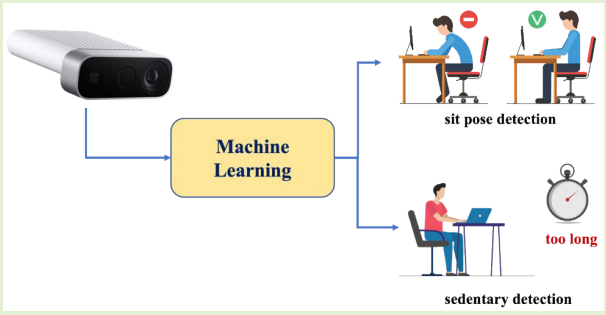
SitPose: Real-Time Detection of Sitting Posture and Sedentary Behavior Using Ensemble Learning With Depth Sensor

Hang Jin, Xin He, *Member, IEEE*, Lingyun Wang, Yujun Zhu, Weiwei Jiang, and Xiaobo Zhou, *Senior Member, IEEE*

***Abstract*—Poor sitting posture can lead to various work-related musculoskeletal disorders (WMSDs). Office employees spend approximately 81.8% of their working time seated, and sedentary behavior can result in chronic diseases such as cervical spondylosis and cardiovascular diseases. To address these health concerns, we present SitPose, a sitting posture and sedentary detection system utilizing the latest Kinect depth camera. The system tracks 3-D coordinates of bone joint points in real-time and calculates the angle values of related joints. We established a dataset containing six different sitting postures and one standing posture, totaling 33 409 data points, by recruiting 36 participants. We applied several state-of-the-art machine learning algorithms to the dataset and compared their performance in recognizing the sitting poses. Our results show that the ensemble learning model based on the soft voting mechanism achieves the highest F_1 score of 98.1%. Finally, we deployed the SitPose system based on this ensemble model to encourage better sitting posture and to reduce sedentary habits.**



***Index Terms*— Azure Kinect, depth camera, ensemble learning, sedentary, sitting posture.**

NOMENCLATURE

$S(t)$	Sinusoidal signal sent by the depth camera.
a	Transmitted amplitude.
f	Modulation frequency.
A	Attenuated amplitude at the receiver.
B	Intensity shift caused by ambient light.
$\Delta\varphi$	Phase difference.
$R(t)$	Signal received at time t.
Δt	Delay.
d	Depth distance.
c	Speed of light.

Received 28 October 2024; accepted 11 February 2025. Date of publication 24 February 2025; date of current version 2 April 2025. This work was supported in part by the Natural Science Foundation of China under Grant 62072004 and Grant 62402229. The associate editor coordinating the review of this article and approving it for publication was Prof. Xiao Wang. *(Corresponding authors: Xin He; Yujun Zhu.)*

Hang Jin, Xin He, Lingyun Wang, and Yujun Zhu are with the School of Computer and Information, Anhui Normal University, Wuhu, Anhui 241002, China (e-mail: hang960104@gmail.com; xin.he@ahnu.edu.cn; lingyunwang@ahnu.edu.cn; zhuyujun@ahnu.edu.cn).

Weiwei Jiang is with the School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, Jiangsu 210044, China (e-mail: weiwei.jiang@nuist.edu.cn).

Xiaobo Zhou is with the School of Computer Science and Technology, Tianjin University, Tianjin 300072, China (e-mail: xiaobo.zhou@tju.edu.cn).

Digital Object Identifier 10.1109/JSEN.2025.3541821

I. INTRODUCTION

OFFICE workers typically remain seated throughout their workday due to the nature of their tasks and various other factors. Consequently, many people experience backaches primarily due to their poor sitting posture and prolonged sedentary habits. Medical research indicates that improper sitting posture can lead to a range of health issues, particularly affecting the cervical and lumbar spine, and significantly affecting respiratory function [1]. Furthermore, prolonged sitting can double the risk of developing diabetes, as well as contribute to the accumulation of abdominal fat, leading to health problems such as overweight, obesity, and cardiovascular disease [2].

In the field of occupational health research, Kallings et al. [3] conducted a comprehensive study demonstrating a correlation between increased durations of sedentary behavior in the workplace and a decline in self-reported general health status. This extensive study encompassed a sample size of 44 978 individuals, with an average participant age of 42.1 years, ranging from 18 to 75 years. Concurrently, Cao et al. [4] provided insights into the health implications of prolonged sedentary lifestyles. Their research, involving a substantial cohort of 360 047 participants from the U.K. Biobank, delved into the relationship between

sedentary behavior (exceeding 6 h per day) and the heightened risk of 12 types of non-communicable diseases (NCDs).

Recognizing these significant health implications, our research aims to mitigate such risks by introducing a novel sitting posture health detection system that utilizes visual detection technology to provide interactive reminders. This system improves the sitting posture and habits of individuals, thereby preventing chronic diseases caused by prolonged sitting, especially in working scenarios. The RoSeFi [5] system adopted WiFi channel state information to monitor sedentary behavior. However, we also aim to examine sitting poses in real-time. Therefore, we have chosen a depth camera as the monitoring sensor.

In 2019, Microsoft released Azure Kinect Developer Kit (DK), which is the third Kinect depth camera after Kinect V1 and Kinect V2. Unlike the previous Kinect V1 and Kinect V2, which are mainly aimed at Xbox game consumers, Azure Kinect is a platform for professional developers and commercial companies. It features advanced AI sensors and accompanying computer vision and speech model development kits. The Azure Kinect camera group consists of an RGB camera and an infrared (IR) camera. It continues the time-of-flight (ToF) depth estimation principle of Kinect V2, which calculates the distance to the target by calculating the time required for the emitted light to reach the target and return to the sensor [6]. Its accuracy is significantly higher than that of other commercial cameras. The latest Kinect camera excels not only in terms of vision but also in hearing and motion perception. It has an array of seven microphones for sensing sound. Kinect also integrates inertial sensors for various types of perception. It is worth mentioning that Microsoft provides an AI algorithm toolkit based on deep learning and convolutional neural networks (CNNs) and develops a set of human body tracking SDK [7].

Ran et al. [8] developed and designed a pressure sensor-based sitting posture detection system and used machine learning for classification. Guduru et al. [9] collected RGB images of upper body poses, preprocessed them with bandpass filters, and finally used deep learning to identify posture changes during sedentary work. Roh et al. [10] deployed pressure sensors on the back of the chair cushion along with the backrest and implemented a sitting posture classification system using a support vector machine (SVM). Existing RGB cameras and pressure sensor seats have limitations in detecting sitting posture and sedentary status.

The Azure Kinect depth camera offers several advantages over traditional RGB cameras, particularly in the context of posture monitoring. Unlike RGB cameras, which have stringent requirements for light sources and are associated with high costs and complexity in sensor-based deployment, the depth camera operates effectively even in suboptimal lighting conditions. This feature not only facilitates easier monitoring of sitting postures but also ensures user privacy is preserved. Furthermore, the ease of deployment of the camera eliminates the need for complex operations like installing pressure sensors on chairs, making it a more convenient and privacy-conscious choice for monitoring sitting states.

Bontrup et al. [11] recruited three trial participants and used the Kinect to collect depth data. Machine learning was utilized to categorize sitting postures into two types: correct and incorrect postures. Paliyawan et al. [12] employed Kinect to gather time and depth coordinates for ten key joints, using machine learning to help office workers avoid the phenomenon of prolonged sitting. Compared to the state-of-the-art studies, SitPose collects posture data from 36 participants and is capable of identifying two harmful behaviors, i.e., prolonged sitting and incorrect sitting postures. The real-time system promptly reminds users to improve their posture through pop-up notifications. Additionally, we categorized the types of incorrect postures, enabling users to understand their habitual incorrect sitting positions. This knowledge empowers users to avoid such postures in the future, contributing to better ergonomic health.

The depth camera works based on the principle of ToF technology. Specifically, when the near-IR light emitted by the sensor encounters a human body or an object, it will be reflected. The distance between the human body or the object can be calculated by computing the difference between the emitted light time and the reflected time or the phase difference, and depth information can be obtained. Then, the human body tracking software development kit (SDK) provided by Kinect can be used to identify human joints and bones.

In this article, we design a sitting posture recognition system, namely *SitPose*, which classifies the sitting postures from the captured 3-D bone data using machine learning. The initial step of the system involves utilizing Azure Kinect as a monitoring tool, which optimizes the data collected from depth images through feature extraction. SitPose also calculates the angles of human bone joints as part of the features. Finally, it employs ensemble learning to classify different sitting postures. To evaluate its classification effect, we use confusion matrices and conduct abnormal experiments. Based on the information of ToF, Kinect's recognition principle, and a soft voting mechanism in ensemble learning, we successfully built a sitting posture recognition system.

The main contributions of this article are summarized as follows.

1) We design a real-time sitting posture monitoring system using a depth camera. The latest Azure Kinect is utilized for processing photographs and presenting bone joint points directly, which offers ease of deployment and addresses privacy concerns.
2) For evaluating the performance, we generate a dataset of 36 participants. The dataset has a total of 33 409 samples, which is a large number. It contains body depth values of skeletal coordinates. We employed k-fold cross-validation to train our model, which offers a more comprehensive evaluation of the model's accuracy and efficiency. To visualize the results more directly, we applied a confusion matrix. Using fivefold cross-validation, the training set is divided into five parts. In each iteration, four parts are used for training and one part for testing, repeating this process five times. Finally, we calculate the average of the results from these five test iterations.

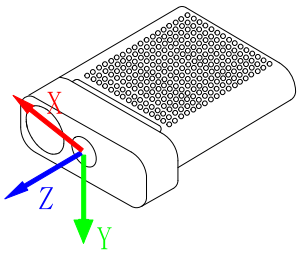


Fig. 1. Azure Kinect's spatial coordinate axis.

3) To recognize sitting postures, we adopted an ensemble learning approach based on a soft voting mechanism, SVM, decision trees (DTs), multilayer perceptron (MLP), gradient boosting DT (GBDT), and TabNet. Comprehensive experiments were conducted to evaluate the recognition accuracy of the respective models. Additionally, we demonstrated the system's latency to showcase its real-time capability.

Therefore, the system has a particular auxiliary effect on improving sitting posture and preventing lousy sitting posture and some chronic diseases caused by sedentary sitting.

The rest of this article is organized as follows. Section II introduces the depth estimation principle of ToF and how Kinect realizes the recognition of human bone joints. In Section III, we discuss the approach of utilizing Azure Kinect as a monitoring tool, as well as incorporating calculated joint angles into the feature set. Section IV presents a comparison of the classification performance of six different algorithms. In this case, we analyze the results of each action through a confusion matrix. In Section V, we summarize the effect of the detection system and propose what can be improved in future research.

II. Preliminary Knowledge

In this section, we introduce the depth estimation principle of ToF, how Kinect realizes the recognition of human bone joints, and the principle of the soft voting mechanism in ensemble learning.

A. Time of Flight

The previous generation Kinect V2, as well as the Azure Kinect, both adopt ToF technology. The basic principle of ToF is to calculate the transmission time of a signal based on its reflection to determine the distance from a person or an object [13]. Specifically, a depth camera emits a continuous pulse of IR light, which would be reflected back when it encounters a human body or an object. By recording the ToF from sending the pulse to receiving the pulse and using continuous wave modulation and phase detection inside the camera, the distance of the human body or object can be calculated [14]. The modulation measurement principle of the continuous sine wave is as follows: $S(t)$ is a sinusoidal signal sent by the depth camera, a is the amplitude, the modulation frequency is f, and $S(t)$ can be expressed as

$$S(t) = a(1 + \sin(2\pi f t)). \tag{1}$$

The signal received after being reflected by the object is $R(t)$, specifically

$$R(t) = A(1 + \sin(2\pi f t - \Delta\varphi)) + B \tag{2}$$

where A represents the amplitude of the attenuated signal after reflection, B represents the intensity shift caused by ambient light, and $\Delta\varphi$ represents the phase difference. To set these parameters, $R(t)$ is sampled four times at intervals of $(T/4)$, and a system of equations containing four equations is constructed using the four sampling values, as follows:

$$R_i = A\sin(2\pi f(t_i - \Delta t)) + (A + B), \quad i = 0, 1, 2, 3 \tag{3}$$

where Δt represents the delay between the time when the sinusoidal signal is emitted and the time when the reflected signal is received. The phase difference $\Delta\varphi$ can be obtained by constructing a system of equations using the four samples, as follows:

$$\Delta\varphi = \arctan 2(R_2 - R_0, R_1 - R_3). \tag{4}$$

Finally, based on the calculated phase difference, the depth distance d is obtained as

$$d = \frac{c}{4\pi f}\Delta\varphi \tag{5}$$

where c is the speed of light.

B. Human Skeleton Joints

The detection and recognition of human skeleton joints has a wide range of applications in motion and posture perception. Azure Kinect Sensor SDK initially processes the IR image through a CNN. This step is crucial for identifying the positions of 2-D joints and segmenting various body parts within the image. Following this information, the depth image, in conjunction with the previously determined 2-D joint locations, is integrated into a skeleton-based tracking model. This sophisticated approach yields several key outputs: the precise locations of 3-D joints, their respective orientations, and the tracking of their movements over time [15], [16], [17]. Microsoft has also developed a new body-tracking SDK for the Azure Kinect, which is based on deep learning and CNN. However, the Azure Kinect Body Tracking SDK is closed-source. The Azure Body Tracking SDK, utilizing the power of AI, intelligently estimates the coordinates of obscured extremities. This is achieved through the application of decision forests, which have been trained on an extensive dataset [18]. We can utilize the API interface provided by Microsoft to capture depth data for human body joint points.

As shown in Fig. 1, the Kinect skeletal joint coordinate system takes the depth camera focus as the coordinate origin, which is at $[0, 0, 0]$. The positive X-axis extends to the right along the depth camera focus, the positive Y-axis is vertically downward, and the Z-axis extends directly in front of the camera focus.

C. Voting Classifier

To enhance the accuracy of sitting posture detection and mitigate the issue of poor generalization when using a single

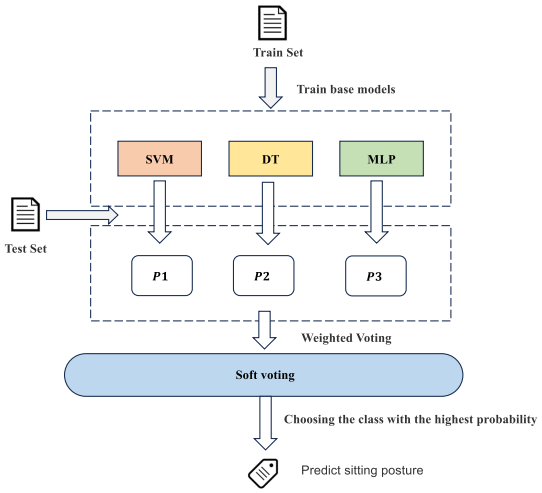


Fig. 2. Framework of the sitting posture and sedentary habits detection system.

model, we propose an ensemble learning model based on a voting classifier. Rather than being a standalone machine learning algorithm, this approach combines multiple machine learning models to complete the learning task. The voting classifier works by aggregating the predictions of several models through a voting mechanism, thereby reducing variance and improving the robustness of the model.

There are two types of voting methods: hard voting and soft voting. The key difference between the two is that hard voting predicts the class with the most votes, while soft voting predicts the class with the highest sum of predicted probabilities from all models. In other words, hard voting relies on the direct class label outputs of each model, whereas soft voting utilizes the probability outputs from each model to make predictions.

As shown in Fig. 2, this article employs three well-established models: SVM, DT, and MLP. These models are chosen because they use different learning strategies and learn features from different perspectives and spaces, complementing each other to enhance the overall performance of the voting ensemble learning model. For the voting method, we opted for soft voting, which uses the predicted probabilities from each model. This approach suits the three selected machine learning models more than hard voting.

III. System Design

In this section, we introduce the method of adopting Azure Kinect to monitor sitting posture detection and the functional modules of the detection system. The adopted notations are summarized in the Nomenclature. The detection system is based on the human skeleton coordinates collected from the depth image. It first optimizes the data through feature extraction, then calculates the angles of relevant bone joints, and adds them to the features. Ensemble learning models are used to classify different sitting postures. Machine learning models are used to classify different sitting postures. Finally, we deploy the trained model, take real-time depth images from Kinect, and input bone data into the model for classification.

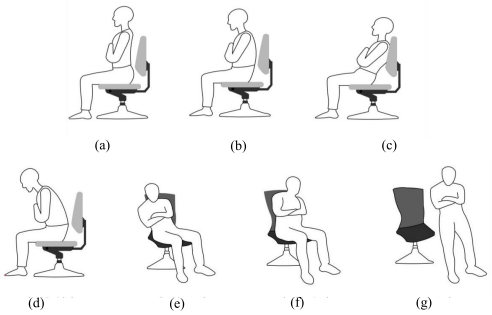


Fig. 3. Illustration of six types of sitting and standing postures examined in this work. (a) Sitting straight. (b) Hunched over. (c) Lying. (d) Lean forward. (e) Left sitting. (f) Right sitting. (g) Standing.

In this work, we classify seven postures as follows: sitting straight, hunching over, left sitting, right sitting, leaning forward, lying, and standing. Except for standing, the other six sitting postures are selected based on comprehensive studies on sitting positions [19], [20]. These postures represent a broad spectrum of typical office postures, as shown in Fig. 3. We include a standing event as a detection feature to determine if the user is seated on an office chair. This addition significantly enhances the accuracy of identifying sedentary behaviors.

A. System Overview

As shown in Fig. 2, we present a posture detection framework based on heterogeneous ensemble learning techniques, specifically utilizing a soft voting mechanism to enhance prediction accuracy and robustness.[1] The framework comprises three core modules: data collection and preprocessing, soft voting ensemble learning, and prediction.

First, the data collection and preprocessing module gathers depth camera data of specified skeletons using the Azure Kinect and extracts critical features such as skeletal angles.

The core of the framework is the soft voting ensemble learning module, which combines predictions from various algorithms, including DT, SVM, and MLP. This ensemble approach effectively integrates the strengths of individual models by averaging their weighted predictions, thereby improving overall performance and reliability compared to traditional single models.

Finally, the prediction module uses the output from the ensemble learning module to make the final posture assessment, providing real-time feedback and recommendations to help users improve their posture and reduce health issues caused by prolonged poor sitting habits. The framework aims to deliver a comprehensive and accurate solution for posture detection through efficient data processing, advanced ensemble learning techniques, and precise prediction capabilities.

B. Collection of Joint Coordinate Data

Although many studies have made Kinect-related datasets publicly available, there is still a notable lack of depth data specifically dedicated to sitting postures. To address this gap,

[1]The dataset and code are available at https://github.com/TaylorDurden1114/SitPose

TABLE I
PARTICIPANT SUMMARY STATISTICS FOR HEIGHT, WEIGHT, AND AGE, CATEGORIZED BY GENDER

Parameter	Gender (count)	Mean ± Std	Min	Max
Height (cm)	Male (18)	174.28 ± 4.82	167	185
	Female (18)	162.06 ± 4.49	155	171
	Overall (36)	168.17 ± 7.71	155	185
Weight (kg)	Male (18)	68.36 ± 8.74	50	87
	Female (18)	53.39 ± 6.02	45	66
	Overall (36)	60.88 ± 10.6	45	87
Age	Male (18)	23.67 ± 2.95	20	29
	Female (18)	25.83 ± 10.12	20	57
	Overall (36)	24.75 ± 7.43	20	57

we gathered sitting posture data from 36 participants, ensuring an identical gender ratio, and gathered a total of 33 409 instances of raw depth data. Table I presents information about participants. This dataset aims to facilitate future research by allowing additional features extraction from the raw data for posture analysis purposes, thereby advancing the field.

We installed two development kits, Azure Kinect sensor SDK and Body Tracking SDK, on a Win10 system computer. Table II outlines the components and specifications of the experimental setup. The experimental environments are illustrated in Fig. 4. Experiments were conducted in four different settings: a hardware laboratory, a home, a company office, and an office at our university. Additionally, the depth camera was positioned on both the right and left sides of the table to provide flexibility for real-world deployment scenarios. Participants performed six sitting postures (sitting straight, hunching over, leaning left, leaning right, leaning forward, and reclining), as well as a standing posture in front of the table. By collecting sitting posture data from various social roles and environments, we aim to ensure a certain level of generalizability for the model.

While the Azure Kinect can capture 32 skeletal points, this study focuses on only nine key skeletal points for several reasons. In a desk-based office setting, leg joints are frequently obstructed by the desk, and although algorithms can estimate leg positions, the accuracy of such data is limited. Hand joints were further excluded due to potential inaccuracies caused by frequent hand movements during office activities, e.g., drinking, writing, using a mouse, or browsing documents. Furthermore, certain head joints—such as the nose, left eye, and right eye—show minimal positional variation, making them redundant for our purposes. Therefore, only nine key skeletal points were selected for data collection based on these considerations.

Through Kinect's SDK, we can obtain human body depth images and bone joint data in real-time. Each joint has a unique index value, which allows us to obtain specific joint data [21]. Azure Kinect can track and accurately extract 32 bone coordinates of the human body in real-time, which is seven more than Kinect V2. We select nine joints of the subject to identify per second while also saving their coordinates of them. These nine joints are the subject's head, left clavicle, right clavicle, neck, right shoulder, left shoulder, spine of the chest, spine of the navel, and pelvis. Therefore, each sampling will obtain the raw data of 1×24 joint coordinate data. The height of the experimenters ranged from 155 to 185 cm, and experimenters of different heights were tested to avoid systematic classification errors caused by height differences. In addition, during the experiment, we slightly changed the seat position and allowed the experimenter to adjust the range of sitting posture movements to enhance the robustness of the sitting posture detection system and make the system deployment more flexible.

C. Joint Angle Feature Extraction

The angles between the bones and joints of different sitting postures are various, and these angles are an important factor in distinguishing sitting postures. To this end, we can calculate the spatial angle between the relevant bone joints, add it to the features of the dataset, and use the angle to distinguish the sitting posture. We consider the angle information between joints an important parameter for training the model. First, the angle calculation is performed on the 3-D bone coordinates, and the angle data related to the 3-D coordinates of the joints of the neck, head, left shoulder, right shoulder, chest, and pelvis are obtained. Here, taking the joint angle between the thoracic spine, cervical spine, and head $(\overrightarrow{ED}, \overrightarrow{DC})$ as an example, the calculation steps of the relevant angles are introduced. Assuming that the 3-D coordinates of the head are $p_1 = (x_1, y_1, z_1)$ and the 3-D coordinates of the neck are calculated by $p_2 = (x_2, y_2, z_2)$ then the spatial vector from the neck to the head $\overrightarrow{P_1P_2}$ is

$$\overrightarrow{P_1P_2} = (x_2 - x_1, y_2 - y_1, z_2 - z_1). \tag{6}$$

Further, we assume that the 3-D coordinates of the chest are $p_3 = (x_3, y_3, z_3)$ then the spatial vector $\overrightarrow{P_2P_3}$ from the neck to the chest is

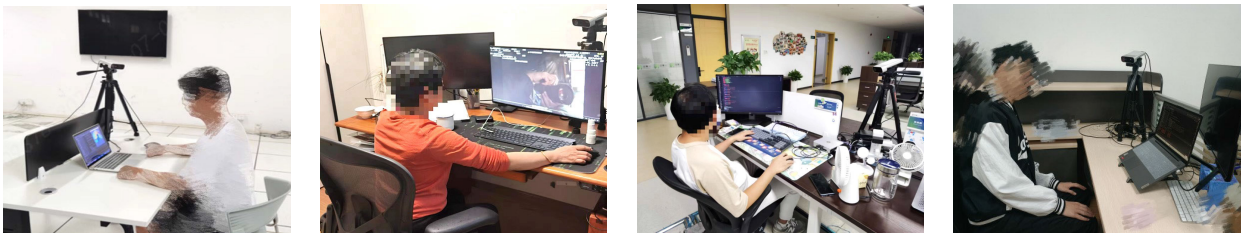
$$\overrightarrow{P_2P_3} = (x_3 - x_2, y_3 - y_2, z_3 - z_2). \tag{7}$$

Let $|\cdot|$ stand for the norm of a vector. Due to two vectors and their norms being known, the angle θ between the spatial vectors $\overrightarrow{P_1P_2}, \overrightarrow{P_2P_3}$ is calculated as

$$\theta = \arccos\left(\frac{\overrightarrow{P_1P_2} \cdot \overrightarrow{P_2P_3}}{|\overrightarrow{P_1P_2}| \cdot |\overrightarrow{P_2P_3}|}\right) \times \frac{180^\circ}{\pi}. \tag{8}$$

To further enhance the accuracy and precision of sitting posture analysis, we introduced several auxiliary points based on skeletal points to assist in measuring and describing the angles between key skeletal points. These auxiliary points include point J, parallel to the x-axis of the pelvis (point A), point K, parallel to the x-axis of the neck (point D), and points M and N, parallel to the x-axis of the left clavicle (point G) and right clavicle (point H), respectively. For example, the spatial angle $\angle JAB$ was calculated from the pelvis to the abdomen, which reflects the degree of spinal curvature in different sitting postures. The letters marked by the human body joints are shown in Fig. 5, after experiments we select $(\overrightarrow{JA}, \overrightarrow{AB})$, $(\overrightarrow{JA}, \overrightarrow{AC})$, $(\overrightarrow{JA}, \overrightarrow{AD})$, $(\overrightarrow{AB}, \overrightarrow{BC})$, $(\overrightarrow{BC}, \overrightarrow{CE})$, $(\overrightarrow{NI}, \overrightarrow{IH})$, $(\overrightarrow{MG}, \overrightarrow{GF})$, $(\overrightarrow{HI}, \overrightarrow{IC})$, $(\overrightarrow{FG}, \overrightarrow{GC})$. The corresponding angles between selected joint spatial vectors have been included in the feature for further processing.

We conducted a preliminary study to show the variations in the angles of our selected skeletal points, where the results



(a) (b) (c) (d)

Fig. 4. Experimental environments in our study. (a) Hardware laboratory. (b) Home. (c) Company. (d) Office.

are shown in Fig. 6. From the results, the distribution of skeletal angle information can be demonstrated. For instance, the data distribution from the joint points KDE shows that the leaning-back posture can be clearly distinguished from other sitting postures. In contrast, healthy sitting, right-leaning sitting, and standing postures can be differentiated from forward-leaning, hunchback, and left-leaning sitting postures. This indicates that different sitting postures exhibit significant differences in the distribution characteristics of skeletal angles, which helps in more accurate posture classification and assessment analysis.

D. Composition of the Dataset

The sitting posture detection dataset has a total of 33 409 items, including sitting straight, hunched over, left sitting, right sitting, leaning forward, lying, and standing. Among them, 5299 are sitting straight, 4985 are hunched over, 4715 are leaning forward, 4259 are left sitting, 4337 are right sitting, 4340 are lying, and 5474 are standing.

By analyzing depth images and depth values, we found that the depth data of the head is very useful for distinguishing between left-sitting and right-sitting postures. Therefore, we included the head's depth data as a feature in the dataset. The dataset comprises the head's depth values and nine extracted key skeletal angle features.

To ensure unbiased and broader model performance, we utilized k-fold cross-validation to evaluate the model, aggregating the evaluation metrics through averaging. The parameters of the model are trained based on the data of the training set, and the performance of the trained model is evaluated using the test set. We evaluated the performance of our collected human sitting posture dataset using an ensemble learning model with a soft voting mechanism.

E. Parameter Settings of Classification Models

We adopted SVM, DT, and MLP as base learners and combined them using an ensemble learning model with a soft voting mechanism to classify and evaluate the performance of the sitting posture dataset. MobileNet and SqueezeNet are exceptional lightweight CNN crafted primarily for handling image data, particularly RGB images. However, these models are not the best choice for 1-D data; traditional machine-learning methods are more suitable for this dataset.

For the SVM, we selected a polynomial kernel function due to its ability to capture nonlinear relationships between features, thus enhancing modeling capabilities. To balance robustness and simplicity, we set the hyperparameter C to its default value of 1.0. A lower C value reduces the risk of overfitting, ensuring that the model does not become overly tailored to the training data, thereby improving its generalization on unseen datasets. This choice allows us to harness the advantages of the polynomial kernel in nonlinear modeling while maintaining computational efficiency, making it suitable for complex datasets.

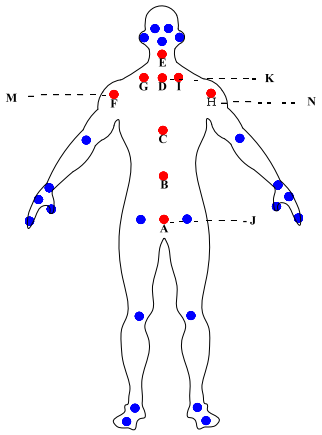


Fig. 5. Recognized joints by the depth camera where the selected joints are marked in red.

TABLE II
COMPONENTS AND SPECIFICATIONS OF THE EXPERIMENTAL SETUP

Component	Specification
Camera	Azure Kinect
Body Tracking SDK	V1.1.2
Desktop Depth	72 cm
Camera Elevation	55 cm
Camera Angle	45° (left & right)

For the DT, we employed the C4.5 algorithm, which not only constructs effective DT but also includes post-pruning techniques to minimize complexity. This optimization simplifies the model structure, leading to improved generalization performance.

In deep learning, various algorithms can handle complex data effectively. In our study, we conducted extensive comparative analysis and successfully reduced the dimensionality of posture features to 13. This reduction allowed us to use a smaller yet accurate model, particularly valuable in scenarios with limited computational resources. Given these conditions, we chose MLP, a classic deep learning architecture well-suited for non-sequential data with intricate inter-feature relationships. In our setup, the MLP has three hidden layers

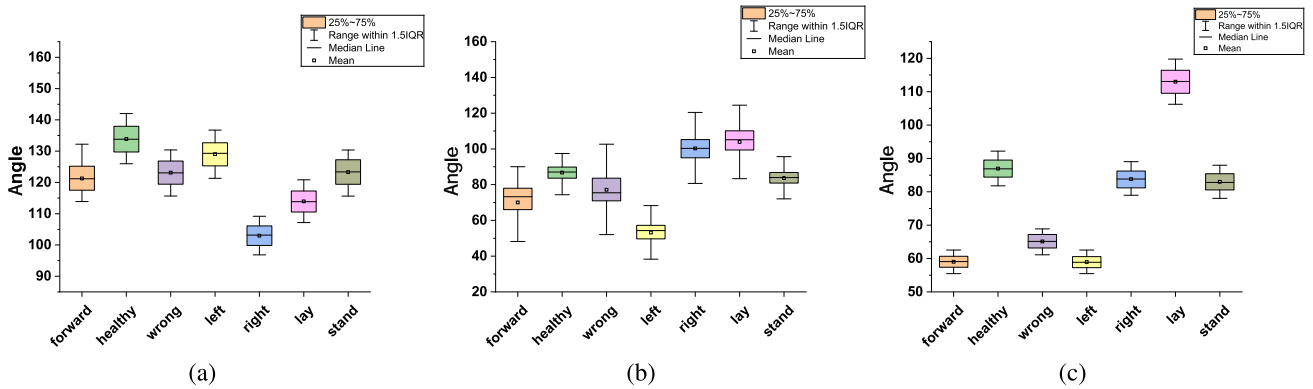


Fig. 6. Illustration the distribution of skeletal angle information across different postures for three methods: (a) points B–C–E (BCE), (b) points J–A–B (JAB), and (c) points K–D–E (KDE).

with configurations of 200, 100, and 25 neurons, respectively. The maximum number of iterations (`max_iter`) was set to 200 to ensure convergence. We employed the ReLU activation function due to its well-documented performance benefits in MLPs. To enhance the learning process, we selected the Adam optimizer, renowned for its adaptability, which automatically manages learning rate adjustments, optimizing our model's training efficiency.

To demonstrate the effectiveness of ensemble learning algorithms, we compared our model with GBDT [22] and TabNet [23] using a soft voting ensemble method. The `GradientBoostingClassifier`, a widely used ensemble learning model, enhances predictive accuracy by training multiple DT incrementally, thus reducing model loss progressively. We set `n_estimators` to 25, striking a balance between preventing overfitting and maintaining strong performance. To ensure reproducibility, `random_state` was fixed at 42.

For tabular data, we applied the `TabNetClassifier`, a deep learning model known for its effective feature selection and classification abilities. The model employs attention-inspired mechanisms, with `n_steps` controlling the number of feature selection iterations, allowing for multiple perspectives on the data. Parameters `n_d` and `n_a` define the embedding sizes for decision and attention steps, set to 16 to balance complexity and speed. A higher `gamma` value, set to 1.5, intensifies the attention mechanism, allowing the model to emphasize important features in each step. Sparse regularization, controlled by `lambda_sparse` (set to 1e-3), mitigates overfitting. The Adam optimizer, with a relatively high learning rate of 0.02, accelerates initial convergence. We used `mask_type='sparsemax'` for feature selection, promoting interpretability and reducing noise.

In particular, our experimental setup and software configurations are shown in Table III. The operating system is Windows 10. The programming language used is Python, version 3.8.12, with PyCharm 2021.2.2 serving as the integrated development environment. The system is powered by an Intel[2] Core[3] i7-10710U CPU, clocked at 1.10 GHz, boosting to 1.61 GHz. An NVIDIA GeForce MX250 GPU handles graphics processing. For deep learning tasks, we utilize the Pytorch framework, version 2.0.1.

TABLE III
EXPERIMENTAL DATA PROCESSING ENVIRONMENT

Component	Specification
GPU	NVIDIA GeForce MX250
CPU	Intel(R) Core(TM) i7-10710U
Python	V3.8.12
Operating system	Windows 10
PyCharm	2021.2.2
Pytorch	V2.0.1

IV. SYSTEM PERFORMANCE ANALYSIS AND DEMONSTRATION

In this section, we compare the classification performance of the base learners SVM, DT, and MLP, as well as the ensemble learning model using soft voting based on these base learners, alongside the comparison algorithms GBDT and TabNet on the posture dataset. We analyze the relationship between each action through the confusion matrix by comparing the degree of misjudgment. Finally, the precision, recall, and F_1 scores are compared across six algorithms.

A. Model Performance Evaluation Indicators

To measure the results, there are many indicators that can be used to evaluate the performance of the model. Precision is the proportion of correctly predicted positives to all positive predictions. Then, there are two possibilities for prediction: one is to predict the positive class as a positive class (TP), and the other is to predict the negative class as a positive class (FP), as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{9}$$

Recall is the proportion of correctly predicted positive classes to all positive ones. There are two possibilities, one is to predict the positive class as a positive class (TP), and the other is to predict the positive class as a negative class (FN), as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{10}$$

[2]Registered trademark.
[3]Trademarked.

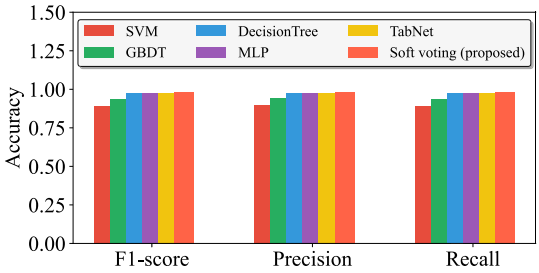


Fig. 7. Recognition rate of different learning methods.

The proportions of the six sitting and standing postures in the dataset we collected are different; that is, the dataset is highly imbalanced. Hence, overall model accuracy is not an appropriate metric for measuring performance. For this, we use the F_1 score to evaluate the performance. The F_1 score takes into account both false positives (FP) and false negatives (FN) [24]. The F_1 score combines precision and recall and is the harmonic mean of the two. The value of F_1 is obtained by

$$F_1 = \sum_i 2 \cdot w_i \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}. \tag{11}$$

Among them, $w_i = (n_i/N)$, which represents the proportion of the ith type of samples in the total samples. n_i is the number of samples of type i, and N is the number of total samples. Precision_i and Recall_i are the precision rate and recall rate of the ith sample.

B. Model Performance Comparison

Next, we compare the precision, recall, and F_1 scores of the six models. As shown in Fig. 7, SVM had the lowest performance, with an F_1 score of 89.1%. GBDT performed well, achieving a score of 93.9%. The F_1 scores for the DT, MLP, and TabNet were all around 97%. The ensemble learning model, which utilized a soft voting mechanism among these base classifiers, achieved the best performance with an F_1 score of 98.1%. Overall, the ensemble learning model demonstrated the most effective classification performance on our sitting dataset.

We use six artificial intelligence algorithms to recognize sitting and standing states: SVM, DT, MLP, GBDT, TabNet, and an ensemble learning approach incorporating a soft voting mechanism that integrates the aforementioned base learners. To compare the accuracy of these six algorithms, the obtained confusion matrix is illustrated in Fig. 8. Across the board, there is a notable trend where several algorithms tend to misclassify "hunched over" and "leaning forward" postures, indicating that similarities in joint angles create challenges in distinguishing these categories. For instance, SVM and DT models displayed higher misclassification rates for these similar postures, though they performed well with distinct categories like "sitting straight" and "standing."

MLP and GBDT showed better performance by capturing more nuanced patterns in the data, with GBDT particularly excelling in reducing errors for complex sitting postures such as "left sitting" and "right sitting." TabNet leveraged its attention mechanisms to improve classification accuracy, effectively handling subtle posture variations. The ensemble learning approach, which integrated all these base learners, significantly reduced misclassification rates. By combining the strengths of each model, the ensemble method delivered the highest overall accuracy, effectively managing posture categories that posed difficulties for individual algorithms.

C. Model Deployment and Real-Time Sitting Posture Recognition

The performance comparison reveals that the ensemble learning model with soft voting achieves the best classification results on this dataset among the four algorithms. Therefore, we have chosen to deploy this model in practice. First, we save the trained model parameters to the actual sitting posture detection system, capture the human bone joints through the Kinect real-time camera, and put the 3-D coordinates of the selected nine joints along with the calculated joint angles as features into the trained classification model, and finally input the classification results to the screen in real-time. Fig. 9 shows the depth skeleton images of various actions acquired by the system, and the implemented Sitpose system operated in Windows operating system is depicted in Fig. 10. Considering the recognition error and time delay, we do not judge all the classification results of the data collected every second but count the data within 1 min. If the abnormal sitting posture accounts for more than 75%, we will judge the sitting posture for this minute. If it is terrible, a pop-up window will be given to remind the users to change their sitting posture to a healthier one.

The sedentary test detects that if the users have been sitting for a cumulative hour, the system will give a pop-up window reminder to remind the users that the users are in a sedentary state and should stand up and move around to improve the sedentary problem. Real-time monitoring is essential for detecting and ameliorating poor posture habits that could culminate in spinal disorders and chronic back pain [25]. Our innovative system proffers instantaneous feedback on the user's posture, significantly diminishing the propensity for poor posture and fostering the adoption of healthier sedentary behaviors.

To assess the latency of the entire system, we conducted latency testing while using commonly used office software. In the process, Azure Kinect data was collected and underwent preprocessing and angle feature extraction, followed by category prediction and total delay time measurement. The results shown in Fig. 11(a) indicate that the time spent on a series of processing steps leading to the final prediction was 97 ms. Furthermore, when running only the posture monitoring system without any other office software, the latency can be reduced to within 30 ms. It demonstrates the real-time capability of the system.

D. Testing of Sitting and Sedentary Systems

Due to variations in the depth data and joint angle information resulting from different camera angles and positions, it is necessary to place the camera on the right side of the desk at a depth of 72 cm. The camera should be positioned at

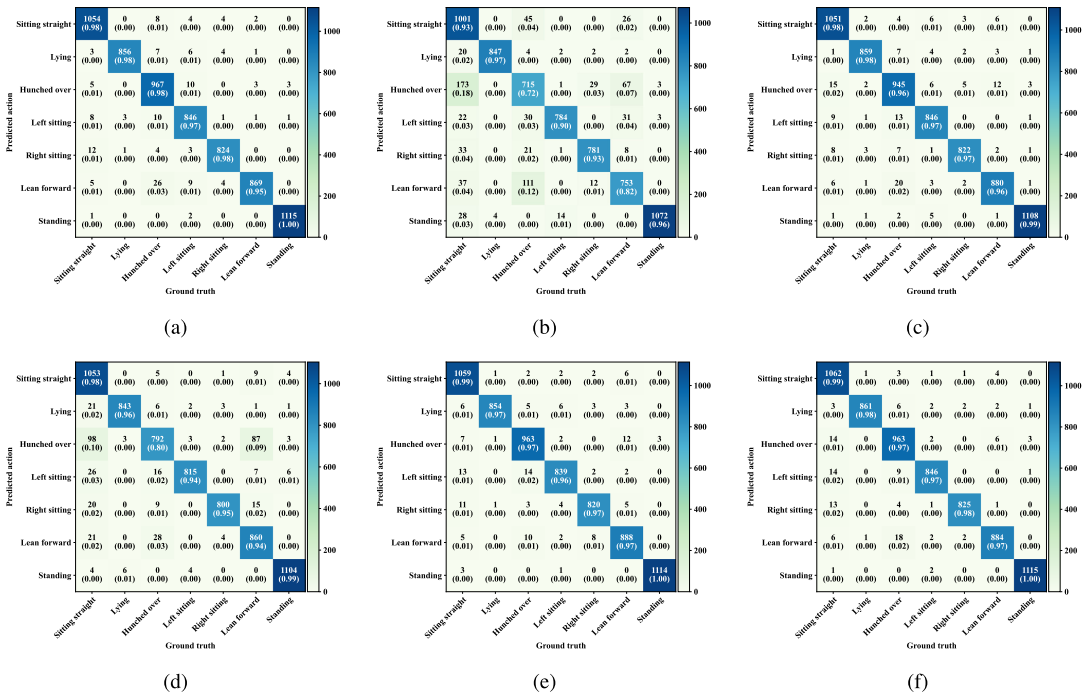


Fig. 8. Confusion matrix of our considered six algorithms. (a) MLP. (b) SVM. (c) DT. (d) GBDT. (e) TabNet. (f) Voting classifier (proposed).

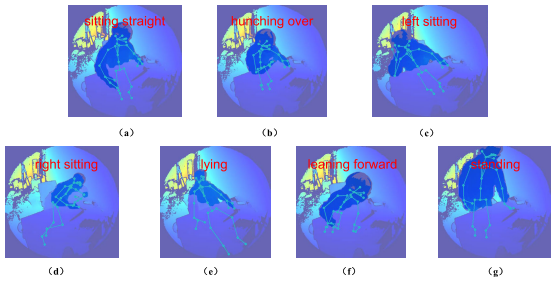


Fig. 9. Skeletal depth picture for sitting position detection with detected results in our system. (a) Sitting straight. (b) Hunching over. (c) Left sitting. (d) Right sitting. (e) Lying. (f) Leaning forward. (g) Standing.

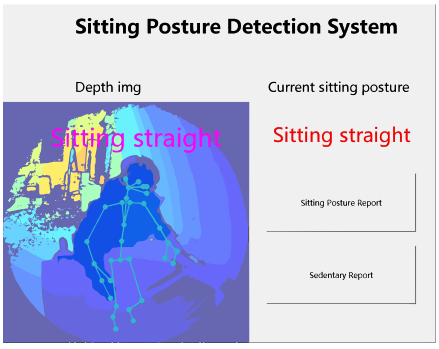


Fig. 10. Implemented sitting posture detection system in Windows operating system.

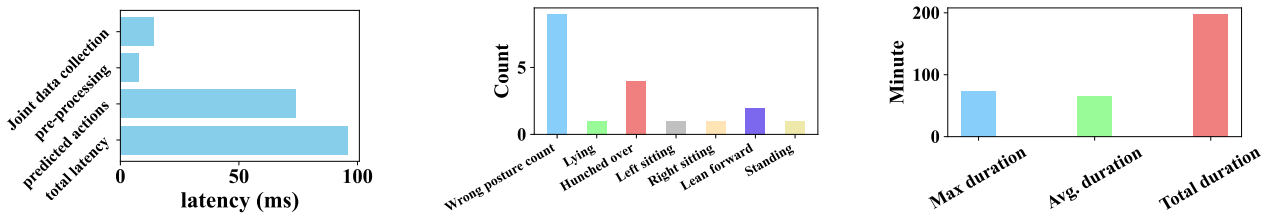
a downward angle of 45°. This setup ensures that the entire desk area falls within the range for posture detection. We also selected two testers to participate in the test of the abnormal sitting posture and sedentary system, and these two testers did not participate in collecting the dataset. With the depth camera and monitoring system turned on, we asked the tester to perform seven actions: sitting straight, hunched over, left sitting, right sitting, leaning forward, lying, and standing.

Each pose was completed 30 times to ensure robust data collection. The quantitative results revealed that when collecting data for a normal sitting posture (sitting straight), there were three instances where it was misjudged as leaning forward and two as hunching over. In a leaning forward sitting posture, there were four instances where it was mistaken for a sitting straight. While in a hunching-over posture, there were three instances of being misjudged as sitting straight and one as leaning forward. No other postures were misidentified.

The system can correctly identify these sitting and standing states in real-time. When the tester sits in an abnormal posture for more than 1 min, the system gives an interactive prompt in a pop-up window to remind the tester to change his current sitting posture. We let the tester sit at the workstation and work normally for the sedentary test. After working continuously for more than 1 h, the system gives interactive reminder information in the form of a pop-up window in the sedentary state so that the tester gets up and moves around to avoid the sedentary state. Tests have proved that our monitoring system can monitor people's sitting posture and efficiently identify the sedentary states in real-time while giving interactive prompts so that people can improve their sitting posture and avoid sedentary states.

E. Implementation of SitPose

Fig. 10 illustrates the user interface of our Sitpose system, which monitors sitting posture and sedentary in a real-time manner. On the left side of the display is a real-time depth image. This depth image showcases a skeletal structure, tracked using Azure Kinect's built-in body tracking feature and rendered using CV2. On the right side of Fig. 10, the system displays the current sitting posture status, monitored in a real-time manner. This system encompasses two key real-time features: monitoring of the sitting posture and sedentary report.



(a) (b) (c)

Fig. 11. Latency and generated reports of our Sitpose system. (a) Latency. (b) Posture. (c) Sedentary.

The posture report includes the wrong posture count and the frequency of five incorrect sitting and standing postures. This report helps users analyze the number and types of incorrect postures they adopted after activating the monitoring system. By examining the frequency of incorrect postures, users can understand which wrong postures they are more prone to, helping them to avoid these postures in the future, as shown in Fig. 11(b).

The sedentary report tallies the total sitting time, average sitting duration, and maximum sitting duration since the Sitpose was activated. It allows users to analyze their daily sedentary patterns. When sedentary behavior is detected, the system alerts the user, suggesting they stand up or move around, shown in Fig. 11(c).

Based on the above, we have developed a visual interface for the system that can display depth images and current sitting postures in real-time. When users exhibit poor sitting postures or sedentary behaviors, the system will prompt a pop-up reminder and record the posture information. This information is then visually represented through bar charts. With the integration of these reporting features, users can gain a comprehensive understanding of their sitting patterns and make informed decisions to improve their health and well-being.

V. Conclusion

This article adopted the Kinect depth camera to implement a sitting posture detection system, namely SitPose. Using the human body tracking SDK, SitPose identifies the coordinates of human bone joints and calculates the angles of related joint points. We then established a dataset containing 3-D joint coordinates of the human body and the angle between the key points, a total of 33 409 pieces. We designed a machine learning-based algorithm to classify the sitting postures and long-term sitting behavior in working scenarios. A comprehensive performance evaluation was conducted to validate the effectiveness of SitPose, demonstrating its accuracy. The ensemble learning model based on the soft voting mechanism achieved an impressive F_1 score of 98.1%, indicating outstanding performance. Finally, we deployed the trained ensemble learning model to implement the real-time SitPose system.

Our study now includes a participant pool of 36 individuals. While this number provides a more substantial dataset, it may still not fully capture the diversity of body types and postures, potentially affecting the generalizability of the findings. In addition, the system‘s evaluation in a controlled environment might not mirror its effectiveness in actual, dynamic workplaces. Our future research will focus on enlarging the participant group to enrich data diversity and conducting tests in varied settings to validate the system‘s effectiveness beyond controlled conditions.

Since sitting posture recognition based on the 3-D coordinates and angles of bones requires the distance, direction, and height of the camera, in the future, we plan to recognize sitting postures at various angles and orientations, comparing the best distance, direction, and angle for sitting posture recognition while also realizing multiple human sitting posture detection.

References

[1] E. Szczygiel, K. Zielonka, S. Metel, and J. Golec, “Musculo-skeletal and pulmonary effects of sitting position–A systematic review,” *Ann. Agricult. Environ. Med.*, vol. 24, pp. 1–7, Dec. 2017.

[2] H. K. Jang, H. Han, and S. W. Yoon, “Comprehensive monitoring of bad head and shoulder postures by wearable magnetic sensors and deep learning,” *IEEE Sensors J.*, vol. 20, no. 22, pp. 13768–13775, Nov. 2020.

[3] L. V. Kallings et al., “Workplace sitting is associated with self-reported general health and back/neck pain: A cross-sectional analysis in 44,978 employees,” *BMC Public Health*, vol. 21, no. 1, pp. 1–9, Dec. 2021.

[4] Z. Cao, C. Xu, P. Zhang, and Y. Wang, “Associations of sedentary time and physical activity with adverse health conditions: Outcome-wide analyses using isotemporal substitution model,” *eClinicalMedicine*, vol. 48, Jun. 2022, Art. no. 101424.

[5] C. Peng, L. Gui, B. Sheng, Z. Guo, and F. Xiao, “RoSeFi: A robust sedentary behavior monitoring system with commodity WiFi devices,” *IEEE Trans. Mobile Comput.*, vol. 23, no. 5, pp. 6470–6489, May 2024.

[6] M. Tölgyessy, M. Dekan, L. Chovanec, and P. Hubinský, “Evaluation of the Azure Kinect and its comparison to Kinect V1 and Kinect V2,” *Sensors*, vol. 21, no. 2, p. 413, Jan. 2021.

[7] M. Tölgyessy, M. Dekan, and L. Chovanec, “Skeleton tracking accuracy and precision evaluation of Kinect v1, Kinect v2, and the Azure Kinect,” *Appl. Sci.*, vol. 11, no. 12, p. 5756, Jun. 2021.

[8] X. Ran, C. Wang, Y. Xiao, X. Gao, Z. Zhu, and B. Chen, “A portable sitting posture monitoring system based on a pressure sensor array and machine learning,” *Sens. Actuators A, Phys.*, vol. 331, Nov. 2021, Art. no. 112900.

[9] R. K. R. Guduru, A. Domeika, M. Dubosiene, and K. Kazlauskiene, “Prediction framework for upper body sedentary working behaviour by using deep learning and machine learning techniques,” *Soft Comput.*, vol. 26, no. 23, pp. 12969–12984, Dec. 2022.

[10] J. Roh, H.-J. Park, K. Lee, J. Hyeong, S. Kim, and B. Lee, “Sitting posture monitoring system based on a low-cost load cell using machine learning,” *Sensors*, vol. 18, no. 2, p. 208, Jan. 2018.

[11] C. Bontrup et al., “Low back pain and its relationship with sitting behaviour among sedentary office workers,” *Appl. Ergonom.*, vol. 81, Nov. 2019, Art. no. 102894.

[12] P. Paliyawan, C. Nukoolkit, and P. Mongkolnam, “Prolonged sitting detection for office workers syndrome prevention using Kinect,” in *Proc. 11th Int. Conf. Elect. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON)*, 2014, pp. 1–6.

[13] V. Castaneda and N. Navab, “Time-of-flight and Kinect imaging,” in *Proc. Kinect Program. Comput. Vis.*, 2011, pp. 1–53.

[14] C. N. G. Yu and Y. Hongchen, “Research on 3D trajectory feature stability based on Kinect bone information,” *J. People’s Public Secur. Univ. China (Science Technology)*, vol. 28, no. 1, pp. 26–35, Mar. 2022.

[15] Z. Liu, "3D skeletal tracking on Azure Kinect–Azure Kinect body tracking SDK," in *Proc. IEEE , Int. Conf. Image Process. (ICIP), Microsoft Ind. Workshop*, Taipei, Taiwan, Sep. 2019. [Online]. Available: https://www.microsoft.com/en-us/research/uploads/prod/2020/01/AKBTSDK.pdf

[16] Microsoft. (2019). *Skeletal Tracking on Azure Kinect*. Accessed: Dec. 28, 2023. [Online]. Available: https://www.microsoft.com/en-us/research/project/skeletal

[17] Microsoft. (2020). *Azure-Kinect-sensor-sdk*. Accessed: Dec. 28, 2023. [Online]. Available: https://github.com/microsoft/Azure-Kinect-Sensor-SDK

[18] S. Jo, S. Song, J. Kim, and C. Song, "Agreement between Azure Kinect and marker-based motion analysis during functional movements: A feasibility study," *Sensors*, vol. 22, no. 24, p. 9819, Dec. 2022.

[19] R. Zemp et al., "Application of machine learning approaches for classifying sitting posture based on force and acceleration sensors," *BioMed Res. Int.*, vol. 2016, no. 1, 2016, Art. no. 5978489. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2016/5978489

[20] E. Grandjean and W. Hünting, "Ergonomics of posture—Review of various problems of standing and sitting posture," *Appl. Ergonom.*, vol. 8, no. 3, pp. 135–140, Sep. 1977. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0003687077900023

[21] Y. Liu, P. Xue, H. Li, and C. Wang, "A review of action recognition using joints based on deep learning," *J. Electron. Inf. Technol.*, vol. 43, no. 6, pp. 1789–1802, 2021.

[22] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Dec. 2017, pp. 1–9. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[23] S. O. Arík and T. Pfister, "TabNet: Attentive interpretable tabular learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 6679–6687. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16826

[24] K. Xia, J. Huang, and H. Wang, "LSTM-CNN architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56855–56866, 2020.

[25] B. El-Sayed, N. Farra, N. Moacdieh, H. Hajj, R. Haidar, and Z. Hajj, "A novel mobile wireless sensing system for realtime monitoring of posture and spine stress," in *Proc. 1st Middle East Conf. Biomed. Eng.*, Feb. 2011, pp. 428–431.

Hang Jin received the B.Eng. degree in computer science from the School of Artificial Intelligence and Big Data, Hefei University, Hefei, China, in 2020, and the M.Eng. degree in electronic information from the School of Computer and Information, Anhui Normal University, Wuhu, China, in 2024.

His research interests include deep learning, wireless sensing, and computer vision.

Xin He (Member, IEEE) received the M.Sc. degree in information science from the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan, in 2013, and the Ph.D. degree from JAIST and the University of Oulu, Oulu, Finland, in 2016.

He is currently an Associate Professor with the School of Computer and Information, Anhui Normal University, Wuhu, China. He is also a Postdoctoral Researcher with the School of Computer Science and Technology, University of Science and Technology of China (USTC), Hefei, China. His research interests include cooperative wireless communications, network information theory, energy harvesting, backscatter communications, wireless sensing, and human-computer interaction.

Dr. He won the Best Paper Runner Up Award on IEEE/ACM IWQoS 2020.

Lingyun Wang is pursuing the B.Eng. degree in computer science, majoring in Internet of Things engineering with the School of Computer and Information, Anhui Normal University, Wuhu, China. She has been recommended for direct Ph.D. admission to the School of Electronic and Information Engineering, Tongji University, Shanghai, China.

Her research interests primarily include wireless sensing, VR applications, and wireless communications.

Yujun Zhu received the B.Sc. degree in computer science and technology from Qingdao University, Qingdao, China, in 2006, the M.Sc. degree in computer application technology from Anhui Polytechnic University, Wuhu, China, in 2009, and the Ph.D. degree in computer software and theory from Tongji University, Shanghai, China, in 2014.

He is currently an Associate Professor at the School of Computer and Information, Anhui Normal University, Wuhu. His research interests include wireless sensor network and the Internet of Things (IoT) applications.

Weiwei Jiang received the B.Eng. degree from the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China, in 2014, the M.Sc. degree from the School of Information Science, Japan Advanced Institute of Science and Technology, Nomi, Japan, in 2016, and the Ph.D. degree from the School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC, Australia, in 2023.

His research interests include wireless sensing, ubiquitous computing, human-computer interaction, and digital fabrication.

Xiaobo Zhou (Senior Member, IEEE) received the B.Sc. degree in electronic information science and technology from the University of Science and Technology of China (USTC), Hefei, China, in 2007, the M.E. degree in computer application technology from the Graduate University of Chinese Academy of Science (GUCAS), Beijing, China, in 2010, and the Ph.D. degree in information science from the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan, in 2013.

From April 2014 to March 2015, he was a Researcher with the Department of Communications Engineering, University of Oulu, Oulu, Finland. Currently, he is a Professor with the School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin, China. His research interests include the Internet of Things, cloud computing, data center networks, vehicular networks, and mobile edge computing.