IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 August 2023, accepted 18 September 2023, date of publication 25 September 2023, date of current version 4 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3318563

RESEARCH ARTICLE

Accuracy Comparison of CNN, LSTM, and Transformer for Activity Recognition Using IMU and Visual Markers

MARÍA FERNANDA TRUJILLO-GUERRERO[1], (Student Member, IEEE), STADYN ROMÁN-NIEMES[2], MILAGROS JAÉN-VARGAS[1], ALFONSO CADIZ[3], RICARDO FONSECA[3], AND JOSÉ JAVIER SERRANO-OLMEDO[1,4]

[1]Center for Biomedical Technology (CTB), Universidad Politécnica de Madrid, 28223 Madrid, Spain
[2]School of Mathematical and Computational Sciences, Yachay Tech University, Urcuquí, Imbabura 170522, Ecuador
[3]Digevo, Santiago 7560941, Chile
[4]Centro de Investigación Biomédica en Red para Bioingeniería, Biomateriales y Nanomedicina, Instituto de Salud Carlos III, 28029 Madrid, Spain

Corresponding author: María Fernanda Trujillo-Guerrero (mariafernanda.trujillo.guerrero@alumnos.upm.es)

This work was supported in part by the Ministerio de Ciencia, Innovación y Universidades of Spain under Grant PGC2018-097531-B-I00; and in part by Cloud Computing funded by Microsoft Founders Hub.

ABSTRACT Human activity recognition (HAR) has applications ranging from security to healthcare. Typically these systems are composed of data acquisition and activity recognition models. In this work, we compared the accuracy of two acquisition systems: Inertial Measurement Units (IMUs) vs Movement Analysis Systems (MAS). We trained models to recognize arm exercises using state-of-the-art deep learning architectures and compared their accuracy. MAS uses a camera array and reflective markers. IMU uses accelerometers, gyroscopes, and magnetometers. Sensors of both systems were attached to different locations of the upper limb. We captured and annotated 3 datasets, each one using both systems simultaneously. For activity recognition, we trained 8 architectures, each one with different operations and layers configurations. The best architectures were a combination of CNN, LSTM, and Transformer achieving test accuracy from 89% to 99% on average. We evaluated how feature selection reduced the sensors required. We found IMU and MAS data were able to distinguish correctly the arm exercises. CNN layers at the beginning produced better accuracy on challenging datasets. IMU had advantages over other acquisition systems for activity recognition. We analyzed the relations between models accuracy, signal waveforms, signals correlation, sampling rate, exercise duration, and window size. Finally, we proposed the use of a single IMU located at the wrist and a variable-size window extraction.

INDEX TERMS Human activity recognition, IMU, movement analysis system, visual marker, CNN, LSTM, Transformer, arm exercises.

I. INTRODUCTION

Human activity recognition (HAR) studies the capture systems and algorithms to recognize activities performed by people in any situation. Data for this task can be captured by inertial sensors, cameras with visual markers, and cameras using human pose estimation, EEG, or EMG. All these sensors produce time series data, in consequence, the computer algorithms able to classify these activities are: Recurrent neural networks, Convolutional neural networks, Long Short Term Memory networks, and Transformer networks [1], [2]. HAR is an essential field of study in computer vision and artificial intelligence. It has many potential applications in various industries including security, surveillance, and healthcare [3], [4]. HAR systems are used to monitor the movements of elderly individuals in care facilities and to alert caregivers if they fall or exhibit other signs of distress [5], [6], [7], [8], [9]. HAR technology is being used in sports training to help athletes measure their performance by providing real-time feedback on their exercises [10].

The associate editor coordinating the review of this manuscript and approving it for publication was Dost Muhammad Khan.

One of the key challenges in human activity recognition is the high variability of human movements and activities. Due to this variability, traditional machine learning algorithms often struggle to accurately classify movements and activities [11]. To overcome this challenge, researchers have developed a range of techniques and approaches, including deep learning and other advanced machine learning methods. Our work aims to identify deep learning architectures that achieve higher accuracies.

One capture system is the Movement Analysis System (MAS) which detects the movement of human joints and limbs. MAS uses infrared cameras and visual markers to measure the physical world and obtain the dimensions and positions of objects. Data is generated using specialized software to analyze images taken from different angles and positions to create a 3D model of the object [12]. MAS takes the video streaming as input and estimates the positions of the visual markers worn by a person [13].

Another capture system is the use of inertial measurement units (IMUs). An IMU is a system of sensors that measures three axes acceleration, angular velocity, and magnetic field [14]. IMUs are attached to different parts of the human body to identify activities such as walking, running, or jumping. Making IMU suitable for healthcare applications such as rehabilitation programs to provide feedback on movements and monitor progress during therapy sessions [15]. Our work aims to identify the advantages and disadvantages of IMU and MAS capture methods.

HAR has three levels of abstraction in exercise recognition: full body exercise, single limb exercise, and stages inside a single exercise. We captured data and built models to classify upper limb exercises and to distinguish the flexion or extension stage inside an exercise. Most HAR studies [7], [10], [16], [17], [18], [19], [20], and [21], perform exercises that involve the entire human body and differ one from each other such as walking, climbing stairs, sitting, jumping, or running. Our work studies 6 specific exercises of the upper limb where they share common behavior making them harder to distinguish.

In this work, we compared the accuracies of two acquisition systems: Inertial Measurement Unit (IMU) vs Movement Analysis System (MAS). Studied the time series data using plots and labeling the signal according to each exercise. Understand graphically which sensor discriminates exercises better (acc, gyro, mag, visual marker). We trained 8 SOA deep learning architectures to recognize arm exercises and compared their accuracy. Designed 8 deep learning architectures using different layers and operations. Compared the performance for every architecture and capturing method to identify the best combination. Applied feature selection to identify minimum sensors and locations to correctly recognize exercises. We studied the effect of different sampling rates, exercise duration, and window size. Finally, we proposed the use of a single IMU located at the wrist and a variable-size window extraction.

II. RELATED WORKS

A. ACTIVITY RECOGNITION USING IMU

The use of IMUs for human activity recognition has become popular in recent years due to the widespread availability of sensors in wearable devices such as smartphones, wrist watches, and fitness trackers. Advances in machine learning and deep learning allowed the development of sophisticated algorithms for recognizing human activities from IMU data. These algorithms involve all kinds of architectures from artificial neural networks which process data with spatial and temporal characteristics. Neural networks are effective in recognizing complex patterns in time series data and can be trained on large datasets to achieve high accuracy.

Monitoring and analyzing human motion can provide valuable information for various applications. In 2014, Ronao and Cho [16] proposed a multi-task learning approach for human activity recognition. Using accelerometer data, the system uses a single CNN to learn multiple tasks, including activity recognition and pose estimation. The authors found that this approach improved the overall accuracy of the activity recognition system. Then, in 2017 Yarnan et al. proposed a framework for detecting arm and human activities based on data fusion from inertial measurement units (IMUs) and surface electromyography (EMG) sensors [22]. Supervised and unsupervised machine learning algorithms were used to train the models and obtain evaluation indicators. The combined IMU and EMG data outperformed the IMU data alone and the EMG data alone, significantly reducing the error in determining activities for supervised algorithms.

In 2018, Xiong et al. [17] proposed a two-stage model for recognizing activities from accelerometer data. The first stage of the model uses a CNN to extract spatial and temporal features from the data. The second stage then uses a recurrent neural network (RNN) to combine features from the CNN with contextual information to recognize the activity.

In 2019, Sarcevic et al. developed a system to detect arm and body movements using wrist sensors that contained an accelerometer, a gyroscope, and a magnetometer [6]. Multiple datasets were tested using various feature extractive approaches, sampling frequencies, processing window widths, and sensor combinations. The authors achieved almost 90% accuracy on validation data.

Lu and Tong [18] worked on HAR using a single 3-axis accelerometer, focused on movement monitoring using wearable sensors and devices. Their method consists of encoding 3-axis signals as 3-channel images using a modified recurrence plot. Then, residual neural networks were used to classify images and, thus, signals. As a result, the authors obtained highly competitive accuracies and good efficiencies on the ASTRI motion dataset, which contains data on human hand movements, and the ADL Dataset from wrist-worn accelerometer data.

The work of Avilés et al. presented a framework to recognize user movement using a smartphone equipped with a tri-axial accelerometer and a tri-axial gyroscope sensor.

The framework used three parallel CNNs for local feature extractive, later fused in the classification stage. The whole CNN scheme is based on a feature fusion of a fine CNN, a medium CNN, and a coarse CNN [10]. The algorithm successfully classified six human activities: walking, walking upstairs, walking downstairs, sitting, standing, and laying.

Yen et al. proposed a wearable device capable of recognizing six basic activities using deep learning and data from a gyroscope and an accelerometer [19]. They used waist devices worn by dialysis patients, whose activities could not be accurately determined using wrist devices. The model achieved recognition rates of 95.99% and 93.77%. That same year, Lemieux and Noumeir proposed a hierarchical CNN model for human activity recognition [23]. The model consists of two levels of CNNs. The first level extracts spatial features from the accelerometer data and the second level combines the spatial features with temporal information to recognize the activity.

In 2020, Clouthier et al. analyzed the movement of athletes [20]. They collected optical motion data on 417 athletes performing 13 athletic movements. The authors trained an existing deep neural network architecture that combines convolutional and recurrent layers. They obtained classification accuracies of 90.1 and 90.2% for full body measurements. The authors concluded that classifying athletic movements using wearable sensors was feasible.

More recent research is the work of Uddin and Soylu [7], focused on the well-being of elderly people using wearable sensors to detect unprecedented events such as falls or other health risks. The authors proposed a "body sensor-based activity modeling and recognition system using time-sequential information-based deep Neural Structured Learning (NSL)" [7]. The algorithm is powered by data from multiple wearable sensors, which then undergo statistical feature processing. The framework is powered by kernel discriminant analysis (KDA) and long short-term memory (LSTM) based models. The authors achieved around 99% recall on the mobile health application dataset (MHEALTH) [24]. The framework also surpassed the recall rate of other algorithms, such as deep belief networks, convolutional neural networks, and recurrent neural networks.

Another recent research work is the paper of Han et al. which focused on enhancing the convolution capacities of CNNs instead of modifying the architectures [25]. The authors proposed the idea of heterogeneous convolution for activity recognition tasks. All filters within a specific convolutional layer are separated into two uneven groups. The authors examined the effectiveness of the framework on several benchmark HAR datasets, finding that the heterogeneous convolution is simple to integrate into convolutional layers without increasing extra parameters and computational overhead. In the same year, Luwe et al. proposed a "hybrid deep learning model that amalgamates a one-dimensional Convolutional Neural Network with a bidirectional long short-term memory (1D-CNN-BiLSTM) model for wearable sensor-based human activity recognition" [26]. This one-dimensional neural network transforms the time series information from the sensor into representative features, which are then encoded by the bidirectional LSTM. The authors found the approach outperformed the existing methods, obtaining a recognition rate of 95.48% on the UCI-HAR dataset, 94.17% on the Motion Sense dataset, and 100% on the Single Accelerometer dataset.

B. ACTIVITY RECOGNITION USING MOVEMENT ANALYSIS SYSTEM

In 1999, Ramsey and Wretenberg published a research paper reporting on the use of intracortical pins to measure knee movement as an alternative to the use of reflective markers [27]. The authors found that their method allowed them to take more precise readings with low error.

In 2005, Cutti et al. proposed an experiment to test the error when using reflective markers for photogrammetric measurements [28]. The authors put the markers on different subjects and made them execute different movements. Then, the readings affected by the error were compared with normal readings. The authors concluded that the error has a strong influence and should not be ignored, opening the way for new research that could compensate for this error.

Tokarczyk and Mazur compiled different Movement Analysis System techniques and methods [8]. They presented the advancements of two Movement Analysis System techniques. The first is Moiré's method of stripes, which involves overlaying two sets of parallel stripes with slightly different spacings to create a moiré pattern, which can be used in the human body to detect conditions such as scoliosis. The second method uses multiple video cameras around the body, in conjunction with physical markers, that the camera system can easily detect and read.

In 2008, Van Andel et al. carried out research to determine a standardization protocol for the clinical application of upper extremity movement analysis [9]. The authors developed measurement methods for hand orientation in different movements, using a stereophotogrammetric recording of active LED markers with a camera system. The wrist, elbow, shoulder, and scapula joint angles were analyzed, and minimum/maximum angles were determined. This way, the authors determined the trajectories and angles of all the movements, cementing the basis for developing more precise and standardized reports on movements that would allow for future comparisons with pediatric and/or pathologic movement patterns. MAS generates easy-to-understand reports on movements because it outputs markers position [9].

Jaén-Vargas et al. used wearable sensors and two reflective markers (mocap) to recognize the activities of walking, sit-to-stand, and squatting [21]. The authors evaluated the performance of four deep learning networks: deep neural network, CNN, LSTM, and a combination of CNN and LSTM. The authors found that a hybrid network (CNN-LSTM) was better than an individual network. The hybrid approach accounted for class imbalance, making it more versatile, obtaining 99% accuracy in both datasets and an F1 score of 99% and 87% with wearable sensors and reflective markers, respectively. The authors also find that the use of wearable sensors yielded better results.

Jaén-Vargas et al. 2022 analyzed the performance of deep neural networks, CNN, LSTM, and CNN-LSTM when variating the sliding window size [29]. The sliding window is a technique in which a fixed-size window is moved over a time series, processing the information in divided sections. The intention was to find an optimal window size for HAR using a sampling rate of 100 Hz. Windows of small sizes 5, 10, 15, 20, and 25 frames and long ones of sizes 50, 75, 100, and 200 frames were compared. The results showed that windows from 20 to 25 frames were optimal, obtaining an accuracy of 99,07% and an F1-score of 87,08% on sensor data and an accuracy of 98,8% and an F1-score of 82,80% on MOCAP data.

A particular field of research that involves the use of cameras for movement recognition is the study of human gait using silhouettes and skeletons. In this field, an important work is the paper of Cicirelli et al. [30], which is a review of human gait analysis with application in neurodegenerative diseases. They compared sensors, features, and processing methodologies where deep networks such CNNs or LSTMs achieved the best results. In another gait-related work [31], the authors performed gait analysis classification for neurodegenerative diseases using support vector machines (SVMs) in optical motion capture data and achieved 99.1% accuracy. This year, three relevant works in gait analysis were published. The first one is the paper of Shayestegan et al. [32], in which they implemented Dual-Head Attentional Transformer-LSTM (DHAT-LSTM) in kinetic data to classify stages of gait disorders and achieved an accuracy of 81%. In the work of Cheriet et al. [33], they applied a Multi-Speed Transformer Network in video data to classify stages of neurodegenerative diseases and achieved an accuracy of 96.9%. Finally, Cosma, Catruna, and Radoi [34] published a paper where they used Self-Supervised Vision Transformers in human gait video data as a biometric authentication method.

A summary of the papers considered in the Related Works section is presented in Table 1.

III. EQUIPMENT AND DATASETS ACQUISITION

A. EQUIPMENT

Three different Inertial Measurement Units were used. Different sensor manufacturers were used to capture variability in the sensor's sensitivity and sampling rate. The first two devices, the MPU9250, and Trigno Avanti [35] inertial sensors are composed of an accelerometer, gyroscope, and magnetometer, each one with 3 axes. The third IMU device used was the Metawear, composed of a 3D accelerometer and a 3D gyroscope. The three systems send data to a PC using Bluetooth communication, using their software to capture data simultaneously from all devices at all locations.

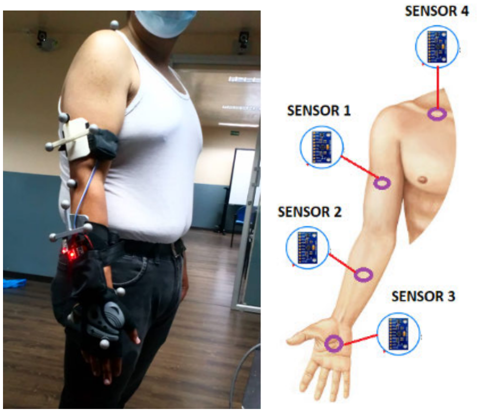


FIGURE 1. IMU sensors distribution.

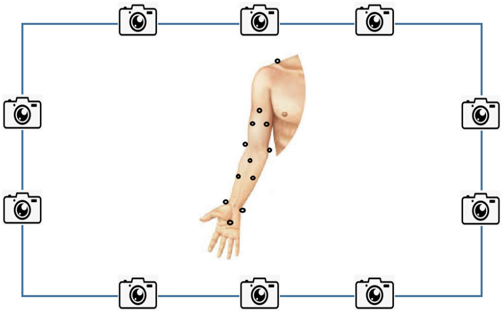


FIGURE 2. Visual markers and cameras distribution.

The sensors were mounted in the shoulder, forearm, arm, and hand, as described in Figure 1 and in the work of Tobar et al. [36].

The equipment used for movement analysis was a Kinescan/IBV. It is a movement analysis system based on visual markers that use rigid segment models. It captures PAL video at 25 frames per second. The system has 10 cameras located at a height of 2.4 m and distributed around the person. The analysis of object movement is performed by tracking the position of markers. These markers are small spheres covered with reflective material attached to the subject. The system provides position and speed for all markers. The markers are placed as follows: one on the shoulder, three on the forearm, two on the elbow, three on the arm, two on the wrist, and one on the hand. Figure 2 shows the distribution of the markers and the cameras used.

B. ARM EXERCISES DESCRIPTION

The datasets were recorded while the subjects performed the follow-arm exercises. Each exercise is described as:

- **Elbow Flexion-Extension:** During elbow flexion the angle formed by the elbow joint decreases. The forearm approaches the arm. During elbow extension the angle formed by the elbow joint increases. The arm separates the forearm. Figure 3.a shows the exercise.

TABLE 1. Summary of related works on HAR.

Reference	Year	Authors	Capture Method	AI Framework	Dataset	Results
[16]	2014	Ronao and Cho	IMU	CNN	UCI-HAR	91.76% accuracy
[22]	2017	Yarnan et al.	IMU and EMG	Unsupervised and supervised machine learning	Made by authors	10% normal data variation rate and 0.85 determination coefficient unsupervised, 6.55% normal data variation rate supervised.
[17]	2018	Xiong et al.	IMU	CNN and RNN	KTH and UCF101	From 96% to 99.66%
[6]	2019	Sarcevic et al.	IMU	NCC, MLP, Bayesian classifier, SVM	Made by authors	99.1% accuracy
[18]	2019	Lu and Tong	IMU	Residual neural networks	ASTRI and ADL	93.9% and 99.9% accuracy
[10]	2019	Avilés et al.	IMU	Three CNNs	UCI-HAR and WIDSM	100% accuracy
[19]	2020	Yen, Liao and Huang	IMU	Three CNNs	UCI-HAR and made by authors	95.99% and 93.77% recognition rate
[23]	2020	Lemieux and Noumeir	IMU	Two CNNs	UTD-MHAD	90.8% accuracy
[20]	2020	Clouthier et al.	IMU	Deep neural network	Made by authors	90.1% and 90.2% classification accuracy
[25]	2022	Han et al.	IMU	CNN	OPPORTUNITY, PAMAP2, UCI-HAR, USC-HAD and made by authors	Accuracies higher than 91%
[26]	2022	Luwe et al.	IMU	CNN and BiLSTM	UCI-HAR, Motion Sense and Single Accelerometer	95.48%, 94.71% and 100% recognition rate
[27]	1999	Ramsey and Wretenberg	Movement Analysis System	Does not apply	Does not apply	Data on the angles of a wide array of knee-related movements
[28]	2005	Cutti et al.	Movement Analysis System	Does not apply	Does not apply	A method to retrieve upper-body movement data with good precision and angle consideration
[8]	2007	Tokarczyk and Mazur	Movement Analysis System	Does not apply	Does not apply	Review paper on data retrieval using physical markers
[9]	2008	Van Andel et al.	Movement Analysis System	Does not apply	Made by authors	A comprehensive dataset of readings, angles and other information on 3D upper body movement
[31]	2020	Dentamaro, Impedovo and Pirlo	Movement Analysis System	SVM	Does not apply	An extensive review of methods and protocols for analyzing human gait
[21]	2022	Jaén-Vargas et al.	Movement Analysis System	CNN, LSTM, and CNN-LSTM	Made by authors	99% accuracy, F1 scores of 99% and 87%
[29]	2022	Jaén-Vargas et al.	Movement Analysis System	CNN, LSTM, and CNN-LSTM	Made by authors	99.07% accuracy, F1 scores of 99.8%, 87.08% and 82.80%
[30]	2022	Cicirelli et al.	Movement Analysis System	Does not apply	Does not apply	An extensive review of methods and protocols for analyzing human gait
[32]	2023	Shayestegan et al.	Movement Analysis System	DHAT-LSTM	Made by authors	81% accuracy on gait disorder detection
[33]	2023	Cheriet et al.	Movement Analysis System	Multi-speed Transformer	Made by authors	96.9% on gait disorder detection
[34]	2023	Cosma, Catruna and Radoi	Movement Analysis System	ViT, CaiT, CrossFormer, Token2Token and TwinsSVT	GREW and DenseGait	Accuracies from 75% to 85% in best cases

- **Hand Pronation-Supination** Starts with the hand and forearm aligned, the palm facing upwards with the thumb facing outwards. A rotation results in the palm facing downward with the thumb facing inwards. The exercise is shown in Figure 3.b.
- **Shoulder Abduction-Adduction:** Abduction is a lateral movement of the entire upper limb away from the trunk until the arm forms a 90-degree angle with the trunk. Adduction is the lateral movement that brings the upper limb closer to the trunk. Figure 4 shows the exercise.
- **Horizontal Shoulder flexion-extension:** Begins with the entire upper limb forming 90° with the trunk and located laterally. During extension, the arm moves horizontally forward and in flexion, it returns to its starting position. Figure 5 shows the exercise.

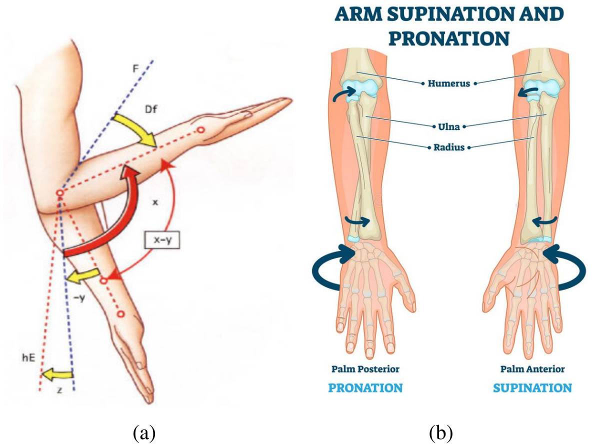


FIGURE 3. Exercises (a) flexion-extension [37], (b) pronation-supination.

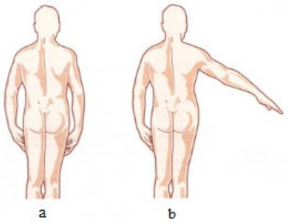


FIGURE 4. Shoulder Abduction-Adduction. (a) neutral position, (b) abduction from 0° to 60° [37].

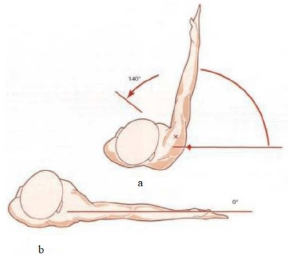


FIGURE 5. Horizontal Shoulder Flexion-extension. (a) flexion with an adduction of 140°, (b) 90° abduction in the frontal plane [37].

- **Vertical Shoulder flexion-extension:** This movement begins with the upper limb close to the trunk. During flexion, the arm moves frontally in a vertical manner until the upper limb reaches a horizontal position. In extension, it returns to the starting position. Figure 6 shows the exercise.
- **Internal and External Shoulder Rotation:** This exercise begins with the arm next to the trunk, arm, and forearm making a 90-degree angle in a L shape. The forearm begins next to the stomach and then moves away horizontally from the body. Figure 7 shows the exercise.

C. DATASET ACQUISITION

A group of 10 people without arm diseases, between 20 and 25 years old, were involved in the data acquisition. The acquisition was separated into 3 sessions of people performing arm exercises. For each session, a person starts in a neutral position and performs 10 repetitions of each exercise. At the end of the exercise, they return to a neutral position and repeat the process for the next exercise.

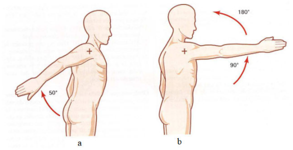


FIGURE 6. Vertical Shoulder Flexion-extension. (a) low amplitude (45° - 50°), (b) high amplitude (180°) [37].

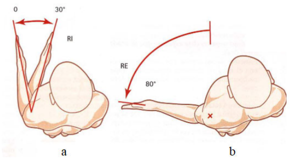


FIGURE 7. Internal and external rotation of the shoulder. (a) internal rotation of 30°, (b) external rotation of 80° [37].

TABLE 2. Datasets and instruments.

Dataset	IMU	Movement analysis system
Dataset 1	MPU9250 1Hz	Kinescan/IBV 200Hz
Dataset 2	Trigno Avanti 370hz	Kinescan/IBV 200Hz
Dataset 3	MPU9250 1Hz	Kinescan/IBV 50Hz
	Metawear 50Hz	

Each session was recorded with Movement Analysis System (MAS) and IMU instruments simultaneously. Exercises were performed one after the other, and data for each exercise was recorded and labeled as a whole. Within the 3 sessions, 7 sub-datasets were created: 3 for MAS and 4 for IMU. Within the last session, 2 different IMU equipment were used. Dataset 1 has 2 elbow exercises, dataset 2 has 3 shoulder exercises, and Dataset 3 has 3 elbow and 3 shoulder exercises.

Table 2 shows the names of the datasets, instruments used, and their capture frequencies. Tables 3 and 4 describe the features captured from IMUs and Movement Analysis System. Each feature is described by its type (Acc, Gyro, Mag, Visual marker), location, and sensor axis. For IMU data sensors were located at the shoulder, forearm, arm, and hand. For Movement Analysis System data, markers were located as follows: one on the shoulder, three on the forearm, two on the elbow, three on the arm, two on the wrist, and one on the hand.

D. DATASET DESCRIPTION

Each sample was built as a matrix where the columns represent the features and rows represent the length given by the window size. The label for the sample was the most common label from the array of data points.

IMU data was captured at frequencies of 1, 25, and 50 Hz with 18, 24, and 27 features. IMU data used window sizes of 10, 20, 40, and 200 data points. A single IMU dataset has at most 500k timesteps and at most 4500 samples. According to the window size, we have data tensors e.g. 2511 × 200 × 24 (samples × window size × features). The acquisition

TABLE 3. **IMU features.**

Sensor	x-axis	y-axis	z-axis
ArmAcc	AAx	AAy	AAz
ArmGyro	AGx	AGy	AGz
ArmMag	AMx	AMy	AMz
ForearmAcc	FAx	FAy	FAz
ForearmGyro	FGx	FGy	FGz
ForearmMag	FMx	FMy	FMz
HandAcc	HAx	HAy	HAz
HandGyro	HGx	HGy	HGz
HandMag	HMx	HMy	HMz
ShoulderAcc	SAx	SAy	SAz
ShoulderGyro	SGx	SGy	SGz
ShoulderMag	SMx	SMy	SMz

TABLE 4. **Movement analysis system features.**

Marker	x-axis	y-axis	z-axis
Hand	Hx	Hy	Hy
Wrist internal	WIx	WIy	WIz
Wrist external	WEx	WEy	WEz
Forearm lower	FLx	FLy	FLz
Forearm middle	FMx	FMy	FMz
Forearm upper	FUx	FUy	FUz
Elbow internal	EIx	EIy	EIz
Elbow external	EEx	EEy	EEz
Arm lower	ALx	ALy	ALz
Arm middle	AMx	AMy	AMz
Arm upper	AUx	AUy	AUz
Shoulder	Sx	Sy	Sz

sampling rate determines the amount of data for an exercise duration. A lower sampling rate will generate less data, the movement will not be correctly captured, and the exercise will not be correctly recognized.

Visual markers data was captured at frequencies of 50, and 200 Hz with 18, 24, and 36 features. Visual Markers data used window sizes of 100 and 200 data points. A single Visual Marker dataset has at most 545k timesteps and 5450 samples. According to the window size, we have data tensors e.g. 1350 × 200 × 24 (samples × window size × features).

For training, we used 100 epochs and a batch size of 64 samples for both acquisition systems. Available data was split using 75% for training, 10% for validation, and 15% for testing. All information related to datasets is shown in table 5.

E. TIME SERIES SIGNALS

The dataset is a multichannel time series. Each channel relates to a location, sensor, and axis. The first row in Fig 8 is from the arm, gyroscope, and x-axis.

IMU signals are shown on the left of Figure 8 and visual markers signals on the right. Signals from both sensors are pseudo-periodic with square-like waveforms. Some channels allow us to distinguish movements easily. Some channels are highly correlated. All channels are height limited and do not show outliers.

Different exercises show responses on different sensors, locations, and axis. During elbow flexion-extension, the arm remains locked, and the forearm moves. The sensors located at the forearm show the most movement while the arm sensors remain still. To recognize different kinds of movements, it is important to have independent information sources, e.g., gyro and accelerometer, located in at least two different positions. The datasets can be accessed through the GitHub repository of this work.[1]

Each exercise repetition depicts a square-like waveform and builds up a pseudo periodic signal as in figure 9. All studied arm exercises have two stages: flexion and extension. These two stages are reflected in a square-like waveform with two levels. A low-level signal for extension and an upper level for flexion. When the limb moves from flexion to extension a slope is visible in the signal. The two levels and the slope are easily recognizable in the signals of IMU and visual markers.

Given the square-like waveform, we can identify the movement duration. For example, in dataset 1 the time needed to perform an exercise was 90 seconds while in dataset 2 the time was 2 seconds. The window size has to be chosen correctly according to the movement duration and sampling rate.

Figure 10, shows IMU signals of ten repetitions of elbow flexion-extension. The imu is located at the arm with a gyro, accelerometer, and magnetometer. Each movement repetition depicts a square-like waveform in the gyro and magnetometer. The square-like waveform has a lower level for flexion and an upper level for extension. Gyro and magnetometer data are easily interpretable to recognize an exercise. The accelerometer shows peaks going up and down, being harder to distinguish the movement being performed.

X and z arm gyroscope axes are highly correlated. Y and z arm magnetometer axes are correlated. X and y arm accelerometer axes are correlated.

The IMU signals show fewer amplitude differences than visual markers, presented in figures 10 and 11. Gyroscope and magnetometer data show a cleaner square than the visual marker data. Accelerometer data is centered at zero.

Figure 11, shows arm markers signals of ten repetitions of elbow flexion-extension. There are three markers located at the arm, each one with three axes. Position signals from markers at the X axis at the lower, middle, and upper arm are highly correlated. The same behavior happens for y and z-axis markers at the lower, middle, and upper arm.

Position signals have different offsets at each repetition, and the offset behaves as a moving offset. It is harder to distinguish the exercises vs the IMU signals. The offsets describe the relative position of the person to the camera system. When the person moves inside the measured area, the position signal will be different. The height and arm size of the person will affect the signal values.

Similar to IMU data, markers data show two levels. We can observe a signal higher level when the forearm is up and a lower level when the arm is down.

Markers signals are affected by occlusion, the angle and distance to the cameras, and the posture and height of the person. IMU is not affected by these variables.

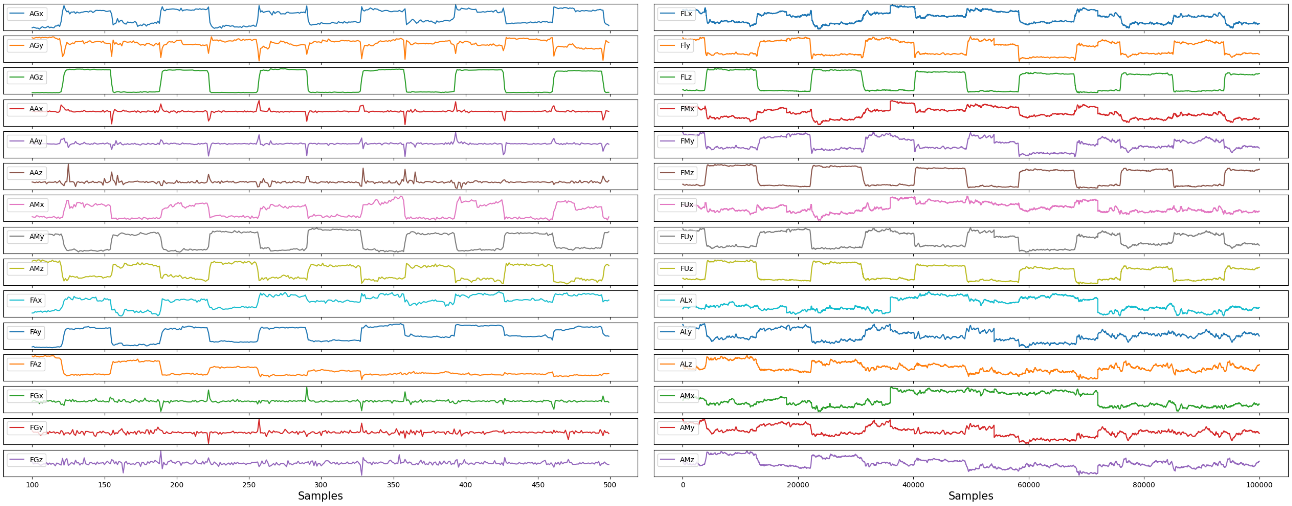
[1] https://github.com/StadynR/HAR-imu-photogrammetry

TABLE 5. Summary of datasets capture conditions.

Dataset	Exercises	Sampling [Hz]	Points	Window Size	Samples	Features	Data Tensor
D1 MPU9250	2	1	1360	10	135	27	135 × 10 × 27
D1 MAS	2	200	313649	200	1568	36	1568 × 200 × 36
D2 Trigno Avanti	3	370	502260	200	2511	24	2511 × 200 × 24
D2 MAS	3	200	270015	200	1350	12	1350 × 200 × 12
D3 MPU9250	6	1	8279	40	206	27	206 × 40 × 27
D3 Metawear	6	50	136226	30	4540	18	4540 × 30 × 18
D3 MAS	6	50	545024	100	5450	36	5450 × 100 × 36

TABLE 6. Summary of sampling rate, exercise duration, and window size.

Dataset	Single exercise duration[sec]	Sampling rate[samp/sec]	Single exercise size[points]	Window Size[points]	Window/ exercise size
D1 - MPU9250	90	1	90	10	0.11
D1 - MAS	90	200	18000	200	0.01
D2 - Trigno	2	370	740	200	0.27
D2 - MAS	2	200	400	200	0.5
D3 - MPU9250	4	1	4	40	10
D3 - Metawear	4	50	200	30	0.15
D3 - MAS	4	50	200	100	0.5



(a) Dataset 1 - MPU9250

(b) Dataset 1 - Movement Analysis System

FIGURE 8. Signals from dataset 1 MPU9250 and visual markers.

We can observe that lower, middle, and upper markers signals at the arm and forearm are similar because it is the same solid section. To show the signal correlation of IMU and markers data, we computed the correlation matrices using all the available features. From the figure 12, we note that MAS signals are highly correlated.

IV. METHODOLOGY

A. PREPROCESSING

Before training, all datasets were preprocessed following these steps: normalization, reshaping into 3D arrays with dimensions (window size, samples, features), and label encoding. Additionally, in every training process, the dataset was split into 5 parts using k-fold cross validation [38] to alleviate the effects of small datasets and class imbalance.

B. ARCHITECTURES

For training and testing, eight neural network architectures were considered and evaluated. The main components of the architectures are LSTM, 1D convolutional layers, and transformer encoders. All architectures used categorical cross entropy as their loss function and Adam as their optimizer.

Long Short-Term Memory (LSTM) is a recurrent neural network layer used in deep learning architectures. It comprises memory cells and gates that allow the network to store or discard information over time selectively [39]. The basic LSTM cell consists of three gates. The input gate determines how much new information is added to the memory cell, the forget gate decides how much old information should be removed, and the output gate regulates the amount of information outputted from the cell. Additionally, each gate

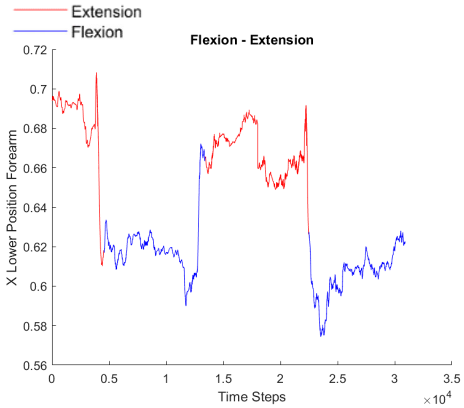


FIGURE 9. **Visual marker at the forearm. A single period of elbow flexion-extension movement. The blue line is the flexion state, and the red line is the extension state.**

has its own set of learnable parameters, allowing the network to adaptively adjust the amount of information stored or discarded based on the input data, making LSTM effective in processing sequential data [40].

One-dimensional (1D) convolution is a mathematical operation frequently used in signal processing and deep learning, which involves sliding a small window or kernel over a one-dimensional input signal, computing the dot product between the kernel and the signal at each position, and generating a new output signal. The output signal is a compressed representation of the input signal, highlighting patterns and features relevant to the task at hand [41]. 1D convolutional layers can also be used to process sequential data by learning a set of filters [42].

The Transformer architecture is a framework used typically for natural language processing (NLP), but can also be used for sequential data because of its attention mechanism. The attention mechanism functions by extracting information from the entire sequence, by using a weighted sum of all the past states of the encoder, generating a matrix. This means that all parts of the sequence are treated by their real importance, and the overall context is considered, prioritizing words with higher weight, allowing the model to focus on the right element of the input to predict the next element of the output [43], [44]. This attention mechanism is improved by using multi-head attention, which applies self-attention to different segments of the input, allowing the transformer to have better discrimination capabilities. As each head will produce its resulting matrix, all matrices are concatenated and multiplied by an additional weight matrix, generating an output matrix that contains information from all the heads [43], [45]. The code can be found in our GitHub repository. The summary of the structures of the architectures is presented in Figure 13. The architectural description used in this work is presented below.

- **Architecture 1 (LSTM + Dropout + Dense + Dense):** This architecture was taken from the web article "Implementing LSTM for Human Activity Recognition using Smartphone Accelerometer data" [46]. The architecture comprises an LSTM layer of 128 neurons, a dropout layer, and two fully connected layers. The LSTM layer allows for time series data analysis due to its ability to handle variable-length input sequences, noisy data, and missing data. Thus, the network can make accurate predictions based on past observations. This network was originally evaluated with the Wireless Sensor Data Mining (WISDM) dataset, obtaining an accuracy of 96.20%.
- **Architecture 2 (LSTM + Dropout + LSTM + Dropout + Dense):** This architecture was based on the GitHub repository "Human-Activity-Recognition" [47]. This architecture is an extension of the last one, adding the LSTM layer after the first dropout layer, adding another dropout, and then a fully connected layer. This network was originally evaluated with the UCI-HAR dataset, obtaining an accuracy of 93.17%.
- **Architecture 3 (Conv1D + Conv1D + Dropout + Max Pooling + Flatten + Dense + Dense):** This architecture was taken from the GitHub repository "ETFA-Workshop" [48]. In contrast to the previous architectures, this network focuses on using convolutional layers. The architecture comprises two 1D convolutional layers of 64 neurons, a dropout layer to avoid overfitting, a max pooling layer to reduce dimensionality, and a final section of a flatten and two fully connected layers. Convolutional networks can be effective for time series analysis, as they are good at extracting features from the input data, which can be useful for identifying patterns and trends in the data. Also, convolution allows downsampling data, reducing the computational complexity of the model and making it easier to train and run. This network was originally evaluated with the UCI-HAR dataset, obtaining an accuracy of 89.89%.
- **Architecture 4 (Conv1D + Max Pooling + LSTM + Dropout + Dense + Dense):** This architecture was assembled empirically as a combination of convolution and LSTM, to analyze the effectiveness of putting convolutional layers at the start of an LSTM network. As mentioned previously, convolution is useful for extracting features and reducing the complexity of the data, along with helping to reduce the amount of noise and irrelevant information. This way, the LSTM layer can work with more refined data, and get better results.
- **Architecture 5 (Conv1D + Conv1D + Max Pooling + Bidirectional LSTM + Dropout + Dense + Dense):** This architecture was based on the proposed model and findings of the paper "Wearable sensor-based human activity recognition with hybrid deep learning model" [26], which used 1D convolution and bidirectional LSTM as an improvement for HAR. The network is composed of two convolutional networks, a max pooling layer, a bidirectional LSTM layer, and a

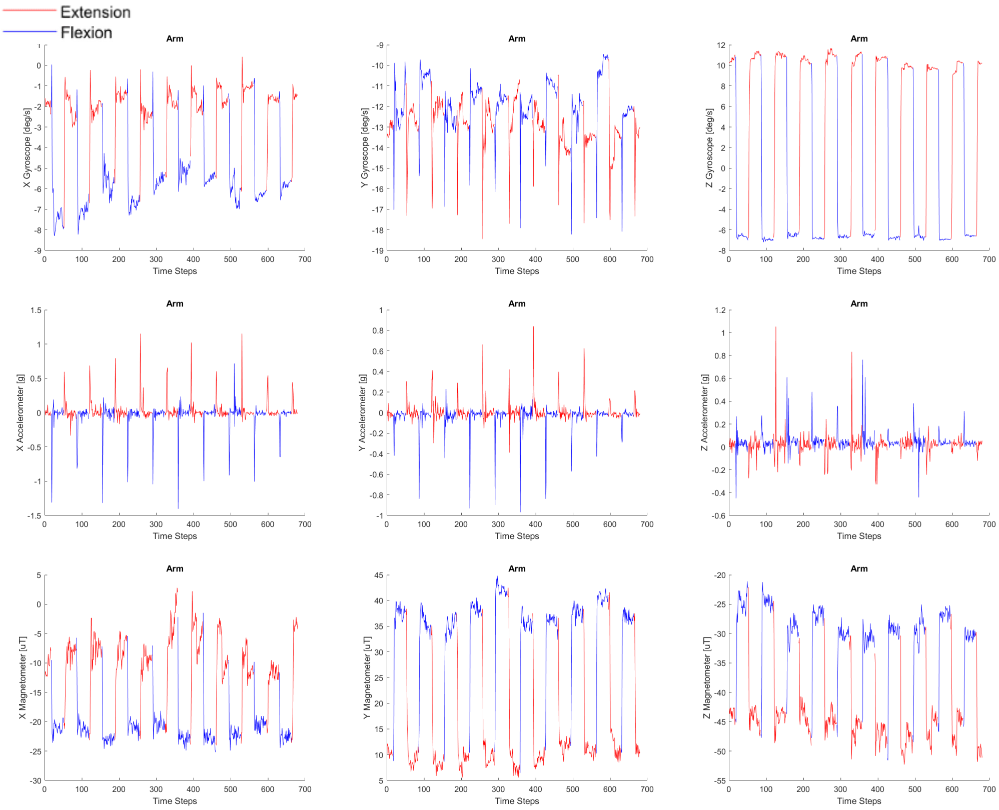


FIGURE 10. Dataset 1. MPU9250 is located at the arm during elbow flexion-extension. Plot of accelerometer, gyro, and mag signal with x,y, and z axes.

dropout and two dense layers. Additionally, from the benefits of using convolutional layers at the start of the architecture, the bidirectional LSTM allows the network to capture information from past and future inputs, helping it better capture dependencies and relationships between different parts of the sequence. This network was originally evaluated with the UCI-HAR dataset, obtaining an accuracy of 95.48%.

- **Architecture 6 (LSTM + Dropout + Reshape + Conv1D + Dropout + Dense + Dense + Dense):** This architecture was assembled empirically to test the performance and effects of placing convolutional layers at the end of the architecture instead of at the start.
- **Architecture 7 (LSTM + Dropout + Dense + Dense + Dense):** A simple LSTM architecture with 64 neurons in the first layer, a dropout layer, and three dense layers. The purpose of this network is to test how LSTM performs singlehandedly, without any particular enhancements.
- **Architecture 8 (Normalization + Position Embedding + Transformer Encoder + Normalization + Dense):** This architecture was based on the proposed model and findings of the paper "Wearable Sensor-Based Human Activity Recognition with Transformer Model" [49], which used a unidirectional Transformer-based architecture as an improvement for HAR. The network is composed of a normalization layer, a positional embedding layer coupled with a sum of weights, a transformer encoder, another normalization layer, and a fully connected layer. The encoder itself contains more layers: first, there is a normalization layer, then a multi-head attention layer that does the main work, and a dropout layer. Then, a sum of weights processes the results obtained previously, which go to a normalization layer, a feed-forward network, and a dropout layer. Before exiting the encoder, a final sum of weights is performed. Using this architecture, the authors take advantage of the benefits of using multi-head attention, described previously. This network was originally evaluated with the KU-HAR dataset, obtaining an accuracy of 99.20%.

C. FEATURE SELECTION

Due to multiple correlated features available, at most 36 features, we propose to find the most important features using random forest feature selection. After the feature selection, we evaluated the 8 architectures with the reduced number of features. It is relevant to find the most important features for IMU and MAS because it allows to reduce the computational complexity, reduce the physical sensors required, sensor information redundancy, and possible

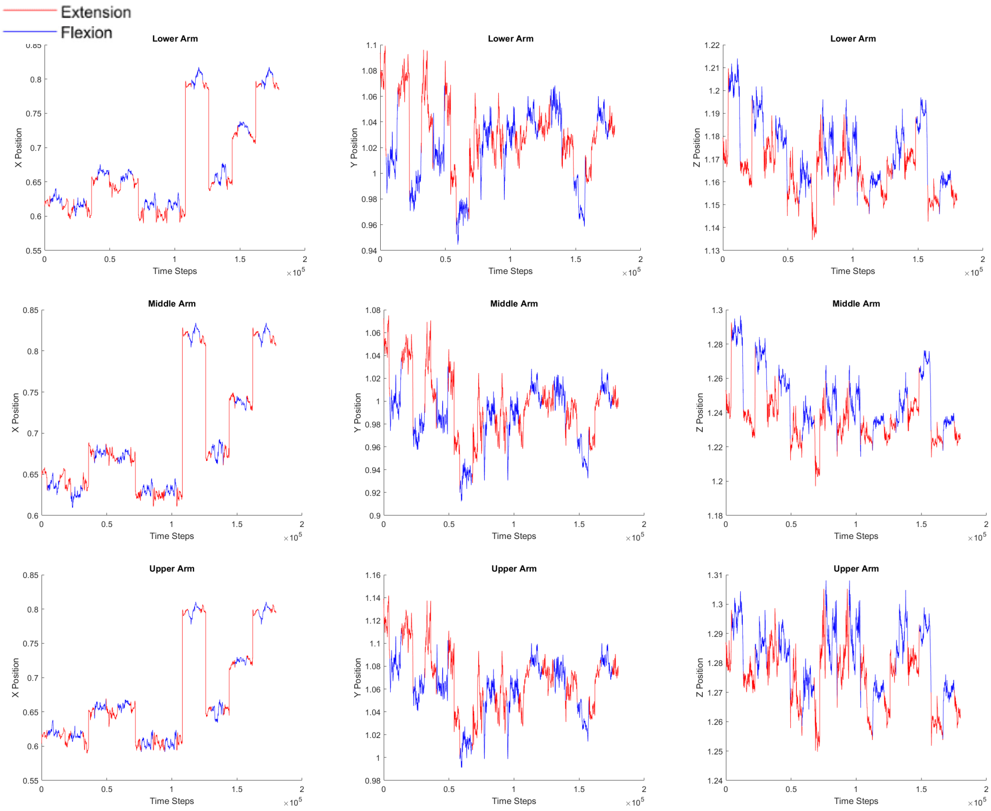
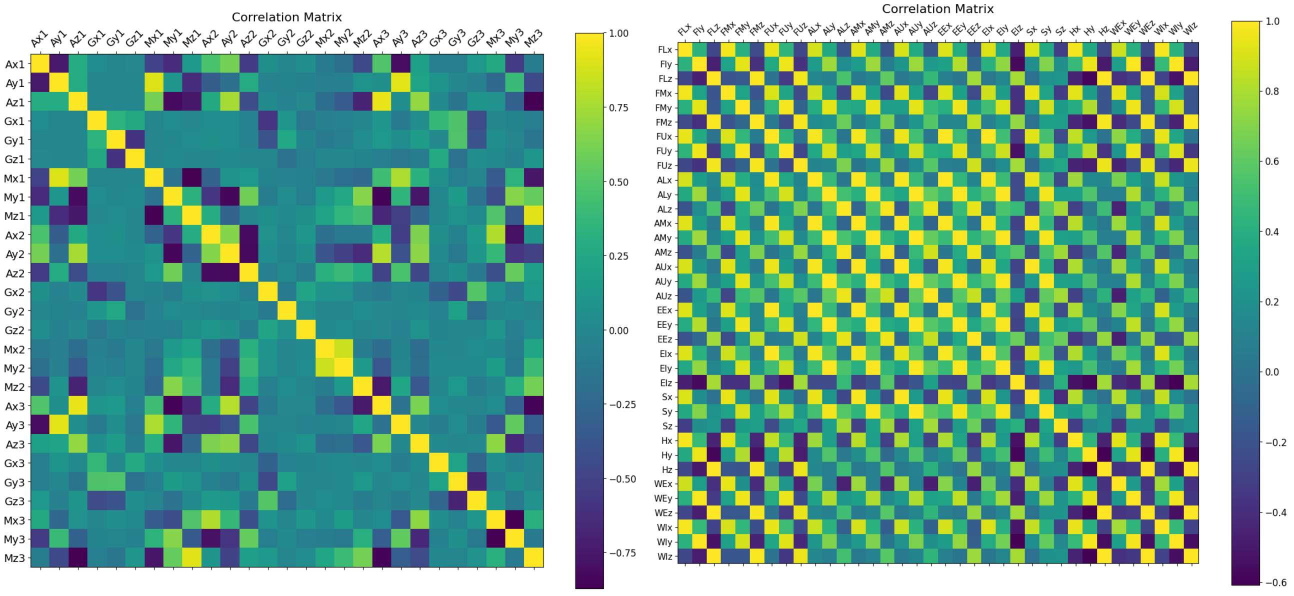


FIGURE 11. Dataset 1. Visual markers located at the lower, middle, and upper arm during elbow flexion extension. The plot of marker position signal in the x,y, and z axes.



(a) Dataset 1 - MPU9250

(b) Dataset 1 - Movement Analysis System

FIGURE 12. Correlation matrices between features of Dataset 1.

overfitting. Figure 12 shows a higher correlation in MAS than in IMU signals because of the three markers located at each position.

The algorithm used for feature selection, random forest (RF) [50], [51], is an ensemble learning method combining multiple decision trees to improve the accuracy of the model.

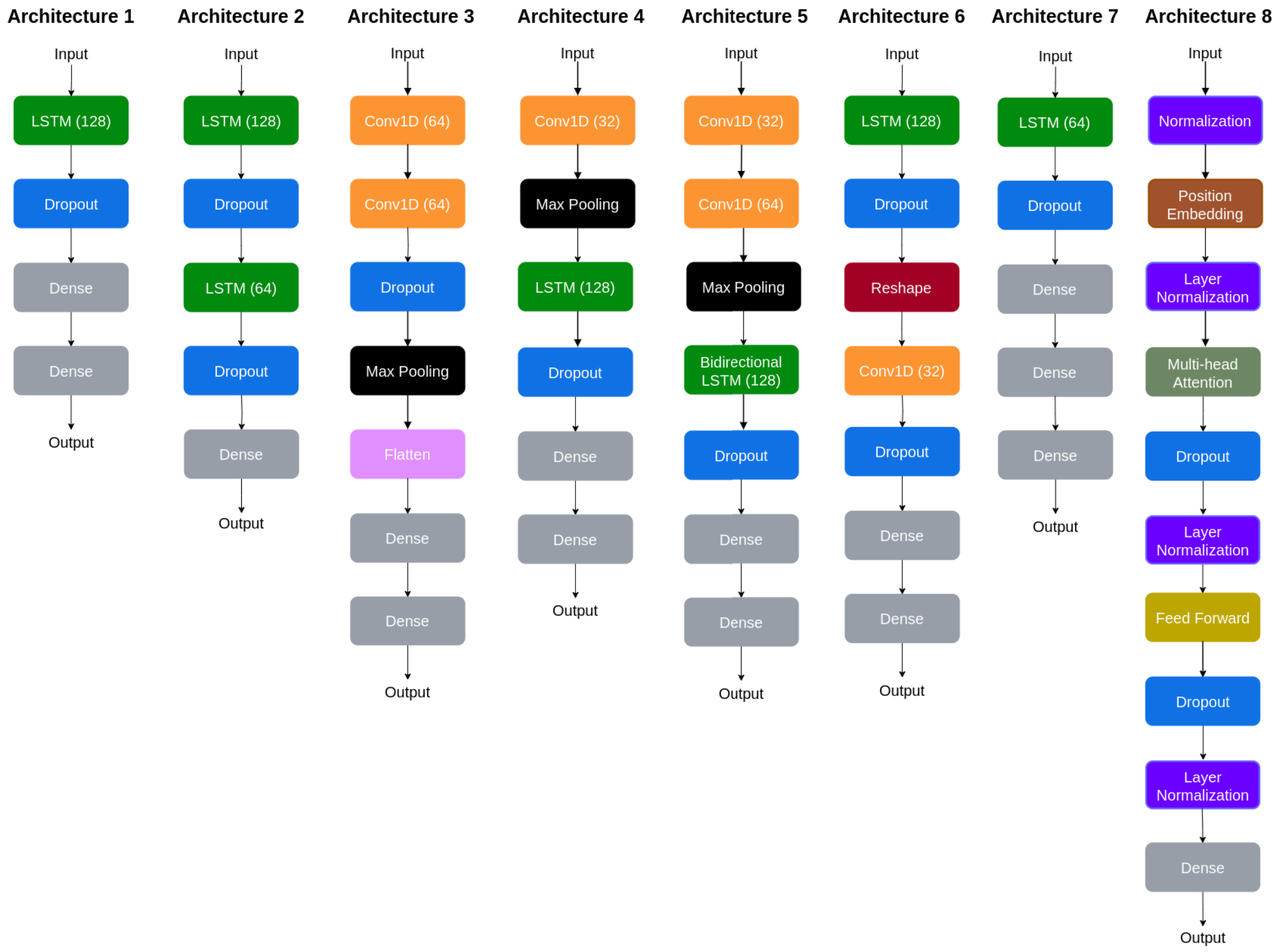


FIGURE 13. Graphical representation of the eight architectures.

The algorithm creates an ensemble of decision trees during training, where each tree is trained on a different subset of data and features, i.e., bootstrap and bagging. The majority vote of the individual trees determines the final output of the algorithm. Each node in a tree makes a binary decision according to a single feature. RF identifies the features which were the best to split the data, then are organized as most important features according to their score [52].

Random forest was trained on every dataset. Feature selection was based on the importance score from a random forest. To select the amount number of important features for training, we began with a small number of features according to their importance. Then, we added features until the model no longer demonstrated any enhancement in accuracy or showed signs of overfitting. Tables 8 to 10 show the best sensors and locations found and their score for each dataset. Figure 15 shows the training curves of the best architectures for every dataset using only the best features.

V. RESULTS

We trained models using the 8 architectures and the 7 sub-datasets. Our metrics were: test accuracy, precision, recall, and F1 score. Each metric was the average of 5 repetitions using k-fold cross validation. The result analysis needs to consider the exercise duration, sampling rate, and window size.

Table 7 shows the summary of the test accuracy. Confusion matrices figure 12. Training curves on figures 14, 15. Best features in tables 8 to 10. Best architectures for each dataset in table 11, and best architectures using only best features in table 12.

VI. DISCUSSION

In this section, we present the most relevant results and how they relate to previous works.

Test accuracy summary: Table 7 includes the capture conditions and test accuracy for all the experiments. We discuss how the sampling rate, window size, and exercise duration affect the test accuracy.

In Dataset 1, we have IMU sampling at 1 Hz and MAS at 200 Hz. IMU data at 1 Hz can distinguish exercises because the people took 90 seconds to complete each exercise repetition. Approximately 40 seconds stay at flexion and 40 seconds at extension. The change from flexion to extension took 3 seconds. The exercises are elbow flexion-extension and elbow pronation-supination. These exercises involve

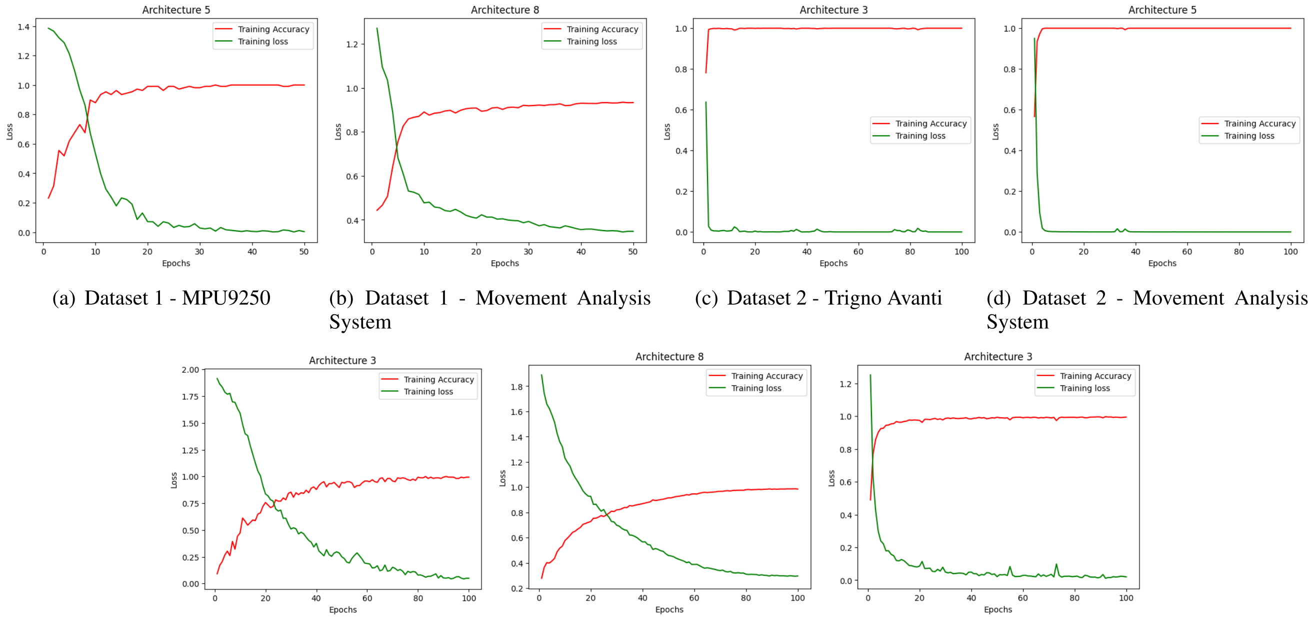
TABLE 7. Mean test accuracy for all models. All features and best features. Datasets description.

	Dataset 1		Dataset 2		Dataset 3			
	MPU9250	Visual Markers	Trigno Avanti	Visual Markers	MPU9250	Metawear	Visual Markers	
Timesteps	1360	313000	502000	270000	8200	136000	545000	
Samples	136	1568	2500	1350	205	4500	5450	
Window size	10	200	200	200	40	30	100	
Sampling rate	1	200	370	200	1	50	50	
Exercise duration seconds	90	90	2	2	4	4	4	
Exercise size points	90	18000	740	400	4	200	200	
window / exercise size	0.11	0.01	0.27	0.5	10	0.15	0.5	
Acc mean	0.961	0.876	0.612	0.958	0.464	0.837	0.981	
All features, Mean Test Accuracy								
Architectures	MPU9250	Visual Markers	Trigno Avanti	Visual Markers	MPU9250	Metawear	Visual Markers	Mean
A1 - LSTM 128 neuron	0.956 (± 0.036)	0.877 (± 0.015)	0.314 (± 0.023)	1.000 (± 0.000)	0.189 (± 0.039)	0.847 (± 0.017)	0.982 (± 0.004)	0.738
A2 - Double LSTM layer	0.956 (± 0.036)	0.881 (± 0.012)	0.317 (± 0.028)	1.000 (± 0.000)	0.257 (± 0.032)	0.843 (± 0.022)	0.980 (± 0.004)	0.748
A3 - Two layer conv	0.963 (± 0.047)	0.886 (± 0.012)	0.998 (± 0.002)	1.000 (± 0.000)	0.845 (± 0.032)	0.886 (± 0.009)	0.989 (± 0.004)	0.938
A4 - Conv + LSTM	0.963 (± 0.033)	0.872 (± 0.010)	0.996 (± 0.003)	1.000 (± 0.000)	0.611 (± 0.082)	0.871 (± 0.013)	0.986 (± 0.003)	0.900
A5 - 2 Conv + BiLSTM	0.970 (± 0.036)	0.885 (± 0.019)	0.998 (± 0.003)	1.000 (± 0.000)	0.738 (± 0.032)	0.882 (± 0.006)	0.985 (± 0.002)	0.923
A6 - LSTM + 1DConv	0.956 (± 0.043)	0.843 (± 0.055)	0.304 (± 0.011)	1.000 (± 0.000)	0.131 (± 0.043)	0.680 (± 0.024)	0.969 (± 0.014)	0.698
A7 - LSTM 64 neuron	0.970 (± 0.036)	0.875 (± 0.021)	0.304 (± 0.011)	1.000 (± 0.000)	0.189 (± 0.032)	0.770 (± 0.013)	0.974 (± 0.006)	0.726
A8 - Transformer encoder	0.954 (± 0.030)	0.892 (± 0.008)	0.666 (± 0.017)	0.665 (± 0.019)	0.757 (± 0.090)	0.918 (± 0.009)	0.988 (± 0.002)	0.834
Best features, Mean Test accuracy								
Architectures	MPU9250	Visual Markers	Trigno Avanti	Visual Markers	MPU9250	Metawear	Visual Markers	Mean
A1 - LSTM 128 neuron	0.948 (± 0.044)	0.874 (± 0.021)	0.869 (± 0.249)	0.992 (± 0.005)	0.703 (± 0.073)	0.834 (± 0.014)	0.966 (± 0.010)	0.884
A2 - Double LSTM layer	0.948 (± 0.055)	0.882 (± 0.022)	0.992 (± 0.001)	0.776 (± 0.179)	0.771 (± 0.051)	0.820 (± 0.025)	0.965 (± 0.011)	0.879
A3 - Two layer conv	0.963 (± 0.033)	0.890 (± 0.011)	0.998 (± 0.001)	0.999 (± 0.001)	0.864 (± 0.030)	0.882 (± 0.012)	0.984 (± 0.004)	0.940
A4 - Conv + LSTM	0.963 (± 0.033)	0.882 (± 0.007)	0.998 (± 0.001)	0.999 (± 0.002)	0.796 (± 0.029)	0.853 (± 0.009)	0.970 (± 0.009)	0.923
A5 - 2 Conv + BiLSTM	0.963 (± 0.033)	0.886 (± 0.016)	0.996 (± 0.003)	0.999 (± 0.001)	0.825 (± 0.042)	0.898 (± 0.005)	0.983 (± 0.003)	0.936
A6 - LSTM + 1DConv	0.948 (± 0.044)	0.870 (± 0.019)	0.991 (± 0.005)	0.979 (± 0.020)	0.564 (± 0.098)	0.718 (± 0.033)	0.924 (± 0.015)	0.856
A7 - LSTM 64 neuron	0.941 (± 0.055)	0.874 (± 0.013)	0.992 (± 0.007)	0.865 (± 0.099)	0.699 (± 0.058)	0.775 (± 0.021)	0.927 (± 0.013)	0.868
A8 - Transformer encoder	0.970 (± 0.043)	0.892 (± 0.017)	0.666 (± 0.017)	0.666 (± 0.019)	0.891 (± 0.038)	0.924 (± 0.012)	0.974 (± 0.002)	0.855

principally the movement of the hand and forearm. The best features were accelerometers located at the hand. This suggests we only need a wrist smartwatch equipped with an accelerometer and gyroscope as in [26] where they suggest using a single accelerometer.

Dataset 1 IMU has 27 features and 4 selected features, both with 97% acc. Dataset 1 MAS has 36 features and 15 selected features, both with 89% acc. Reduced accuracy in MAS is related to occlusion during pronation-supination exercise. MAS system cannot correctly detect small movements. There is no improvement in using fewer features because of the problems with occlusion and small movements. More features were needed from MAS than from IMU because MAS do not have enough information to distinguish the exercises.

In Dataset 2, we have IMU sampling at 370 Hz and MAS at 200 Hz. Both rates achieve a test accuracy of 99%. We have a good amount of data and used a window size



(a) Dataset 1 - MPU9250

(b) Dataset 1 - Movement Analysis System

(c) Dataset 2 - Trigno Avanti

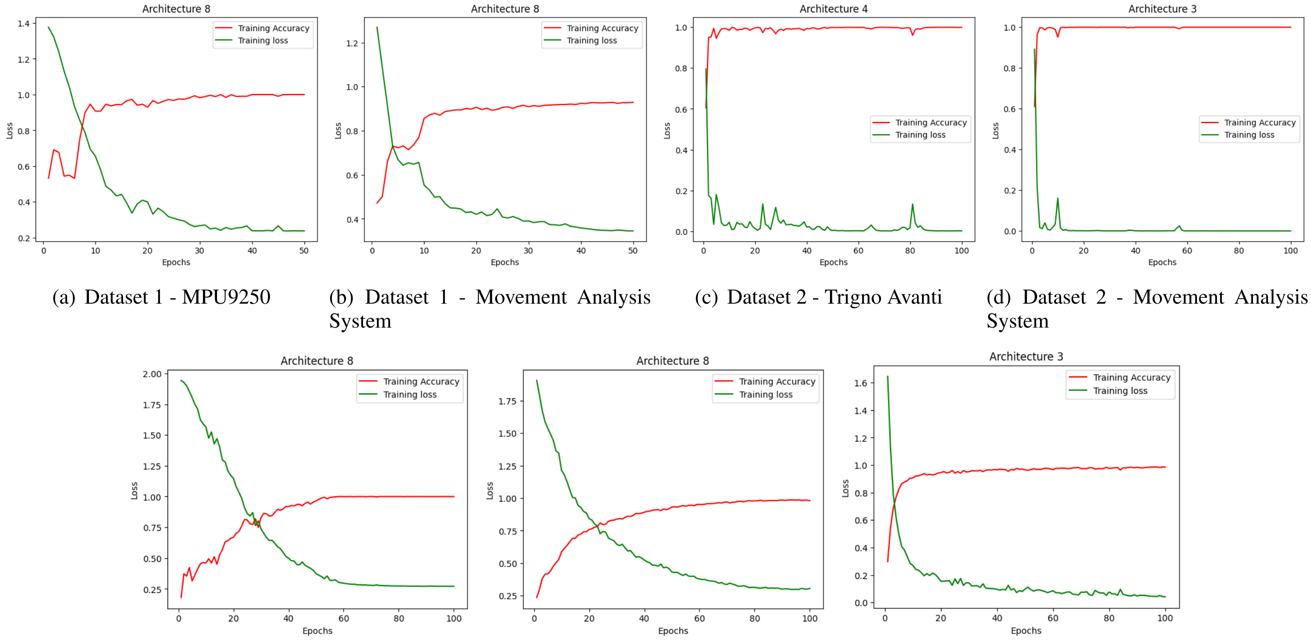
(d) Dataset 2 - Movement Analysis System

(e) Dataset 3 - MPU9250

(f) Dataset 3 - Metawear

(g) Dataset 3 - Movement Analysis System

FIGURE 14. Plot of best training curves for every dataset using all features.



(a) Dataset 1 - MPU9250

(b) Dataset 1 - Movement Analysis System

(c) Dataset 2 - Trigno Avanti

(d) Dataset 2 - Movement Analysis System

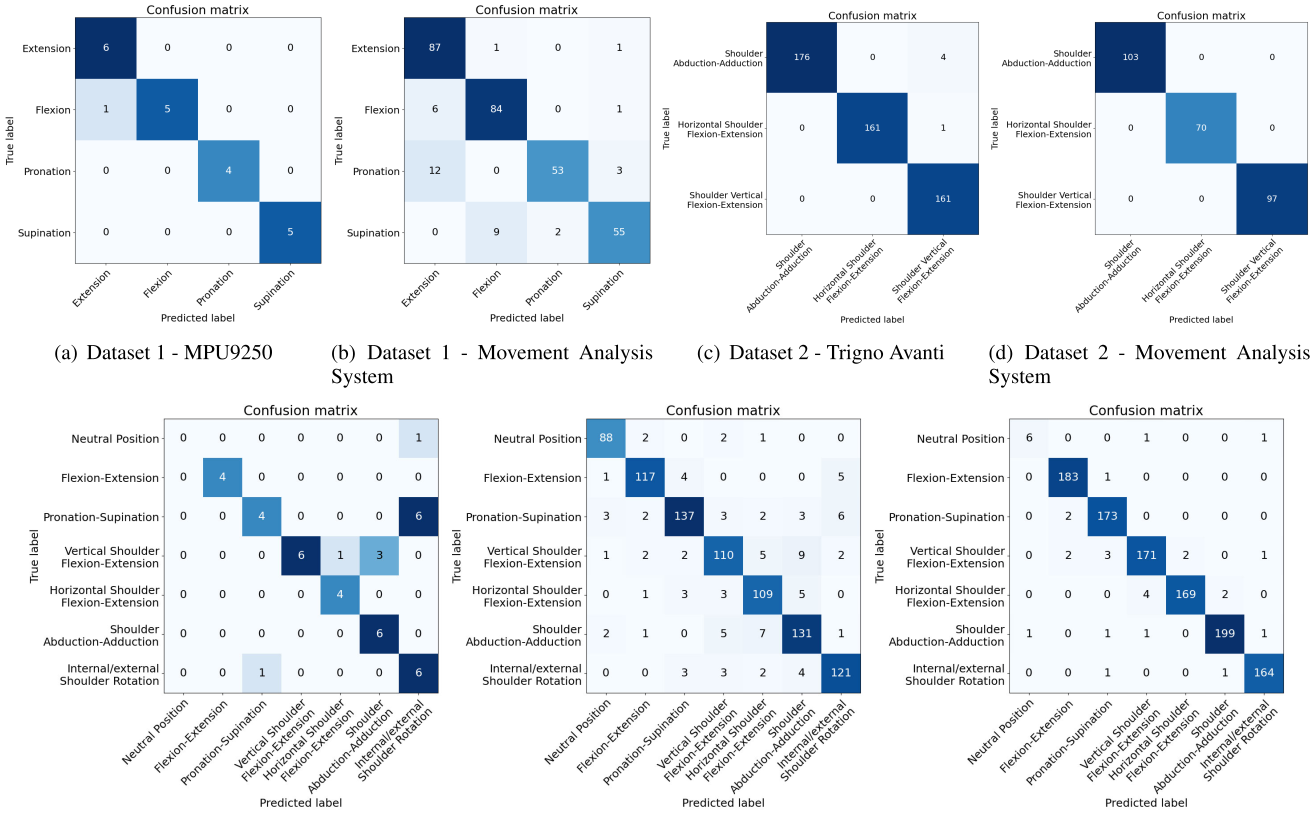
(e) Dataset 3 - MPU9250

(f) Dataset 3 - Metawear

(g) Dataset 3 - Movement Analysis System

FIGURE 15. Plot of best architectures for every dataset using best features.

of 200 points. The exercises in dataset 2 performs fully extended arm exercises, which are easily distinguished due to visible markers and large movements. This is the only dataset with acceleration estimated from the MAS system. Using only 3 position features we achieve 99% accuracy, the good performance is related to the exercise but not the use of acceleration from MAS. MAS data could be augmented with the velocity markers as an additional feature. IMU data could be augmented with additional features from the accumulation and derivative.



(a) Dataset 1 - MPU9250

(b) Dataset 1 - Movement Analysis System

(c) Dataset 2 - Trigno Avanti

(d) Dataset 2 - Movement Analysis System

(e) Dataset 3 - MPU9250

(f) Dataset 3 - Metawear

(g) Dataset 3 - Movement Analysis System

FIGURE 16. Confusion matrix of the best-performing architecture of every dataset (best features).

TABLE 8. Best features for dataset 1, ordered by score.

Dataset 1 - MPU9250			Dataset 1 - Movement Analysis System		
Max Accuracy with all features (27): 100% **Max Accuracy with best features (4):** 100%			**Max Accuracy with all features (36):** 87.6% **Max Accuracy with best features (15):** 87.2%		
Feature	**Description**	**Score**	**Feature**	**Description**	**Score**
HAy	hand accelerometer	0.136	FLz	forearm lower	0.100
HAx	hand accelerometer	0.125	FMz	forearm middle	0.090
HAz	hand accelerometer	0.122	Hz	hand	0.086
HMz	hand magnetometer	0.095	WIz	wrist internal	0.072
			FUz	forearm upper	0.069
			WEz	wrist internal	0.067
			FUy	forearm upper	0.040
			EIy	elbow internal	0.034
			WEx	wrist external	0.034
			AUy	arm upper	0.030
			FMx	forearm middle	0.026
			FLy	forearm lower	0.026
			Hy	hand	0.025
			WIx	wrist internal	0.024
			Hx	hand	0.020

Dataset 2 IMU has 24 features and 6 selected features, both with 99% acc. Dataset 2 MAS has 24 features and 3 selected features, both with 99% acc. The selected features were located at the arm, forearm, and shoulder. IMU at hand features were not selected because those sensors were affected by large accelerations and were not able to distinguish the exercises. Hand acceleration due to different arm lengths didn't show any effect on the exercise classification.

In Dataset 3, we have MPU9250 sampling at 1 Hz, Metawear at 50 Hz, and MAS at 50 Hz. This dataset performs the same exercises from Dataset 1 plus Dataset 2. D3 MPU9250 at 1 Hz is the first dataset that shows overall bad performance, 13% to 85% accuracy. This is because sampling

TABLE 9. **Best features for dataset 2, ordered by score.**

Dataset 2 - Trigno Avanti			Dataset 2 - Movement Analysis System		
Max Accuracy with all features (24): 99.6% **Max Accuracy with best features (6):** 99.8%			**Max Accuracy with all features (24):** 100% **Max Accuracy with best features (3):** 100%		
Feature	**Description**	**Score**	**Feature**	**Description**	**Score**
SAy	shoulder accelerometer	0.266	ALz	arm lower	0.211
AAz	arm accelerometer	0.132	AMz	arm middle	0.132
FAz	forearm accelerometer	0.118	FLz	forearm lower	0.118
SAx	shoulder accelerometer	0.111			
AAx	arm accelerometer	0.111			
AAy	arm accelerometer	0.073			

TABLE 10. **Best features for dataset 3, ordered by score.**

Dataset 3 - MPU9250			Dataset 3 - Metawear			Dataset 3 - Movement Analysis System		
Max Accuracy with all features (27): 81.0% **Max Accuracy with best features (5):** 80.1%			**Max Accuracy with all features (18):** 89.7% **Max Accuracy with best features (12):** 90.8%			**Max Accuracy with all features (36):** 98.7% **Max Accuracy with best features (15):**: 98.6%		
Feature	**Description**	**Score**	**Feature**	**Description**	**Score**	**Feature**	**Description**	**Score**
AGx	arm gyroscope	0.124	AAy	arm accelerometer	0.119	Sz	shoulder	0.123
AAy	arm accelerometer	0.101	AAz	arm accelerometer	0.098	Sy	shoulder	0.103
AGy	arm gyroscope	0.090	HAx	hand accelerometer	0.091	AMx	arm middle	0.080
AGz	arm gyroscope	0.062	HAz	hand accelerometer	0.087	AUx	arm upper	0.078
AMy	arm magnetometer	0.048	AAx	arm accelerometer	0.086	Sx	shoulder	0.074
			HAy	hand accelerometer	0.081	ALx	arm lower	0.069
			FAz	forearm accelerometer	0.080	AMy	arm middle	0.058
			FAx	forearm accelerometer	0.073	AMz	arm middle	0.054
			FAy	forearm accelerometer	0.068	ALy	arm lower	0.044
			HGy	hand gyroscope	0.031	AUy	arm upper	0.034
			FGy	forearm gyroscope	0.030	AUz	arm upper	0.034
			AGx	arm gyroscope	0.028	EEx	elbow external	0.027
						FMx	forearm middle	0.023
						EIx	elbow internal	0.021
						EIz	elbow internal	0.020

at 1 Hz is too low for a movement that takes 4 seconds to complete. We used a window size of 40 points capturing multiple exercise repetitions, which produced bad accuracy. Metawear showed low performance, 68% to 92% because the system presented data transmission problems. MAS system did not show any problems with a sampling of 50 Hz and a window size of 100 points. This result shows we can have low sampling rates around 50 Hz.

Dataset 3 MPU9250 has 27 features with 84% acc and 5 selected features with 89%. D3 Metawear has 18 features and 12 selected features with 92% acc. D3 MAS has 36 features with 99% acc and 15 selected features with 98% acc. MAS requires more features than IMU because of the complexity of the exercises, markers occlusion, and difficulty in detecting small movements.

Training curves: During training, it takes at most 100 epochs to achieve a steady state of accuracy and loss. All datasets show loss reduction and accuracy increase over time.

Failure cases: The most challenging datasets were D2-Trigno and D3-MPU9250. D3-MPU9250 showed bad test accuracy, this is related to a sampling rate of 1 Hz and a window size of 40 points. The sampling rate is slow for a movement that takes 4 seconds to complete. The window size is too long, capturing multiple repetitions in a single window. Having multiple exercise repetitions captured in a single window show bad performance because it is ideal to have aligned signals to recognize them. The accuracy of the challenging datasets increased using selected features. The best architectures stand out by achieving better performance on challenging datasets.

Confusion Matrices: Reflects the correct and incorrect classification for each dataset using the best features and the best model. We can see the hardest exercise to recognize was elbow pronation-supination for the MAS dataset.

Effect of window size, sampling rate, and exercise duration: We found the window size should be variable because it is proportional to the sampling rate and exercise duration. For example, D1-MPU9250 with a window size of 10 points is enough to recognize an exercise sampled at 1 Hz with a duration of 90 points. In this case, the ratio from window size to exercise duration is 10/90=0,11. From D3-MPU9250, a window size of 40 points was too big to recognize an exercise sampled at 1 Hz with a duration of 4 points. In this case, the ratio from window size to exercise duration is 40/4=10. A window should capture less than a single repetition to achieve good accuracy. From our results, ratios less than 50% achieved high accuracy.

We found the sampling rate should be proportional to the exercise duration. In daily activities, a muscular-focused exercise or a full-body exercise takes around 2 seconds to perform. To capture a well-defined shape of the movement, we found the sampling rate should be around 60 Hz. A sampling rate of 1 Hz was insufficient to correctly

TABLE 11. Results for all datasets using all features.

Dataset	Samples	Exercise duration [s]	Sampling rate	Window Size	Window /exercise Size	Total Features	Best mean test accuracy	Best mean precision	Best mean recall	Best mean F1 score	Best architecture
D1 - MPU9250	135	90	1	10	0.11	27	97%	97.6%	97%	97.1%	2Conv + BiLSTM, LSTM 64 neuron
D1 - MAS	1568	90	200	200	0.01	36	89.2%	89.3%	89.2%	89%	Transformer encoder
D2 - Trigno Avanti	2511	2	370	200	0.27	24	99.8%	99.8%	99.8%	99.8%	2 Layer conv, 2Conv + BiLSTM
D2 - MAS	1350	2	200	200	0.5	24	100%	100%	100%	100%	2 Layer conv, 2Conv + BiLSTM
D3 - MPU9250	206	4	1	40	10	27	84.5%	87.5%	84.5%	84.7%	2 Layer conv
D3 - Metawear	4540	4	50	30	0.15	18	91.8%	91.9%	91.9%	91.9%	Transformer encoder
D3 - MAS	5450	4	50	100	0.5	36	98.9%	98.9%	98.9%	98.9%	2 Layer conv

TABLE 12. Best results for all datasets using best features.

Dataset	Samples	Excercise duration [s]	Sampling rate	Window Size	Window /exercise Size	Best features (number)	Best features (names)	Best mean test accuracy	Best mean precision	Best mean recall	Best mean F1 score	Best architecture
D1 - MPU9250	135	90	1	10	0.11	4/27	hand	97%	97%	97%	97%	Transformer encoder
D1 - MAS	1568	90	200	200	0.01	15/36	forearm, hand, wrist	89.2%	89.2%	89.1%	88.9%	Transformer encoder
D2 - Trigno Avanti	2511	2	370	200	0.27	5/24	arm, forearm	99.8%	99.8%	99.8%	99.8%	Conv+LSTM
D2 - MAS	1350	2	200	200	0.5	3/24	arm, forearm	99.9%	99.9%	99.9%	99.9%	2 Layer conv, Conv+LSTM, 2Conv+BiLSTM
D3 - MPU9250	206	4	1	40	10	5/27	arm	89.1%	89.1%	88.9%	88.8%	Transformer encoder
D3 - Metawear	4540	4	50	30	0.15	12/18	arm, hand, forearm	92.4%	92.4%	92.5%	92.5%	Transformer encoder
D3 - MAS	5450	4	50	100	0.5	15/36	arm	98.4%	98.4%	98.4%	98.4%	2 Layer conv

recognize exercises. The sampling rate of 200 and 370 Hz didn't show any improvement and generated too much information.

Best models architectures: The best architectures were the same using all and selected features. The 3 best architectures were: 2LayerConv 93.8% acc, 2LayerConv+BiLSTM 92.3% acc, Conv+LSTM 90% acc. The networks analyzed are divided into LSTM first, CNN first, and Transformer.

Best architectures have two convolutional layers at the beginning. The networks are learning filters with the shape of the multiple-channel signals. During the signals analysis performed in the time series section, we noted the exercises show well-defined shapes for each exercise.

LSTM first networks achieved lower test accuracy on the challenging datasets.

Convolution + BiLSTM has good performance showing the importance of convolution and the importance of bidirectional learning on LSTM. Transformer didn't achieve good performance alone. The transformer was a unidirectional encoder. Accuracy could be improved using a bidirectional decoder architecture.

Feature reduction: Using as low as 4 features, IMU and MAS systems achieved high accuracy above 89%. Feature reduction was feasible without affecting accuracy.

To achieve good accuracy with few features, a single IMU with accelerometer and gyroscope should be located in the wrist. Therefore we recommend a single IMU in a smartwatch. Feature selection shows the IMU acceleration variable is the most representative similar to [16], [17], [18], [23], and [26]. The test accuracies above 89% found for our datasets using the 8 models were similar to the accuracies found in [26], [46], [48], [53], and [49].

VII. CONCLUSION

The duration of daily exercises is variable, then we need a variable window size. The window size is proportional to the equipment sampling rate and exercise duration. From our experiments sampling rate of 60 Hz is able to distinguish arm movements.

State of the art deep learning models was able to correctly classify exercises. Best architectures were 2LConv 93.8%, 2LConv+BiLSTM 92.3%, Conv+LSTM 90%. Selected features lead to a reduction of 4 features for IMU or MAS, with accuracies higher than 89%.

IMU allows measuring the activity of people simultaneously in any environment, even on water. IMU is portable and can be used in all kinds of daily activities. MAS needs an equipment room and complex video capture system and wearable markers. The use of cameras interferes with people's privacy inside the everyday environment.

The IMUs achieve high accuracy, low cost, and noise reduction compared to other instruments such as EEG or EMG where the signals are naturally mixed from the source. The accuracy improves using independent sources, such as independent axes in accelerometers and gyroscopes [6], [22]. In MAS, the visual markers have problems related to occlusion, distance, size of movements, and out-of-plane movements [36], e.g. pronation-supination, shoulder rotation.

From our experiments detailed in table 7, we found a relation to estimating the window size given by equation 1. Two-second exercises captured at 60Hz will require window sizes between 30 and 60 points.

$$\frac{\text{window size}}{\text{exercise duration points}} < 0.5 \tag{1}$$

$$\frac{\text{window size}}{\text{exercise duration seconds} \times \text{sampling rate}} < 0.5 \tag{2}$$

VIII. FUTURE WORK

A limitation of this work was the number of persons involved which do not reflect the behavior of the signals for a sample of population. The 10 persons chosen for this study represent only a test group to validate the capture methodology and evaluate the performance of deep learning architectures. The people involved had different arm lengths and the differences in acceleration didn't show misclassifications. A limitation of this work was the reduced number of samples captured, at most 540k timesteps for an exercise. It is important to capture data about elderly in their daily lives. The data needs to have multiple people and data for several weeks. To evaluate overfitting and to perform robust statistical accuracy tests, we need larger datasets with more people, more exercises, and include full-body exercises. With larger datasets, we can identify better algorithms for data extraction and models to recognize a wider range of movements. To achieve robust models we need to perform data augmentation adding noise, scaling, offset in time, and random signal erase. To have a better window extraction we propose replacing the fixed sliding window with dynamic time warping, wavelet transform, and spectrogram for scale and location, anchors like in image segmentation (different sized sliding windows), sliding window align by regression as in image segmentation. Feature reduction suggested it is possible to achieve high accuracy with few sensors. We propose to use a single IMU located at the wrist being worn as a smartwatch. The challenge is to recognize exercises with a single IMU and implement variable window size. To classify exercises with a single IMU we need to capture the variations in signal amplitudes, shapes, and durations. A single IMU smartwatch solution with HAR capabilities is interesting for elderly care, athletes, and healthcare. The solution involves a smartwatch to capture and send data to a processing unit. The data necessary would be at 60 Hz and 6 features, making this solution feasible.

Our main contributions were:

- Capture methodology and comparison for IMU and MAS acquisition systems.
- We analyzed the relations between 8 architectures accuracies, signal waveforms, signals correlation, sampling rate, exercise duration, and window size.
- Feature reduction analysis.
- Deep learning models were able to recognize human exercises. The next challenge is to classify using a single IMU and use variable window size.

REFERENCES

[1] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, Feb. 2019.

[2] N. Ma, Z. Wu, Y.-M. Cheung, Y. Guo, Y. Gao, J. Li, and B. Jiang, "A survey of human action recognition and posture prediction," *Tsinghua Sci. Technol.*, vol. 27, no. 6, pp. 973–1001, Dec. 2022.

[3] M. Mokari, H. Mohammadzade, and B. Ghojogh, "Recognizing involuntary actions from 3D skeleton data using body states," 2017, *arXiv:1708.06227*.

[4] S. Khan, M. A. Khan, M. Alhaisoni, U. Tariq, H.-S. Yong, A. Armghan, and F. Alenezi, "Human action recognition: A paradigm of best deep learning features selection and serial based extended fusion," *Sensors*, vol. 21, no. 23, p. 7941, Nov. 2021.

[5] M. Jian, S. Zhang, L. Wu, S. Zhang, X. Wang, and Y. He, "Deep key frame extraction for sport training," *Neurocomputing*, vol. 328, pp. 147–156, Feb. 2019.

[6] P. Sarcevic, Z. Kincses, and S. Pletl, "Online human movement classification using wrist-worn wireless sensors," *J. Ambient Intell. Humanized Comput.*, vol. 10, no. 1, pp. 89–106, Jan. 2019.

[7] M. Z. Uddin and A. Soylu, "Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning," *Sci. Rep.*, vol. 11, no. 1, p. 16455, Aug. 2021.

[8] R. Tokarczyk and T. Mazur, "Photogrammetry-principles of operation and application in rehabilitation," *Med. Rehabil.*, vol. 10, pp. 31–38, Jan. 2006.

[9] C. J. van Andel, N. Wolterbeek, C. A. M. Doorenbosch, D. E. J. Veeger, and J. Harlaar, "Complete 3D kinematics of upper extremity functional tasks," *Gait Posture*, vol. 27, no. 1, pp. 120–127, Jan. 2008.

[10] C. Avilés-Cruz, A. Ferreyra-Ramírez, A. Zúñiga-López, and J. Villegas-Cortéz, "Coarse-fine convolutional deep-learning strategy for human activity recognition," *Sensors*, vol. 19, no. 7, p. 1556, 2019.

[11] M. Ramanathan, W.-Y. Yau, and E. K. Teoh, "Human action recognition with video data: Research and evaluation challenges," *IEEE Trans. Hum.-Mach. Syst.*, vol. 44, no. 5, pp. 650–663, Oct. 2014.

[12] E. M. Mikhail, J. S. Bethel, and J. C. McGlone, *Introduction to Modern Photogrammetry*. Hoboken, NJ, USA: Wiley, 2001.

[13] A. Bux, P. Angelov, and Z. Habib, "Vision based human activity recognition: A review," in *Advances in Computational Intelligence Systems*. Lancaster, U.K.: Springer, 2017, pp. 341–371.

[14] M. Yoneyama, Y. Kurihara, K. Watanabe, and H. Mitoma, "Accelerometry-based gait analysis and its application to Parkinson's disease assessment— Part 2 : A new measure for quantifying walking behavior," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 21, no. 6, pp. 999–1005, Nov. 2013.

[15] P. Bonato, "Wearable sensors/systems and their impact on biomedical engineering," *IEEE Eng. Med. Biol. Mag.*, vol. 22, no. 3, pp. 18–20, May 2003.

[16] C. A. Ronao and S.-B. Cho, "Human activity recognition using smartphone sensors with two-stage continuous hidden Markov models," in *Proc. 10th Int. Conf. Natural Comput. (ICNC)*, Aug. 2014, pp. 681–686.

[17] Q. Xiong, J. Zhang, P. Wang, D. Liu, and R. X. Gao, "Transferable two-stream convolutional neural network for human action recognition," *J. Manuf. Syst.*, vol. 56, pp. 605–614, Jul. 2020.

[18] J. Lu and K.-Y. Tong, "Robust single accelerometer-based activity recognition using modified recurrence plot," *IEEE Sensors J.*, vol. 19, no. 15, pp. 6317–6324, Aug. 2019.

[19] C.-T. Yen, J.-X. Liao, and Y.-K. Huang, "Human daily activity recognition performed using wearable inertial sensors combined with deep learning algorithms," *IEEE Access*, vol. 8, pp. 174105–174114, 2020.

[20] A. L. Clouthier, G. B. Ross, and R. B. Graham, "Sensor data required for automatic recognition of athletic tasks using deep neural networks," *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 473, Jan. 2020.

[21] M. Jaén-Vargas, K. Rivas, F. Fernandes, S. Gonçalves, M. T. Silva, D. Lopes, and J. S. Olmedo, "A deep learning approach to recognize human activity using inertial sensors and motion capture systems," *Frontiers Artif. Intell. Appl.*, vol. 340, pp. 250–256, 2021.

[22] Y. Li, X. Zhang, Y. Gong, Y. Cheng, X. Gao, and X. Chen, "Motor function evaluation of hemiplegic upper-extremities using data fusion from wearable inertial and surface EMG sensors," *Sensors*, vol. 17, no. 3, p. 582, Mar. 2017.

[23] N. Lemieux and R. Noumeir, "A hierarchical learning approach for human action recognition," *Sensors*, vol. 20, no. 17, p. 4946, Sep. 2020.

[24] O. Banos, R. Garcia, J. A. Holgado-Terriza, M. Damas, H. Pomares, I. Rojas, A. Saez, and C. Villalonga, "mHealthDroid: A novel framework for agile development of mobile health applications," in *Ambient Assisted Living and Daily Activities*. Belfast, U.K.: Springer, 2014, pp. 91–98.

[25] C. Han, L. Zhang, Y. Tang, W. Huang, F. Min, and J. He, "Human activity recognition using wearable sensors by heterogeneous convolutional neural networks," *Exp. Syst. Appl.*, vol. 198, Jul. 2022, Art. no. 116764.

[26] Y. J. Luwe, C. P. Lee, and K. M. Lim, "Wearable sensor-based human activity recognition with hybrid deep learning model," *Informatics*, vol. 9, no. 3, p. 56, Jul. 2022.

[27] D. K. Ramsey and P. F. Wretenberg, "Biomechanics of the knee: Methodological considerations in the in vivo kinematic analysis of the tibiofemoral and patellofemoral joint," *Clin. Biomechanics*, vol. 14, no. 9, pp. 595–611, Nov. 1999.

[28] A. G. Cutti, G. Paolini, M. Troncossi, A. Cappello, and A. Davalli, "Soft tissue artefact assessment in humeral axial rotation," *Gait Posture*, vol. 21, no. 3, pp. 341–349, Apr. 2005.

[29] M. Jaén-Vargas, K. M. R. Leiva, F. Fernandes, S. B. Goncalves, M. T. Silva, D. S. Lopes, and J. J. S. Olmedo, "Effects of sliding window variation in the performance of acceleration-based human activity recognition using deep learning models," *PeerJ Comput. Sci.*, vol. 8, p. e1052, Aug. 2022.

[30] G. Cicirelli, D. Impedovo, V. Dentamaro, R. Marani, G. Pirlo, and T. R. D'Orazio, "Human gait analysis in neurodegenerative diseases: A review," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 1, pp. 229–242, Jan. 2022.

[31] V. Dentamaro, D. Impedovo, and G. Pirlo, "Gait analysis for early neurodegenerative diseases classification through the kinematic theory of rapid human movements," *IEEE Access*, vol. 8, pp. 193966–193980, 2020.

[32] M. Shayestegan, J. Kohout, K. Trnková, M. Chovanec, and J. Mareš, "Motion tracking in diagnosis: Gait disorders classification with a dual-head attentional transformer-LSTM," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, p. 98, Jun. 2023.

[33] M. Cheriet, V. Dentamaro, M. Hamdan, D. Impedovo, and G. Pirlo, "Multi-speed transformer network for neurodegenerative disease assessment and activity recognition," *Comput. Methods Programs Biomed.*, vol. 230, Mar. 2023, Art. no. 107344.

[34] A. Cosma, A. Catruna, and E. Radoi, "Exploring self-supervised vision transformers for gait recognition in the wild," *Sensors*, vol. 23, no. 5, p. 2680, Mar. 2023.

[35] DELSYS. (Sep. 2022). *Documentation*. [Online]. Available: https://delsys.com/support/documentation/#usersguide

[36] A. Tobar, B. Lopez, M. F. Trujillo, and A. Rosales, "Machine learning to determine upper extremity motion from inertial measurement unit signals," in *Proc. IEEE 6th Ecuador Tech. Chapters Meeting (ETCM)*, Oct. 2022, pp. 1–6.

[37] A. Kapandji, *Fisiología Articular, Tomo I, Tomo II, Tomo III*. Madrid, Spain: Panamericana, 2010.

[38] S. Yadav and S. Shukla, "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification," in *Proc. IEEE 6th Int. Conf. Adv. Comput. (IACC)*, Feb. 2016, pp. 78–83.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[40] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space Odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[41] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[42] E. Akleman, "Deep learning," *Computer*, vol. 53, no. 9, p. 17, Sep. 2020.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[44] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, "A comparative study on transformer vs RNN in speech applications," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 449–456.

[45] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, and M. Funtowicz, "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Processing: Syst. Demonstrations*, 2020, pp. 38–45.

[46] P. Nabriya. (Jul. 2021). *Implementing LSTM for Human Activity Recognition Using Smartphone Accelerometer Data*. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/07/implementing-lstm-for-human-activity-recognition-using-smartphone-accelerometer-data/

[47] S. Das. (2021). *Human-Activity-Recognition*. [Online]. Available: https://github.com/srvds/Human-Activity-Recognition

[48] P. Rathnayakas. (2021). *ETFA-workshop*. [Online]. Available: https://github.com/CDAC-lab/ETFA-Workshop

[49] I. D. Luptáková, M. Kubovčík, and J. Pospíchal, "Wearable sensor-based human activity recognition with transformer model," *Sensors*, vol. 22, no. 5, p. 1911, 2022.

[50] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[51] T. Kam Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, 1995, pp. 278–282.

[52] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[53] S. Das, A. Chaudhary, F. Bremond, and M. Thonnat, "Where to focus on for human action recognition?" in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 71–80.

MARÍA FERNANDA TRUJILLO-GUERRERO (Student Member, IEEE) received the B.S. degree in electronic engineering from Politécnica Nacional, Quito, Ecuador, in 2014, and the M.Sc. degree in biomedical engineering from Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2016.

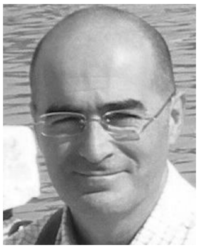
She has been a Lecturer with Escuela Politécnica Nacional, since 2017. She has been a part of the Bioinstrumentation and Nanomedicine Research Group, Center for Biomedical Technology. Her research interests include wearable sensing, human activity recognition, and deep learning applied to the health domain.

STADYN ROMÁN-NIEMES is currently pursuing the B.E. degree in information technology with Yachay Tech University. He was an Intern and a Research Assistant with Digevo, an AI solutions business. He is also an Independent Researcher with Harkay. He has also published two conference papers that covered topics, such as artificial intelligence, image processing, and computer vision. His main research interests include artificial intelligence, data science, computer vision, and natural language processing.

MILAGROS JAÉN-VARGAS was born in Panamá, in 1988. She received the B.S. degree in electronic and telecommunication engineering from Universidad Tecnológica de Panamá, in 2011, and the master's degree in project management from Universidad Interamericana de Panamá, in 2013. In 2018, she was granted an IFARHU-SENACYT scholarship to pursue the Ph.D. degree in biomedical engineering with Universidad Politécnica de Madrid, Spain. Currently, she is with the Assistive Technologies Group, Bioinstrumentation and Nanomedicine Laboratory (LBN), Technology Biomedical Center (CTB), Madrid, Spain. Hence, she is doing research on aims focused on limited vision or visually impaired people to develop technologies for its safe navigation, guiding, and rehabilitation. In addition, she is working with AI algorithms focused on recognizing human activities.

ALFONSO CADIZ received the degree in electronic engineering in digital systems from Federico Santa María University, Chile, in 1997, and the master's degree in artificial intelligence from Universidad Internacional Menéndez Pelayo, España, in 2018. He is currently a serial entrepreneur, an investor, and the director of multiple startups. He has founded over 25 technology companies and completed three successful with its partners. He is also the Technology Director of Digevo, a Chilean group with operations in 15 countries. He is also responsible for the technology of the group and all associated companies, with vigorous activity in Latam, focused on the development of technology to support the entrepreneurship, innovation, and evolution of the group in the technological and management areas.

RICARDO FONSECA received the Ph.D. degree in electronic engineering from Universidad Federico Santa Maria, Chile, in 2021. He is currently the CTO of Computer Vision with Digevo, where they develop solutions for smart retail and conversational AI. He works on large scale deploys of deep learning models for person and vehicle analytics, speech to text, text to speech, and large language models.



JOSÉ JAVIER SERRANO-OLMEDO received the Graduate and Ph.D. degrees in telecommunication engineering from the Engineering School on Telecommunication (Escuela Técnica Superior de Ingeniería de Telecomunicación), Technical University of Madrid [Universidad Politécnica de Madrid, (UPM)], in 1990 and 1996, respectively. He has been teaching Electronic Instrumentation, Bioinstrumentation, Biosensors, Technologies for Nanomedicine, Human Computer Interfaces, Electronic Health Records, and Clinical Engineering as an Associate Professor, since 1998, with the Technical University of Madrid, and a Full Professor, since 2023. He is the Coordinator of the Doctorate Program in Biomedical Engineering. He is a fellow of the Networking Center for Biomedical Research on Bioengineering, Biomaterials and Nanomedicine, Center for Biomedical Technologies, UPM (CTB-UPM), and Spanish Society of Biomedical Engineering. He has devoted his researcher career to the instrumentation field having worked on semiconductor materials characterization, sensor networking, and seismic instrumentation among others, although for the last 20 years, he has been focused on biomedical technology development. He is heading the Laboratory of Bioinstrumentation and Nanomedicine, a CTB-UPM, where he is also a member of the Life Supporting Technologies Group. The main research lines being followed at this laboratory are the development of technologies for nanomedicine, mainly for new anticancer therapies based on nanoparticle mediated hyperthermia, the development of gravimetric biosensors and electromedicine instrumentation, and the development of accessible and assistive technologies based on serious games, virtual and augmented reality technology, and the use human activity recognition technologies. He has published more than 100 and 50 papers and conference contributions, participated in or headed more than 60 projects, and he has supervised more than 20 doctoral theses.