIII. Conclusion

This work presents a complete, end-to-end framework for multi-modal human activity recognition (HAR) in the fitness domain, centered around the development of the novel **EJUST-GYM-3** dataset. Collected from 128 diverse participants, the dataset includes synchronized recordings of four foundational exercises—sit-ups, jumping jacks, high knees, and steam engine—captured via frontal and side RGB video, thermal imaging, and inertial measurement units (IMUs). The dataset is further enriched with detailed participant metadata (e.g., height, weight, sleep patterns, smoking and dietary habits), enabling research in personalized modeling and user-specific feedback.

A robust BiLSTM-based deep learning architecture was designed to process temporal pose sequences for dual-task learning: exercise classification and per-repetition correctness scoring. The model achieved **99.9% classification accuracy** on a cross-subject test set, with near-perfect precision across all classes, validating the expressiveness of 3D pose dynamics and the effectiveness of our temporal modeling strategy.

The data pipeline incorporated advanced preprocessing methods, including GEI generation from thermal views, IMU signal analysis, pose synchronization, and per-repetition segmentation. Data augmentation techniques such as noise injection, horizontal flipping, and frame shuffling were used to improve generalization. The complete system was optimized and deployed on mobile devices using TensorFlow Lite, enabling real-time, offline inference and user feedback without the need for cloud services.

Together, these contributions demonstrate a scalable, privacy-conscious solution for AI-driven fitness coaching and open new avenues for research in multimodal HAR, biomechanics, and digital rehabilitation. EJUST-GYM-3 serves not only as a dataset but also as a benchmark platform for the next generation of intelligent, mobile, and personalized activity recognition systems.

Acknowledgment

The authors would like to express their sincere gratitude to **Dr. Walid Gomaa** for his continuous guidance, insightful feedback, and unwavering support throughout the development of this work. His expertise and mentorship were instrumental in shaping both the theoretical foundation and practical implementation of this research.

References

References

[1] I. U. Khan, S. Afzal, and J. W. Lee, "Human activity recognition via hybrid deep learning based model," *Sensors*, vol. 22, no. 1, p. 323, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/1/323

[2] Z. Zhuang and Y. Xue, "Sport-related human activity detection and recognition using a smartwatch," *Sensors*, vol. 19, no. 22, p. 5001, 2019.

[3] N. A. Rahmad et al., "A survey of video based action recognition in sports," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 11, no. 3, pp. 987–993, 2018.

[4] A. Lees, "Technique analysis in sports: A critical review," *J. Sports Sci.*, vol. 20, no. 10, pp. 813–828, 2002.

[5] M. D. Hughes and R. M. Bartlett, "The use of performance indicators in performance analysis," *J. Sports Sci.*, vol. 20, no. 10, pp. 739–754, 2002.

[6] H. Kaur et al., "Physical fitness and exercise during the COVID-19 pandemic: A qualitative enquiry," *Front. Psychol.*, vol. 11, p. 2943, 2020.

[7] K. M. Tsiouris et al., "A Review of Virtual Coaching Systems in Healthcare: Closing the Loop With Real-Time Feedback," *Front. Digit. Health*, vol. 2, p. 567502, 2020.

[8] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion," *Int. J. Comput. Vision*, vol. 87, no. 1, pp. 4–27, 2010.

[9] A. Jain et al., "Learning Human Pose Estimation Features with Convolutional Networks," in *Proc. Int. Conf. Learn. Represent.*, 2014.

[10] C. Ionescu et al., "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, 2013.

[11] A. Sharshar et al., "MM-DOS: A Novel Dataset Of Workout Activities," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2022, pp. 1–8.

[12] S. Ghorbani et al., "MoVi: A large multi-purpose human motion and video dataset," *PLOS ONE*, vol. 16, no. 6, p. e0253157, 2021.

[13] A. Fayez et al., "ValS: A Leading Visual and Inertial Dataset of Squats," in *Proc. Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM)*, IEEE, 2022, pp. 1–8.

[14] D. Strömbäck et al., "MM-Fit: Multimodal Deep Learning for Automatic Exercise Logging across Sensing Devices," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 4, pp. 1–22, 2020. [Online]. Available: https://doi.org/10.1145/3432701

[15] R. Ogata et al., "Temporal Distance Matrices for Squat Classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019, pp. 2533–2542.

[16] V. R. Reddy and T. Chattopadhyay, "Human Activity Recognition from Kinect Captured Data Using Stick Model," in *Human-Computer Interaction*, LNCS, vol. 8511, Springer, 2014. doi:10.1007/978-3-319-07230-2$_3$0.

[17] D. Pagliari and L. Pinto, "Calibration of Kinect for Xbox One and Comparison Between the Two Generations of Microsoft Sensors," *Sensors*, vol. 15, pp. 27569–27589, 2015.