IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 August 2023, accepted 7 September 2023, date of publication 12 September 2023,
date of current version 21 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3314492

RESEARCH ARTICLE

An Efficient Human Activity Recognition Using Hybrid Features and Transformer Model

OUMAIMA SAIDANI[1], MAJED ALSAFYANI[2], ROOBAEA ALROOBAEA[2], NAZIK ALTURKI[1], RASHID JAHANGIR[3], AND LEILA JAMEL[1]

[1]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[2]Department of Computer Science, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia
[3]Department of Computer Science, COMSATS University Islamabad, Vehari Campus, Vehari 61100, Pakistan

Corresponding author: Nazik Alturki (namalturki@pnu.edu.sa)

This research is supported via funding from Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R333), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

ABSTRACT Human activity recognition is a challenging and active research topic in computer science due to its applications in video surveillance, health monitoring, rehabilitation, human-robot interaction, robotics, gesture and posture analysis, and sports. In the past, various studies have utilized manual features to identify human activities and obtained good accuracy. Nonetheless, the performance of such features degraded in complex situations. Therefore, recent research used deep learning (DL) techniques to capture the local features automatically from given activity instances. Though automatic feature extraction overcomes the problems of manual features, there is still a need to enhance the efficiency and accuracy of existing techniques. The motivation behind this research is to improve the efficiency and accuracy of HAR systems. This research proposed a HAR system, which applies data enhancement techniques before capturing robust and discriminative features set from each activity instance. The captured feature set is given to the transformer model for activities recognition using the PAMAP2, UCI HAR, and WISDM datasets. The achieved results revealed that the proposed HAR model outperformed the baseline methods. Specifically, the proposed HAR achieved 98.2% accuracy for PAMAP2 with all instances in 12 activities, 98.6% accuracy for UCI HAR with all instances in 6 activities, 97.3% for WISDM with all instances in 6 activities. The advantage of the proposed hybrid features is the capability to capture both low-level and high-level information from the sensor data, potentially enhancing the discriminative power of the system. In addition, this study employed a transformer a model due to its ability to capture long-range dependencies, which are beneficial in recognizing complex human activities patterns.

INDEX TERMS Human activity recognition, HAR, transformer, hybrid features, PAMP2.

I. INTRODUCTION

Human activity recognition (HAR) is the process of automatically identifying actions and behaviors based on sensors data [1]. HAR has become a rapidly growing field of research due to its of applications in several areas, including healthcare [2], sports [3], robotics [4], and security [5] as shown in Figure 1. The aim of HAR system is to propose and develop accurate and robust algorithms that can identify and recognize human activities with high precision, regardless of the environment or context. The importance of HAR lies in its ability to monitor and track human activities in real-time, which can provide valuable insights and support for several applications. For instance, in healthcare, HAR system can be employed to monitor the activities of elderly or disabled individuals and provide assistance when needed. In sports, HAR system can be utilized to monitor and analyze the movements of athletes to identify areas for improvement. In robotics, HAR can be used to enable robots to interact with humans in a more natural and intuitive manner. In security, HAR system can be utilized to detect and monitor suspicious activities in public places. The process of HAR typically

The associate editor coordinating the review of this manuscript and approving it for publication was Cesar Vargas-Rosales.

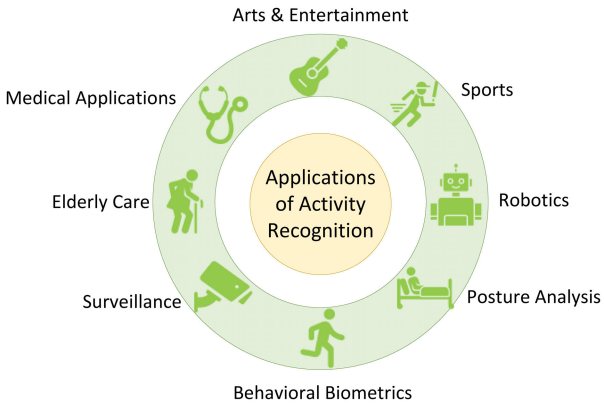


FIGURE 1. Applications of human activity recognition system.

involves three main stages: data collection, feature retrieval, and classification [6] as shown in Figure 2. In the first stage, sensor data is collected from various sources, such as accelerometers, gyroscopes, and video cameras. In the second stage, relevant features are retrieved from the raw signal data, such as frequency-domain, time-domain, and statistical features. In the last stage, classification algorithms are employed to classify the retrieved features into different activity classes, such as walking, running, sitting, and standing. Deep learning (DL) methods have shown great potential in HAR due to their ability to learn complex and high-level features from raw sensors data. DL methods, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have shown great potential in accurately identifying activities from raw sensor data, such as accelerometer and gyroscope readings. By leveraging the power of DL, researchers aim to improve the reliability and accuracy of HAR systems, ultimately leading to more efficient and effective applications in various domains.

The key challenge in HAR is to deal with the variability and complexity of human activities [7]. Human activities can vary greatly in terms of duration, intensity, and context, making it difficult to build robust and accurate recognition algorithms. Additionally, the choice of sensors and their placement can also impact the performance of HAR algorithms. For example, wearable sensors may provide more accurate data but may be more intrusive for the user, while non-wearable sensors may be less accurate but more convenient for the user. Recently, there has been a growing interest in developing HAR algorithms that can work in real-world environments, where the activities may be unstructured and unpredictable. One approach to address this challenge is to use multi-modal data fusion, where raw data from various sensors is merged to offer a more detailed view of the activity being performed [8].

In this study, several features are captured from smartphone time-series sensors data for HAR. These features include: tonnetz representations, mel-spectrogram, spectral contrast, chromagram and MFCCs and its variants. Subsequently, these features were given to the transformer model to recognize human activities. To evaluate the performance, three public datasets (PAMAP2, UCI, WISDM) were utilized. The proposed HAR technique achieved better recognition results for all three datasets. Following are the four contributions of this study.

- Captured robust and relevant features for HAR.
- The data enhancement methods are employed to enhance the size of training data.
- A transformer model was employed to increase the identification rate of human activities and lessen the training time of model under limited computational resources.
- The HAR technique using a transformer model achieved better identification results compared to the existing HAR techniques.

The remaining paper is organized as follows. Section II reports the related work on HAR methods. Section III describes the datasets, extraction of relevant features, transformer model employed for HAR and performance evaluation metrics. Section IV presents the results of different experiments and discussion of the findings. Finally, the conclusions and future directions are presented in Section V.

II. RELATED WORK

Numerous research has been conducted for HAR from the mobile and wearable sensors data using DL approaches. Nevertheless, the efficiency and performance of HAR systems depends on the retrieval of relevant feature and the selection of the suitable model that recognize the human activities. This section reviews the latest studies published for the development of HAR systems using DL approaches.

A. MOBILE ACCELEROMETER-BASED MODELS

Wan et al. [9] designed a mobile accelerometer-based model for HAR. The proposed model collected the sensory data from the daily activities of various individuals. The collected data was preprocessed by normalization, denoising and segmentation. In second phase, the discriminative and robust feature vectors were extracted from the three-axis accelerometers preprocessed data. These feature vectors were fed as input to a CNN for the retrieval of local features. Finally, five DL models were evaluated on UCI HAR and PAMP2 datasets. The achieved results revealed that the proposed approach outperformed the existing approaches. In [10] proposed a novel DL model by fusing RNN and inception neural network. The multi-channel sensors data was given to the HAR model. The authors derived multi-dimensions features by using several kernel-based convolutional layers. The results verified that the proposed method obtained the consistent better results on widely used HAR datasets, when compared with baseline methods.

Another study [34] proposed a DL-based technique for HAR with smartphone gyroscope and accelerometer data. The authors combine three DL models named CNN,

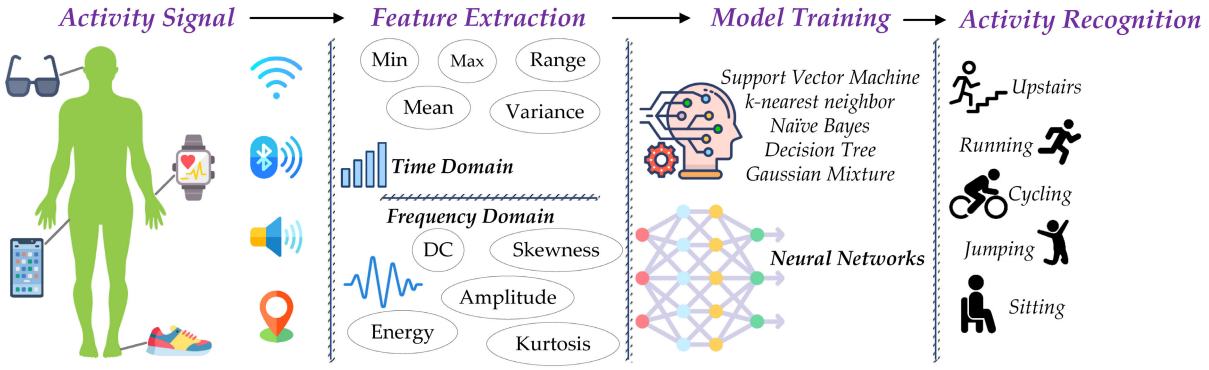


FIGURE 2. An illustration of HAR process using conventional approaches.

autoencoders, and LSTM. The CNN model was used to extract automatic features, autoencoders were employed to reduce the dimensionality and LSTM was adopted for temporal modeling. The proposed ConvAE-LSTM architecture was evaluated on four benchmark datasets (PAMAP2, WISDM, OPPORTUNITY and UCI). The results revealed that the proposed technique improved the performance in terms of accuracy and computational time over baseline methods.

B. DEEP CONVOLUTIONAL NEURAL NETWORKS

In another research [11], authors presented a deep CNN model to identify human activities efficiently from mobile sensors data. The authors provided a method to extract automatic robust features from raw time-series data. Experiments revealed that the deep convnets retrieved complex and relevant features with each additional layer. The proposed model also achieved optimal results on moving activities and overall accuracy of 95% on the test set. Similarly, Hassan et al. [12] proposed a mobile sensors-based technique for HAR. In first phase, the authors extracted the efficient features (median, mean, autoregressive coefficients etc.) from raw data. Secondly, the features were handled by principal component analysis and linear discriminant analysis to make them more relevant and robust. Lastly, the feature were utilized to train the Deep Belief Network for identification of activities. The authors' proposed technique was compared with baseline methods such as artificial neural network and support vector machine (SVM). The experiments showed that the proposed technique outperformed traditional methods. To choose the optimum parameters of the CNN, Raziani and Azimbagirad [30] introduced a one-dimensional CNN for HAR and investigated seven metaheuristic techniques. The UCI HAR dataset was used to evaluate the optimization algorithms. The achieved performance revealed the robustness of metaheuristic algorithms to optimize the parameters of CNN.

To capture wide range of receptive fields of HAR in each feature layer, Tang et al. [31] proposed a CNN that used the concept of hierarchical-split (HS). The experiments were conducted on benchmarks datasets and results demonstrated that the HS method achieved impressive performance compared to the baseline models on WISDM, UCI-HAR, PAMAP2, and UNIMIB-SHAR. Finally, the authors demonstrated the multiscale receptive fields to derive the discriminative features.

C. HYBRID DL MODELS AND FEDERATED LEARNING

To investigate the efficiency of integration of DL models in recognizing human activities, [13] applied four different hybrid DL models. Each model integrated a CNN with a variant of RNN. A PAMAP2 dataset was used to analyse the results of the hybrid DL models. The achieved results showed an optimal performance of each hybrid DL model as compared to the RNN and CNN individually. Nevertheless, this performance was achieved at the cost of high computational time. To facilitate the informed and data-driven decision making, [14] proposed a deep neural network for HAR using multiple sensor data. Exclusively, the reported method encoded the time-series signal data to ensure the relevant features for HAR. A residual network was adopted by combining two deep networks and training diverse sensor data. Moreover, distinct layers were utilized to deal with the variations in dataset size. The model was tested on two HAR datasets, which comprised several heterogeneous smartphone sensor combinations. The findings demonstrated that the HAR model outperformed other competing models. The traditional machine learning approaches for HAR have failed to protect users' sensitive information and privacy in the process of achieving high performance. To handle this, [15] designed a federated learning model for HAR. This model enabled all users to recognize its activity safely. To retrieve suitable features from data, this model designed a perceptive extraction network. This network was responsible to discover local features from data and the relation network focused on extracting global relationships in sensors data. The authors used four datasets, i.e., WISDM, UCI-HAR, PAMAP2, and OPPORTUNITY to evaluate the performance. The results demonstrated that perceptive extraction network outperformed 14 baseline HAR techniques on these datasets.

In another study [32], the authors explored environmental contexts, such as noise level and illumination, to support raw

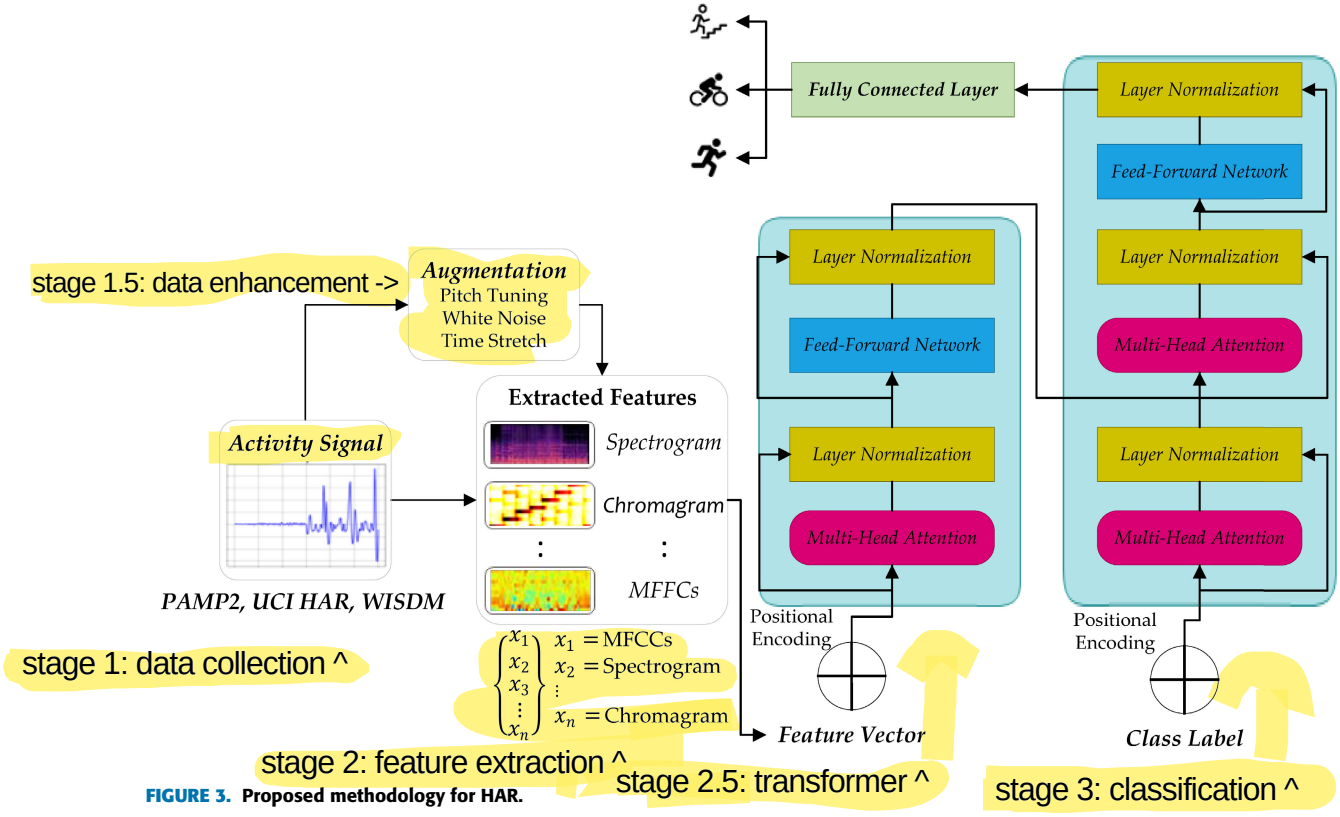


FIGURE 3. Proposed methodology for HAR.

sensors data using a hybrid CNN-LSTM model. The proposed approach performed fusion of rich contextual data with low-level sensor data to enhance the generalization and recognition accuracy. In first experiment, the authors employed triaxial sensing signals for training the baseline models. While in second experiment, the triaxial sensing signals were combined with contextual information. The results revealed that contextual information (noise and light) achieved better accuracy compared to the baseline HAR models.

D. USER-INDEPENDENT DL TECHNIQUE

To overcome the issue of computational cost and handcrafted feature engineering, [16] presented a user-independent DL technique online HAR. This study employed CNN to capture local feature fused with statistical features to retain information regarding global form of time-series. Moreover, this study examined the effect of the length of time-series on model performance and limited it to 1 second for real-time HAR. The performance of the technique was evaluated on two widely used UCI HAR and WISDM datasets. The findings showed that the proposed technique outperformed baseline techniques.

III. PROPOSED METHODOLOGY

This section reports the comprehensive research methodology used to recognize human activities from sensors raw data. In the proposed method, various features including spectrograms, tonnetz, chroma, MFCCs, and spectral contrast are retrieved from each data sample and used for training the transformer model. Lastly, the HAR model is evaluated employing three datasets (PAMAP2, UCI, WISDM) based on training time, accuracy, and robustness. The description of proposed HAR methodology is illustrated in Figure 3. The details of various phases is presented in subsequent subsections.

A. DATASETS

This research utilized three public datasets: PAMAP2, UCI and WISDM. These datasets include a wide range of physical activities, including postures and both repetitive and non-repetitive movements. Furthermore, the Pamap2 dataset contains 27 signals collected in a laboratory, whereas the Opportunity dataset includes 113 signals collected in the wild environment. Each dataset used different number of sensors: 9 and 39 sensors, respectively. The detail of each dataset is presented below.

1) WISDM DATASET

This dataset recorded by the Wireless Sensor Data Mining Lab, is a publicly available benchmark HAR dataset. This dataset was collected by performing specific set of daily activities from 36 subjects. The participants placed an Android mobile in front pocket of their pants and performed various activities including sitting, jogging, upstairs, downstairs, standing, and walking for specific duration. An embedded 3-axial (x, y, and z) accelerometer was utilized

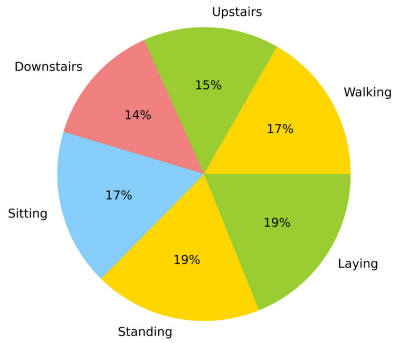


FIGURE 4. % of activities in UCI HAR dataset.

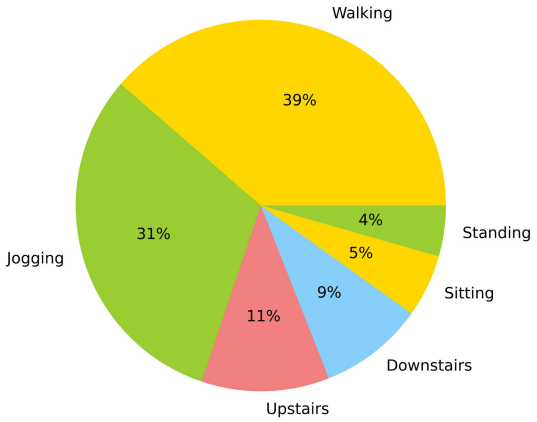


FIGURE 5. % of activities in WISDM dataset.

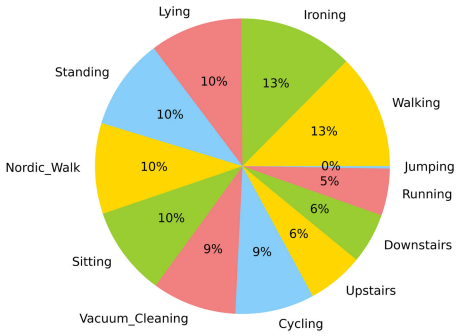


FIGURE 6. % of activities in PAMP2 dataset.

to measure changes in linear acceleration, which provides valuable information about human movement and activity patterns after every 50 ms.

WISDM dataset contains 1098209 activities instances as shown in Figure 5. A designated subject monitored the data collection process to ensure the data of high quality. The signals of all activities are depicted in Figure 7.

2) UCI HAR DATASET

This balance dataset as shown in Figure 4 is widely utilized in human activity recognition research and is collected by recording the 6 daily movements of 30 subjects. Strapping a Samsung Galaxy SII on the waist, all the subjects performed 6 activities including sit, stand, walk, lying, upstairs, and downstairs. The smartphone embedded gyroscope and an accelerometer sampling at 50 Hz are used to record 3-axial (x, y, z) angular velocity as well as linear acceleration. The angular velocity and linear acceleration collected separately from a smartphone are utilized as activity data. The instances of each activity have been video-recorded and labelled manually.

3) PAMAP2 DATASET

The Pamap2 dataset contains recordings of 12 distinct physical activities from 9 subjects (8 male and 1 female) wearing 3 Inertial Measurement Units (IMUs). These IMUs were placed on the body at three different locations: wrist, chest, and ankle. Each IMU includes a 3D magnetometer, a 3D gyroscope, and two 3D accelerometers. The magnetometer measures the magnetic field, the gyroscope measures angular velocity, and the accelerometer measures linear acceleration. The magnetometer measures the magnetic field, which can be useful in certain scenarios. For instance, it can help to determine the orientation or direction of motion. However, the impact of using a magnetometer for activity recognition may vary depending on the context and the specific activities being recognized. Considering the typology of physical activity formerly described, Pamap2 dataset contains 9 repetitive activities (running, walking, cycling, ascending stairs, vacuum cleaning, Nordic walk, descending stairs, rope jumping, and ironing) and 3 postures (standing, lying, and sitting). Most of the activity instances in this dataset lasted four minutes except rope jumping, descending stairs, and ascending stairs because of the constraints of the building or to avoid the fatigue of the participants as shown in Figure 6. Table 1 presents the summary about these datasets.

B. DATA AUGMENTATION

This technique is a widely used in HAR systems to enhance generalization and robustness while preventing overfitting. This approach not only improves performance, but also promotes data invariance by improving the distribution of the data [17]. Since DL techniques require a large data size, three methods including time stretch, pitch shift and Gaussian noise are used to increase the training data, and to improve the generalization of the model. Time stretch was used to change the temporal duration of an activity sequence without altering its content. This method simulates variations in the speed at which activities are performed, which can occur naturally in real-world scenarios. Time stretch helps the model handle activities performed at different speeds, making it more robust to variations in the execution tempo and enables the model to recognize the same activity, even if it occurs faster or slower than during training. Secondly, pitch shift method was used to alter the frequency or pitch of an activity without changing its duration. By applying pitch shift, the model becomes more capable of recognizing activities even when the pitch of activity is altered. Finally, the

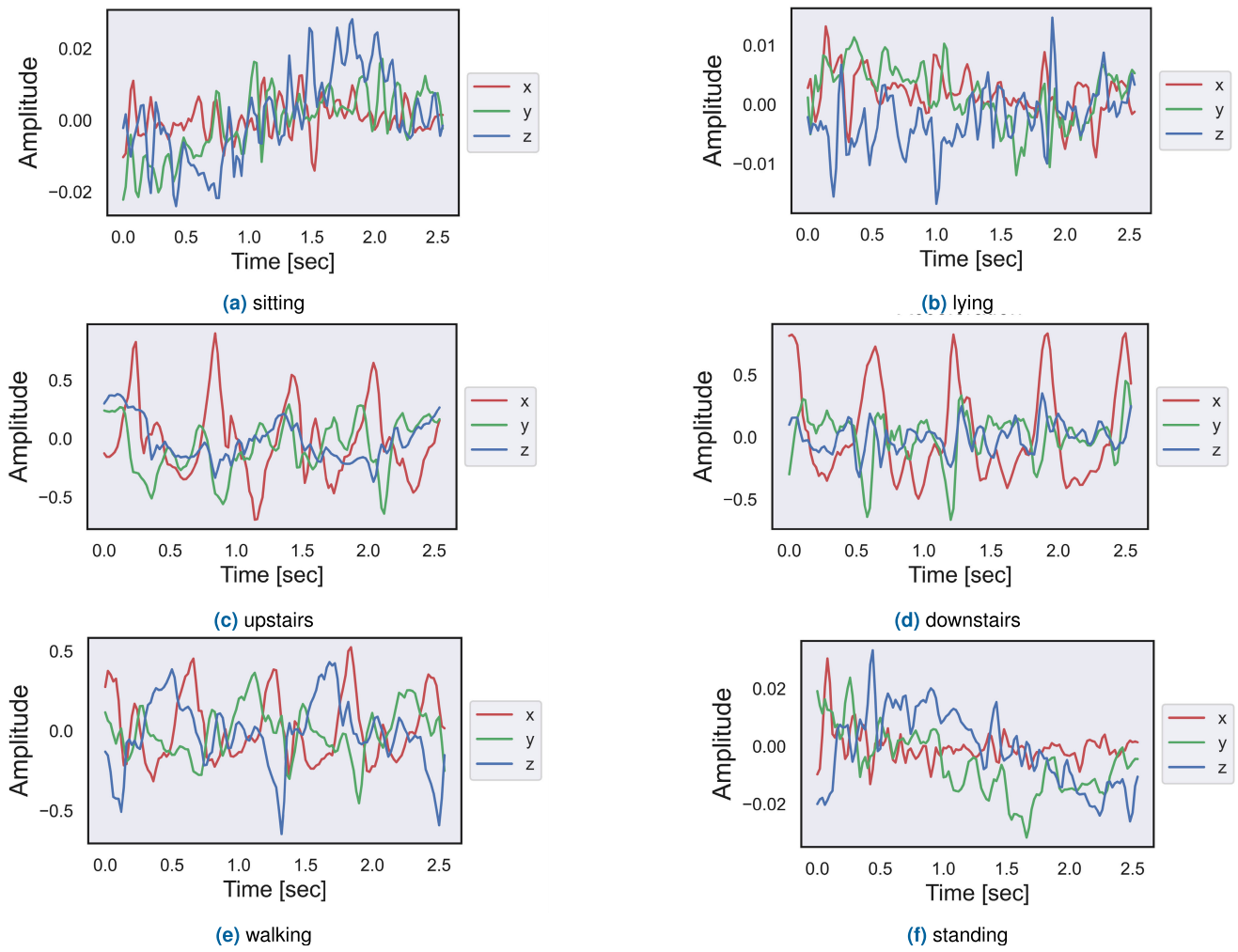


FIGURE 7. The signals of six activities from WISDM datase.

Gaussian Noise technique was applied to introduce random variations to the data points, which helps mimic noise and uncertainties present in real-world sensor measurements and data acquisition. By including Gaussian noise in the data, the model becomes less sensitive to small fluctuations or measurement errors that might be present in the input data during actual deployment [18]. It is significant to select the correct value of "σ" parameter for the amplitude of noise.

C. FEATURE EXTRACTION

Raw sensor data can be quite high-dimensional which can lead to challenges in training deep neural networks efficiently, as the models need to process and learn from vast amounts of data. Secondly, Raw sensor data can be massive, which could lead to longer training times and potentially require more advanced hardware infrastructure. Lastly, DL models are prone to overfitting, when trained on high-dimensional raw data. Converting raw signals to classical feature sets can help reduce the dimensionality and make the data more compact and representative for DL models. Therefore, the retrieval of robust and discriminative features that correctly recognize activities from sensors time series data is an important phase to achieve the high performance of the HAR system [19]. The relevant features can significantly affect the accuracy and efficiency of HAR system, while irrelevant features can enhance the model training time [20]. This research used Librosa python library to capture the seven features from each sample of time series sensor data. These features include:

- Tonnetz representation
- Mel-spectrogram
- Spectral contrast
- Chromagram
- MFCCs
- delta MFCCs
- delta-delta MFCCs

MFCCs and its variants are widely utilized in the area of signal processing and HAR [21], [22]. To capture MFCCs feature, the signal is split into frames of fixed length. Secondly, an operation called windowing is performed to reduce the silence from each frame of a signal. The *t-domain*

TABLE 1. Information of datasets employed in experiments.

		PAMAP2	UCI HAR	WISDM
Total Instances		18664	10299	1098209
Subjects		9	30	36
Activities		12	6	6
Device		IMUs	Samsung Galaxy S II	Android Mobile
Magnetometer		✓	×	×
Gyroscope		✓	✓	×
Accelerometer		✓	✓	✓
Sample Rate		100Hz	50Hz	20Hz
Activity	Label	count	count	count
Downstairs	A_1	981	986	100427
Jogging	A_2	×	×	342177
Sitting	A_3	1783	1286	59939
Standing	A_4	1832	1374	48395
Upstairs	A_5	1102	1073	122869
Walking	A_6	2321	1226	424400
Lying	A_7	×	1407	×
Vaccum	A_8	1685	×	×
Running	A_9	931	×	×
Nordic Walk	A_{10}	1821	×	×
Jumping	A_{11}	449	×	×
Ironing	A_{12}	2317	×	×
Cycling	A_{13}	1585	×	×

shows the common activity in each dataset

activity signal is then transformed into the *f-domain* by computing the Fast Fourier Transform (FFT). The Mel scale filter is used to measure all values calculated from the FFT. Afterwards, powers logs are calculated at every Mel-frequency and lastly each log Mel spectrum is converted back to t-domain by applying DCT (Discrete Cosine Transform). Finally, derived amplitudes from resulting spectrum are known as MFCCs. This research captured 40 MFCCs feature along with its variants (40 Δ MFCCs and 40 $\Delta\Delta$ MFCCs). Though, MFCCs are proven very successful in monitoring and identifying the timbre variations in activity signal, they strive to differentiate the pitch and representations of harmony [23]. To resolve this problem, chroma features are derived from the activity signal through binning techniques and short-time Fourier Transform (STFT). In this research, the chromagram information is captured for each frame of the activity signal and converted it to one coefficient.

To extract melspectrogram feature, a signal was split into a number of frames and then FFT was calculated for all frames. Afterwards, a Mel-scale was produced by first splitting the spectrum of frequency into equally spaced frequencies. Lastly, the frequencies were computed on the Mel-scale for each frame of the activity signal. Tonnetz feature is a 6-dimensions pitch space, which shows the pitch relationships in the rise and fall of activity signal. In this paper, the tonal centroid representations were computed for each frame of the activity signal. Spectral contrast measures the RMS (root mean square) variation between spectral peak and spectral depression for each frame. This research extracted 273 features (6 tonnetz, 128 melspectrogram, 7 spectral contrast, 12 chromagram, 40 MFCCs, 40 delta-MFCCs, 40 delta-delta MFCCs) to integrate the diverse characteristics of signal such as pitch, harmony, timber etc. into one training sample. The correlation between extracted features is shown in Figure 8.

D. TRANSFORMER MODEL

Transformer models are a type of DL techniques that have gained popularity in recent years for their ability to model sequential data. This technique has demonstrated great potential in various natural language processing and computer vision tasks, including HAR. Transformer models are particularly well-suited for this task because they can capture long-range temporal dependencies between activities and their associated sensor data. In traditional HAR models, such as CNNs and RNNs, the input data is typically processed in a fixed-length sequence, where each sample is represented by a fixed number of features. However, HAR data is inherently sequential and varies in length and complexity, making it challenging to model effectively. Transformer models address this challenge by allowing for the modeling of long-range temporal dependencies between the sensor data and activity labels. Additionally, transformer models can be trained on data with varying lengths, making them more flexible and adaptable to different types of data.

TABLE 2. Proposed transformer model structure.

Layer (type)	Output Shape	Param #	Connected to
inputs (InputLayer)	[(None, None, 273)]	0	[]
tf.math.reduce_sum(TFOpLambda)	(None, None)	0	['inputs[0][0]']
tf.cast (TFOpLambda)	(None, None)	0	['tf.math.reduce_sum[0][0]']
enc_padding_mask (Lambda)	(None, 1, 1, None)	0	['tf.cast[0][0]']
encoder (Functional)	(None, None, 273)	5239008	['inputs[0][0]', 'enc_padding_mask[0][0]']
tf.reshape (TFOpLambda)	(None, 273)	0	['encoder[0][0]']
outputs (Dense)	(None, 12)	2466	['tf.reshape[0][0]']
Total params: 5,241,474			
Trainable params: 5,241,474			
Non-trainable params: 0			
Epochs: 100			
Learning Rate: 0.001			
Batch Size: 16			
Dropout: 0.1			
Number of Layers: 6			
Attention Heads: 1			
Optimizers: Adam			
Loss Function: Categorical Crossentropy			

TABLE 3. Training performance of HAR models utilizing the PAMAP2, UCI HAR, and WISDM.

features	Dataset	Precision	Recall	F1-score	Accuracy
273	PAMAP2	1.00	1.00	1.00	1.00
	UCI HAR	1.00	1.00	1.00	1.00
	WISDM	0.98	0.98	0.98	0.98

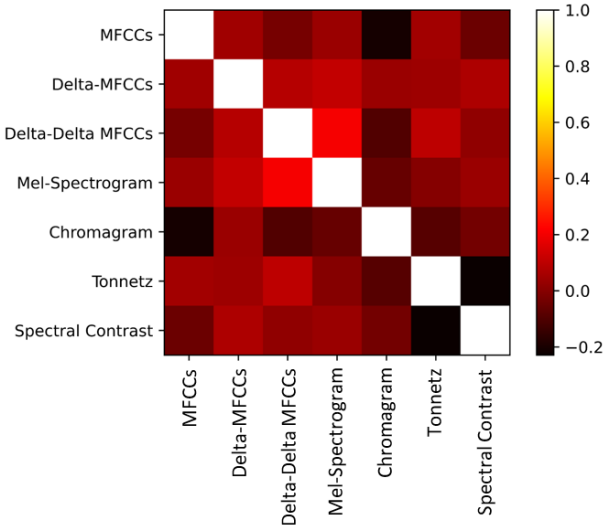


FIGURE 8. Correlation heatmap between audio features.

In the context of HAR, transformer models can be used to model the sequential nature of the sensor data and activity labels. The transformer model comprises of an encoder and a decoder. The encoder takes in the sensor data as input and generates a set of latent representations that capture the temporal dependencies between the sensor readings. The decoder then takes these latent representations and predicts the corresponding label.This technique has been shown to achieve optimum performance on various HAR benchmarks and has potential to improve the effectiveness of HAR systems. The structure of the proposed transformer is given in Table 2.

E. PERFORMANCE METRICS

To compute the performance of HAR system, this study utilized four distinct metrics including precision, recall, F1-value, and accuracy. These metrics have been commonly employed to evaluate the various HAR and related applications [24]. The performance of every daily activity was calculated by utilizing the confusion matrix which comprise of TN (true-negative), TP (true-positive) when classification algorithm accurately recognizes and FN (false-negative), FP (false-positive) when classification algorithm recognizes incorrectly. Accuracy computes the accuracy recognize instances of activity class from the total instances of that activity class using Eq (1). The high value of accuracy indicates that the proposed algorithm has better performance and efficiency.

$$Accuray = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{TP+TN}{TP+TN+FP+FN}\right)_i \quad (1)$$

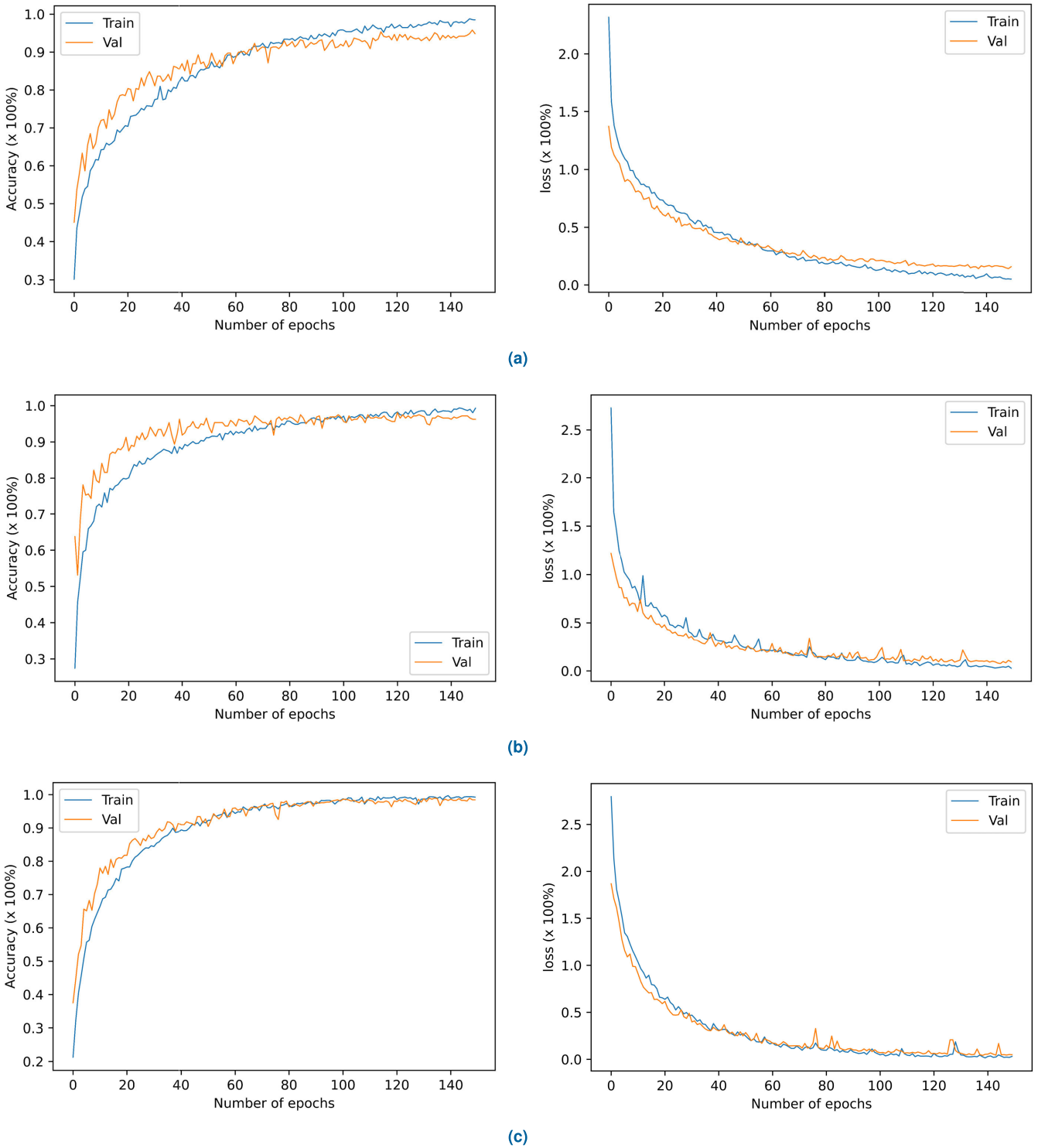


FIGURE 9. The validation and training performance of HAR models for the (a) UCI HAR (b) WISDM (c) and PAMAP2.

where N shows the total of instances of activity class. Recall is determined by the ratio of correctly recognized positive instances of a class and the sum of correctly recognized positive instances and incorrect recognized negative instances as expressed in Eq (2).

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{TP}{TP+FN}\right)_i \tag{2}$$

TABLE 4. Performance of HAR system for WISDM, UCI HAR, and PAMAP2.

Dataset	Matric	Performance of each activity in all datasets of HAR (%)												
		A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
PAMAP2	Precision	0.97	×	0.99	0.99	0.98	0.99	0.99	0.98	0.98	0.99	0.91	0.99	0.99
	Recall	0.99	×	0.98	0.99	0.98	0.99	0.99	0.98	0.99	0.98	0.94	0.99	0.98
	F1-score	0.98	×	0.99	0.99	0.98	0.99	0.99	0.98	0.98	0.98	0.93	0.99	0.99
	Accuracy	0.99	×	0.98	0.99	0.98	0.99	0.99	0.98	0.99	0.98	0.94	0.99	0.98
	G-mean	0.97	×	0.97	0.98	0.96	0.98	0.98	0.96	0.97	0.97	0.98	0.97	0.98
UCI HAR	Precision	0.98	×	0.97	0.98	0.99	0.99	0.98	×	×	×	×	×	×
	Recall	0.98	×	0.98	0.99	0.97	0.98	0.99	×	×	×	×	×	×
	F1-score	0.98	×	0.97	0.98	0.98	0.98	0.99	×	×	×	×	×	×
	Accuracy	0.98	×	0.98	0.99	0.97	0.98	0.99	×	×	×	×	×	×
	G-mean	0.96	×	0.95	0.97	0.96	0.97	0.98	×	×	×	×	×	×
WISDM	Precision	0.95	0.99	0.93	0.90	0.95	0.99	×	×	×	×	×	×	×
	Recall	0.99	0.97	0.96	0.96	0.98	0.98	×	×	×	×	×	×	×
	F1-score	0.97	0.98	0.95	0.93	0.97	0.98	×	×	×	×	×	×	×
	Accuracy	0.99	1.00	0.98	0.98	0.99	1.00	×	×	×	×	×	×	×
	G-mean	0.95	0.97	0.91	0.90	0.95	0.98	×	×	×	×	×	×	×
A_1=Downstairs, A_2=Jogging, A_3=Sitting, A_4=Standing, A_5=Upstairs, A_6=Walking, A_7=Lying, A_8=Vaccum, A_9=Running, A_{10}=Nordic Walk, A_{11}=Jumping, A_{12}=Ironing, A_{13}=Cycling														

TABLE 5. Comparison of HAR model and existing HAR models utilizing the PAMAP2, UCI HAR and the WISDM datasets.

Study	Dataset	Accuracy (%) of all Activities													
		A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	Avg
[16]	UCI HAR	99	×	94	95	99	99	99	×	×	×	×	×	×	**98**
	WISDM	87	98	83	93	72	98	×	×	×	×	×	×	×	**93**
[9]	PAMAP2	94	×	90	91	90	91	96	84	89	95	96	86	89	**91**
	UCI HAR	86	×	88	94	91	98	98	×	×	×	×	×	×	**93**
[28]	PAMAP2	94	×	100	97	82	98	100	93	99	100	82	87	98	**94**
	UCI HAR	96	×	90	98	96	98	100	×	×	×	×	×	×	**96**
	WISDM	89	98	96	98	87	98	×	×	×	×	×	×	×	**96**
[29]	PAMAP2	93	×	90	99	98	99	99	98	94	100	100	97	96	**97**
	WISDM	91	98	98	100	88	99	×	×	×	×	×	×	×	**97**
[33]	PAMAP2	96	×	97	100	90	100	97	97	98	95	75	97	100	**95**
	UCI HAR	95	×	98	93	98	99	100	×	×	×	×	×	×	**97**
	WISDM	98	99	95	99	99	94	×	×	×	×	×	×	×	**97**
Our model	PAMAP2	99	×	98	99	98	99	99	98	99	98	94	99	98	**98**
	UCI HAR	98	×	98	99	97	98	99	×	×	×	×	×	×	**99**
	WISDM	99	100	98	98	99	100	×	×	×	×	×	×	×	**97**
A_1=Downstairs, A_2=Jogging, A_3=Sitting, A_4=Standing, A_5=Upstairs, A_6=Walking, A_7=Lying, A_8=Vaccum, A_9=Running, A_{10}=Nordic Walk, A_{11}=Jumping, A_{12}=Ironing, A_{13}=Cycling															

The precision is calculated by the ratio of correctly recognized positive instances of activity class and the sum of correctly and incorrectly recognized positive instances as shown in Eq (3).

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{TP}{TP+FP}\right)_i \tag{3}$$

F1-value is commonly used when the dataset is imbalanced to find the accuracy of single class. As the datasets utilized in this paper are imbalanced thus, to calculate the performance of each activity class, F1-value metric was used to validate the completeness of HAR models. F1-value is a mean of precision and recall as provided in Eq (4).

$$F1 - value = \frac{1}{N}\sum_{i=1}^{N} 2 \times \left(\frac{recall \times precision}{recall + precision}\right)_i \tag{4}$$

IV. EXPERIMENTS

For the HAR, this research evaluated the proposed approach using three datasets. The percentage spilt method was

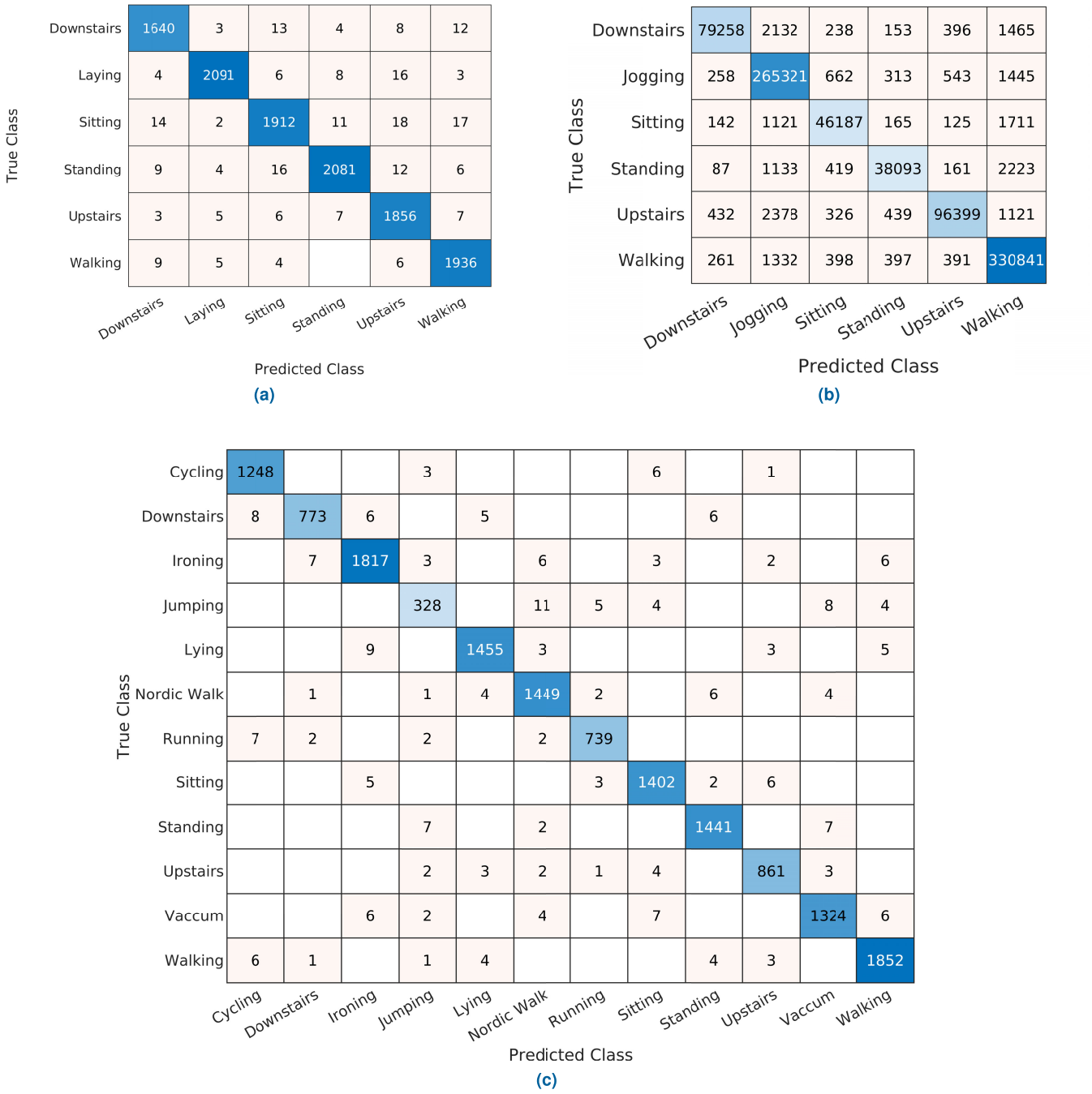


FIGURE 10. **Confusion matrixes of the PAMAP2, WISDM, UCI HAR. (a) The UCI HAR; (b) the WISDM; (c) the PAMAP2. The y-axes show the actual/true classes and x-axis show the predicted classes.**

employed to evaluate the proposed HAR models, where a set of 80% features was utilized to train the transformer model for HAR and remaining 20% features were utilized to train the model [25], [26]. According to existing literature [27], the optimal performance is obtained when 20-30% of the data is used to test the model while remaining data is utilized to train the deep learning model. For this dataset split, the deep learning model obtain an accurate and valid performance and do not achieve the overestimated performance. The training accuracy of proposed transformer model for the WISDM, PAMAP2, and UCI HAR that used the 273 extracted feature as input is given in Table 3. The transformer model achieved the optimal performance for PAMAP2 and UCI HAR datasets except WISDM where the proposed approach attained 98% accuracy which is very close to optimal performance. In this research, the proposed transformer technique reduced the loss

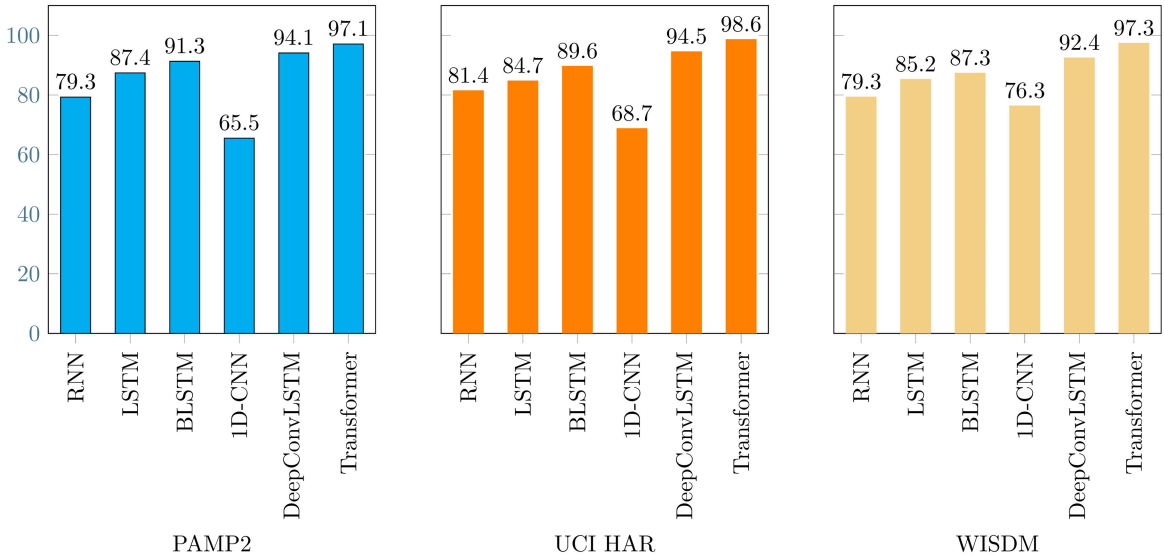


FIGURE 11. Comparison of proposed transformer model with DL classifiers.

and improved the accuracy for both testing and training data samples, which reveal the effectiveness and significance of the transformer model for all datasets as shown in Figure 9.

A. MODELS PREDICTION PERFORMANCE

This research used three datasets that have distinct human activities including WISDM, PAMAP2, and UCI HAR etc. The transform model was evaluated on these datasets, and its recognition performance is given in Table 4. These tables show the performance reports of the WISDM, PAMAP2, and UCI HAR transformer to the recall, precision, F1-score, and recognition accuracy for all human activities in the dataset, which reveals the reliability of the transformer model compared to the existing techniques [9], [16], [28], [29].

The proposed HAR transformer model for PAMAP2 dataset achieved an outstanding performance (98% accuracy) and outperformed the baseline method for PAMAP2 HAR methods [9] for all human activities. The baseline HAR models for PAMAP2 dataset achieved low accuracy for the vacuum cleaning activity due to the mix-up information with other human activities, nevertheless the proposed model recognized all human activities from relevant features with enhanced performance. The activity-wise performance and the inter-activity confusion matrix for HAR model of PAMAP2 dataset is demonstrated in Figure 10c. The figure demonstrates the actually recognized instances diagonally and inter-activity confusion in the corresponding rows.

The confusion matrix of proposed HAR model for UCI HAR shows reveals that the highest accuracies obtained for lying, standing, and downstairs were 99%, 99%, and 98% respectively. The walking and upstairs activities were also recognized with improved accuracy as compared to existing methods for UCI HAR [9], [28] and the average recognition accuracy of transformer also outperformed the baseline models [9], [28]. In the WISDM confusion matrix (Figure 10b), the highest accuracy of 100%, 100%, 99%, and 99% were achieved for jogging, walking, downstairs, and upstairs respectively. The lowest accuracy of 97% was achieved for sitting. The high F1-measure of 98%, 98%, 97% and 97% were obtained for jogging, walking, downstairs and upstairs respectively, and standing obtained the lowest F1-measure of 93%. Similarly, the highest precision of 99%, 99% and 95% were for jogging, walking, and downstairs respectively, and lowest precision of 93% was achieved for sitting. Our proposed HAR models recognized all activities in each dataset with an enhanced recognition rate and low computational time due to the robust and relevant feature and simple transformer model architecture.

B. COMPARISON WITH DEEP LEARNING ALGORITHMS

To build the human activity recognition system, the master feature was fed to distinct DL models: RNN, Long Short-term Memory (LSTM),BLSTM, 1D CNN, and DeepConvLSTM. In addition, 15 experiments (5 DL techniques x 3 datasets) were performed to assess how well the extracted features and DL models performed together.

Figure 11 demonstrates how the proposed transformer model outperformed all DL models, with weighted accuracy for PAMAP2, UCI HAR, and WISDM of 98.2%, 98.6%, and 97.3%, respectively. Using the PAMAP2 dataset, the DeepConvLSTM, and BLSTM models obtained high average accuracy of 93.6% and 90.5%, respectively, as opposed to 84.8% for LSTM, 81.1% for RNN, and 68.4% for 1D-CNN. Moreover, employing UCI HAR and WISDM, DeepConvLSTM beat the other four ML models. The LSTM, RNN, and 1D-CNN models obtained the lowest average accuracy. In conclusion, utilizing all three datasets, the proposed model for HAR outperformed the DL models and achieved the highest weighted accuracy.

C. COMPARATIVE ANALYSIS WITH BASELINE TECHNIQUES

To demonstrate the significance and reliability of proposed model for HAR, this research employed UCI HAR, WISDM, and PAMAP2 to compare the performance with the baseline techniques. An inclusive summary of comparative examination is given in Table 5. The accuracy of the HAR technique is considerably better than existing techniques, which demonstrate the reliability of proposed technique. Nevertheless, in few cases the recognition rate of proposed technique for specific activity is less compared to the baseline techniques. For instance, the HAR system for PAMAP2 dataset in [28] recognized the sitting activity with a 100% accuracy, while the recognition accuracy of proposed technique for PAMAP2 dataset for sitting activity is 98%. However, the proposed technique outperformed baseline techniques by obtaining an average accuracy of 98%. The proposed technique for HAR recognized each activity with low computation time, high accuracy, and is helpful for the real-time healthcare applications. Therefore, it can be concluded that the proposed HAR technique is more generic, accurate, and reliable than the baseline techniques.

V. CONCLUSION

The extraction of discriminative and robust features and accurate recognition are the key issues that make HAR a challenging task. This research extracted a diverse set of features to capture a wide range of patterns and information from the sensor data, making the model more robust. The extracted features were used as input to Transformer model, constructed using a fewer number of layers to derive long-range dependencies leading to improved accuracy. The proposed technique was evaluated on three datasets: WISDM, PAMAP2, and UCI HAR, obtaining accuracy of 97.3%, 98.2%, and 98.6% respectively. The comparison of performance with stat-of-the-art HAR techniques revealed the significance and robustness of the transformer model. The findings also revealed that transformer is specifically suited to all three datasets. The drawback of this approach is that it requires domain expertise to extract and combine a large number of diverse features. While this research shows promising results on all three datasets, the performance on other datasets remains to be seen. The model's effectiveness might vary depending on the characteristics of different datasets. Moreover, a comparative investigation of HAR systems based on DL approaches using other datasets is also scheduled for future work. To further enhance performance, combining device-based activity recognition and device-free activity recognition can be explored. This approach can leverage the strengths of both methods to improve accuracy and robustness. Another aspect that can be addressed is the imbalanced classification problems in activity recognition. Imbalanced datasets, where certain activity classes have significantly fewer samples than others, can lead to biased models that perform poorly on minority classes.

ACKNOWLEDGMENT

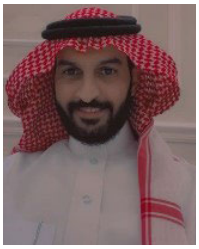
This research is supported via funding from Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R333), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

REFERENCES

[1] N. A. Choudhury and B. Soni, "An adaptive batch size-based-CNN-LSTM framework for human activity recognition in uncontrolled environment," *IEEE Trans. Ind. Informat.*, vol. 19, no. 10, pp. 10379–10387, Oct. 2023.
[2] A. Subasi, K. Khateeb, T. Brahimi, and A. Sarirete, "Human activity recognition using machine learning methods in a smart healthcare environment," in *Innovation in Health Informatics*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 123–144.
[3] M. A. Khatun, M. A. Yousuf, S. Ahmed, M. Z. Uddin, S. A. Alyami, S. Al-Ashhab, H. F. Akhdar, A. Khan, A. Azad, and M. A. Moni, "Deep CNN-LSTM with self-attention model for human activity recognition using wearable sensor," *IEEE J. Transl. Eng. Health Med.*, vol. 10, pp. 1–16, 2022.
[4] M. Tammvee and G. Anbarjafari, "Human activity recognition-based path planning for autonomous vehicles," *Signal, Image Video Process.*, vol. 15, no. 4, pp. 809–816, Jun. 2021.
[5] A. Sunil, M. H. Sheth, E. Shreyas, and Mohana, "Usual and unusual human activity recognition in video using deep learning and artificial intelligence for security applications," in *Proc. 4th Int. Conf. Electr., Comput. Commun. Technol. (ICECCT)*, Sep. 2021, pp. 1–6.
[6] B. Zhang, H. Xu, H. Xiong, X. Sun, L. Shi, S. Fan, and J. Li, "A spatiotemporal multi-feature extraction framework with space and channel based squeeze-and-excitation blocks for human activity recognition," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 7, pp. 7983–7995, Jul. 2021.
[7] I. Jegham, A. B. Khalifa, I. Alouani, and M. A. Mahjoub, "Vision-based human action recognition: An overview and real world challenges," *Forensic Sci. Int., Digit. Invest.*, vol. 32, Mar. 2020, Art. no. 200901.
[8] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowl.-Based Syst.*, vol. 223, Jul. 2021, Art. no. 106970.
[9] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep learning models for real-time human activity recognition with smartphones," *Mobile Netw. Appl.*, vol. 25, no. 2, pp. 743–755, Apr. 2020.
[10] C. Xu, D. Chai, J. He, X. Zhang, and S. Duan, "InnoHAR: A deep neural network for complex human activity recognition," *IEEE Access*, vol. 7, pp. 9893–9902, 2019.
[11] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Syst. Appl.*, vol. 59, pp. 235–244, Oct. 2016.
[12] M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning," *Future Gener. Comput. Syst.*, vol. 81, pp. 307–313, Apr. 2018.
[13] S. Abbaspour, F. Fotouhi, A. Sedaghatbaf, H. Fotouhi, M. Vahabi, and M. Linden, "A comparative analysis of hybrid deep learning models for human activity recognition," *Sensors*, vol. 20, no. 19, p. 5707, Oct. 2020.
[14] Z. Qin, Y. Zhang, S. Meng, Z. Qin, and K.-K.-R. Choo, "Imaging and fusing time series for wearable sensor-based human activity recognition," *Inf. Fusion*, vol. 53, pp. 80–87, Jan. 2020.
[15] Z. Xiao, X. Xu, H. Xing, F. Song, X. Wang, and B. Zhao, "A federated learning system with enhanced feature extraction for human activity recognition," *Knowl.-Based Syst.*, vol. 229, Oct. 2021, Art. no. 107338.
[16] I. Andrey, "Real-time human activity recognition from accelerometer data using Convolutional Neural Networks," *Appl. Soft Comput.*, vol. 62, pp. 915–922, Jan. 2017.
[17] S. Chen, E. Dobriban, and J. H. Lee, "A group-theoretic framework for data augmentation," *J. Mach. Learn. Res.*, vol. 21, no. 245, pp. 1–71, Sep. 2020.
[18] J. Zhang, F. Wu, B. Wei, Q. Zhang, H. Huang, S. W. Shah, and J. Cheng, "Data augmentation and dense-LSTM for human activity recognition using WiFi signal," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4628–4641, Mar. 2021.
[19] D. Thakur and S. Biswas, "Feature fusion using deep learning for smartphone based human activity recognition," *Int. J. Inf. Technol.*, vol. 13, no. 4, pp. 1615–1624, Aug. 2021.
[20] S. Ahmed, K. K. Ghosh, S. Mirjalili, and R. Sarkar, "AIEOU: Automata-based improved equilibrium optimizer with U-shaped transfer function for feature selection," *Knowl.-Based Syst.*, vol. 228, Sep. 2021, Art. no. 107283.

[21] Y. Zhan, J. Nishimura, and T. Kuroda, "Human activity recognition from environmental background sounds for wireless sensor networks," *IEEJ Trans. Electron., Inf. Syst.*, vol. 130, no. 4, pp. 565–572, 2010.
[22] R. San-Segundo, J. M. Montero, R. Barra-Chicote, F. Fernández, and J. M. Pardo, "Feature extraction from smartphone inertial signals for human activity segmentation," *Signal Process.*, vol. 120, pp. 359–372, Mar. 2016.
[23] R. Jain, B. Jain, and M. Puri, "Learning theory (supervised/unsupervised) for signal processing," in *Machine Learning in Signal Processing*. London, U.K.: Chapman & Hall, 2021, pp. 17–53.
[24] H. F. Nweke, Y. W. Teh, M. A. Al-garadi, and U. R. Alo, "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges," *Expert Syst. Appl.*, vol. 105, pp. 233–261, Sep. 2018.
[25] E. Garcia-Ceja, M. Riegler, A. K. Kvernberg, and J. Torresen, "User-adaptive models for activity and emotion recognition using deep transfer learning and data augmentation," *User Model. User-Adapted Interact.*, vol. 30, pp. 365–393, Oct. 2019.
[26] W. Nie, M. Ren, J. Nie, and S. Zhao, "C-GCN: Correlation based graph convolutional network for audio–video emotion recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 3793–3804, 2021.
[27] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation," Tech. Rep., 2018.
[28] K. Xia, J. Huang, and H. Wang, "LSTM-CNN architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56855–56866, 2020.
[29] Y. Li and L. Wang, "Human activity recognition based on residual network and BiLSTM," *Sensors*, vol. 22, no. 2, p. 635, Jan. 2022.
[30] S. Raziani and M. Azimbagirad, "Deep CNN hyperparameter optimization algorithms for sensor-based human activity recognition," *Neurosci. Inform.*, vol. 2, no. 3, Sep. 2022, Art. no. 100078.
[31] Y. Tang, L. Zhang, F. Min, and J. He, "Multiscale deep feature learning for human activity recognition using wearable sensors," *IEEE Trans. Ind. Electron.*, vol. 70, no. 2, pp. 2106–2116, Feb. 2023.
[32] A. Omolaja, A. Otebolaku, and A. Alfoudi, "Context-aware complex human activity recognition using hybrid deep learning models," *Appl. Sci.*, vol. 12, no. 18, p. 9305, Sep. 2022.
[33] W. Ding, M. Abdel-Basset, and R. Mohamed, "HAR-DeepConvLG: Hybrid deep learning-based model for human activity recognition in IoT applications," *Inf. Sci.*, vol. 646, Oct. 2023, Art. no. 119394.
[34] D. Thakur, S. Biswas, E. S. L. Ho, and S. Chattopadhyay, "ConvAE-LSTM: Convolutional autoencoder long short-term memory network for smartphone-based human activity recognition," *IEEE Access*, vol. 10, pp. 4137–4156, 2022.

OUMAIMA SAIDANI received the M.Sc. degree in computer sciences from Paris Dauphine University, France, and the Ph.D. degree in computer sciences from Paris 1-Panthéon Sorbonne University, France. She is currently an Assistant Professor with the Information Systems Department, College of Computer and Information Sciences (CCIS-IS), Princess Nourah bint Abdulrahman University (PNU), Saudi Arabia. Her research interests include information systems engineering, business process engineering, process mining, context-aware computing, deep learning, and artificial intelligence.



MAJED ALSAFYANI received the bachelor's degree (Hons.) in computer science and the master's degree (Hons.) in advance computer science from the University of Hertfordshire (UH), U.K., in 2013 and 2014, respectively, and the Ph.D. degrees in computer science from the University of Hertfordshire, U.K., in 2015 and 2020, respectively. He is currently an Assistant Professor with the College of Computers and Information Technology, Taif University, Saudi Arabia. His research interests include image processing, machine learning, applications of Internet of Things, artificial intelligence, and software engineering.

ROOBAEA ALROOBAEA received the bachelor's degree (Hons.) in computer science from King Abdul-Aziz University (KAU), Saudi Arabia, in 2008, and the master's degree in information systems and the Ph.D. degree in computer science from the University of East Anglia, U.K., in 2012 and 2016, respectively. He is currently an Associate Professor with the College of Computers and Information Technology, Taif University, Saudi Arabia. His research interests include human–computer interaction, software engineering, cloud computing, the Internet of Things, artificial intelligence, and machine learning.

NAZIK ALTURKI received the Ph.D. degree in information systems from The University of Melbourne. She is currently an Assistant Professor with the Information Systems Department, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Her research interests include health informatics, big data, data analytics, and mining.

RASHID JAHANGIR received the bachelor's degree in computer engineering from the University of Engineering and Technology (UET), Lahore, Pakistan, the master's degree from the University of New South Wales (UNSW), Sydney, Australia, and the Ph.D. degree from the Faculty of Computer Science and Information Technology, University of Malaya (UM), Kuala Lumpur, Malaysia. He has been an Assistant Professor with COMSATS University Islamabad, Vehari Campus, Pakistan, since 2014. He was a Software Engineer with Software House, Lahore, for two years. He has vast experience in teaching and research. He has published several articles in academic journals indexed in well-reputed databases, such as ISI and Scopus. He is working on digital signal processing and deep learning. His research interests include deep learning, pattern recognition, machine learning, and data mining. He is an Active Reviewer of various journals, including *Artificial Intelligence Reviews*, *Expert Systems with Applications*, *Multimedia Tools and Applications*, IEEE Access, and *Social Network Analysis and Mining*.

LEILA JAMEL received the engineering degree in computer sciences and the Ph.D. degree in computer sciences and information systems. She was the Program Leader of the IS Program and the ABET and NCAAA accreditation committees in CCIS, PNU. She was the Head of the Information Systems Security Department, Premier Ministry, Tunisia. She is currently an Assistant Professor with the Information Systems Department, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University (PNU), Saudi Arabia. She is a Researcher with the RIADI Laboratory, Tunisia. Her research interests include business process reengineering, process modeling, BPM, data sciences, ML, process mining, e-learning, and software engineering. She was a member of scientific/steering committees of many international conferences. She is a reviewer of many international journals and conferences.