

Automatic Recognition of Construction Worker Activities Using Deep Learning Approaches and Wearable Inertial Sensors

Sakorn Mekruksavanich¹ and Anuchit Jitpattanakul^{2,*}

¹Department of Computer Engineering, School of Information and Communication Technology, University of Phayao, Phayao, 56000, Thailand

²Department of Mathematics, Intelligent and Nonlinear Dynamic Innovations Research Center, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand

*Corresponding Author: Anuchit Jitpattanakul. Email: anuchit.j@sci.kmutnb.ac.th

Received: 20 June 2022; Accepted: 22 September 2022

Abstract: The automated evaluation and analysis of employee behavior in an Industry 4.0-compliant manufacturing firm are vital for the rapid and accurate diagnosis of work performance, particularly during the training of a new worker. Various techniques for identifying and detecting worker performance in industrial applications are based on computer vision techniques. Despite widespread computer vision-based approaches, it is challenging to develop technologies that assist the automated monitoring of worker actions at external working sites where camera deployment is problematic. Through the use of wearable inertial sensors, we propose a deep learning method for automatically recognizing the activities of construction workers. The suggested method incorporates a convolutional neural network, residual connection blocks, and multi-branch aggregate transformation modules for high-performance recognition of complicated activities such as construction worker tasks. The proposed approach has been evaluated using standard performance measures, such as precision, F1-score, and AUC, using a publicly available benchmark dataset known as VTT-ConIoT, which contains genuine construction work activities. In addition, standard deep learning models (CNNs, RNNs, and hybrid models) were developed in different empirical circumstances to compare them to the proposed model. With an average accuracy of 99.71% and an average F1-score of 99.71%, the experimental findings revealed that the suggested model could accurately recognize the actions of construction workers. Furthermore, we examined the impact of window size and sensor position on the identification efficiency of the proposed method.

Keywords: Complex human activity recognition; wearable inertial sensors; deep learning; construction workers; automatic recognition

1 Introduction

Since the 1980s, studies in the domain of human activity recognition (HAR) have progressed significantly. Exposure to high-quality computers and intelligent devices such as smartphones and



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

smartwatches is a contributing factor. Smartphones are widely utilized by the majority of the world's people in various nations. The necessity to replace humans with technology to oversee the actions of other humans has sustained HAR exploration. There is also widespread usage of sensing technology in Ambient Assisted Living (AAL) and intelligent houses [1]. Available options include physiological, environmental, and visual sensors that commonly contribute to identifying human behavior and activities for healthcare monitoring. Recognizing activities is one of the essential elements of an intelligent home [2]. The elderly and the mentally disabled suffer from a variety of medical conditions. Consequently, this might result in incorrect execution of everyday activities in these groups. Therefore, identifying anomalous behaviors becomes essential for ensuring that old and dependent individuals do their responsibilities with slight variation [3]. This may protect their safety and good health [4]. In addition, relevant sensor types are utilized for various applications, including falls, indoor location, identification of activities of daily living (ADL), and anomaly detection. As sensing modalities, wearable inertial sensors such as particular accelerometers are used for all these applications except for indoor localization [5]. This demonstrates that accelerometers are frequently utilized as action recognition sensors.

Human workers are a valuable asset in any project in the construction sector, and workers are frequently subject to different safety concerns and possible mishaps. Monitoring the process of workers' actions, behaviors, and mental conditions might reduce the incidence of worksite accidents, and avoid injuries, falls, or work-related musculoskeletal disorders (WMSDs) [6,7]. Hence, measuring and assessing workers' actions is vital to control the time and cost in the construction stage and identify possible safety issues [8].

In the previous decades, developments in new technologies have allowed academics and practitioners to progress toward establishing automated, real-time performance surveillance systems for the construction sector [9–11]. Recent innovations in sensor-based solutions (from both hardware and software viewpoints) have allowed managers and supervisors to easily engage with other organizations on the worksite and obtain enhanced productivity and safety performance [12]. Based on the kind of sensors deployed, automated action observation and identification techniques might be split into three broad groups, including 1) computer vision-based approaches, 2) audio-based approaches, and 3) sensor-based aspects.

Utilizing two-dimensional (2D) image/video cameras and three-dimensional (3D) range cameras, computer vision-based techniques collect visual data from building sites for subsequent analysis. These technologies analyze photos or videos recorded from a building site by various cameras, such as depth cameras [13], which must be positioned in strategic areas with clear lines of sight to collect all active construction activity. Even though the actual picture data improved identification accuracy, it needed more resources at a higher expense.

In addition, most audio-based approaches depend on capturing the sound patterns of devices executing specified activities. Recently, these strategies have been used as a practical compromise to kinematic-based procedures. Standard microphones, contact microphones, and microphone arrays are the most popular equipment used in this method. Microphones may be classified based on their pick-up pattern and transducer design [14]. Signal processing methods are then used to gather audio data to identify various sorts of field activity.

Lastly, sensor-based approaches use many sensors, such as accelerometers and gyroscopes, to identify the distinct movement patterns of construction personnel and equipment. These sensors may be microfabricated on to an embedded device, such as an inertial measurement unit (IMU), to gather data that, when processed, might reveal the rotational speed and orientation of the devices. In addition, some IMUs are based on microelectromechanical systems (MEMS), which are the most common sensor type owing to their tiny size and affordability. MEMS sensors are highly prevalent in various sectors and have

several applications in multiple fields, including healthcare systems, unsupervised home surveillance (home telecare), fall monitoring, identification of athlete movement patterns, asset management, and factory automation [15,16].

In recent years, human motion identification based on wearable inertial sensors has expanded to include more complicated everyday tasks such as driving, smoking, cooking, and other non-repetitive behaviors to enlarge the research area [17–19]. Nevertheless, studies and research into human behavior and conditions are extensively employed in wellness, sports, and medical treatment, but construction management studies are still limited. In the domain of the construction process, a preliminary investigation has been undertaken into employees' activity management, and the study is at an early level. The majority of the preliminary study was conducted using visual technologies. With the design and implementation of sensor-based technologies, several academics have investigated these strategies to achieve the identifiable management of construction workers' actions. Unfortunately, the few building tasks examined in prior research were straightforward. Consequently, a research gap exists between HAR and construction management. The interest in looking for a novel and efficient sensor-based HAR model to identify construction worker activities was motivated by our study in this work. Therefore, this research presents an activity recognition technique based on wearable inertial sensors for construction worker activities. The following are the contributions of this article:

- 1) The WorkerNeXt model, a unique DL architecture, is proposed since this model: 1) can automatically extract the spatial characteristics of raw signals of wearable inertial sensors using the residual connection and multi-branch aggregation processing, and 2) can increase the efficiency and accuracy of detecting construction worker activities.
- 2) We illustrate the efficacy of our proposed WorkerNeXt network by conducting empirical investigations on a common public wearable sensor-based HAR dataset, including various construction worker actions.
- 3) Moreover, we compare the experimental findings achieved using our suggested technique to those of various cutting-edge methodologies from the HAR research.

The content of this study is structured as follows. The second section examines the literature and prior research on sensor-based human activity recognition in the construction stage. This section further compares historical HAR datasets and deep learning techniques. In Section 3, a HAR approach and a new deep learning model for construction workers' activity identification are presented. The fourth section explains the experimental setup and presents the outcomes for several situations, and the fifth section addresses the observed effects. Lastly, Section 6 summarizes the key results of this study and offers recommendations for further studies.

2 Related Works

2.1 Human Activity Recognition in Construction

The construction industry's HAR studies have typically focused on three methodologies [20]: the audio-based, vision-based, and sensor-based approaches.

Audio-based solutions rely primarily on microphones for data collection and for signal processing methods used for data cleaning and feature extraction [21]. Hence, machine-learning algorithms such as support vector machines (SVM) [22] and deep neural networks [23] are used to identify construction activities. However, audio-based approaches are more adapted to distinguish equipment actions than construction worker motions since the latter significantly depends on workplace sounds.

Improvements in processing capacity and computer-vision methods have allowed researchers and practitioners to utilize camera sensors to offer semi-real-time data on worker actions at a relatively low

cost in tracking worker operations. Object detection, object tracking, and behavior identification were engaged in developing camera-based approaches. Some studies have applied cameras as the base truth against which the methodology and output of other types of sensors could be evaluated [24,25], while others use them as the primary sensor for gathering data [26,27]. For example, both [26] and [28] employed color and depth data collected from the Microsoft Kinect (camera-based) sensor to distinguish worker behaviors and measure their performance; the sensor outcomes were then analyzed and passed into an algorithm. The algorithm [28] distinguishes between distinct site personnel's professional ranks and localizes their 3D placement, while the algorithm [26] follows employees' activities and detects those activities based on their stance. Utilizing ActiGraph GT9X Link and Zephyr BioHarnessTM3, the study of trunk posture for construction tasks was also investigated.

Nevertheless, camera-based systems have several technical flaws that restrict their dependability and applicability despite their usability. Except for thermal cameras, cameras are susceptible to environmental conditions such as lighting, dust, snow, rain, and fog; they cannot work effectively in total darkness or direct sunshine. Moreover, when the collected images are packed or congested, these techniques may not be able to manage the excessive noise [20,29]. Moreover, camera-based systems need substantial storage capacities to accommodate the number of images and video data, making them computationally demanding compared to other alternatives [29]. Additionally, they can present privacy and ethical concerns since employees may feel uneasy being continually scrutinized. A network of cameras is necessary [30] to provide complete coverage of building sites, which may be rather costly.

Researchers [31–35] have also evaluated the benefit of kinematic-based approaches, including accelerometers and gyroscopes, since they may offer a more trustworthy data source for worker movement monitoring. Initial kinematic-based labor activity identification studies can be traced back to a survey [33] in which single-wired triaxial accelerometers were utilized to determine five operations of construction laborers: collecting bricks, twist placing, fetching mortar, spreading mortar, and trimming bricks. More recently, [34] gathered information from accelerometers and gyroscopes placed in the upper arms of laborers in the process of designing and assessing a construction activity identification system for seven sequential functions: hammering, turning a wrench, idling, packing wheelbarrows, pushing loaded wheelbarrows, offloading wheelbarrows, and pressing unloaded wheelbarrows. Using accelerometers connected to the wristbands of construction laborers, [31] identified four of their operations: spreading mortar, hauling and placing bricks, correcting blocks, and discarding residual mortar. Using a smartphone accelerometer and aroscope sensors attached to the wrists and legs of rebar laborers, [35] developed a technique for recognizing eight operations, including standing, strolling, squatting, washing templates, setting rebar, lashing rebar, welding rebar, and cutting rebar.

Generally, HAR is a time-series classification issue comprising sequential procedures such as sensor readings, data preparation, feature extraction, feature selection, feature annotation, supervised learning, and classifier model evaluation [20]. Various machine-learning algorithms, including Naive Bayes [33], decision trees [31,34], logistic regression [31,33], KNN [31,34], SVM [31,35], artificial neural networks [34], and others, were analyzed to obtain the highest activity classification performance from the data obtained by kinematic-based sensors.

As indicated before, despite significant study attempts to produce innovative ways to automate the action detection of laborers, no complete solution has been deployed. The industry needs a comprehensive and effective alternative that considers the variability of building activities and the complexity of elements regulating the differentiation between one activity and another, which machine learning classifiers could be utilized to provide the correct results. According to the researchers' experience, no prior study has examined the use of physiological signals to train classifiers for recognizing construction worker activities. Some researchers have incorporated wearable sensors to monitor laborers' energy consumption,

metabolic equivalence, physical activity levels, and sleep quality [36]. However, studies investigating using such a rich data source for construction activity detection are insufficient. This work investigates the viability of employing worker physiological characteristics (BVP, RR, HR, GSR, and TEMP) as an alternative data source to enhance the accuracy of machine learning classifiers in recognizing industrial work activities.

2.2 Deep Learning Approaches for Sensor-Based HAR

Many sensor-based HAR research studies have explored deep learning with significant results. Convolutional neural networks (CNNs) have become a prominent framework for deep learning in HAR because of their ability to extract the unique properties of the signal data automatically. Yang et al. [35] provide one of the earliest and more intriguing applications of CNN in HAR. Raw sensor data can be represented at a higher level, according to the researchers, if CNN's structure is sufficiently deep. Additionally, substantially enhancing feature learning and categorization inside a single model makes the learning features more discriminatory. Ronao et al. [37] have demonstrated that a deep CNN with 1D convolutional procedures outperforms conventional pattern recognition techniques for classifying smartphone sensor-based activity data. Huang et al. [38] present a CNN model with two stages to increase the recognition performance of actions with complicated characteristics and limited training data.

Recent years have seen the practical application of numerous cutting-edge algorithms introduced in computer vision to HAR. The study presented in [39] provided a HAR model based on U-Net [40] to conduct sampling point-level forecasting, thereby resolving the multi-class issue. Mahmud et al. [41] converted 1D time-series sensor data to 2D data before using residual block-based CNN to extract features and categorize actions. Ronald et al. [42] employed inception components [43] and residual connections [44] to designate a CNN-based neural network to push the boundaries of predictive performance in HAR.

2.3 Available Dataset for Construction Worker Activities

According to recent literature assessments, the research community lacks generally utilized standard datasets representing numerous individuals' realistic and variable construction worker behaviors. Even though there are a set number of datasets [45–47] for investigating construction worker activity recognition based on wearable sensors, a construction worker activity dataset is publicly available. It may be used to develop and evaluate HAR classifiers.

Mäkela et al. [48] introduced the VTT-ConIoT dataset, a publicly accessible dataset for evaluating HAR from inertial sensors in professional construction contexts. The collection comprises data from 13 people and 16 distinct actions and is gathered from three particular sensor sites on wearable devices. The dataset is provisionally accessible on the Zenodo website [48]. It was compiled under controlled settings to portray actual construction operations and was built in collaboration with important building construction stakeholders. The data set is designed using a procedure that can be used immediately to gather activity data from building sites. The system is based on IoT devices embedded into essential workwear and complies with construction industry requirements and laws.

The VTT-ConIoT dataset includes high-resolution movements from several IMU sensors, supplementary person stance, and critical points from a stationary camera. Each participant's data were meticulously labeled using a six-class general protocol and a sixteen-class protocol that was more detailed. The dataset is ideal for benchmarking and evaluating systems for human movement identification in professional contexts, particularly construction labor.

Consequently, this work trains and evaluates detection algorithms using the VTT-ConIoT dataset. The fundamental objective of this research is to precisely identify construction worker actions, as shown in Table 1.

Table 1: A summary of existing datasets that collected construction worker activities based on wearable sensor data

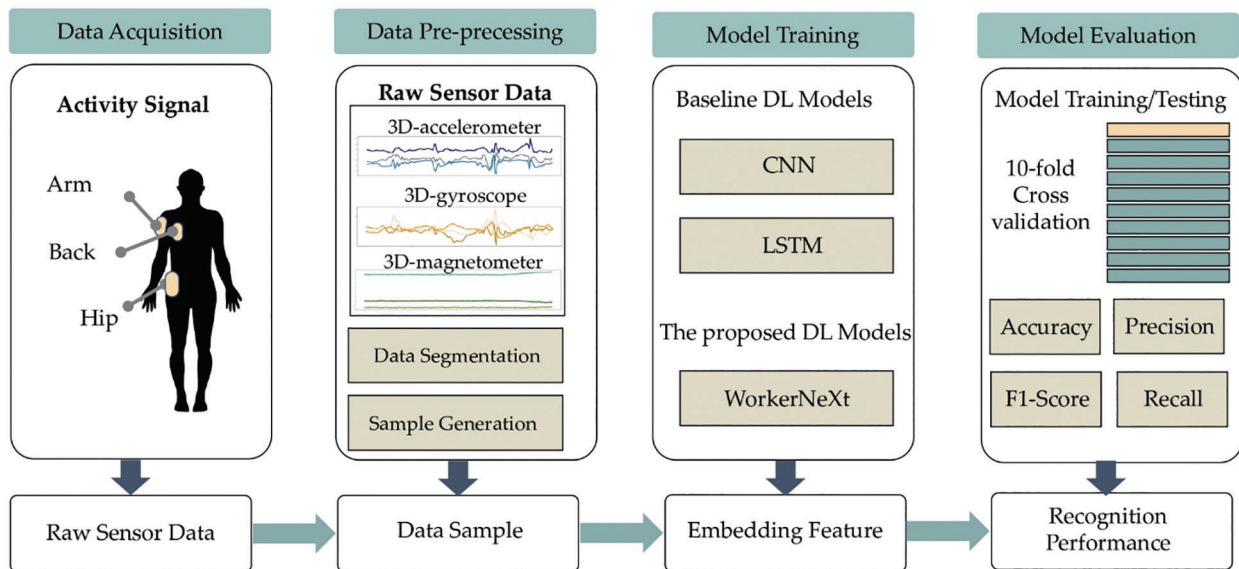
References	Year	Data availability	Sensors ¹	Locations	Sample frequency	Number of workers	Types of activities
Akhavian et al. [34]	2018	Private	A, G	Arm	100 Hz	4	10
Ryu et al. [47]	2018	Private	A	Wrist	22 Hz	10	4
Zhang et al. [45]	2018	Private	A, G	Wrist, Leg	10 Hz	9	8
Yang et al. [46]	2019	Private	A, G	Wrist, Leg	10 Hz	25	8
Mäkela et al. [48] (VTT-ConIoT dataset)	2022	Public	A, G, M	Arm, Back, Hip	103 Hz (A) 97 Hz (G, M)	16	13

Note: ¹Sensors used for recording worker activities, where A = Accelerometer, G = Gyroscope, M = Magnetometer.

3 Methodology

3.1 Overview of the Proposed Approach

This research studied construction worker activity recognition based on DL approaches. As shown in Fig. 1, the HAR framework used in this study comprises four key processes: data acquisition, data pre-processing, model training, and model evaluation.

**Figure 1:** The HAR framework used in this study

We began by acquiring a HAR dataset, which included data from wearable sensors on the behaviors of construction workers. This study has selected a publicly available dataset, VTT-ConIoT, to investigate. Three-dimensional sensor data include accelerometer, gyroscope, and magnetometer measurements. Using a sliding window approach, the sensor data were denoised, normalized, and segmented to provide sample data for training and assessing DL models. Using a 10-fold cross-validation procedure, these samples

have been generated. The trained models were then evaluated using the four most important HAR measures. Each step is discussed in depth in the subsections that follow.

3.1.1 VTT-ConIoT Dataset

The VTT-ConIoT dataset [48] is a newly suggested public dataset for evaluating HAR from inertial sensors in skilled construction environments. The dataset consists of raw sensor data for 16 specific actions (described in Table 2) gathered from 13 construction workers employing three IMU-sensored clothing items. One sensor is placed at the hip close to the non-dominant hand, and two are placed near the shoulder close to the non-dominant hand (one sensor on the upper part of the arm and another on the back part of the shoulder). The researchers collected data using Aistin iProxoxi sensors that include an IMU consisting of a 3D-accelerometer, 3D-gyroscope, 3D-magnetometer, and a barometer. Accelerometer samples were taken at 103 Hz, and the sampling rate of the gyroscope and magnetometer was 97 Hz.

Table 2: Description of construction worker activities in the VTT-ConIoT dataset

Activity	Description
Roll painting	A participant paints a wall with a paint roller
Spraying paint	A participant utilizes a tube that simulates a machine to conduct gestures portraying paint being sprayed on a wall
Leveling paint	A participant uses an implement to simulate the application of screed or paint to a wall
Vacuum cleaning	A participant vacuums the floor using a vacuum cleaner
Picking objects	A participant manually collects things off the floor and puts them in a container
Climbing stairs	A participant ascends three stairs, turns around, and descends three stairs
Jumping down	A participant ascends three steps on a staircase, turns around, and descends three steps by leaping
Laying back	A participant imitates operating with their hands up while reclining on a platform of moderate height
Hands up high	A subject imitates operating on tubes by raising their hands over their head
Hands up low	A participant imitates operating on tubes by bringing their hands to their head or shoulders
Crouch on floor	A participant works on the floor while squatting, laying tiles
Kneel on floor	A participant places tiles on the floor while kneeling
Pushing cart	A participant carries a cart 20 meters down a hallway, then turns around and brings it back
Walking straight	A participant moves 20 meters in a straight line down a hallway, then turns and comes back
Walking winding	A participant circumnavigates seven cones, then goes 20 meters, turns around, and comes back
Stairs up and down	A participant ascends the steps for thirty seconds, then descends them

3.1.2 Data Preprocessing

The VTT-ConIoT dataset's unprocessed sensor data were preprocessed to ensure the effectiveness and consistency of the interpretations. Initially, the sensor data resampled the signals by applying linear interpolation to synchronize the gyroscope and magnetometer data (97 Hz) with the accelerometer data (103 Hz) by correlating both signals to the higher sampling rate [48].

The resampled data were then separated using 2, 5, and 10 s sliding windows with a 50% overlap. In IMU-based HAR solutions, the size of these windows is determined by the application [49,50].

3.2 The Proposed WorkerNext Network and Model Training

3.2.1 The Proposed WorkerNext Network Architecture

The proposed WorkerNeXt network is an end-to-end deep-learning model built on convolutional blocks with multi-kernel residual blocks of the deep residual structure. Fig. 2 depicts the suggested model's general design.

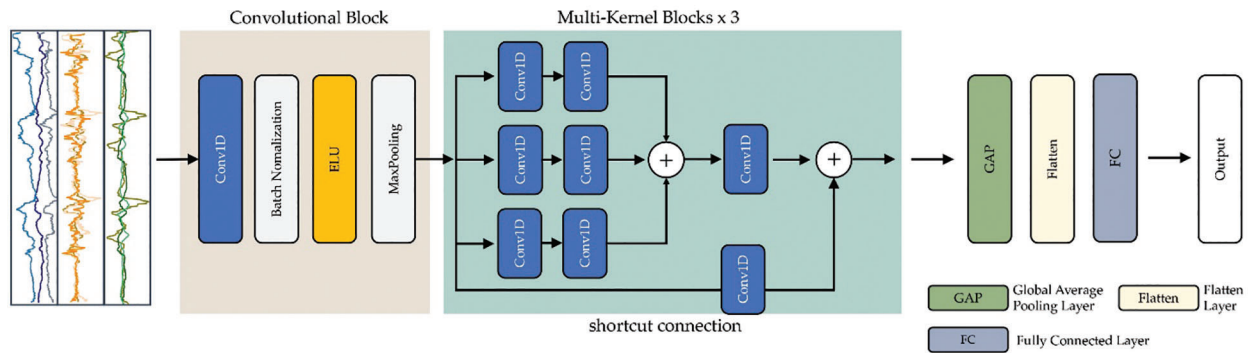


Figure 2: The architecture of the proposed WorkerNeXt model

In this work, Convolutional Blocks (ConvB) are used to extract low-level characteristics from raw sensor data. ConvB consist of four layers, as shown in Fig. 2: 1D-convolutional (Conv1D), batch normalization (BN), exponential linear unit (ELU), and max-pooling (MP). Multiple convolutional kernels that are trainable obtain particular features in the Conv1D, and each kernel generates a feature map. To accelerate and stabilize the training process, the BN layer was chosen. The ELU layer was applied to enhance the expressiveness of the model. The MP layer was utilized to condense the feature map while preserving its most vital components.

The Multi-Kernel Blocks (MK) have three components with convolutional kernels of various sizes: 1×3 , 1×5 , and 1×7 . Each part employs 1×1 convolutions while employing these kernels to reduce the system complexity of the proposed network and the number of parameters.

Utilizing the Global Average Pooling algorithm and flattened layers, the averages of each feature map were transformed into a 1D vector in the classification block (GAP). The outcome of the utterly connected layer was translated into probabilistic reasoning employing the Softmax function. To calculate the network's losses, the cross-entropy loss function, frequently used in classification tasks, was utilized.

3.2.2 Hyperparameters

In DL, the hyperparameter configuration is used to manage the learning experience. The proposed model uses the following hyperparameters: 1) epochs, 2) batch size, 3) learning rate, 4) optimization, and 5) loss function. To determine these hyperparameters, the number of epochs was set to 200, and the batch size was set to 128. After 30 epochs, if there was no improvement in validation loss, we introduced an early stopping call to terminate the training phase. Initially, we defined the learning rate at 0.001. If the proposed model validation accuracy did not improve throughout seven epochs, its value was reduced to 75% of its initial value. To decrease error, $\beta_1 = 0.90$, $\beta_2 = 0.999$, and $\varepsilon = 1 \times 10^{-8}$ were input into the Adam optimizer [51]. The error is determined by the optimizer using the categorical cross-entropy algorithm. Cross-entropy overcomes classification error and mean squared error [52]. The overall hyperparameters of the proposed WorkerNeXt network are detailed in Table 3.

Table 3: The summary of hyperparameters for the WorkerNeXt network used in this work

Stage	Hyperparameters		Values
Convolutional block	Conv1D	Kernel Size	5
	Multi-kernel block x 3	Filters	64
	Batch Normalization		-
	Activation		ReLU
	Max Pooling		2
	Branch_1_1		
	Conv1D	Kernel Size	1
		Filters	16
	Conv1D	Kernel Size	3
		Filters	16
	Branch_1_2		
	Conv1D	Kernel Size	1
		Filters	16
	Conv1D	Kernel Size	5
		Filters	16
	Branch_1_3		
	Conv1D	Kernel Size	1
		Filters	16
	Conv1D	Kernel Size	7
		Filters	16
	Branch_1		
		Kernel Size	1
	Conv1D	Stride	1
		Filters	64
	Branch_2		
		Kernel Size	1
	Conv1D	Stride	1
		Filters	64
Classification block	Global Average Pooling		-
	Flatten		-
	Dense		128
Training	Loss Function		Cross-entropy
	Optimizer		Adam
	Batch Size		128
	Number of Epochs		200

3.3 Model Training and Evaluation

3.3.1 Model Training

Following the setting of the modeling hyperparameters in the previous part, the VTT-ConIoT dataset was used to train the hybrid deep residual network. Instead of utilizing a fixed train-test split, we used five-fold cross-validation (CV) to assess the efficacy of the proposed model. The five-fold CV procedure divided the dataset into five equal-sized, unique, nonoverlapping folds. It supported the models with a four-fold CV, leaving the fifth fold for measuring assessment.

3.3.2 Evaluation Metrics

The result of an N -class classification can be visualized in a confusion matrix of size $l \times l$. Each row in a confusion matrix represents an actual class, while each column represents a predicted class. Each entry C_{ij} in a confusion matrix C denotes the number of observations (segments) from class i predicted to be of class j .

The recall, precision, and F1-scores are used as performance metrics to evaluate the correctness of a classification. The F1-score metric is particularly interesting for its robustness to class imbalance [53]. The recall (R_i), precision (P_i), and F1-measure ($F1_i$) for class i in a multiclass problem can be defined by the following equations,

$$P_i = \frac{TP_i}{TP_i + FP_i} \quad (1)$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \quad (2)$$

$$F1_i = 2 \times \frac{P_i \times R_i}{P_i + R_i} \quad (3)$$

where TP_i is the number of objects from class i assigned correctly to class i , FP_i is the number of objects that do not belong to class i but are assigned to class i , and FN_i is the number of objects from class i predicted to another class.

Typically, two methods are used to evaluate the total categorization effectiveness: macro-averaging and micro-averaging. Firstly, the indicator is computed independently for each class, and the unweighted mean is calculated. To correct for class imbalance, a weighted average might be calculated by guidance (i.e., the number of valid cases for each label). The second component globalizes the calculation by calculating the total number of true positives, false negatives, and false positives.

In the following equations, the κ and μ indices refer to the macro- and micro-averaging, respectively. P , R , and $F1$ are the total precision, recall, and $F1$ measures.

$$P_K = \frac{1}{N} \sum_{i=1}^N P_i, \quad R_K = \frac{1}{N} \sum_{i=1}^N R_i, \quad \text{and} \quad F1_K = 2 \times \frac{P_K \times R_K}{P_K + R_K} \quad (4)$$

$$P_\mu = \frac{\sum_{i=1}^N TP_i}{\sum_{i=1}^N (TP_i + FP_i)}, \quad R_\mu = \frac{\sum_{i=1}^N TP_i}{\sum_{i=1}^N (TP_i + FN_i)} \quad (5)$$

4 Experiments and Results

This section provides the description of the experimental setup, implementation, and obtained results.

4.1 Experimental Setup and Implementation

All tests were carried out using the Google Colab Pro+ framework. The GPU module (Tesla V100-SXM2-16GB) was used to accelerate the training of deep learning models. The proposed model and several basic deep learning models were built in Python using Tensorflow (version 3.9.1) and CUDA (version 8.0.6) backends. GPUs are utilized to accelerate the training and testing of deep learning techniques. The following Python libraries were used to execute the experimentations:

- Numpy and Pandas were applied for data organization to read, manipulate, and analyze sensor data.
- Matplotlib and Seaborn were applied for charting and presenting the results of data exploration and model evaluation.
- As a library, Scikit-learn was utilized in investigations for sampling and data production.
- Deep learning models were developed and trained using TensorFlow, Keras, and TensorBoard for model construction.

4.2 Results

Several experiments were performed on the VTT-ConIoT dataset to evaluate the optimal approach under different window sizes of 2, 5, and 10 s. The experiments employed a 10-fold cross-validation methodology, with four distinct physical activity scenarios in the VTT-ConIoT dataset as follows:

- Scenario I: Using only sensor data from the hip location
- Scenario II: Using only sensor data from the shoulder location
- Scenario III: Using only sensor data from the wrist location
- Scenario IV: Using only sensor data from all locations (hip, shoulder, wrist)

The results in [Tables 4–7](#) showed that the proposed WorkerNeXt models gave optimal accuracy and F1-scores for construction worker activities in all experimental scenarios. The highest accuracy was 99.71%, with an F1-score of 99.72%. When comparing the two baseline models, LSTM performed better than CNN in every design.

Table 4: Performance evaluation of baseline deep learning models and the proposed models using wearable sensors from the hip location

Window		Performance		
Sizes	Model	Accuracy	Loss	F1-score
2 s	CNN	12.97% ($\pm 13.390\%$)	2.62 (± 0.297)	8.23% ($\pm 14.985\%$)
	LSTM	75.21% ($\pm 0.890\%$)	0.97 (± 0.023)	75.19% ($\pm 0.969\%$)
	WorkerNeXt	82.26% ($\pm 1.677\%$)	0.84 (± 0.098)	82.30% ($\pm 1.653\%$)
5 s	CNN	6.27% ($\pm 0.017\%$)	2.77 (± 0.000)	0.74% ($\pm 0.002\%$)
	LSTM	89.38% ($\pm 1.251\%$)	0.39 (± 0.042)	89.39% ($\pm 1.259\%$)
	WorkerNeXt	97.54% ($\pm 0.678\%$)	0.10 (± 0.031)	97.55% ($\pm 0.680\%$)
10 s	CNN	6.26% ($\pm 0.002\%$)	2.77 (± 0.000)	0.74% ($\pm 0.000\%$)
	LSTM	95.82% ($\pm 0.354\%$)	0.13 (± 0.010)	95.82% ($\pm 0.355\%$)
	WorkerNeXt	98.85% ($\pm 1.610\%$)	0.05 (± 0.068)	98.85% ($\pm 1.604\%$)

Table 5: Performance evaluation of baseline deep learning models and the proposed models using wearable sensors from the shoulder location

Window		Performance		
Sizes	Model	Accuracy	Loss	F1-score
2 s	CNN	26.03% ($\pm 24.195\%$)	2.88 (± 0.13780)	22.68% ($\pm 26.872\%$)
	LSTM	78.45% ($\pm 0.745\%$)	0.92 (± 0.04688)	78.46% ($\pm 0.706\%$)
	WorkerNeXt	85.29% ($\pm 2.686\%$)	0.72 (± 0.12339)	85.31% ($\pm 2.675\%$)
5 s	CNN	56.92% ($\pm 3.546\%$)	3.05 (± 0.87563)	56.61% ($\pm 3.804\%$)
	LSTM	91.44% ($\pm 1.334\%$)	0.31 (± 0.05180)	91.42% ($\pm 1.339\%$)
	WorkerNeXt	98.51% ($\pm 0.129\%$)	0.08 (± 0.01682)	98.51% ($\pm 0.127\%$)
10 s	CNN	37.56% ($\pm 25.916\%$)	3.51 (± 1.703)	35.06% ($\pm 28.385\%$)
	LSTM	97.33% ($\pm 1.044\%$)	0.10 (± 0.040)	97.32% ($\pm 1.048\%$)
	WorkerNeXt	99.71% ($\pm 0.144\%$)	0.01 (± 0.005)	99.72% ($\pm 0.143\%$)

Table 6: Performance evaluation of baseline deep learning models and the proposed models using wearable sensors from the arm location

Window		Performance		
Sizes	Model	Accuracy	Loss	F1-score
2 s	CNN	21.73% ($\pm 18.938\%$)	2.51 (± 0.334)	17.92% ($\pm 21.052\%$)
	LSTM	76.91% ($\pm 1.447\%$)	0.98 (± 0.032)	76.81% ($\pm 1.477\%$)
	WorkerNeXt	88.91% ($\pm 0.408\%$)	0.66 (± 0.054)	88.91% ($\pm 0.403\%$)
5 s	CNN	37.35% ($\pm 25.393\%$)	3.65 (± 0.856)	35.06% ($\pm 28.025\%$)
	LSTM	91.00% ($\pm 0.908\%$)	0.32 (± 0.043)	90.99% ($\pm 0.914\%$)
	WorkerNeXt	98.31% ($\pm 0.242\%$)	0.06 (± 0.011)	98.31% ($\pm 0.240\%$)
10 s	CNN	34.94% ($\pm 24.046\%$)	3.43 (± 1.444)	32.31% ($\pm 26.410\%$)
	LSTM	97.31% ($\pm 0.453\%$)	0.09 (± 0.013)	97.31% ($\pm 0.451\%$)
	WorkerNeXt	99.70% ($\pm 0.129\%$)	0.01 (± 0.007)	99.71% ($\pm 0.128\%$)

Table 7: Performance evaluation of baseline deep learning models and the proposed models using wearable sensors from all locations

Window		Performance		
Sizes	Model	Accuracy	Loss	F1-score
2 s	CNN	57.19% ($\pm 26.089\%$)	2.37 (± 0.345)	56.14% ($\pm 28.306\%$)
	LSTM	88.64% ($\pm 0.755\%$)	0.41 (± 0.019)	88.66% ($\pm 0.769\%$)
	WorkerNeXt	94.89% ($\pm 0.726\%$)	0.31 (± 0.030)	94.89% ($\pm 0.719\%$)

(Continued)

Table 7 (continued)

Window		Performance		
Sizes	Model	Accuracy	Loss	F1-score
5 s	CNN	72.91% ($\pm 3.988\%$)	3.93 (± 1.203)	72.71% ($\pm 4.266\%$)
	LSTM	95.50% ($\pm 0.602\%$)	0.15 (± 0.017)	95.51% ($\pm 0.606\%$)
	WorkerNeXt	99.31% ($\pm 0.157\%$)	0.02 (± 0.005)	99.31% ($\pm 0.155\%$)
10 s	CNN	51.01% ($\pm 28.795\%$)	4.24 (± 2.404)	50.42% ($\pm 30.123\%$)
	LSTM	98.55% ($\pm 0.258\%$)	0.05 (± 0.007)	98.55% ($\pm 0.258\%$)
	WorkerNeXt	99.91% ($\pm 0.037\%$)	0.00 (± 0.002)	99.91% ($\pm 0.037\%$)

Based on the experimental results, the proposed WorkerNeXt model outperforms the other two baseline CNN and LSTM models for every scenario. The convolutional block works by extracting local spatial features from raw sensor data. Then, the multi-kernel block can extract relational information from the local spatial features. Comparing the results in [Tables 4–7](#), we notice that the sensor data from arm and shoulder locations are more helpful in identifying construction worker activities. The reason is that most worker activities relate to upper limb movement.

5 Discussion

In this section, we explored and discussed the practical consequences in various aspects, such as the effect of different window sizes and the effect of the multi-kernel module. The discussion demonstrates that the proposed WorkerNeXt model is appropriate for efficiently identifying construction worker activities.

5.1 Effect of Different Window Sizes

We investigate the impact of various window widths on WorkerNeXt's effectiveness. [Fig. 3](#) describes the F1-score achieved from each investigation based on the findings presented in [Tables 4–7](#). The comparison result demonstrates that the F1-score values are lower when the window size is set to 2 s. Construction worker activities are complicated human actions [18] that require lengthy monitoring durations to obtain the defining characteristics of motion [54].

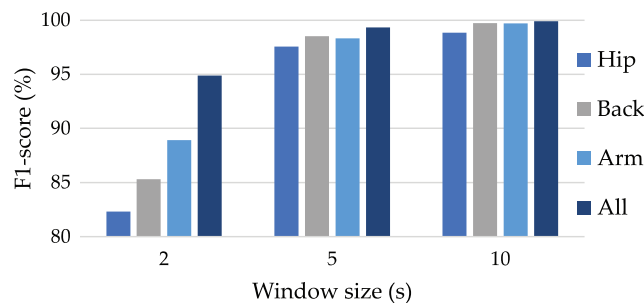


Figure 3: Performance evaluation of the WorkerNeXt model using different window sizes of sensor data recorded from different locations

5.2 Effect of Sensor Locations

Fig. 3 demonstrates that the WorkerNeXt obtained a high F1-score utilizing sensor data collected from upper body areas such as the shoulder and wrist. Regarding the construction worker tasks listed in Table 2, the VTT-ConIoT consists of 12 hand-oriented operations out of 16. Fig. 4 depicts confusion matrices derived from analyzing the WorkerNeXt model with 10 s window-sized sensor data captured at various locations (hip, back, arm, and all places).

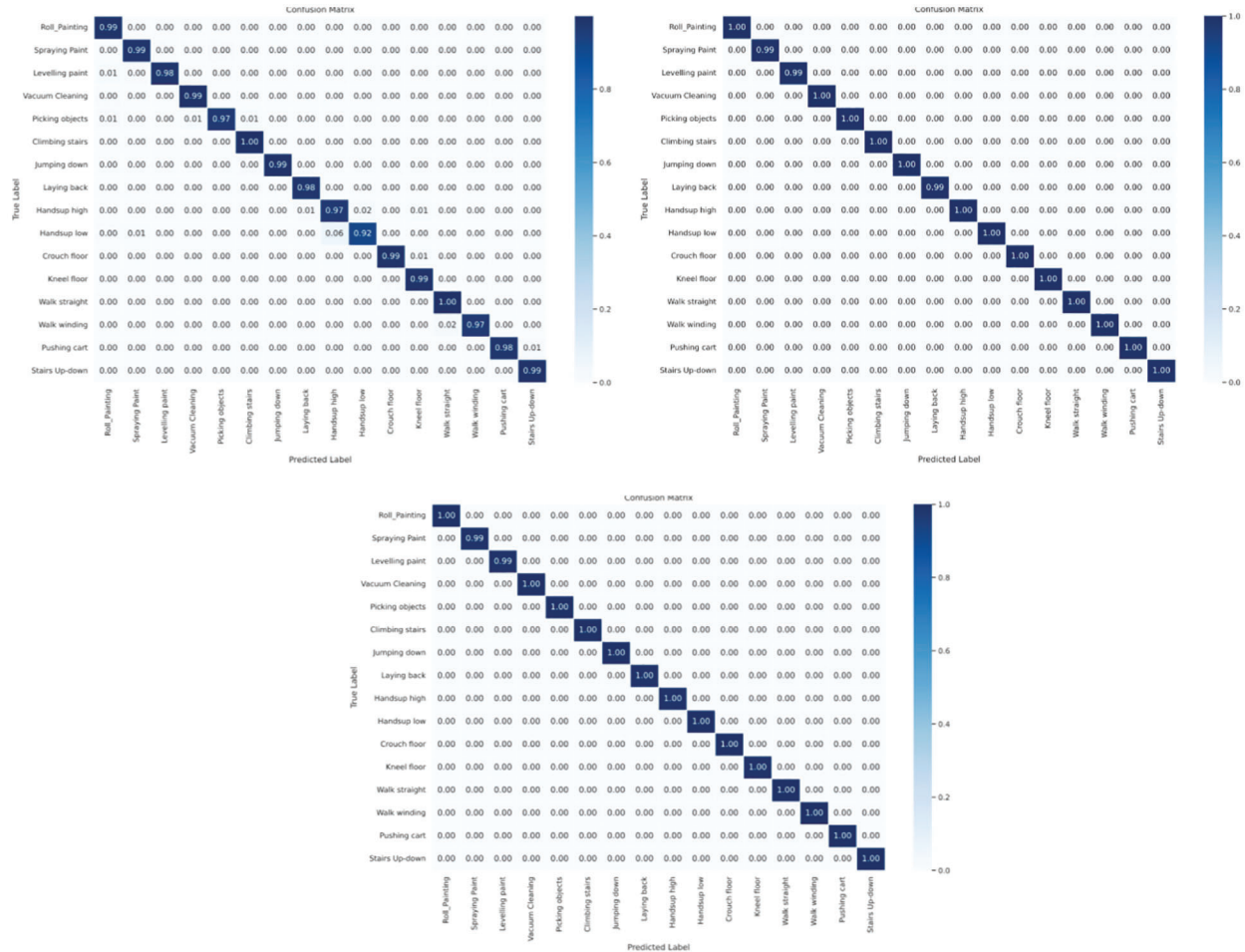


Figure 4: Confusion matrices obtained from testing the WorkerNeXt model on sensor data from different locations (from left to right: hip, back, and arm)

Considering the confusion matrices in Fig. 4, the findings suggest that micro F1-scores (acquired from the WorkerNeXt using sensor data from the hip location) of hand-oriented actions such as Picking up objects, Hands up high, and Hands up are lower than those obtained from sensor data from the back and arm locations.

6 Conclusion and Future Works

Administration of the construction process has become difficult for the construction sector and academics. Process management is performed by strategically placing personnel, materials, and

equipment to regulate cost, efficiency, productivity, and security. Nonetheless, studies on the management and supervision of construction workers are often too broad and neglectful of specifics. In addition, the activity recognition research of construction employees selects just a few categories of predetermined operations, providing construction managers with restricted expected values. This research provides a deep learning approach for identifying the activities of construction workers based on triaxial acceleration, gyroscope, and magnetometer information obtained by wearable sensors attached to three body regions (hip, back, and arm). The WorkerNeXt model, a proposed deep residual network, was selected to categorize sixteen construction worker tasks. The average classification accuracy of all samples using a 10-fold cross-validation process achieved 99.91%.

The experimental findings demonstrated that 16 building tasks could be identified and distinguished based on three-axis acceleration and angle data taken from the human participants' hip, back, and arm regions. Furthermore, the feasibility of deploying sensors incorporated in wearable devices to capture data produced by construction management has been confirmed. Compared to standard DL models, the suggested model offers noteworthy improvements in simplicity and effectiveness. Over ninety percent identification accuracy demonstrated that an approach combining a deep residual network and wearable sensors might be utilized to identify the complicated construction worker operations.

Future research will focus on the necessity for fixed-size input signals and the misclassification of specific recordings associated with certain worker tasks, such as painting-related operations. To solve the first constraint, we consider building a variable-size model that operates on global pooling rather than a flattened layer. With this strategy, we might propose a novel way to calculate the input vector length using data from additional sensors. Even though we produced good results with 1D CNN, we are exploring ensemble classifiers or multi-input deep learning models for identifying a wider variety of actions in the future.

Funding Statement: This research project was supported by University of Phayao (Grant No. FF66-UoE001); Thailand Science Research and Innovation Fund; National Science, Research and Innovation Fund (NSRF); and King Mongkut's University of Technology North Bangkok with Contract No. KMUTNB- FF-65-27.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

- [1] J. Cahill, R. Portales, S. McLoughlin, N. Nagan, B. Henrichs *et al.*, "Iot/sensor-based infrastructures promoting a sense of home, independent living, comfort and wellness," *Sensors*, vol. 19, no. 3, pp. 1–39, 2019.
- [2] H. Yu, G. Pan, M. Pan, C. Li, W. Jia *et al.*, "A hierarchical deep fusion framework for egocentric activity recognition using a wearable hybrid sensor system," *Sensors*, vol. 19, no. 3, pp. 1–28, 2019.
- [3] O. Aran, D. Sanchez-Cortes, M. T. Do and D. Gatica-Perez, "Anomaly detection in elderly daily behavior in ambient sensing environments," in *Lecture Notes in Computer Science*, Cham, Switzerland: Springer, pp. 51–67, 2016.
- [4] M. Zerkouk and B. Chikhaoui, "Long short-term memory based model for abnormal behavior prediction in elderly persons," in *How AI Impacts Urban Living and Public Health*, Cham, Switzerland: Springer, pp. 36–45, 2019.
- [5] N. Zhu, T. Diethe, M. Camplani, L. Tao, A. Burrows *et al.*, "Bridging e-health and the internet of things: The sphere project," *IEEE Intelligent Systems*, vol. 30, no. 4, pp. 39–46, 2015.
- [6] N. D. Nath, T. Chaspari and A. H. Behzadan, "Automated ergonomic risk monitoring using body-mounted sensors and machine learning," *Advanced Engineering Informatics*, vol. 38, pp. 514–526, 2018.

- [7] J. Ryu, J. Seo, H. Jebelli and S. Lee, "Automated action recognition using an accelerometer-embedded wristband-type activity tracker," *Journal of Construction Engineering and Management*, vol. 145, no. 1, pp. 1–15, 2018.
- [8] L. Sanhudo, D. Calvetti, J. P. Martins, N. M. Ramos, P. Mêda *et al.*, "Classification using accelerometers and machine learning for complex construction worker activities," *Journal of Building Engineering*, vol. 35, pp. 1–16, 2021.
- [9] K. Asadi, R. Jain, Z. Qin, M. Sun, M. Noghabaei *et al.*, "Vision-based obstacle removal system for autonomous ground vehicles using a robotic arm," in *ASCE Int. Conf. on Computing in Civil*, Baltimore, MD, 2019.
- [10] K. Asadi, H. Ramshankar, M. Noghabaei and K. Han, "Real-time image localization and registration with BIM using perspective alignment for indoor monitoring of construction," *Journal of Computing in Civil Engineering*, vol. 33, no. 5, pp. 1–15, 2019.
- [11] I. Awolusi, E. Marks and M. Hallowell, "Wearable technology for personalized construction safety monitoring and trending: Review of applicable devices," *Automation in Construction*, vol. 85, pp. 96–106, 2018.
- [12] A. Kamišalić, I. Fister, M. Turkanovic and S. Karakatič, "Sensors and functionalities of non-invasive wrist-wearable devices: A review," *Sensors*, vol. 18, no. 6, pp. 1–33, 2018.
- [13] A. Khosrowpour, J. C. Niebles and M. Golparvar-Fard, "Vision-based workplace assessment using depth images for activity analysis of interior construction operations," *Automation in Construction*, vol. 48, pp. 74–87, 2014.
- [14] C. F. Cheng, A. Rashidi, M. A. Davenport and D. V. Anderson, "Evaluation of software and hardware settings for audio-based analysis of construction operations," *International Journal of Civil Engineering*, vol. 17, pp. 1469–1480, 2019.
- [15] S. H. Khan and M. Sohail, "Activity monitoring of workers using single wearable inertial sensor," in *2013 Int. Conf. on Open Source Systems and Technologies*, New York, USA, IEEE, pp. 60–67, 2013.
- [16] H. Kim, C. R. Ahn, D. Engelhaupt and S. Lee, "Application of dynamic time warping to the recognition of mixed equipment activities in cycle time measurement," *Automation in Construction*, vol. 87, pp. 225–234, 2017.
- [17] C. Pham, S. Nguyen-Thai, H. Tran-Quang, S. Tran, H. Vu *et al.*, "Senscapsnet: Deep neural network for non-obtrusive sensing based human activity recognition," *IEEE Access*, vol. 8, pp. 86934–86946, 2020.
- [18] S. Mekruksavanich and A. Jitpattanakul, "Deep convolutional neural network with RNNs for complex activity recognition using wrist-worn wearable sensor data," *Electronics*, vol. 10, no. 14, pp. 1–33, 2021.
- [19] H. Kwon, G. D. Abowd and T. Plötz, "Complex deep neural networks from large scale virtual imu data for effective human activity recognition using wearables," *Sensors*, vol. 21, no. 24, pp. 1–26, 2021.
- [20] B. Sherafat, C. R. Ahn, R. Akhavian, A. H. Behzadan, M. Golparvar-Fard *et al.*, "Automated methods for activity recognition of construction workers and equipment: State-of-the-art review," *Journal of Construction Engineering and Management*, vol. 146, no. 6, pp. 1–19, 2020.
- [21] C. Sabillon, A. Rashidi, B. Samanta, M. A. Davenport and D. V. Anderson, "Audio-based Bayesian model for productivity estimation of cyclic construction activities," *Journal of Computing in Civil Engineering*, vol. 34, no. 1, pp. 04019048, 2020.
- [22] C. Cho, Y. Lee and T. Zhang, "Sound recognition techniques for multi-layered construction activities and events," in *ASCE Int. Workshop on Computing in Civil Engineering 2017*, Reston, VA, ASCE, pp. 326–334, 2017.
- [23] O. Gencoglu, T. Virtanen and H. Huttunen, "Recognition of acoustic events using deep neural networks," in *22nd European Signal Processing Conf. (EUSIPCO)*, New York, NY, IEEE, pp. 506–510, 2014.
- [24] T. Cheng, J. Teizer, G. C. Migliaccio and U. C. Gatti, "Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data," *Automation in Construction*, vol. 29, pp. 24–39, 2013.
- [25] R. Akhavian and A. H. Behzadan, "Smartphone-based construction workers' activity recognition and classification," *Automation in Construction*, vol. 71, pp. 198–209, 2016.
- [26] V. Escorcía, M. A. Davila, M. Golparvar-Fard and J. C. Niebles, "Automated vision-based recognition of construction worker actions for building interior construction operations using RGBD cameras," in *Construction Research Congress 2012*, Indiana, US, ASCE, pp. 879–888, 2012.

- [27] N. Mani, K. Kisi, E. Rojas and E. Foster, "Estimating construction labor productivity frontier: Pilot study," *Journal of Construction Engineering and Management*, vol. 143, no. 10, pp. 04017077, 2017.
- [28] I. Weerasinghe, J. Ruwanpura, J. Boyd and A. Habib, "Application of microsoft kinect sensor for tracking construction workers," in *Construction Research Congress 2012*, Indiana, US, ASCE, pp. 858–867, 2012.
- [29] R. Akhavian and A. H. Behzadan, "Knowledge-based simulation modeling of construction fleet operations using multimodal-process data mining," *Journal of Construction Engineering and Management*, vol. 139, no. 11, pp. 04013021, 2013.
- [30] C. F. Cheng, A. Rashidi, M. A. Davenport and D. V. Anderson, "Activity analysis of construction equipment using audio signals and support vector machines," *Automation in Construction*, vol. 81, pp. 240–253, 2017.
- [31] J. Ryu, J. Seo, H. Jebelli and S. Lee, "Automated action recognition using an accelerometer-embedded wristband-type activity tracker," *Journal of Construction Engineering and Management*, vol. 145, no. 1, pp. 04018114, 2019.
- [32] E. Valero, A. Sivanathan, F. Bosché and M. Abdel-Wahab, "Musculoskeletal disorders in construction: A review and a novel system for activity tracking with body area network," *Applied Ergonomics*, vol. 54, pp. 120–130, 2016.
- [33] L. Joshua and K. Varghese, "Construction activity classification using accelerometers," in *Construction Research Congress 2010: Innovation for Reshaping Construction Practice*, Indiana, US, ASCE, pp. 61–70, 2010.
- [34] R. Akhavian and A. H. Behzadan, "Coupling human activity recognition and wearable sensors for data-driven construction simulation," *ITcon*, vol. 23, pp. 1–15, 2018.
- [35] S. Yang, J. Cao and J. Wang, "Acoustics recognition of construction equipments based on LPCC features and SVM," in *2015 34th Chinese Control Conf. (CCC)*, New York, NY, IEEE, pp. 3987–3991, 2015.
- [36] W. Lee, K. Lin, E. Seto and G. Migliaccio, "Wearable sensors for monitoring on-duty and off-duty worker physiological status and activities in construction," *Automation in Construction*, vol. 83, pp. 341–353, 2017.
- [37] C. A. Ronao and S. B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Systems with Applications*, vol. 59, pp. 235–244, 2016.
- [38] J. Huang, S. Lin, N. Wang, G. Dai, Y. Xie *et al.*, "Tse-cnn: A two-stage end-to-end cnn for human activity recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 292–299, 2020.
- [39] Y. Zhang, Z. Zhang, Y. Zhang, J. Bao, Y. Zhang *et al.*, "Human activity recognition based on motion sensor using u-net," *IEEE Access*, vol. 7, pp. 75213–75226, 2019.
- [40] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany: Springer, pp. 234–241, 2015.
- [41] T. Mahmud, A. Q. Sayyed, S. A. Fattah and S. Y. Kung, "A novel multi-stage training approach for human activity recognition from multimodal wearable sensor data using deep neural network," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 1715–1726, 2021.
- [42] M. Ronald, A. Poulose and D. S. Han, "iSPLInception: An inception-resnet deep learning architecture for human activity recognition," *IEEE Access*, vol. 9, pp. 68985–69001, 2021.
- [43] H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier and D. F. Schmidt *et al.*, "Inceptiontime: Finding alexnet for time series classification," *Data Mining Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.
- [44] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, US, pp. 770–778, 2016.
- [45] M. Zhang, S. Chen, X. Zhao and Z. Yang, "Research on construction workers' activity recognition based on smartphone," *Sensors*, vol. 18, no. 8, pp. 1–18, 2018.
- [46] Z. Yang, Y. Yuan, M. Zhang, X. Zhao and B. Tian, "Assessment of construction workers' labor intensity based on wearable smartphone system," *Journal of Construction Engineering and Management*, vol. 145, no. 7, pp. 04019039, 2019.

- [47] J. Ryu, J. Seo, H. Jebelli and S. Lee, “Automated action recognition using an accelerometer-embedded wristband-type activity tracker,” *Journal of Construction Engineering and Management*, vol. 145, no. 1, pp. 04018114, 2018.
- [48] S. -M. Mäkela, A. Lämsä, J. S. Keränen, J. Liikka, J. Ronkainen *et al.*, “Introducing VTT-ConIoT: A realistic dataset for activity recognition of construction workers using IMU devices,” *Sustainability*, vol. 14, no. 1, pp. 1–20, 2022.
- [49] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger and K. V. Laerhoven, “Introducing wesad, a multimodal dataset for wearable stress and affect detection,” in *The 20th ACM Int. Conf. on Multimodal Interaction*, Boulder, CO, US, pp. 400–408, 2018.
- [50] L. McInnes, J. Healy and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” arXiv:1802.03426, 2018.
- [51] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *The Int. Conf. on Learning Representations*, Banff, AB, Canada, 2014.
- [52] K. Janocha and W. Czarnecki, “On loss functions for deep neural networks in classification,” arXiv:1702.05659, 2017.
- [53] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Information Processing & Management*, vol. 45, pp. 427–437, 2009.
- [54] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten and P. J. Havinga, “Complex human activity recognition using smartphone and wrist-worn motion sensors,” *Sensors*, vol. 16, no. 4, pp. 1–24, 2016.