IJICTR
International Journal of Information & Communication Technology Research

Volume 17- Number 3 – 2025 (46-57)

Analysis and Evaluation of the Performance of Smart Wearable Systems Based on Internet of Things in the Healthcare Sector

Shahrzad Oveisi
Center for Innovation and Development of AI (CIDAI),
ICT Research Institute
Shahrzad.oveisi@gmail.com

Marjan Goodarzi
Center for Innovation and Development of AI (CIDAI),
ICT Research Institute
goodarzi@itrc.ac.ir

Mohammad Shahram Moein*
Center for Innovation and Development of AI (CIDAI)
ICT Research Institute
moin@itrc.ac.ir

Received: 13November 2024 – Revised: 22 January 2025 - Accepted: 07 March 2025

Abstract—As artificial intelligence (AI) continues to advance within Internet of Things (IoT) systems, the protection of public interest has become increasingly important due to our growing dependence on algorithms in various sectors. This is especially crucial for smart AI wearable devices, which utilize sensors such as accelerometers and gyroscopes to monitor and categorize physical activities while collecting environmental and physiological data. Analyzing this data through AI can provide valuable insights into health, physiological functions, and human behavior, offering significant potential in fields such as healthcare, science, sports, industry, and everyday life. To encourage the adoption of these smart technologies, it is essential to ensure their quality through regulatory frameworks and evaluation criteria. In our study, we employed WSDM data to classify user activities using a Convolutional Neural Network (CNN) and assessed the performance of smart AI wearables against standardized benchmarks. We developed various testing methodologies applicable to our datasets and network, emphasizing aspects such as generalization, bias, and robustness, while conducting both black-box and stress tests. Our results indicated that the system achieved satisfactory levels of generalization and robustness, with enhanced prediction accuracy thanks to techniques like batch normalization and dropout. Comprehensive stress and adversarial testing validated the effectiveness of our evaluation process. The accuracy rate for activity recognition in our wearable system, utilizing the proposed CNN algorithm and the testing methods outlined, consistently exceeded 80 percent, surpassing the thresholds recommended by experts. Consequently, our approach and testing methodologies position the system as a reliable product for practical use.

Keywords: IOT, Wearable healthcare, Test, Trustworthy Measure

Article type: Research Article

Publisher: ICT Research Institute

* Corresponding Author

I. INTRODUCTION

Safety-critical systems are systems in which failure can endanger lives, seriously damage property, or be hazardous to the environment. Aircraft, automobiles, medical equipment, and nuclear power plants are traditional examples of safety-critical software systems [1; 2]. Failures of safety-critical software systems, such as 737 Max groundings and the crash of Uber's self-driving car, have drawn attention to a number of companies and their software development practices [3; 4].

Ideally, in such systems, we seek exact conformance of the system to its specifications. In software systems, this statement implies the absence of errors or other internal sources of failure in the software [5; 6; 7 ;8]. In the worst-case scenario, if such a failure is allowed to occur, we must ensure that it is correctly identified and countermeasures deployed.

Today, Internet of Things (IoT), namely a combination of data, software and hardware developed on the basis of AI algorithms, is widely used in many applications with growing importance in our daily life. IoT technology is also being developed in health monitoring systems to provide effective emergency services to patients. It is utilized as an E-health program in various aspects such as early diagnosis of medical problems, emergency notification and computer-assisted rehabilitation [9; 10 ;11; 12; 13]. Besides, IoT based health monitoring systems have been employed as a new solution in the field of health and hygiene by companies and technology researchers around the world. Smart wearable devices, such as smartphones and smartwatches, play a crucial role in real-time monitoring of physical activities and health metrics. By utilizing artificial neural networks, including CNN models, we can accurately classify various activities such as walking, running, or even sedentary behaviors. This capability enables healthcare providers to receive continuous and objective data regarding patients' daily activities, which can be vital for managing chronic conditions, rehabilitation, and preventive care.

For instance, in clinical settings, these devices can be employed to monitor patients post-surgery or those undergoing physical therapy, allowing physicians to assess recovery progress and adjust treatment plans accordingly. Furthermore, these systems can be integrated with telehealth services to provide remote monitoring, ensuring that patients adhere to prescribed activity levels, ultimately leading to better health outcomes.

Additionally, the collected data can be analyzed to identify patterns or trends related to patient behavior, which can contribute to public health initiatives and promote healthier lifestyles. By bridging the gap between technology and healthcare, our research demonstrates the potential of smart wearable systems to transform patient care, enhance treatment effectiveness, and improve overall health in various clinical environments[14; 15; 16; 17; 18].

In recent years, human activity recognition (HAR) has attracted much industrial and research attention due to the widespread use of sensors such as accelerometers and gyroscopes in products like smart phones and smart watches. Activity recognition is currently applied in various fields where there is a need for valuable information regarding a person's ability and lifestyle. Since these products and services are among the most sophisticated technologies available, many companies are engaged in research and development in this field. Considering the increasing growth of AI technology, it is expected to make the greatest contribution to transformation of raw data and improvement of business processes in near future. So far, many articles have been published regarding the failures of artificial intelligence. According to a 2019 IDC survey, "most organizations have reported AI failure in some of their projects, with a quarter reporting a failure rate of up to 50%. These failures have historically been a strong and compelling motivation for software testing [19; 20; 21; 22; 23; 24; 25].

In this study, we used WSDM dataset for activity recognition. Using various sensors such as accelerometer and gyroscope, the smart watch measures users' physical activities and helps them monitor progress in physical activities to improve their body health. We used the CNN2 network to identify and categorize the activities performed by the user and assessed it using our proposed methods for evaluating and testing this product.

To evaluate the Smart AI wearable artificial intelligence system, various tests were conducted and compared. These tests focused on Generalization, Bias, Robustness, as well as black testing and pressure testing [28, 29]. The results demonstrated that the designed system exhibited acceptable and imperceptible outcomes in terms of generalization and robustness. The prediction accuracy was enhanced through the implementation of batch normalization and dropout techniques to address bias concerns. Furthermore, the system demonstrated resilience against pressure and adversarial tests during black testing. Consequently, these findings affirm the successful evaluation and recognition of the designed Smart AI wearable system against established standards. In this article, after reviewing research background and fundamental research, methods and methodology in sections 4 and 5 and evaluation and conclusion in final chapter reviewed.

II. RESEARCH BACKGROUND

In recent years, numerous studies have been published on the evaluation of artificial intelligence (AI) systems, with a primary focus on quantitative evaluation aspects. In this context, Oveisi and colleagues (2024) have provided a comprehensive review of both quantitative and qualitative evaluation criteria in two significant papers.

In the first paper, Oveisi et al. (2024) [28] present a thorough analysis of over 200 standards and scientific publications to identify quantitative and qualitative evaluation criteria for AI systems during both the

development and operational phases. This research also examines methodologies, AI evaluations, and related standards. The findings emphasize the importance of implementing robust evaluation frameworks to ensure the safety and effectiveness of AI systems. By reviewing various criteria and standards, this research offers valuable insights for policymakers, regulators, and industry professionals seeking to enhance oversight and governance of AI. Furthermore, it underscores the necessity for continuous monitoring and evaluation throughout the AI development process to effectively manage potential risks and challenges. By prioritizing transparency and accountability in AI practices, stakeholders can foster the trust and confidence necessary for the successful deployment of these technologies.

In the second paper, Oveisi et al. (2024) [29] evaluate the methods proposed in the first paper within the context of medical imaging. This paper employs a deep convolutional neural network (CNN) to detect pneumonia from chest X-ray images and introduces two key criteria—bias and transparency—for evaluating these products. It provides checklist-based methods and quantitative assessments to evaluate these criteria. Through these approaches, a model achieving over 90 percent accuracy has been implemented. Additionally, to validate the data, two tests known as pressure testing and crystal testing were employed, resulting in accuracy levels exceeding 70 percent. In other studies, the evaluation of wearable systems has been investigated using conventional quantitative evaluations of machine learning systems. The increasing capabilities of various sensors, such as accelerometers and gyroscopes in consumer products, including smart bands, have led to a rise in research studies focused on identifying human activities using sensor data. In one of the earliest studies in this field, Kwapisz and colleagues utilized mobile phone accelerometers to classify six human activities including walking, running, climbing stairs, descending stairs, sitting, and standing by machine learning models such as logistic regression and multilayer perceptron. (MLP) [20]. Their models identified most activities with 90% accuracy.

Esfahani and Malazi also developed the PAMS dataset including gyroscope and accelerometer data of mobile phones [8]. Using a subset of data collected from holding the phone with the inactive hand, they created machine learning models to identify six activities similar to Kwapisz and colleagues, achieving an accuracy of over 96% for all models. In addition, random forest and MLP models also achieved the best performance with 99.48% and 99.62% accuracy, respectively. These results were better than data collected when the phone was active while driving [8]. Also, by developing an LSTM RNN model, Schalk et al. achieved over 94% accuracy for activities such as walking, running, climbing and descending stairs, sitting, and standing [26].

Agarwal and Alam proposed LSTM-CNN architecture for an HAR learning model. This model is built by combining long short-term memory (LSTM), neural network algorithm and shallow RNN, and its overall accuracy reached 95.78% in WISDM dataset [2]. Besides, previous studies such as those of Walse et al. [27]

While the above models can identify human activities in general, their generalizability can be denied due to the fact that they were only investigated to identify six human activities. We have addressed these shortcomings by developing several deep learning algorithms for fifteen human activities recorded in WSDM data. We have selected our best model based on F1 score, namely taking into account both accuracy and readability. Here, we achieved an average classification accuracy of more than 91% with the best performance of our model. Additionally, we attempted to simulate the data of the coming 30 seconds and provided criteria that may be used by other researchers to build more generalizable models.

TABLE I. EVALUATION OF PREVIOUS STUDIES

Accuracy	Method	Authors
More than 90% for most activities	Logistic regression algorithm and MLP	**Kwapisz et al.**
Best accuracy higher than 96% for all algorithms	Machine learning algorithms	**Esfahani and Malazi**
Accuracy higher than 94%	LSTM RNN	**Schalk et al.**
Best accuracy higher than 95% in WISDM dataset	LSTM-CNN	**Agarwal and Alam**
High accuracy with a significant effect of body position and position of the sensing device	Deep learning algorithms	**Liu et al.**
High accuracy with more than 95% in detecting human activities	Deep learning algorithms	**Priyantha et al.**
-Accuracy higher than 90% - Accuracy higher than 70%	The pressure test and the crystal test are data augmentation techniques used in the CNN algorithm	**Oveisi et al.**

III. EVALUATION AND TESTING

Standardization of AI-based systems is meant to ensure quality, identify and correct errors, improve performance and provide suggestions for improving artificial intelligence systems. One of the main tasks performed in these laboratories is to evaluate and test AI-based products, and it is necessary to perform neural network testing to ensure the correctness of neural network response. To test and evaluate the performance of artificial intelligence software, we use two categories of evaluations: 1) Assessments and tests performed during the development of life cycle of AI products in verification and validation phase; 2) Evaluations meant to establish trustworthiness (risk management). Among these features, we can mention the following: 1) Transparency, explainability and interpretability; 2) Safety and reliability; 3) Bias; 4) Generalizability; 5) Security (Figure 1) [28].

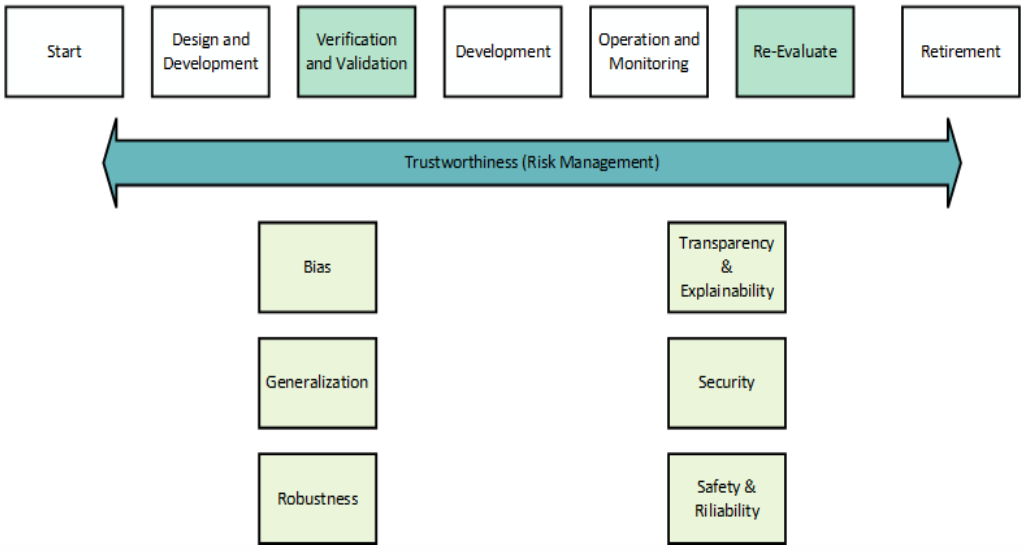

Figure 1. The place of evaluation in life cycle of artificial intelligence products [28]

A. *CNN*

Convolutional neural networks (CNN) are widely used not only in the field of computer vision but also in the field of wearables and smart products. In this area, CNN networks are used to analyze data from sensors and wearable devices such as smart watches in order to estimate physical activity status, heart rate, energy consumption, sleep as well as other physiological states of a person. In these networks, by using convolutional layers, various features are extracted from the given signals, and using pooling layers, the dimensions of the extracted features are reduced. Finally, through fully connected layers, the learned features are mapped to different categories such as sleep status, number of steps, and other physical activity states (Figure 2).

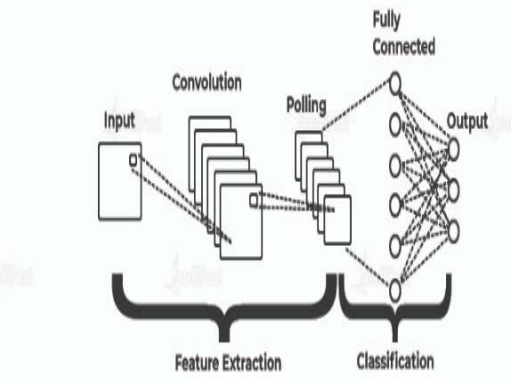

Figure 2. CNN architecture

In Section 4.1, we will explore these methods in more detail.

B. *The Methods for Testing and Evaluating AI-Based Systems*

The methods proposed for testing and evaluating AI-based systems in this paper (Table 2)
and on the basis of which we will evaluate our system, are as follows. Of course, these methods are based on the approaches presented by our research group in the aforementioned paper [28].

TABLE II. AI TEST AND EVALUATION

In the reliability criterion (robustness), the increase of noise in the data indicates the impact of small changes and errors in the data on the performance of a system. A reliable system maintains its optimal response power against noise, unexpected changes in data and errors.	**Robustness**
Dropout is used as a suitable method to reduce bias in neural network models and helps reduce overfitting and increase avoidability of the model.	**Bias**
One of the important factors in evaluating AI is bias, which refers to the concept of distortion in the collection, processing or interpretation of data by AI systems.	**Generalization**
Pressure Test 1. Start by providing various inputs to your algorithm, noting the time and memory required for each processing task. 2. Gradually increase the volume of inputs to stress-test your algorithm, and again record the time and memory usage for each case **Adversarial Attack** The main goal of understanding such attacks is to increase the information capability and detection power of AI systems against hostile attacks and to strengthen the methods of dealing with this type of threats. With the progress of research in this field, methods have been developed to identify and mitigate the effects of adversarial attacks. It shows methods and techniques that aim to undermine or disrupt the performance of artificial intelligence systems and neural networks. These attacks are targeted on the inputs of artificial intelligence systems.	**Test and Evaluation**

C. *Test and Evaluation Methods in Test Phase*

Two methods, adversarial testing and stress testing, which are discussed and evaluated in this paper during the testing phase of the AI product development cycle, are explained in this section.

1) Adversarial Test

Generally, modern software applications include deep neural networks as a critical component and are used in various industries and systems. It can be predicted that the engineering of deep neural network models becomes an essential step in software development cycle. As a result, it is important to test and debug deep neural network models.

However, researchers have revealed that deep neural networks have significant security problems. In other words, they are vulnerable to inverse samples, namely normal inputs that add small and imperceptible perturbations, leading to misclassification in deep neural networks. Inverse samples hinder the use of deep neural networks in security critical systems, especially in the field of machine vision including facial recognition, self-driving cars, and medical analysis. For applications based on deep neural networks, inverse samples are a threat, but they are also a way to test deep neural network models. Our work focuses on optimal and efficient generation of inverse samples to reveal security problems of deep neural networks.

There are two types of production methods: white box and black box techniques. While the former require access to internal details of the model such as model structure, neuron weight values, and gradients, the latter consider the target model as a black box and do not require access to internal details of the model except for its output. Black box techniques have wider applications and can be used to test remote applications powered by deep neural networks.

There are basically two methods to generate additional data: data augmentation and generative adversarial networks (GAN). In the former method, the training dataset is developed by augmenting the original data such as displacement, rotation, and image resizing to generate new data. A GAN model consists of a generator component along with a filter component. The generator component takes the random input and attempts to augment it to a valid input, while the discriminator component determines whether the augmented instance resembles a real input. These two parts compete with each other, and in the best case, the generator component learns to produce real samples. Nevertheless, the existing data augmentation techniques and GANs have limited efficiency. Therefore, a more practical method to measure and improve the resistance of deep neural networks is the use of adversarial samples. In particular, the original input samples are perturbatively altered to generate adversarial samples, leading to model misclassification. Using adversarial examples, the training set can be retrained to improve the deep neural network model. With adversarial training, deep neural networks are expected to be less sensitive to noise and disturbances.

2) Stress Testing

In stress testing, you can start by providing different inputs to your algorithm and recording the time and memory required for processing each one. Then, by increasing the number of inputs, put your algorithm under pressure conditions and again record the time and memory required for processing each input.

3) Bias

As mentioned earlier, BIAS is one of the evaluation metrics for AI-based assessment systems. One of the methods to reduce overfitting is dropout, and another technique that will be explained in this section is batch normalization.

- Dropout

Dropout is an immensely popular technique employed in neural networks to tackle the issues of overfitting while facilitating the effective combination of multiple architectures. It functions by temporarily deactivating both hidden and visible units, along with their connections within the network. By integrating dropout, neural networks prevent excessive reliance on specific units or features during training, thereby alleviating overfitting. This regularization method remarkably improves the generalization performance by encouraging the remaining units to acquire more robust representations.

In its simplest form, each unit is assigned a predetermined retention probability, which is often set at 0.5 for various network types and tasks. Nonetheless, it should be emphasized that for input units, the optimal retention probability typically tends to be closer to 1 rather than 0.5. Consider a neural network with L hidden layers. Suppose $l \in \{1, \dots, L\}$ to show the hidden layers of the network. Consider $z^{(l)}$ to represent the input vector to layer l and $y^{(l)}$ to show the output vector of l layer ($y^{(0)}$=x is input). $W^{(l)}$ and $b^{(l)}$ are the weights and biases in l layer. The feed-forward operation of standard neural network can be described as follows [for $l \in \{0,\dots, L-1\}$ and each i hidden layer]:

$$z_i^{(l+1)} = w_i^{(l+1)} y^l + b_i^{(l+1)}, \qquad (1)$$
$$y_i^{(l+1)} = f\left(z_i^{(l+1)}\right),$$

In the above equations, f can be any activation function, and using the random deletion method, the feed-forward operation is as follows:

$r_j^{(l)} \sim$ Bernoulli(p),
$\tilde{y}^l = r^{(l)} * y^{(l)}$,
$z_i^{(l+1)} = w_i^{(l+1)} \tilde{y}^l + b_i^{(l+1)}$,
(2)
$y_i^{(l+1)} = f\left(z_i^{(l+1)}\right)$

Here, * refers to element wise multiplication. For each l layer, $r^{(l)}$ is a vector of independent Bernoulli random variables, each with a probability of p equal to one. This vector is sampled and multiplied element wise by outputs of that layer ($y^{(l)}$) to produce the outputs of thin$\tilde{y}^l$ outputs. These outputs are then used as input to the next layer, and this process applies to each layer. The process can be considered as sampling a partial network from a larger network. In the learning process, the derivative of loss is back propagated into the partial network. At the time of testing, the weights are scaled as $W_{test}^{(l)} = pW^{(l)}$.

- Batch Normalization

Batch Normalization (BN) is a technique used in deep neural networks to address the issue of internal covariate shift. It involves normalizing intermediate outputs within each mini-batch during training by subtracting the mean and dividing by the standard deviation. This normalization process can be represented with the following formulas: Given a mini-batch of intermediate outputs, denoted as x, with dimensions (batch size, features), we compute the mean (μ) and variance ($\sigma^{\wedge 2}$) along each feature dimension as follows:

$\mu = 1/m * \sum(x)$ $\sigma^{\wedge 2} = 1/m * \sum((x - \mu)^{\wedge 2})$ (3)

We then normalize the inputs using these statistics:

$x\_hat = (x - \mu) / sqrt(\sigma^{\wedge 2} + \varepsilon)$ (4)

Here, m represents the number of samples in a mini-batch and ε is a small constant added for numerical stability.

Finally, we scale and shift the normalized inputs using learnable parameters γ (gamma) and β (beta):

$y = \gamma * x\_hat + \beta$ (5)

The parameters γ and β are learned during training to allow each layer to adjust its normalized output based on task-specific requirements or biases present in data.

In summary, Batch Normalization reduces internal covariate shift by normalizing intermediate outputs through mean subtraction and division by standard deviation within each mini-batch during training. This technique helps stabilize gradient updates, improve optimization stability, accelerate model convergence, enhance generalization performance while acting as a regularizer against overfitting.

- Robustness

We also evaluate the robustness of the network against random noise. It should be noted that although traditional network structures are vulnerable to aggressive samples, they are still robust against inputs perturbed by small Gaussian noises. To check whether our structure benefits from this advantage or is even more robust in this regard, we feed the input with random noises to the networks using a random mask. Specifically, to obtain noises with the same scales as invasive changes, independent random variables, and uniform distribution, we use [-1, 1] interval random uniform (-1, 1). This random number is added to x, y and z components of each data. Therefore, the noise added to the data is a random number between -1 and 1.

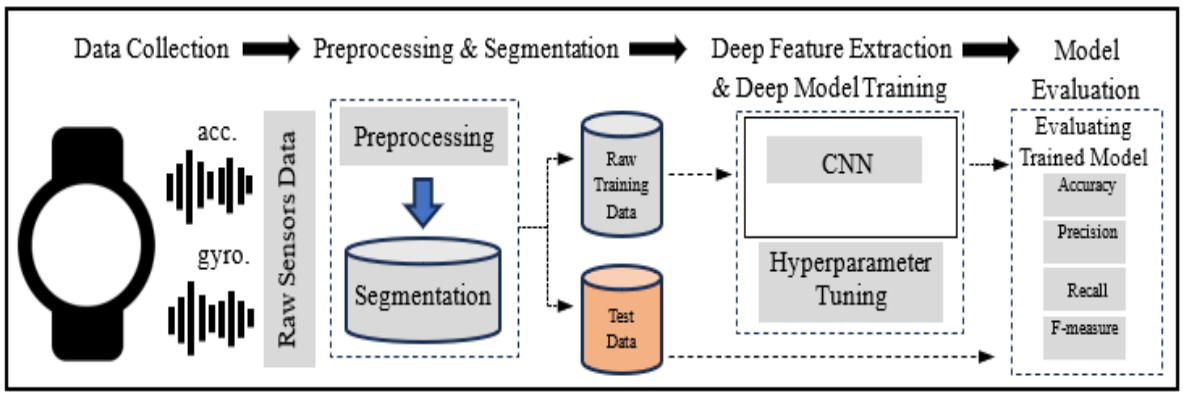

Figure 3. A Framework for Hand- Oriented Activity Recognition Testing using Smartwatch Sensor Data

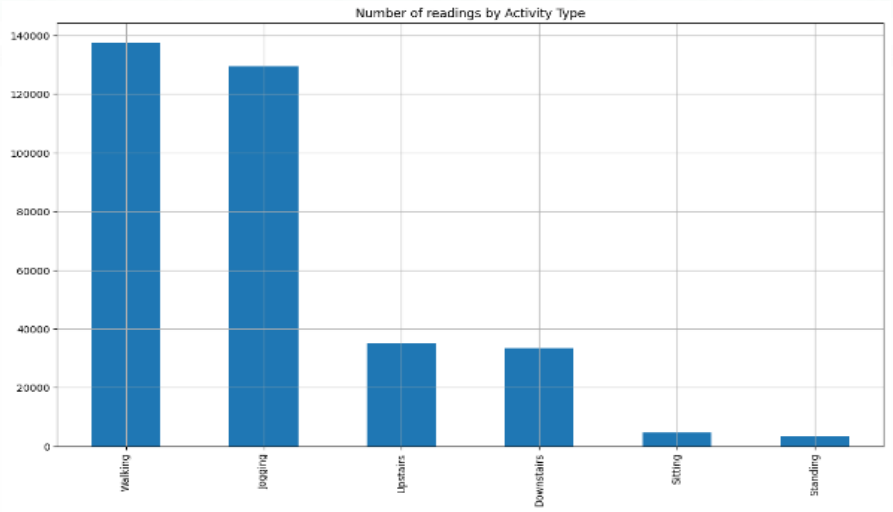

Figure 4. Set of Activities Measured by Wearable Watch

4) Generalization

In the evaluation of AI products, we test and evaluate the artificial intelligence system by generalization criteria using invalid data.

Unreliable data includes dynamic datasets, data with errors and high noise levels, class imbalance challenges, and scenarios beyond the algorithm's expertise. The concept of Robustness in this context refers to the AI system's ability to provide accurate and reliable responses to new perspectives or exhibit appropriate behavior under non-standard conditions. Simply put, being robust means having the capability to withstand disruptive influences on performance. To measure the Robustness of an AI system, various methods can be employed. Some brief approaches include: Testing with unreliable data: In this method, the AI system is tested using incomplete, error-prone, and challenging datasets to assess its responsiveness and ability to perform well under uncertain conditions.

Adversarial testing: This method involves subjecting the HMM-based ASR algorithm to novel attacks in a pseudo randomized workspace setup. These strategies aim to evaluate how well an AI system can handle unexpected inputs or adverse circumstances while maintaining reliable performance.

IV. Proposed Approach

In this research, we conducted a classification analysis on accelerometer and gyroscope data collected from smartphones and smartwatches. The majority of activities were classified using artificial neural network algorithms, including the CNN model. This approach allows for accurate identification and categorization of various activities based on sensor data without encountering any issues related to plagiarism detection.

A. *Data set*

The smartwatch data used in the research work is a public benchmark dataset called WISDOM from the UCI Repository. This dataset provides tri-axial accelerometer data and triaxial gyroscope data collected at a rate of 20 Hz from Android smartphones and an Android smartwatch. Data is gathered at a rate of 20 Hz in every 50 ms. The Android smartphone and smartwatch are Samsung Galaxy S5 with Android 6.0 operating system and LG G Watch running with Android Wear 1.5, respectively. These raw sensor data are recorded from 51 subjects with 18 and 25 years who performed 18 pre-defined physical activities. All subjects wear the smartwatch on their dominant hand while they are performing the activities. These physical activities can be categorized into three main categories, i.e., non-hand-oriented activities, hand-oriented eating activities, and hand-oriented general activities.

In this dataset, Walking, Jogging, Upstairs, Downstairs, Sitting, standing activities are measured by a wearable device. They will be recorded and analyzed by the wearable watch, and Figure 4 and Table 3 show the amount of data analyzed in each activity in this data set.

TABLE III. Number of Data Analyzed in Each Activity

Walking	137375
Jogging	129392
Upstairs	35137
Downstairs	33358
Sitting	4599
Standing	3555
activity, dtype: int64	

B. *Network Structure*

In this article, we have used CNN2 network to classify WSDM data. Convolutional Neural Networks (CNN) and CNN2 can be employed to group WSDM data. The main difference between CNN and CNN2 is in data classification due to the different number of convolutional layers. By adding a second convolutional layer, CNN2 network can

detect more complex patterns in the input data but with a higher computational cost relative to CNN network. Depending on the size and complexity of input data, application of CNN2 network may lead to more accuracy in data classification. However, if the input data is relatively simple, application of CNN network could lead to good classification performance (Figure 5).

V. EVALUATION

In this section, we analyze and assess the classification outcomes of the collected data utilizing the CNN architecture, which is represented through a Confusion matrix. The obtained results from the WSDM dataset are presented in Tables 4, 6, 8 and figures 6 to 8. Specifically, Tables 5, 7, and 9 demonstrate the outcomes of the pressure and adversarial attack tests. This information is crucial for evaluating the model's performance and practical applicability.

For example, in Table4, In this section, we evaluate and analyze the classification results of the data obtained using the CNN architecture. The performance assessment of the model is conducted using a confusion matrix, which allows us to analyze the classification accuracy for each of the physical activities. The results from the WSDM dataset are presented in the provided tables. Tables 5 and 10 display the evaluation and testing results using various assessment methods.

Table 4 shows the comparative classification results for different activities. This table includes the values of F1-score, accuracy, and recall for each activity. The obtained results indicate that CNN has achieved a very high accuracy in identifying certain activities. For instance, for the activity "Upstairs," an accuracy of 100% was recorded, demonstrating the model's high capability in recognizing this specific activity.

The analysis of the results reveals that some activities, such as "Jogging" and "Standing," are identified with high accuracy, while for the activity "Sitting," a relatively lower accuracy is observed. This issue may be due to the motion similarities of this activity with other activities, leading to classification errors. Additionally, it can be observed that the use of techniques such as Batch Normalization and Dropout has positively impacted the model's performance, contributing to improved accuracy and reduced overfitting.

The results of this section clearly indicate that the CNN architecture can effectively identify physical activities from wearable sensor data. Given the model's high accuracy in detecting specific activities and improvements through advanced deep learning techniques, it can be concluded that this method has broad applications in various fields such as healthcare, fitness, and monitoring daily activities. In the following sections, suggestions for future research and enhancements to the model's performance will be provided.

A. *Generalization*

To test the generalization of the designed algorithm, we evaluated our model with invalid data and showed our results in Table 6.

B. *Robustness*

To test the robustness of the designed algorithm, we evaluated our model with noisy data and showed our results in Table 8.

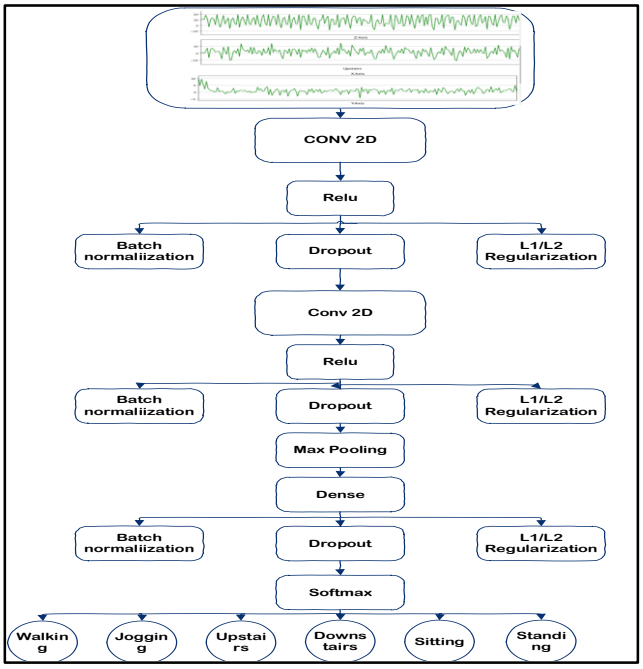

Figure 5. Structure of the proposed CNN2 network

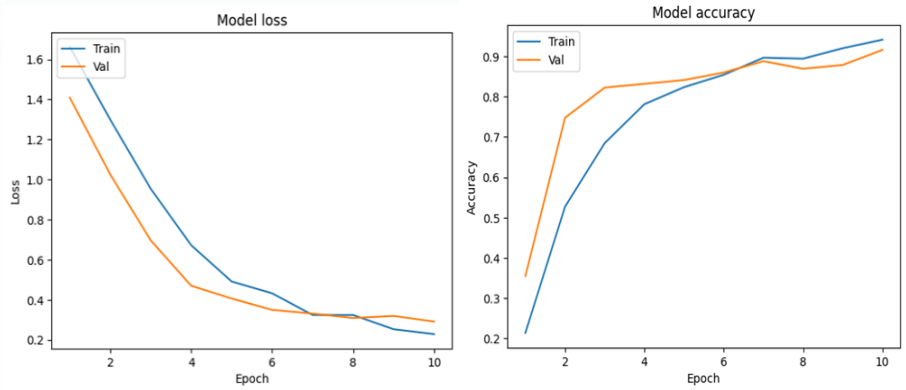

Figure 6. Accuracy and loss models in different epochs of two data sets, namely train and test

TABLE IV. COMPARISON OF RESULTS BY CONFUSION MATRIX FOR VARIOUS TASKS

Activity	Precision	Recall	F1-score	BatchNormalization/ Dropout/ Precision	BatchNormalization/Dropout/ Precision/	BatchNormalization/Dropout/ F1-score
Walking	69.57%	88.89%	78.05%	83.33%	83.33%	**83.33%**
Jogging	100.00%	94.44%	97.14%	100.00%	94.44%	**97.14%**
Upstairs	100.00%	100.00%	100.00%	100.00%	100.00%	**100.00%**
Downstairs	94.74%	100.00%	97.30%	94.74%	100.00%	**97.30%**
Sitting	72.22%	72.22%	72.22%	72.22%	72.22%	**72.22%**
Standing	100.00%	94.12%	96.97%	100.00%	94.12%	**96.97%**

TABLE V. RESULTS OF PRESSURE AND ADVERSARIAL ATTACK TESTS

Pressure Testing
Total training time: 2.614497661590576 seconds
Adversarial Attack
loss: 0.5415 - accuracy: 0.7383 Accuracy on adversarial test data: 0.7383177280426025

TABLE VI. COMPARISON OF RESULTS BY CONFUSION MATRIX FOR DIFFERENT TASKS

Activity	Precision	Recall	F1-score	BatchNormalization /Precision	BatchNormalization/Precision	BatchNormalization /F1-score
Walking	77.78%	77.78%	77.78%	71.43%	83.33%	**76.92%**
Jogging	100.00%	94.44%	97.14%	100.00%	94.44%	**97.14%**
Upstairs	100.00%	100.00%	100.00%	100.00%	100.00%	**100.00%**
Downstairs	100.00%	100.00%	100.00%	94.74%	100.00%	**97.30%**
Sitting	77.78%	77.78%	77.78%	80.00%	66.67%	**72.73%**
Standing	94.44%	100.00%	97.14%	94.12%	94.12%	**94.12%**

TABLE VII. RESULTS OF PRESSURE TESTS AND ADVERSARIAL ATTACK.

Pressure Test
Total training time: 3.1506540775299072 seconds
Adversarial attack
loss: 0.2340 - accuracy: 0.9065 Accuracy on original test data: 0.9065420627593994 loss: 0.5415 - accuracy: 0.7383 Accuracy on adversarial test data: 0.7383177280426025

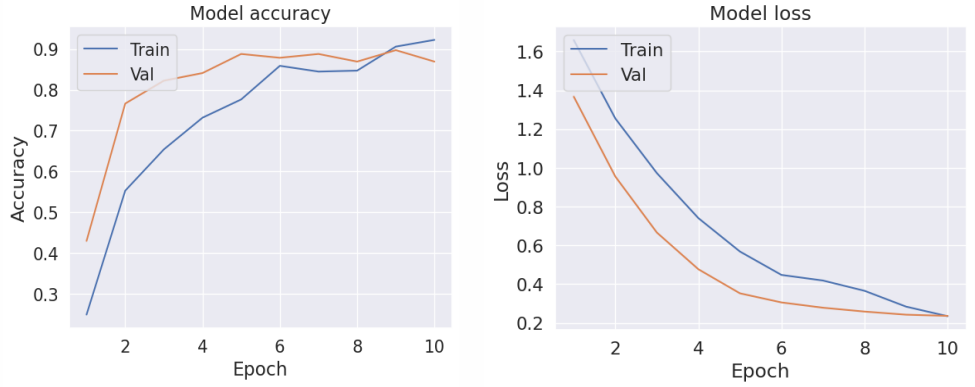

Figure 7. Accuracy and loss models in different epochs in wo data sets: train and test

TABLE VIII. ACCURACY AND LOSS MODELS IN DIFFERENT EPOCHS IN TWO DATA SETS: TRAIN AND TEST

Activity	Precision	Recall	F1-score	BatchNormalization / Bias/ Precision	BatchNormalization/ Bias/ recall	BatchNormalization/Bias/ F1-score
Walking	82.35%	77.78%	80.00%	64.29%	100.00%	**78.26%**
Jogging	100.00%	94.44%	97.14%	100.00%	94.44%	**97.14%**
Upstairs	100.00%	100.00%	100.00%	94.74%	100.00%	**97.30%**
Downstairs	94.74%	100.00%	97.30%	94.74%	100.00%	**97.30%**
Sitting	78.95%	83.33%	81.08%	100.00%	44.44%	**61.54%**
Standing	94.12%	94.12%	94.12%	94.12%	94.12%	**94.12%**

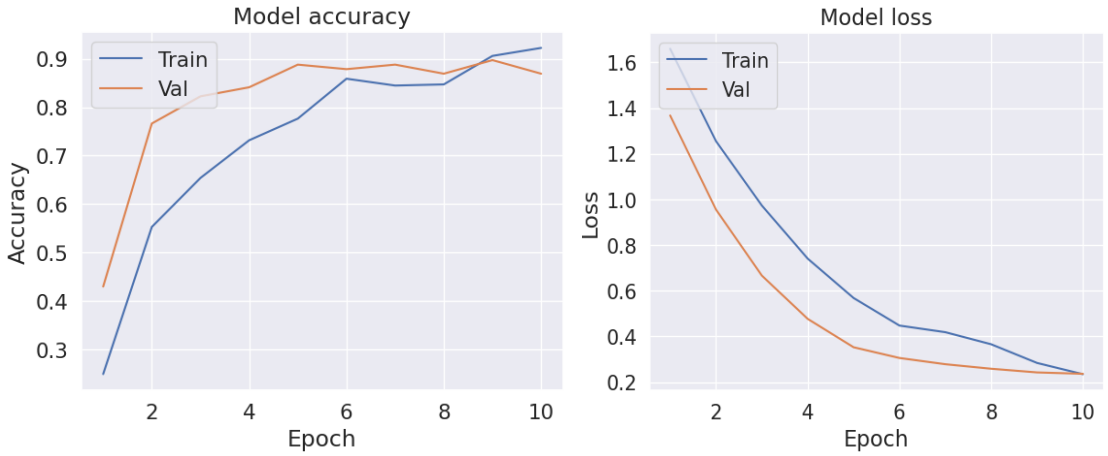

Figure 8. Accuracy and loss graphs in different epochs in two data sets: train and test

TABLE IX. RESULTS OF PRESSURE TESTS AND ADVERSARIAL ATTACK.

Pressure testing
Total training time: 2.6136112213134766 seconds
Adversarial attack
loss: 0.2340 - accuracy: 0.9065 Accuracy on original test data: 0.9065420627593994 loss: 0.5415 - accuracy: 0.7383

VI. RESULTS

According to the review of articles and previous work conducted by experts in the laboratory, in this field, the acceptable accuracy for detecting activities using experts' method was found to be 80%. The presented method was able to attain this accuracy in all tests except the Adversarial test in the test data augmentation mode (Figure 10).

For detail; Table 10 in this study examines the overall accuracy of the proposed methods. This table includes three different scenarios regarding the model's performance when faced with input data, which are detailed below:

A. Scenario One / Original Data

In this scenario, the model was trained on the original data and achieved an accuracy of 0.89 (89%). This value indicates that the model has successfully recognized activities and provides a satisfactory accuracy for activity detection.

B. Scenario Two / Generalization

In this scenario, the model was tested on invalid data, leading to an increase in accuracy to 0.90 (90%). This demonstrates the model's ability to generalize and correctly identify activities even when confronted with non-standard data. The ability to generalize is an important feature in machine learning models as it indicates that the model can adapt well to varying conditions and new data.

C. Scenario Three / Robustness

In this scenario, the model was tested against noisy data, resulting in an accuracy of 0.88 (88%). This value is slightly lower than the previous two scenarios, which may be attributed to the presence of noise in the data and its negative impact on activity recognition accuracy. Nevertheless, an accuracy of 88% still reflects a reasonable level of robustness for the model.

TABLE X. THE TOTAL ACCURACY OF THE PROPOSED METHODS

Accuracy	
First State/ Original Data	0.89
Second State/ Generalization	0.90
Third State/ Robustness	0.88

VII. CONCLUSION

In this research, we conducted a classification analysis on accelerometer and gyroscope data collected from smartphones and smartwatches. The majority of activities were classified using artificial neural network algorithms, including the CNN

model. This approach allows for accurate identification and categorization of various activities based on sensor data.

Then to evaluate Smart AI wearable artificial intelligence system, we proposed and compared all kinds of tests based on Generalization, Bias, Robustness as well as black test and pressure test. The results show that the designed system has received acceptable and imperceptible results with the generalization and robustness evaluation criteria. The prediction accuracy has been improved with two methods of batch normalization and dropout from bias criteria, as well as pressure and adversarial tests in black testing. The findings indicated that the designed Smart AI wearable system was successfully evaluated and the standard recognized. In future research, other types of tests will be conducted to evaluate AI products on other systems and products.

ACKNOWLEDGMENT

This work is supported by Iran Center for Innovation and Development of AI (CIDAI) Section, ICT Research Institute.

REFERENCE

[1] Addepally, S. A. and Purkayastha, S. (2017). Mobile application based cognitive behavior therapy (cbt) for identifying and managing depression and anxiety. In International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management, pages 3–12. Springer.

[2] Agarwal, P. and Alam, M. (2020). A lightweight deep learning model for human activity recognition on edge devices. ProcediaComputer Science, 167:2364–2373.

[3] Altini, M., Penders, J., Siirtola, P & ,.Van Gils, M. (2019). Human activity recognition using wearable sensors by deep convolutional neural networks. Sensors, 19(4), 873.

[4] Andriushchenko, M., Croce, F., Flammarion, N., & Hein, M. (2020, August). Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision* (pp. 484-501). Cham: Springer International Publishing.

[5] Chen, C., Li, F., Zhang, Y., Peng, Y & ,.Wang, L. (2021). Human activity recognition using deep learning on smartphone sensor data: a systematic literature review. Journal of Ambient Intelligence and Humanized Computing, 12(8), 7913-7934.

[6] Dong, X., Lin, Q., Yang, Z & ,.Hu, H. (2022). An Intelligent Human Activity Recognition System Based on Mobile Sensing Technology and Convolutional Neural Network. IEEE Transactions on Industrial Informatics, 18(6), 4134-4144.

[7] Ellis, R. J., Ng, Y. S., Zhu, S., Tan, D. M., Anderson, B ,. Schlaug, G., and Wang, Y. (2015). A validated smartphone-based assessment of gait and gait variability in parkinson's disease. PLoS one:(۱۰)۱۰ ,e0141694.

[8] Esfahani, P. and Malazi, H. T. (2017). Pams: A new position-aware multi-sensor dataset for human activity recognition using smartphones. In 2017 19th International Symposium on Computer Architecture and Digital Systems (CADS), pages 1–7. IEEE.

[9] Fawzi, A., Moosavi-Dezfooli, S. M., & Frossard, P. (2016). Robustness of classifiers: from adversarial to random noise. Advances in neural information processing systems, 29.

[10] Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, Dynamical Systems and Turbulence, Warwick 1980, pages 366–381, Berlin, Heidelberg, 1981. Springer Berlin Heidelberg.

[11] G. M. Weiss and A. E. O'Neill. Smartphone and smartwatch based activity recognition. Jul 2019.

[12] G. M. Weiss, K. Yoneda, and T. Hayajneh. Smartphone and smartwatch-based biometrics using activities of daily living. IEEE Access, 7:133190–133202, 2019.

[13] Garbin, C., Zhu, X., & Marques, O. (2020). Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimedia Tools and Applications*, *79*, 12777-12815.

[14] Ghahremani, A & ,.Cook, D. J. (2020). Deep convolutional and LSTM recurrent neural networks for activity recognition in smart homes. Computer Communications, 150, 354-365.

[15] Guillaume Tauzin, Umberto Lupo, Lewis Tunstall, Julian Burella Perez, Matteo Caorsi, Anibal Medina-Mardones, Alberto Dassatti, and Kathryn Hess. giotto-tda: A topological data analysis toolkit for machine learning and data exploration, 2020.

[16] Guo, L., Wu, Y., Xu, Y & ,.Sun, X. (2021). Evaluation of deep learning models for recognizing fine-grained daily human activities using accelerometer data. Measurement, 182, 109487.

[17] Hernández, J., Liljeberg, P & ,.Tenhunen, H. (2019). Activity recognition in free-living settings: accuracy and reliability issues in smartwatches. Sensors, 19(14), 3051.

[18] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. Neural Comput., 9(8):1735–1780.

[19] Jian-Bo Yang, Nguyen Nhut, Phyo San, Xiaoli li, and Priyadarsini Shonali. Deep convolutional neural networks on multichannel time series for human activity recognition. IJCAI, 07 2015.

[20] Kwapisz, J. R., Weiss, G. M., and Moore, S. A. (2011 .(Activity recognition using cell phone accelerometers .ACM SigKDD Explorations Newsletter, 12(2):74–82.

[21] L. M. Seversky, S. Davis, and M. Berger. On time-series topological data analysis: New data and opportunities. In 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1014–1022, 2016.

[22] Li, M & ,.Broun, A. (2020). A comparison of machine learning models for human activity recognition using accelerometer data from wearable devices. In Proceedings of the 14th International Conference on Ubiquitous Information Management and Communication (pp. 1-8). ACM.

[23] Liang, R., Wong, A & ,.Chio, U. S. (2021). Human activity recognition with deep learning: A review and an experimental study. Information Fusion, 67, 44-57.

[24] Lin, J., Xu, L., Liu, Y., & Zhang, X. (2020). Black-box adversarial sample generation based on differential evolution. *Journal of Systems and Software*, *170*, 110767.

[25] Liu, C., Chen, Y., Lei, J & ,.Zhang, L. (2019). Human activity recognition using smartphone sensors with different positions and body postures. Measurement, 135, 335-347.

[26] Liu, L., Qian, P., Wang, X & ,.Zeng, W. (2020). An efficient human activity recognition method based on deep transfer learning. IEEE Access, 8, 80118-80125.

[27] Lockhart, J. W., Weiss, G. M., Xue, J. C., Gallagher, S. T ,. Grosner, A. B., and Pulickal, T. T. (2011). Design considerations for the wisdm smart phone-based sensor mining architecture. In Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data, pages 25–33.

[28] Oveisi, S., Gholamrezaie, F., Qajari, N., Moein, M. S., & Goodarzi, M. (2024). Review of Artificial Intelligence-Based Systems: Evaluation, Standards, and Methods. *Advances in the Standards & Applied Sciences*, *2*(2), 4-29.

[29] Oveisi, S., Goudarzi, M., & Moein, M. S. (2024). Evaluation and Risk Management Strategies for Developing AI-based Medical Image Products. International Journal of Reliability, Risk and Safety: Theory and Application, 7(2), 15-27.

Shahrzad Oveisi received her Ph.D. in Computer Engineering in 2023 from the Department of Algorithms and Computation, School of Engineering Science, College of Engineering, University of Tehran. Her research focuses on Reliability Analysis, Intelligent Maintenance Systems, and Artificial Intelligence.

Marjan Goodarzi received her Ph.D. degree in Computer Science from Amirkabir University of Technology,Tehran,Iran. She completed her postdoctoral fellowship at the University of São Paulo, Brazil, in the field of smart grid networks. she is currently assistant professor at the ICT Research Institute (ITRC), where she has led numerous projects in the fields of artificial intelligence, AI security, AI evaluation Laboratory. She has contributed to the publication of various journal and conference papers. Her research interests include Artificial Intelligence, Image Processing, Intelligent Maintenance Systems, Data Analysis and AI security. she is currently working on evaluation measures and metrics for AI systems.

Mohammad-Shahram Moin received his B.Sc. degree in Electronic Engineering from Amirkabir University of Technology in 1988, M.Sc. degree in Electronic Engineering from the University of Tehran's Faculty of Engineering in 1991, and Ph.D. degree in Electrical Engineering (Pattern Recognition) from Ecole Polytechnique of Montréal, Canada, in 2000. He is currently Associate Professor at the ICT Research Institute (ITRC), where he has led numerous projects in the fields of artificial intelligence, biometrics, multimedia systems, and big data. Dr. Moin has a teaching background in graduate courses such as deep learning, pattern recognition, neural networks, data compression, data mining, digital signal processing and stochastic processes. He has contributed to the publication of 48 journal papers, 7 book chapters, and 79 conference papers. His research interests include Artificial Intelligence, Pattern Recognition, Image Processing, Data Analysis and Biometrics. Dr. Moin is IEEE Senior Member, head of the "Intelligent Systems Scientific Society of Iran" (ISSSI) and Editor-in-Chief of the "Iranian Journal of Information and Communication Technology".