

Detection of Ergonomic Sitting Postures in Office Environments

Theresia A. Pawitra ^{1,*}, Farida D. Sitania ¹, Aji E. Burhandenny ², Muhammad F. Wijaya ¹,
Anindita Septiarini ³, Hamdani Hamdani ³, and Vauwez S. E. Fareez ⁴

¹ Department of Industrial Engineering, Faculty of Engineering, Mulawarman University, Samarinda, Indonesia

² Department of Electronic Engineering, Faculty of Engineering, Yogyakarta State University, Yogyakarta, Indonesia

³ Department of Informatics, Faculty of Engineering, Mulawarman University, Samarinda, Indonesia

⁴ Department of Artificial Intelligence and Data Science, Institute of Natural Sciences, Sivas Cumhuriyet University, Sivas, Türkiye

Email: triciapawitra@gmail.com (T.A.P.); ida.sitania@gmail.com (F.D.S.); ajiery@gmail.com (A.E.B.);
muh.farhanuddin@gmail.com (M.F.W.); anindita@unmul.ac.id (A.S.); hamdani@unmul.ac.id (H.H.);
vsefareez@gmail.com (V.S.E.F.)

*Corresponding author

Abstract—Prolonged sitting in office environments increases the risk of musculoskeletal disorders. This study develops an automated posture classification system that combines MoveNet Thunder v4 upper-body keypoints with feature engineering and lightweight classifiers: AdaBoost, Multi-Layer Perceptron (MLP), and XGBoost. We curated 8041 multi-view images from 29 participants in controlled settings and expanded variability with synthetic images. Under standard 5-fold cross-validation on combined views, MLP achieved the best overall performance (accuracy 94.59%, precision 96.49%, recall 92.52%, F1-Score 94.46), while XGBoost was close behind (accuracy 94.16%) and attained the highest ROC-AUC (0.986). To assess generalization, we conducted grouped 5-fold cross-subject validation: all pose-based classifiers performed similarly (accuracy ≈ 0.91), with XGBoost showing a modest edge (accuracy 0.921 ± 0.025 ; ROC-AUC 0.973 ± 0.009) and no significant differences among models. A pretrained MobileNetV2 image baseline failed to generalize in this cross-subject setting (accuracy ≈ 0.51 ; AUC ≈ 0.52), indicating degenerate predictions and underscoring the data efficiency of keypoint representations. These results support pose-based methods as accurate and practical for near-real-time ergonomic monitoring, with XGBoost a sensible default and MLP a competitive alternative. Future work will expand real-world data, incorporate temporal modeling from short videos, and explore deeper fine-tuning and transformer backbones under the same cross-subject protocol.

Keywords—ergonomic sitting posture, computer vision, machine learning, MoveNet, musculoskeletal disorder

I. INTRODUCTION

Desk-based tasks dominate most office work environments. This environment forces sedentary behavior, as the average office worker spends approximately 79% of

their time sitting [1]. While sedentary behavior covers all low-level activity, sitting is the most prevalent form of this behavior. Prolonged sitting has been linked to a variety of health problems, such as decreased vision, weight gain, type 2 diabetes, vitamin deficiencies, hypercholesterolemia, muscle and skin changes, cardiovascular conditions, cancer, and musculoskeletal disorders [2–4].

These problems can be solved by taking periodic breaks that comply with ergonomic principles while sitting for long periods [3, 4]. It is also essential to maintain proper posture while sitting, as improper positioning can lead to musculoskeletal disorders [4–6]. Ergonomic integration is essential to reduce the adverse effects of prolonged sitting during work activities [3].

Several body parts need to be measured to accurately determine ergonomic sitting posture. However, the measuring procedure can unnecessarily disturb workers or people in general while working. Therefore, one needs to automatically monitor and classify whether the current sitting posture is ergonomic without a long measuring process, as well as additional human supervision [7].

Several studies have proposed sensor device usage and machine learning to detect and classify sitting postures [2, 8–11]. While such systems were able to provide accurate, real-time feedback to users, they are often cumbersome and require users to wear or use additional equipment.

In recent years, the field of computer vision has advanced significantly, primarily due to the advent of deep-learning techniques [12]. Computer vision has been widely implemented in various industries and workplaces, such as 3D reconstruction from 2D images, robot guidance, part classification, automated inspection [12, 13], automatic attendance management with facial recognition [14–17], traffic accident detection [18], and

sitting posture detection [7, 19]. The use of computer vision eliminates the need for manual measurement and constant human monitoring to achieve an ergonomic sitting posture for workers, as the system is capable of automatically detecting and alerting them to any incorrect postures [7].

This study proposes a method for determining whether a sitting posture is ergonomic based on data from key points of the human body in images. The MoveNet pose estimation model was chosen as a key points extractor for generating features that are used by Machine Learning (ML) algorithms, namely Multi-Layer Perceptron (MLP), AdaBoost, and XGBoost, as it is a compact and less computationally demanding model [19–21]. This research aimed to develop a reliable yet efficient model for classifying ergonomic sitting postures, thereby reducing the risks of musculoskeletal disorders and enhancing workplace health and safety.

II. LITERATURE REVIEW

Work on sitting-posture and ergonomic assessment using computer vision falls into three broad directions: end-to-end video (RGB/RGB-D) models that learn directly from frames or frame sequences, pose-based systems that operate on extracted 2D/3D landmarks, and classical/multimodal approaches that combine vision with other sensors or engineered features.

A. End-to-End Video (RGB/RGB-D) Models

Kulikajevas *et al.* [7] proposed a method for classifying sitting posture across three categories—backward, forward, and straight—using a combination of MobileNetV2 with recurrent layers. The model was trained on 66 video sequences, including augmentation. The model was able to classify videos at 10 frames per second without skeletal data. The accuracy was 91.47%, the sensitivity was 91.85%, and the specificity was 95.95%. End-to-end video models benefit from temporal cues and multimodal signals (e.g., depth) for handling occlusions and achieving richer scene understanding. However, they typically require longer sequences, more annotated video data, and higher compute (often GPU) for training and inference—constraints that complicate lightweight, CPU-only deployment and single-image inference.

B. Pose-Based Method

Another family first extracts body keypoints (2D landmarks) and then applies rules or lightweight classifiers. Ji *et al.* [6] proposed an anomaly-sitting posture detection model that runs on Internet of Things (IoT) devices. It leverages the lightweight pose estimation model, MoveNet Thunder, to extract 17 key body landmarks as well as a shoe position detector as an additional feature to enhance detection. The study employed a 5042 labeled image dataset with three distinct posture categories: normal, crossed leg, and forward head. The features were classified with a neural network model that consists of an embedding layer and several dense layers. The model achieves an overall F1-Score of 97%.

Estrada *et al.* [19] introduced a rule-based classification model for sitting posture. The features were key points on the human body, such as the nose, shoulders, and spine, extracted using the human pose estimation method. There are 7200 captured instances via multi-camera setups. This system achieved 91.5% and 97.05% accuracy on the left and right camera datasets, respectively; rule-based thresholds are interpretable but can be brittle across camera setups and diverse subjects.

OpenPose can also utilize video data for tasks such as maintenance, walking, and running to extract joint angles and key point movements by leveraging another model, such as a Decision Tree, as performed by Lin *et al.* [22] to run ergonomic assessment based on Rapid Entire Body Assessment (REBA), Rapid Upper Limb Assessment (RULA), and Ovako Working Posture Analyzing System (OWAS). The system's result identified a high-risk posture of 10.4% of the working period, which was consistent with the assessment conducted by an ergonomic expert. The random forest model was also tested on the KTH Royal Institute of Technology (KTH) Action Dataset, resulting in the best performance among others, both for long-term and short-term approaches [23]. The model achieved an accuracy of 90.48% at 15 sample rates and a sequence length of 15 subsampled frames. Pose-based pipelines are often more interpretable and lighter-weight at inference time than full video models, but they rely critically on the quality and consistency of landmark extraction.

C. Classical Machine Learning (ML), Boosting, and Sensor-Based Approaches

Works in related motion and intent recognition demonstrate the effectiveness of boosting and tree-based methods on engineered features. Akhter *et al.* [24] have developed an event recognition system using AdaBoost for human activity. The study employed feature representations, including movable body, optical flow, and motion data. The UCF101 and YouTube datasets were used to develop the model, which includes a diverse range of activities, such as cycling, swinging, and walking. The model achieved an accuracy of 75.33% on the UCF101 dataset and 76.66% on the YouTube dataset.

In a further study, Gao *et al.* [25] proposed a method for recognizing human motion intentions using a Bayesian-optimized XGBoost algorithm. The algorithm was designed to analyze lower extremity movements by integrating Electromyography (EMG) and Inertial Measurement Unit (IMU) signals. The study employed data from 10 subjects doing activities such as walking, squatting, and leg extensions. Some preprocessing and feature extraction techniques were used to enhance signal quality. The Bayesian-optimized XGBoost model demonstrated high accuracy across various metrics, with an average precision of 94.42%, a recall of 95.68%, and an F1-Score of 95.33%. These studies show that boosting methods are strong on tabular, engineered feature sets; however, their modalities and problem settings (optical flow, wearables) differ from monocular camera-based sitting posture classification.

Our approach operates on 2D body landmarks (e.g., nose for neck estimation, shoulders, elbows, spine) extracted from single RGB frames using widely adopted pose estimators such as OpenPose [22], MediaPipe [19], and MoveNet [20]. Accurate landmark detection is crucial for distinguishing correct from incorrect sitting postures; our classifiers consume these keypoints, optionally augmented with engineered geometric features [19].

Relative to the above literature, our work targets the pose-based, deployment-centric regime and contributes the following:

- **Deployment focus:** we design a CPU-only pipeline that runs near real-time on single images and extends directly to video streams by applying the same landmark extraction and classifier frame-by-frame. This retains low latency and no-GPU requirements, in contrast to sequence-centric RGB-D or end-to-end video models that leverage temporal/depth cues but demand larger datasets and higher compute.
- **Evaluation rigor:** we adopt grouped, subject-aware cross-validation and report uncertainty (95% credible intervals) and effect sizes (Cohen's *d*) to assess cross-subject generalization.
- **Comparative benchmarking:** we systematically compare multiple lightweight classifiers (MLP, XGBoost, AdaBoost) on the same landmark features and include a Convolutional Neural Network (CNN) end-to-end baseline to quantify the trade-offs.
- **Ablations and view analysis:** we provide ablation studies over data source (real vs synthetic vs combined), engineered feature design (keypoints vs keypoints + engineered) to pinpoint the components that drive performance.

By explicitly contrasting model families, modalities, and practical constraints, and by providing a statistically rigorous, subject-aware evaluation, we position our work as a pragmatic, interpretable, and deployment-oriented alternative to more compute-heavy video approaches while empirically quantifying the benefits of engineered pose features and boosting/MLP classifiers in the sitting-posture domain.

III. MATERIALS AND METHODS

A. Data Collection

1) Participants

The study employed a dataset comprising 29 participants (20 males and 9 females; aged 19–24 years). The inclusion criteria were no history of any musculoskeletal disorders, and all participants had signed informed consent forms.

2) The setup

The data collection process took place in a 6×7 meter well-lit office. A plain gray background was used to eliminate clutter. Participants sat at a steel-framed chair behind a wooden desk. Image data were collected using a

digital camera with a 5472×3648-pixel resolution, an ISO of 250, an aperture of *f*/2.8, and a 1/30 s shutter speed. The camera was positioned 1.67 Meters for the front view and 1.15 m for the side view, measured from the person's body. The height is constant at 1.27 m for all views.

3) Procedures

Each participant was instructed to do 10 poses for both ergonomic and non-ergonomic sitting posture examples with the supervision of an ergonomic expert. Each pose is captured twice, with front and side views, and the table is visible in the frame. The side views were mirrored later to simulate the opposite side.

Indonesia National Standard (SNI) 9011:2021 on Measurement and Evaluation of Ergonomics Hazards in the Workplace was utilized to detect non-ergonomic sitting postures. Based on this standard, a posture regarded as non-ergonomic when it experiences one or more of the following conditions: the neck exceeds 20° of flexion or 5° of extension or it is twisted; the trunk flexed forward more than 20° or extended backward up to 30°; trunk rotation is observed and arm or elbow is not supported. To validate the finding derived from SNI 9011:2021, the result was further reviewed by an ergonomic expert. Descriptions and visual examples of ergonomic and non-ergonomic image samples are presented in Tables I and II, respectively. However, due to limited poses for correctly ergonomic sitting postures, participants were asked to repeat the whole poses specific to ergonomic postures one more time to be captured, resulting in 10 instances.

4) Data augmentation with synthetic images

To improve the model's robustness beyond controlled-environment images, we employed synthetic image data using image generation techniques with mature Stable Diffusion (SD) models. These images served just as original data to train and test the proposed models [26]. Upon generating images, a proper posture needs to be maintained, just as in real images with real supervision, while other mutable factors, such as environment, background, lighting, and the person's outfit, vary. This ensures the detection models are capable of detecting posture in various office environments.

First introduced by Stability AI in 2023, Stable Diffusion XL (SDXL) is a superior SD model compared to the previous versions, such as SD 1.5 and 2.1 [27]. Another open diffusion model introduced by Black Forest Labs is Flux, which offers several model types and purposes available as open weights.1 Dev, Flux.1 Schnell, and Flux.1 Kontext [28]. The code and the weights are openly available on their Hugging Face page. This leads the SD community to use, extend, and fine-tune the models to their liking. The community often uploads the fine-tuned models to their own Hugging Face account or uses AI-generated content social platforms, namely Civitai, Pixai, and Tensor.art.

This study uses three fine-tuned SDXL model, namely GonzaLomo [29], Juggernaut XL [30], and Real Dream [31] as fine-tuned model often gives enhanced performance, finer details, and a lower number of steps (epochs) [32]. These SDXL models can be paired with

Low-Rank Adaptation (LoRA) to further enhance the quality, steer the model towards a specific style, or even generate content with even lower steps. SDXL-Lightning LoRA, developed by ByteDance, allows the SDXL-based

model to have steps as low as 1 step per generation [33]. This is especially important when generating thousands of images at exceptional speed while maintaining quality.

TABLE I. DESCRIPTION AND VISUAL EXAMPLES OF ERGONOMIC SITTING POSTURES





























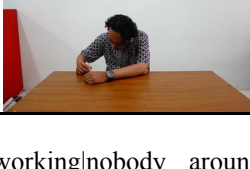

Name	Description	Image Sample	
		Front view	Side view
Upright Position with Neutral or Resting Hand Placement	Sits upright, back straight, shoulders relaxed; hands are either placed comfortably on the desk with elbows at approximately 90 degrees and wrists in neutral position, or resting in the lap. Head and neck aligned, gaze forward, no leaning more than 20° forward or 5° backward.		
Lateral Reaching (Left)	Sits upright, back straight, shoulders relaxed, torso slightly rotated left no more than 20°, left arm reaching laterally (as if to the side of the desk), head and neck aligned, eyes looking in the direction of the reach.		
Lateral Reaching (Right)	Sits upright, back straight, shoulders relaxed, torso slightly rotated right, no more than 20°, right arm reaching laterally (as if to the side of the desk), head and neck aligned, eyes looking in the direction of the reach.		
Diagonal Reaching (Left)	Sits upright, back straight, shoulders relaxed, left hand reaching forward and downward (as if accessing a drawer), right hand near body midline, forward lean from hips no more than 20°, head following hand movement.		
Diagonal Reaching (Right)	Sits upright, back straight, shoulders relaxed, right hand reaching forward and downward (as if accessing a drawer), left hand near body midline, forward lean from hips no more than 20°, head following hand movement.		

TABLE II. DESCRIPTION AND VISUAL EXAMPLES OF NON-ERGONOMIC SITTING POSTURES

Name	Description	Image Sample	
		Front view	Side view
Twisted Slouch (Left)	Sits with torso twisted to the left, back rounded, shoulders hunched, head and gaze downward 20°–45°, arms resting unevenly.		
Twisted Slouch (Right)	Sits with torso twisted to the right, back rounded, shoulders hunched, head and gaze downward 20°–45°, arms resting unevenly.		
Overextended Forward Reach (Left)	Leans far left forward, more than 45°, with the back and neck bent more than 20°, arms fully extended across the desk, shoulders and head dropped towards the table.		

Overextended Forward Reach (Right)	Leans far right forward more than 45° with back and neck bent more than 20°, arms fully extended across the desk, shoulders and head dropped towards the table.		
Rounded Forward Lean	Sits with upper back rounded 20°–45°, shoulders slumped, head bent down more than 20°, arms resting on desk, lacking upright support.		
Forward Slouch with Closed Chest	Sits with head and neck bent forward more than 20°, shoulders rolled in, chest compressed, hands held close together in front of the torso.		
Sideways Bent Slouch (Left)	Sits with torso and head bent to the left side more than 20°, back and shoulders slouched, gaze downward, arms held close to the body.		
Sideways Bent Slouch (Right)	Sits with torso and head bent to the right side more than 20°, back and shoulders slouched, gaze downward, arms held close to the body.		
Asymmetric Forward Lean (Left)	Sits with head and torso leaning forward more than 20° and slightly to the left, shoulders uneven, both forearms resting on the desk, elbows bent, torso rotated, and poor spinal alignment.		
Asymmetric Forward Lean (Right)	Sits with head and torso leaning forward more than 20° and slightly to the right, shoulders uneven, both forearms resting on the desk, elbows bent, torso rotated, poor spinal alignment.		

Another diffusion model used in this study is the Flux.1 Krea [dev] model, an open-weight model from the collaboration of Black Forest Labs and Krea AI that overcomes the oversaturated “AI look” of image generation [34].

Compared to SDXL, the Flux model has 12 billion parameters. The bigger the number of parameters, memory and latency start to become a problem as we scale the generation. A 4-bit quantization technique developed by Li *et al.* [35] allows Flux model speedup for 3.1x. This, combined with Flux.1 Turbo LoRA that makes the image generation as low as eight steps while still maintaining quality makes the image generation even scalable [36].

With these four models, we generate images with the following prompt: “A <view-type> view of a person sitting behind a desk in a {busy|quiet|active|normal} office environment. Person’s desk {empty|full of stuff}, The person sitting with <posture-type> posture. The office has {bright|natural lighting|outdoor|dynamic lighting|well-lit room lighting|little to no light|dark environment}, {large|small} room, {open|closed} space, {people can be

seen working|nobody around}”. The “<view-type>” matched with our view type image data (e.g. front, right, left) and the “<posture-type>” matched with participants posture (ergonomic/non-ergonomic) in our original data. The value between curly brackets is randomized from one generation to the next. For example, “{busy|quiet|active|normal}” randomly generates “busy”, “quiet”, “active”, or “normal”. This way, we have a very diverse image dataset ranging from lighting, room size, open/closed space, and whether there are other people in the background.

To maintain the person’s posture and pose, we employed ControlNet in the image generation workflow. ControlNet, introduced by Zhang *et al.* [37], is a conditional control generation model based on a diffusion model that enables control over the generated image from multimodal input conditions, such as depth maps, human pose, and edges. We employed both pose estimation and depth maps to control the model generation. For strength, we use 0.9 and 0.65 for pose estimation and depth maps, respectively. This ensures the image generation strictly

follows the human pose from the real image and captures the sense of depth information, while still allowing for variations and creativity. Table III illustrates how original images, combined with ControlNet, generate the same

pose for every model. The image generation process was performed using ComfyUI. The generation parameters are shown in Table IV.

TABLE III. IMAGE GENERATION WITH CONTROLNET AND DIFFUSION MODELS

Original Image	Control Input (Pose estimation & Depth map)	GonzaLomo	Juggernaut XL	Real Dream	Flux.1 Krea [dev]
					

TABLE IV. IMAGE GENERATION PARAMETERS

Parameter	Value
Steps	8
Seed strategy	randomize
Cfg	1.0
Sampler	euler
Scheduler	sgm uniform
Openpose controlnet strength	0.9
Depth controlnet strength	0.65
Lora	SDXL-Lightning (for SDXL-based model), Flux-Turbo (for Flux-based model) with strength 1

After the entire image generation process was completed for the four diffusion models, further cleanup was performed. Every generated image was carefully checked one by one to ensure that it accurately represents the posture and does not contain any defects from the generation process, such as a deformed body, incorrect anatomy, or the model failing to generate a human inside the frame. This results in an original and generated combination of 8041 images, with a multi-view image for both ergonomic (4058 images) and non-ergonomic (3983 images) posture datasets.

B. Proposed Sitting Posture Detection Model

The proposed model for detecting ergonomic and non-ergonomic sitting postures is based on the MoveNet pose estimation model. The MoveNet model is used for identifying key points in the human body. The detected key points are utilized as features for classifying sitting postures.

In this study, the key points used for evaluating ergonomic seating posture focused on the upper body, namely the head, neck, trunk, and pelvis. Zhang *et al.* [38] stated keypoints extracted from the upper body (head to pelvis) were sufficient for accurate classification of ergonomic versus non-ergonomic posture.

Cardoso *et al.* [39] further emphasized the importance of trunk alignment by stating that improper spinal

alignment during static sitting significantly increases disc pressure in the lumbar region and contributes to musculoskeletal discomfort. Moreover, Weston *et al.* [40] described that trunk and shoulder posture deviations are the primary indicators of non-ergonomic sitting, focusing on upper body keypoints allows for reliable posture classification without the need for lower limb analysis. In summary, these studies highlighted that an accurate evaluation of ergonomic sitting position can be achieved by focusing on key upper body points. Meanwhile, the position of the legs does not significantly affect posture classification, as long as the feet remain stable on the floor.

The methodology comprises four principal phases: key points extraction, data preparation, classification, and evaluation. Fig. 1 illustrates the comprehensive process of the proposed model. The subsequent subsection provides details on each phase.

C. Feature Extraction

The feature extraction begins with the key points extraction phase, which utilizes the MoveNet pose estimation model. MoveNet was developed by IncludeHealth and Google to identify key points on the human body. Released in 2021, MoveNet offers two versions: Lightning and Thunder, with the latter utilized in this study for its emphasis on accuracy. The model employs a bottom-up approach using TensorFlow's object detection API and MobileNet V2. It simultaneously generates a person-centered heatmap, initializes key points through regression, and filters out background points using distance-weighted multiplication. The final key points are determined by the maximum heatmap values, essential for precise sitting posture analysis. Each key point detected by MoveNet has an x and y coordinate, along with a confidence score ranging from 0.0 to 1.0. MoveNet also supports both single pose and multiple poses [41, 42]. This study only utilizes the former since there is only one person in the frame.

1) Preprocessing

To extract the key points from an image using MoveNet, it is necessary to ensure that the image meets the specified

size requirements. For MoveNet to process the image, it must be resized to a 256×256-pixel size. When the original image is resized to a 1:1 ratio, it can cause the image or the subject within it to appear squeezed. To resolve this issue,

padding is added to the longer side—specifically, the top and bottom—to simulate the square aspect ratio. This process ensures that the longer side of the image is preserved, preventing a squeezed appearance.

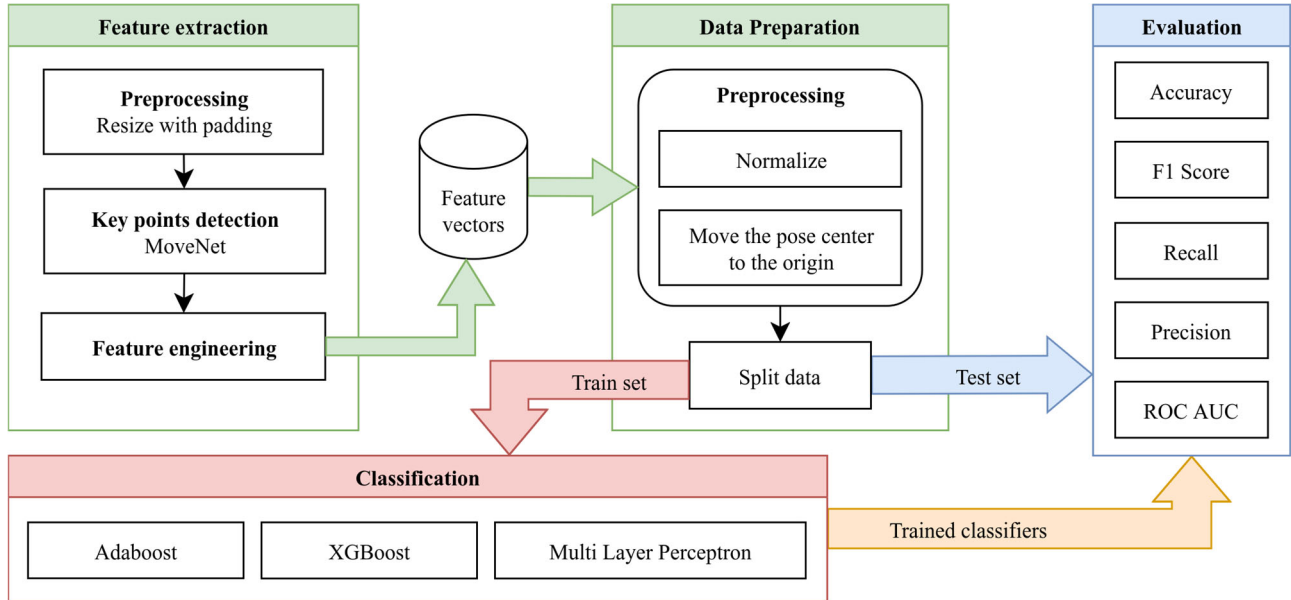


Fig. 1. The process flow of the proposed sitting posture classification model.

2) Key points detection

After resizing and padding the image, it is fed into the MoveNet model, which detects 17 key points on the human body [6, 21, 43] (Fig. 2). However, as pointed out before, this study only considered the key points from head to pelvis (hips), which represent the most critical points for determining sitting posture. This is because these body parts generally provide sufficient information for people who sit behind a desk, as only the top part of the body is visible. This results in only 13 key points to be processed further for feature engineering.

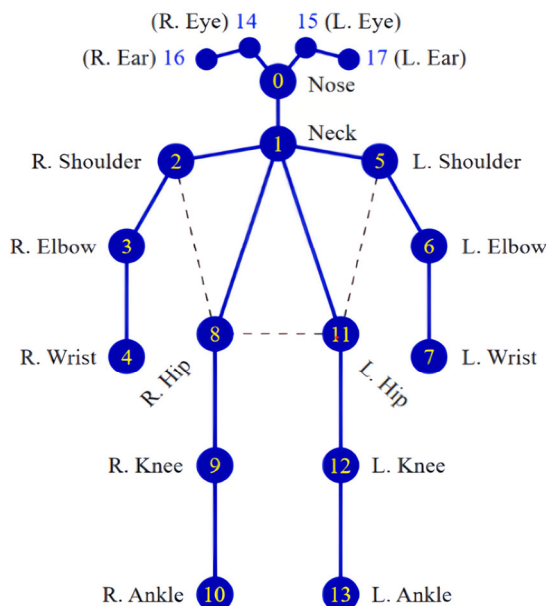


Fig. 2. The 17 key points detected by the MoveNet model.

3) Feature engineering

To enhance feature extraction, we utilized several features described in previous research, leveraging the x and y coordinates of the key points. Following the feature extraction strategy for the side view outlined by Estrada *et al.* [19], we calculated the difference of the y -axis of the three key points that made the spine (i.e., lower, mid, and upper points), the distance of the nose to the right or left shoulder, the distance of the nose to the neck, and all three angles derived from the triangle of shoulder-elbow-wrist, as well as the distance between each other. Furthermore, as described by Lin *et al.* [43], some angle measurements can also be derived from the key points data. These include the angle between the neck and nose relative to the vertical axis, the angles between each shoulder and elbow relative to the horizontal axis, and the angles between each elbow and wrist relative to the vertical axis. These measurements can be obtained for both the left and right arms.

In addition to the features previously mentioned, we also computed the angles of shoulder slope and hip slope relative to the horizontal line, as well as the torso length—the distance between the neck and the midpoint of both hips. When used for front and side views independently, these features offered a unique distinction. For example, an angle between the connection of the neck to the nose, relative to 90°, will represent head tilt in the front view and head lean for the side view. Complete feature and description can be viewed in Table V.

D. Data Preparation

Now that we have a feature vector per image, which consists of the x and y coordinates of 13 key points, 31 hand-engineered features, and the class label, we proceed

to the next stage. The data preparation phase encompasses the preprocessing and splitting of the data.

The key points are normalized so that the largest distance between the key points is equal to one. This

normalization process ensures that all poses are comparable by removing variations due to differences in distance from the camera or body size.

TABLE V. COMPLETE FEATURE SET AND DESCRIPTION

Feature Type	Feature Name	Description
Keypoints	nose	¹ Body keypoints
	eye	
	ear	
	shoulder	
	elbow	
	wrist	
Engineered (angles)	hip	Slope angle between left and right shoulder Slope angle between left and right hip Angle of left wrist from Shoulder-Elbow-Wrist (SEW) triangle Angle of left elbow from SEW triangle Angle of left shoulder from SEW triangle Angle of right wrist from SEW triangle Angle of right elbow from SEW triangle Angle of right shoulder from SEW triangle Angle between neck-to-nose vector and vertical vector. Angle between right elbow-to-shoulder vector and horizontal vector. Angle between right elbow-to-shoulder vector and horizontal vector Angle between right elbow-to-shoulder vector and horizontal vector Angle between right elbow-to-shoulder vector and horizontal vector
	shoulder_angle	
	hip_angle	
	SEWangle_a_left	
	SEWangle_b_left	
	SEWangle_c_left	
	SEWangle_a_right	
	SEWangle_b_right	
	SEWangle_c_right	
	phi1	
	phi2	
	phi3	
	phi4	
	phi5	
Engineered (distance)	torso_length	Distance from shoulder to hips Distance from nose to left shoulder Distance from nose to right shoulder Distance from left shoulder to left elbow Distance from right shoulder to right elbow Distance from left elbow to left wrist Distance from right elbow to right wrist Distance from left shoulder to thoracolumbar (midpoint of the spine) Distance from right shoulder to thoracolumbar Distance from thoracolumbar to middle hip Distance from middle hip to left shoulder Distance from middle hip to right shoulder Distance from nose to middle point between shoulders Difference between thoracic (midpoint between shoulders) and thoracolumbar Difference between thoracolumbar and lumbar (midpoint between hips) Difference between thoracic and lumbar
	nose_to_shoulder_left_distance	
	nose_to_shoulder_right_distance	
	shoulder_to_elbow_left_distance	
	shoulder_to_elbow_right_distance	
	elbow_to_wrist_left_distance	
	elbow_to_wrist_right_distance	
	shoulder_to_mid_left_distance	
	shoulder_to_mid_right_distance	
	mid_to_middle_hip_distance	
	middle_hip_to_shoulder_left_distance	
	middle_hip_to_shoulder_right_distance	
	nose_to_middle_shoulder_distance	
	t_tl_diff_y	
	tl_l_diff_y	
	t_l_diff_y	

¹ Each body keypoint feature has an x and a y point, and each left and right point, except for the nose.

Mathematically, each key points (L_i) are normalized using the Eq. (1):

$$L'_i = \frac{L_i - C}{S} \quad (1)$$

where C is the center point between the left and right hip, and S is the pose size, defined as the maximum of the torso size (shoulders-to-hips distance, scaled by a constant) and the largest distance from the pose center to any landmark. This makes the normalized pose invariant to translation and scale. Meanwhile, the hand-engineered features were scaled using a standard scaler, fit on training folds only and applied to validation/test folds to avoid leakage.

In addition to normalization, the key points are also shifted to the center to standardize poses, regardless of the person's position in the image. This allows the model to focus on the shape and relative positions of the keypoints rather than their absolute distances.

The next stage of the process is splitting the data into training and testing sets. This is done by dividing the data

into an 80:20 ratio, with 80% allocated to the training set and 20% to the testing set [44]. The data in the training set is used to train the model, while that in the testing set is used to evaluate the model's performance.

E. Classification

The classification phase involves training the model using the training data for each view type. The classifiers being used for training are AdaBoost, XGBoost, and MLP. This results in three different models for each classifier and each view type. Each model underwent hyperparameter optimization using the Bayesian method. Bayesian optimization was chosen because, unlike other optimization (search) methods, it considers prior beliefs about the current function and updates them accordingly [45, 46]. With the circumstances of limited data, cross-validation was employed to accompany the optimization algorithm by simulating validation data [47, 48].

AdaBoost shows a prominent result when optimized, while also having efficient resource consumption [45, 49]. AdaBoost is an ensemble learning algorithm that combines

several weak classifiers (n estimators) to create a strong classifier [24, 50]. This “ n estimators” is the hyperparameter to be optimized with the value as in [45].

XGBoost, on the other hand, employs the second-order Taylor expansion of the objective function. It incorporates the function’s second-order derivative to train decision tree models. Furthermore, to enhance the efficiency and robustness of the learning process, XGBoost leverages the complexity of the tree models as a regularization term in the optimization objective [25]. A few hyperparameters that are being optimized, as in [46] can be seen with other classifiers in Table VI.

TABLE VI. HYPERPARAMETERS OF CLASSIFIERS

Classifier	Search Space Hyperparameter	Search Type	Search Space Range
AdaBoost	n estimators	continuous	[10, 2000]
MLP	hidden layer’s neuron	continuous	[128, 256]
	learning rate	discrete	0.01, 0.001, 0.0001
	gamma	continuous	[5.0, 11.0]
	learning rate	continuous	[0.07, 0.6]
	n estimators	discrete	50, 100, 150
XGBoost	regularization alpha	discrete	0.00001, 0.01, 0.75
	regularization lambda	discrete	0.00001, 0.01, 0.45
	min child weight	discrete	1.5, 6, 10
	subsample	discrete	0.6, 0.95
	max depth	discrete	3, 6, 9

Unlike the former algorithms, which generally boost the Tree classifier, MLP is a type of feed-forward artificial neural network that operates as a single model with a single input layer, one or more hidden layers, and an output layer. MLP employs the backpropagation algorithm to adjust weights between neurons and layers based on the difference between expected and actual outputs. This way, the model’s neurons could find the patterns between the input data and the target [51]. The number of neurons and learning parameters are the hyperparameters that are being optimized (Table VI) [52].

F. Evaluation

1) Standard evaluation

The evaluation phase involves assessing the model’s capability to accurately detect ergonomic sitting postures using the testing data. Each of the model’s performances was evaluated using multiple metrics [9], including accuracy, precision, recall, and F1-Score, presented in Eqs. (2)–(5):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1 - Score = 2 \times \frac{Precision + Recall}{Precision \times Recall} \quad (5)$$

where TP , TN , FP , and FN represent true positives, true negatives, false positives, and false negatives, respectively. TP denotes the number of instances that belong to the positive class and are correctly classified as such. TN represents the number of cases that do not belong to the positive class and are correctly classified as the negative class. FP refers to the number of instances that do not belong to the positive class but are incorrectly classified as the positive class. Finally, FN denotes the number of cases that belong to the positive class but are incorrectly classified as the negative class [9].

In addition to the metrics, the model’s performance was also evaluated using the Receiver Operating Characteristic-Area Under the Curve (ROC-AUC). The ROC shows the model’s ability to classify between binary classes as the discrimination threshold is varied. The AUC is a commonly utilized metric that quantifies the area under the ROC curve. A value of 1 marks a perfect classification, whereas a value of 0.5 indicates a random classification performed by the model [9, 43].

2) Validation protocols

We evaluated models under a subject-aware scheme; a grouped 5-fold cross-validation by participant, where each fold holds out disjoint participants to prevent subject-level leakage (Stratified Group K Fold with subject IDs as groups and class balance preserved). For each fold, we compute the metrics defined above and report the mean \pm standard deviation across folds [53].

3) Statistical comparison

In addition to reporting mean \pm standard deviation over the grouped 5-fold splits by participant, we compared classifiers using a Bayesian omnibus test on fold-wise accuracies, followed by pairwise Bayesian comparisons against the best-ranked model. For each model, we report the posterior mean accuracy with 95% credible intervals, Cohen’s d effect size (and magnitude), and posterior decisions based on the probability of being smaller (p_{smaller}) or practically equal (p_{equal}). This subject-aware, Bayesian analysis avoids subject leakage and is appropriate for the small-sample, non-normal fold distributions inherent to grouped cross-validation.

4) Deep learning baseline

We included an image-based baseline by fine-tuning MobileNetV2, which was initialized from ImageNet. The input images were resized to 224×224 as per the MobileNetV2 requirements and further preprocessed using the model’s standard preprocessing; the classification head was replaced with a single sigmoid unit. The head was trained for five epochs while the backbone remained frozen with the Adam optimizer and a learning rate value of 0.001. In the second phase, the backbone was unfrozen, and then the whole model was retrained with a very small learning rate (i.e., 0.00001). Basic data augmentation was applied, such as random rotation, random zoom, random contrast, random brightness, and random horizontal flip. This baseline was evaluated using the same grouped 5-fold cross-validation by participant protocol and analyzed with the same statistical procedures described above.

5) Ablation studies

We conducted a subject-aware ablation using the same grouped 5-fold cross-validation by participant and the same Bayesian analysis as in the main comparison. The ablation matrix crossed data source with feature design: real-only, synthetic-only, and real + synthetic datasets, each evaluated with keypoints-only and keypoints augmented with engineered angles/ratios. For each condition, we computed fold-wise accuracies and summarized posterior mean \pm Standard Deviation (STD); an omnibus Bayesian test and pairwise Bayesian comparisons were run to assess whether any configuration clearly dominated the others.

IV. RESULTS AND DISCUSSION

The experiments were run on a CPU-only machine provided by Lightning.ai with the following specifications:

- Processor: AMD EPYC 7B13 (2.25 GHz, 4 cores);
- Memory: 16 GB.

For AdaBoost and MLP, the Scikit-learn v1.7.0 library was utilized, whereas for the XGBoost classifier, the XGBoost v3.0.2 library was employed. For Bayesian hyperparameter optimization Scikit-optimize v0.8.1 was used. The MoveNet model version 4, obtained from TensorFlow Hub, is used to generate key points. Fig. 3 displays the key point extraction process (note that the image size in the illustration is not to scale).

After performing hyperparameter optimization using a subset of training data derived from cross-validation for each model, each model was trained one more time using the whole training data with the best hyperparameters. Table VII shows the best hyperparameter results for each classifier.

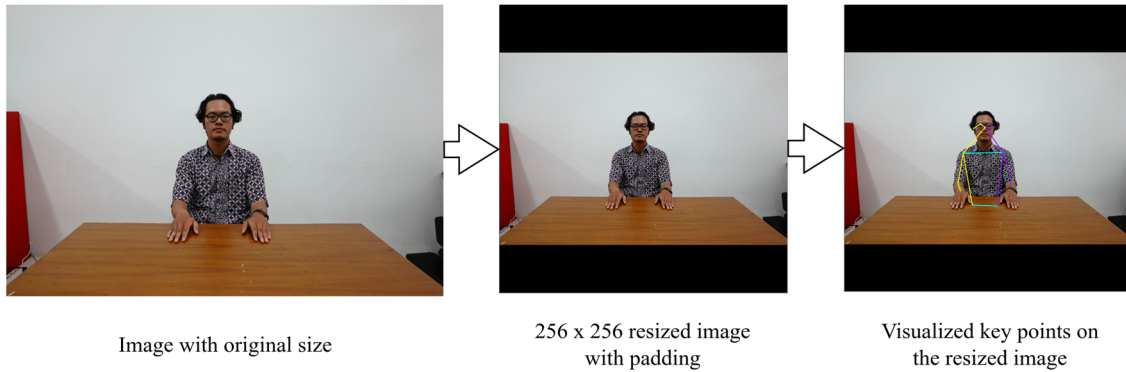


Fig. 3. The process of extracting key points from an image using MoveNet.

TABLE VII. SUMMARY OF CLASSIFIERS' HYPERPARAMETER SEARCH RESULT (BEST TRIALS)

Classifier	Hyperparameter	Values
AdaBoost	n estimators	1999
MLP	hidden layer's neuron	285
	learning rate	0.01
	gamma	5.0
	learning rate	0.07
XGBoost	n estimators	150
	regularization alpha	0.00001
	regularization lambda	0.00001
	min child weight	1.5
	subsample	0.6
	max depth	9

A. Standard Evaluation

The data were split into training and testing sets for each view, resulting in 6432 instances being in the training sets and the remaining 1609 instances becoming the testing set. The performance of each classifier was evaluated based on a range of metrics, including accuracy (acc), precision (prec), recall (rec), F1-Score (F1), and ROC-AUC. Table VIII presents the results of each model's performance for different classifiers.

The results reveal that the MLP classifier attained the highest accuracy, precision, recall, and F1-Score, with values of 94.59%, 96.49%, 95.52%, and 94.46%, respectively. The XGBoost is closely catching up with

94.16% accuracy, while AdaBoost has the lowest performance across all metrics.

TABLE VIII. PERFORMANCE METRICS (%) OF THE PROPOSED SITTING POSTURE DETECTION MODEL

Classifier	Acc	Prec	Rec	F1
AdaBoost	92.91	94.56	91.02	92.76
MLP	94.59	96.49	92.52	94.46
XGBoost	94.16	95.04	93.14	94.08

The confusion matrix of each view of the classifiers presented in Table IX, where ergonomic class is denoted by *E* and non-ergonomic by *NE*. On each confusion matrix, the total amount of instances in a column represents the total number of predicted instances of that class to which the column belongs. Meanwhile, each row represents the number of instances of that class to which the row belongs [25]. Across the test set, all three classifiers achieved high concordance with the reference labels. However, the error rate differs and is suited for different use cases. AdaBoost correctly identified 765 ergonomic and 730 non-ergonomic postures, with 42 ergonomic instances misclassified as non-ergonomic (i.e., unnecessary alerts) and 72 non-ergonomic instances mislabeled as ergonomic (i.e., potentially risky approvals). The MLP yielded the smallest overall error, correctly recognizing 780 ergonomic and 742 non-ergonomic cases, and reducing both types of mistakes relative to AdaBoost,

with 27 ergonomic cases incorrectly flagged as non-ergonomic and 60 non-ergonomic cases incorrectly flagged as ergonomic. As mentioned earlier, the model's performance, as measured by XGBoost, occupied a middle ground in total correct classification (768 ergonomic, 747 non-ergonomic), but had the fewest risky approvals, mislabeling only 55 non-ergonomic postures as ergonomic, at the cost of slightly more missed ergonomic cases than the MLP (39 versus 27).

For a direct visual comparison, model-view ROC curves are presented in Fig. 4. The ROC curve traces out performance at every possible decision threshold, allowing us to choose the optimal point for a crisp ergonomic decision, and any value below that can be considered non-ergonomic. A higher AUC value indicates better performance, as all classifiers achieved this goal. XGBoost traces the upper envelope across most false-positive rates, achieving the highest AUC (0.986), closely followed by

the neural network (0.983) and AdaBoost (0.976). In the low-false-positive region—most relevant when the cost of approving a non-ergonomic posture is high—XGBoost and the neural network attain higher true-positive rates than AdaBoost, with XGBoost exhibiting a slight advantage. At moderate false-positive rates, the curves converge nearly to the top of the plot, suggesting that all models can achieve very high sensitivity under less stringent operating conditions. These patterns are consistent with the previously discussed confusion-matrix profiles: XGBoost affords the most conservative approvals (fewer non-ergonomic instances are labeled as ergonomic), while the neural network offers slightly greater sensitivity to ergonomic cases and can be further extended to identify intermediate postures via threshold adjustment.

TABLE IX. CONFUSION MATRIX OF EACH CLASSIFIER

AdaBoost			MLP			XGBoost		
E (True label)	765	42	E (True label)	780	27	E (True label)	768	39
NE (True label)	72	730	NE (True label)	60	742	NE (True label)	55	747
E (Predicted label)			E (Predicted label)			E (Predicted label)		
NE (Predicted label)			NE (Predicted label)			NE (Predicted label)		

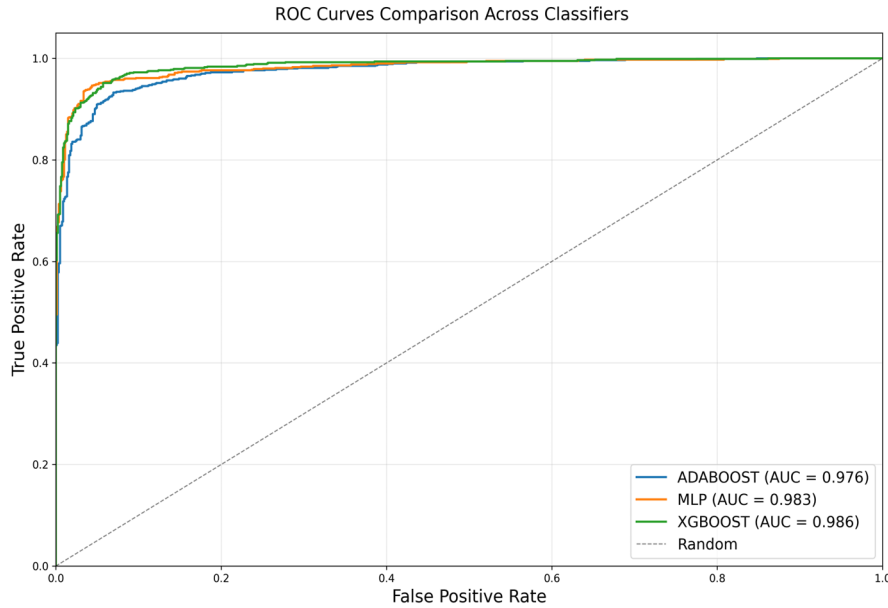


Fig. 4. ROC curves for the proposed sitting posture detection model.

B. Validation Protocols

We evaluated all models using a grouped 5-fold cross-validation by participant (Stratified Group K Fold) and report mean \pm STD across folds.

Overall, the three classifiers produced similar performance: XGBoost achieved the highest mean accuracy ($92.16\% \pm 2.25\%$) and best ROC-AUC ($97.30\% \pm 0.88\%$), while the MLP closely matched XGBoost with mean accuracy $91.96\% \pm 2.21\%$ and ROC-AUC $97.13\% \pm 1.06\%$. AdaBoost yielded a slightly

lower mean accuracy ($91.56\% \pm 2.51\%$) and ROC-AUC $97.00\% \pm 0.91\%$. Precision and F1 are comparable across models (all $\approx 93\%$ and $\approx 91\text{--}92\%$, respectively), but recall shows the largest fold-to-fold variability (std $\approx 4\%$), indicating some subject-dependent sensitivity. Table X provides a comprehensive result of each fold for each classifier.

C. Statistical Comparison

For statistical analysis, we employed auto rank v1.3.0. Under the subject-aware grouped 5-fold protocol, all models clustered around 92% mean accuracy, with

overlapping 95% credible intervals. XGBoost attained the highest mean accuracy (≈ 0.922), serving as the reference model for all statistical comparisons. However, the Bayesian comparison against the MLP was inconclusive (negligible effect size, p_{equal} and p_{smaller} not decisive), supporting practical equivalence between these two. AdaBoost trailed slightly (≈ 0.916) and was judged smaller than XGBoost with a small effect (Cohen's $d \approx 0.23$), indicating a modest but consistent gap. Taken together, these results show that performance differences across models are minor at the subject level; thus, model selection can be guided by deployment considerations (e.g., latency, resource usage) rather than accuracy alone. Table XI summarizes the statistical analysis along with the results from CNN models as deep learning baseline.

TABLE X. GROUPED 5-FOLD CROSS-VALIDATION RESULTS

Classifier	Fold	Acc	Prec	Rec	F1
AdaBoost	¹ 1	0.905	0.930	0.893	0.911
	² 2	0.959	0.944	0.969	0.956
	³ 3	0.895	0.948	0.835	0.888
	⁴ 4	0.891	0.883	0.896	0.889
	⁵ 5	0.928	0.952	0.896	0.923
MLP	1	0.898	0.938	0.870	0.903
	2	0.949	0.935	0.957	0.946
	3	0.901	0.952	0.844	0.895
	4	0.906	0.890	0.921	0.905
	5	0.944	0.965	0.916	0.940
XGBoost	1	0.906	0.918	0.908	0.913
	2	0.956	0.938	0.969	0.953
	3	0.900	0.951	0.842	0.893
	4	0.905	0.893	0.916	0.904
	5	0.941	0.961	0.915	0.937

¹ Training subject id fold 1: 1, 3, 4, 5, 6, 8, 9, 10, 11, 13, 14, 15, 16, 18, 19, 20, 21, 23, 24, 25, 26, 27, 29; test subject id fold 1: 2, 7, 12, 17, 22, 28.

² Training subject id fold 2: 1, 2, 4, 5, 6, 7, 9, 10, 11, 12, 14, 15, 16, 17, 19, 20, 21, 22, 24, 25, 27, 28, 29; test subject id fold 2: 3, 8, 13, 18, 23, 26.

³ Training subject id fold 3: 1, 2, 3, 5, 6, 7, 8, 10, 11, 12, 13, 15, 16, 17, 18, 20, 21, 22, 23, 25, 26, 27, 28; test subject id fold 3: 4, 9, 14, 19, 24, 29.

⁴ Training subject id fold 4: 2, 3, 4, 5, 7, 8, 9, 10, 12, 13, 14, 15, 17, 18, 19, 20, 22, 23, 24, 25, 26, 28, 29; test subject id fold 4: 1, 6, 11, 16, 21, 27.

⁵ Training subject id fold 5: 1, 2, 3, 4, 6, 7, 8, 9, 11, 12, 13, 14, 16, 17, 18, 19, 21, 22, 23, 24, 26, 27, 28, 29; test subject id fold 5: 5, 10, 15, 120, 25.

TABLE XI. SUBJECT-AWARE 5-FOLD BAYESIAN COMPARISON (ACCURACY)

Classifier	Mean \pm STD	95% credible interval	Cohen's d vs top model (magnitude)	Posterior decision (p_{smaller} , p_{equal})
XGBoost	0.921 \pm 0.025	0.860–0.982	0.0	Reference for comparisons
MLP	0.919 \pm 0.024	0.860–0.979	0.079 (negligible)	inconclusive (0.618, 0.002)
AdaBoost	0.916 \pm 0.028	0.847–0.983	0.225 (small)	smaller (0.967, 0.001)
CNN (MobileNetV2)	0.507 \pm 0.031	0.419–0.595	14.728 (large)	smaller (0.999, 0.001)

D. Deep Learning Baseline

Under the subject-aware grouped 5-fold protocol, the CNN baseline performed near chance (mean accuracy 0.507 ± 0.031 ; 95% credible interval 0.419 – 0.595), while all keypoint-based models were clustered around 0.92 mean accuracy with tight credible intervals (XGBoost 0.922, MLP 0.920, AdaBoost 0.916). Bayesian pairwise comparisons were decisive: the posterior probability that each keypoint-based model outperforms the CNN exceeded 0.999, with the CNN judged smaller in every case. Among the keypoint-based models, XGBoost had the highest mean accuracy, but its advantage over the MLP was negligible (posterior decision inconclusive), and the gap to AdaBoost was small. These results show that, in this dataset, learning from landmarks and engineered features is markedly more effective and data-efficient than end-to-end image learning, and that XGBoost and MLP are practically interchangeable in accuracy, allowing deployment decisions to be guided by runtime and resource considerations.

TABLE XII. ABLATION STUDY—ACCURACY (GROUPED 5-FOLD BY PARTICIPANT) FOR DATASET \times FEATURE CONFIGURATIONS

Condition (Dataset \times Features)	Mean \pm STD	Median	Fold Accuracies (5 Folds)	Bayesian Comparison vs Real + Synthetic
Real-only \times Keypoints-only	0.913 \pm 0.015	0.915	0.9153, 0.9167, 0.9341, 0.9105, 0.8878	-
Real-only \times All features (engineered)	0.913 \pm 0.022	0.911	0.9377, 0.9112, 0.9377, 0.8946, 0.8837	-
Synthetic-only \times Keypoints-only	0.923 \pm 0.020	0.911	0.9107, 0.9514, 0.9044, 0.9067, 0.9439	vs Real+Synthetic \times Keypoints-only: $P_{\text{left}} = 0.75948$ (inconclusive); vs Real+Synthetic \times All features: $P_{\text{left}} = 0.85952$ (inconclusive)
Synthetic-only \times All features	0.929 \pm 0.021	0.919	0.9134, 0.9541, 0.9185, 0.9050, 0.9564	vs Real+Synthetic \times Keypoints-only: $P_{\text{left}} = 0.89922$ (inconclusive); vs Real+Synthetic \times All features: $P_{\text{left}} = 0.91260$ (inconclusive)
Real+Synthetic \times Keypoints-only	0.919 \pm 0.023	0.916	0.9163, 0.9565, 0.9025, 0.8901, 0.9305	-
Real+Synthetic \times All features	0.922 \pm 0.022	0.906	0.9061, 0.9559, 0.8998, 0.9054, 0.9408	Reference for comparison

Mean \pm SD, median, and per-fold accuracies for each combination of data source (real, synthetic, combined) and feature set (keypoints only, keypoints + engineered features); Bayesian pairwise comparisons are shown against the Real+Synthetic (all features) reference.

E. Ablation Study

Across conditions, XGBoost, being the best model statistically, was used; synthetic-only training produced

the highest mean accuracies, with synthetic + engineered features performing best overall (0.929 ± 0.021), followed by synthetic keypoints-only (0.923 ± 0.020). Real + synthetic was close behind (0.922 ± 0.022 with engineered

features; 0.919 ± 0.023 keypoints-only), while real-only configurations were modestly lower ($\approx 0.913 \pm 0.015 - 0.022$). Adding engineered features yielded small, consistent gains within each data source (synthetic: $+0.006$; real + synthetic: $+0.002$; real-only: ≈ 0.0001). Bayesian pairwise tests among the top settings (synthetic vs real + synthetic) were proven inconclusive, indicating overlapping uncertainty rather than clear superiority. Despite that, the trend favors synthetic-only data with engineered features. Practically, these results show that synthetic data is a strong supervisory signal for posture classification; when real data is limited, synthetic-only models can match or slightly exceed mixed-data training, and engineered features provide a slight but reliable boost. Table XII summarizes the statistical result from the

ablation study.

F. Web-Based Prototype

We also developed a web-based inference prototype to simulate a deployment scenario. The web interface utilizes the camera connected to the device (for example, a webcam). The resolution of the webcam does not matter, as the prototype preprocesses video feed frames as images in the same manner as we did during the data preparation step. The interface will indicate whether the current sitting position is ergonomic, accompanied by a confidence score, processing time, a small posture analysis history graph, and additional information. The web interface is shown in Fig. 5.

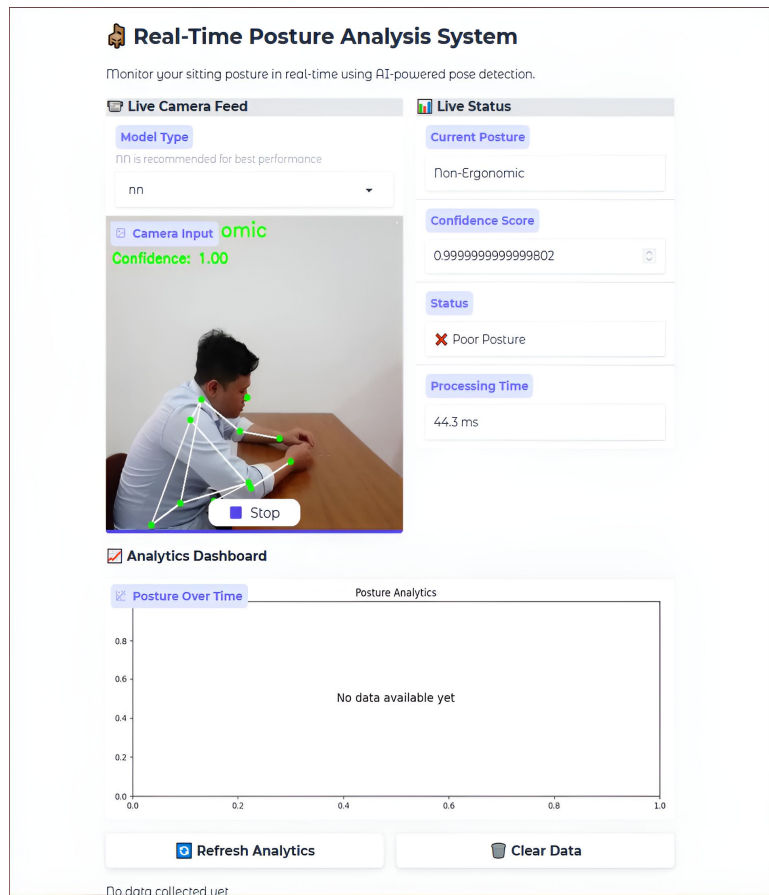


Fig. 5. Posture analysis web UI prototype.

G. Comparison with Previous Study

A comparison of the proposed model with other studies is presented in Table XIII. The proposed model, which implements the XGBoost algorithm, demonstrated an accuracy of 94.16% with grouped cross-validated by participant mean accuracy of $92.16\% \pm 2.25\%$ in classifying ergonomic and non-ergonomic sitting postures. The proposed model showed superior performance compared to the Deep Recurrent Hierarchical Network [7], which achieved an accuracy of 91.47% in spine posture recognition while sitting. It also outperforms the MediaPipe + Decision Tree model [19], which achieved accuracies of 91.5% and 97.05% in the recognition of

correct and incorrect sitting postures from the left and right, respectively. Although the proposed model exhibited only a slight advancement over the studies, it is crucial to acknowledge that the proposed model employs image data from three distinct angles—front, left, and right. This highlights the effectiveness and generalized ability of the proposed model in classifying ergonomic and non-ergonomic sitting postures.

In terms of computational requirements and inference performance, the proposed model shows a favorable distinction. For example, while the Deep Recurrent Hierarchical Network [7] achieved real-time processing capability at 94 milliseconds (ms) per frame on a GPU-equipped system (Intel i7-4790 with NVIDIA 1070), the

proposed model maintained high accuracy while being trained and tested on a CPU-only environment for only 40.2 ms per frame on average, this allows for periodic or near-real-time assessment in an expected low-resource environment where GPU access is expensive or impractical. Other methods, such as OpenPose + Person Detection (PD) + Deep Neural Network (DNN) [43], reported significantly higher processing time (~335 ms/frame) on average, even with a high-end GPU (NVIDIA Titan V). However, this is justified since this was applied to more complex multi-person detection settings.

Furthermore, several prior studies did not explicitly report execution times or ran exclusively on GPU-accelerated machines, making it challenging to directly compare their efficiency. However, the training time of the proposed XGBoost classifier was significantly shorter than that of models using complex deep learning backbones, around 8 seconds. To summarize, the proposed model not only offers a compelling balance of accuracy but also generalizability and computational feasibility. This makes it suitable for potential integration into a system with limited computing infrastructure to detect and reduce musculoskeletal disorders in the office environment.

TABLE XIII. COMPARISON OF THE PROPOSED MODEL WITH OTHER STUDIES

Method	Task	Machine Spec	Dataset/People Density	Number of Training Frames/Images	Training Time (s)	Number of People in Frame	Processing Time (ms/frame)	Accuracy (%)
OpenPose + PD + DNN [43]	Student Behavior Recognition	Intel Core i7 (3.70GHz); 32 GB RAM; NVIDIA Titan V GPU	11,500 images belonging to asking, boring, bowing, and looking classes in classroom settings	8050 (70% of total)	-	9 (average)	335	-
MoveNet + Shoe Detector + Artificial Neural Network [6]	Anomaly Sitting Posture Detection	Training: Intel Xeon E5-2660 v3 (2.60GHz, 16 cores); 64 GB RAM; Titan RTX GPU Inference: Intel Core i7 (2.6Ghz, 6 cores); 16 GB RAM; Intel UHD Graphics 630	5042 of 3 distinct postures—normal, crossed leg, and forward head postures	3529 (70% of total)	-	1	-	-
MLP + Hyperparameter Optimization (HPO) [52]	Human Posture Detection	Intel Core i7-6600U (2.60 Ghz); 8 GB RAM	22,000 images of the MPII human posture dataset	16,496 (75% of total)	420	2 (average)	-	89.9
Deep recurrent hierarchical network [7]	Spine posture recognition while sitting	Intel i7-4790; 16 GB RAM; NVIDIA 1070 8 GB GDDR5 VRAM	17,596 images of a person sitting behind a desk in an office environment	15,747	-	1	94	91.47
MediaPipe + Decision Tree [19]	Recognition of correct and incorrect sitting postures	-	7200 images of a side view of a person using a computer	5040 (70% of total)	-	1	-	91.5 and 97.05
MoveNet + XGBoost (Proposed method)	Ergonomic sitting posture classification	AMD EPYC 7B13 (2.25 GHz, 4 cores); 16 GB RAM	8041 images of a front and side view of a person sitting in an office environment	6432 (80% of the total images per view)	8.42	1	40.2	94.16

H. Limitation

Despite promising results, our study has several limitations. The dataset comprises 29 participants recorded under controlled conditions; however, we expanded to 8041 images with a high-diversity synthetic independent external test set. The pipeline uses monocular RGB and primarily upper-body keypoints [38–40]; performance varies by view and may degrade under occlusion, and does not incorporate lower-limb cues. Accuracy depends on the pose-estimator quality; failure cases occur under poor lighting. The posture taxonomy is limited, and class imbalance may influence metrics. Statistically, we use grouped 5-fold cross-validation with modest samples per fold and no external validation; thus, slight differences among top models remain uncertain despite reporting credible intervals and effect sizes. Finally, latency and memory were measured on a desktop-class CPU;

embedded edge devices and privacy aspects (e.g., on-device processing, storage policies) were not evaluated.

I. Future Work

To address these limitations, future work will include collecting and testing on external, in-the-wild datasets. The reduction of view sensitivity can be achieved through multi-view fusion and the incorporation of lower-limb cues. Lightweight temporal models (e.g., sequence smoothing with tracking or Spatio-Temporal Graph Convolutional Network (ST-GCN)/transformer backbones) can be incorporated under the same subject-aware protocol. The adoption of uncertainty-aware inference and domain adaptation techniques has the potential to mitigate the discrepancy between synthetic and real data. The annotation process will be enhanced by the implementation of clearer taxonomies and the establishment of inter-rater agreement. The assessment of

fairness across demographics could be furthered. Consequently, the final phase of the study will entail the profiling of end-to-end performance and privacy metrics on a selection of representative edge devices.

V. CONCLUSION

This study successfully developed and evaluated a keypoint-based pipeline for classifying ergonomic versus non-ergonomic sitting postures using movenet Thunder v4 and three lightweight classifiers (AdaBoost, XGBoost, MLP). Under standard 5-fold cross-validation, MLP achieved the highest overall performance (accuracy = 94.59%, precision = 96.49%, recall = 92.52%, F1 = 94.46), with XGBoost close behind (accuracy = 94.16%) and the highest AUC (0.986). Crucially, grouped 5-fold cross-subject validation showed that all pose-based classifiers generalize similarly to unseen participants (accuracy \approx 0.91), with XGBoost exhibiting a modest edge in average rank and mean AUC (accuracy = 0.9145 ± 0.0269 ; AUC = 0.9713 ± 0.0168). An image-only deep baseline (MobileNetV2) failed to generalize in the same cross-subject setting (accuracy \approx 0.49; AUC \approx 0.50), underscoring the data efficiency and robustness of pose-keypoint representations for this task. Overall, we recommend XGBoost as a practical default, given its slightly stronger cross-subject performance, while noting that MLP and AdaBoost remain competitive and efficient alternatives.

This work serves as a preliminary step toward real-time or periodic monitoring systems that detect non-ergonomic postures and provide timely feedback to users engaged in prolonged sedentary activity (e.g., students, administrative staff), thereby helping to mitigate musculoskeletal discomfort and its impact on productivity.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

AUTHOR CONTRIBUTIONS

T. A. P. and A. S. designed the methodology and gave extensive supervision throughout the study. M. H. W. focused on data collection and validation. V. S. E. F. curated and applied validation to the data. F. D. S. was responsible for conducting data analysis. H. H. and A. E. B. experimented and implemented the model. All authors involved in interpreting the results and critically reviewing the manuscript have approved the final version for submission.

FUNDING

The Faculty of Engineering at Mulawarman University, Samarinda, Indonesia, funded this work in 2024 (8995/UN17.9/PT.00.03/2024).

REFERENCES

- [1] H. Daneshmandi, A. Choobineh, H. Ghaem, and M. Karimi, "Adverse effects of prolonged sitting behavior on the general health of office workers," *J. Lifestyle Med.*, vol. 7, no. 2, pp. 69–75, 2017. doi: 10.15280/jlm.2017.7.2.69
- [2] F. Feradov, V. Markova, and T. Ganchev, "Automated detection of improper sitting postures in computer users based on motion capture sensors," *Computers*, vol. 11, no. 7, 2022. doi: 10.3390/computers11070116
- [3] C. Soares, S. G. N. Shimano, P. R. Marcacine *et al.*, "Ergonomic interventions for work in a sitting position: An integrative review," *Rev. Bras. Med. Trab.*, vol. 21, no. 1, pp. 01–10, 2023. doi: 10.47626/1679-4435-2023-770
- [4] N. Vincent, M. N. C. K. P. G. D. and S. E. O., "Prolong sitting: A metabolic health risk among white-collar workers: A review article," *IJRASET*, vol. 11, no. 7, pp. 1932–1938, 2023. doi: 10.22214/ijraset.2023.55005
- [5] Y. Zhao, C. Sun, X. Xu, and J. Chen, "RIC-Net: A plant disease classification model based on the fusion of inception and residual structure and embedded attention mechanism," *Computers and Electronics in Agriculture*, vol. 193, 106644, 2022. doi: 10.1016/j.compag.2021.106644
- [6] H. Ji, J. Xie, and P. Sun, "A lightweight IoT device-friendly anomaly sitting posture detector for protecting adolescent bone development," in *Proc. 2023 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCoM) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, 2023, pp. 233–238. doi: 10.1109/iThings-GreenCom-CPSCoM-SmartData-Cybermatics60724.2023.00059
- [7] A. Kulikajevs, R. Maskeliunas, and R. Damaševičius, "Detection of sitting posture using hierarchical image composition and deep learning," *PeerJ Comput. Sci.*, vol. 7, e442, 2021. doi: 10.7717/peerj-cs.442
- [8] H. Jeong and W. Park, "Developing and evaluating a mixed sensor smart chair system for real-time posture classification: Combining pressure and distance sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1805–1813, 2021. doi: 10.1109/JBHI.2020.3030096
- [9] F. Luna-Perejón, J. M. Montes-Sánchez, L. Durán-López *et al.*, "IoT device for sitting posture classification using artificial neural networks," *Electronics*, vol. 10, no. 15, p. 1825, 2021. doi: 10.3390/electronics10151825
- [10] T. A. Gelaw and M. T. Hagos, "Posture prediction for healthy sitting using a smart chair," in *Proc. Advances of Science and Technology*, 2022, pp. 401–411. doi: 10.1007/978-3-030-93709-6_26
- [11] S. M. Lee, H. J. Kim, S. J. Ham, and S. Kim, "Assistive devices to help correct sitting-posture based on posture analysis results," *JOIV: International Journal on Informatics Visualization*, vol. 5, no. 3, pp. 340–346, 2021. doi: 10.30630/joiv.5.3.673
- [12] L. Zhou, L. Zhang, and N. Konz, "Computer vision techniques in manufacturing," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 1, pp. 105–117, 2023. doi: 10.1109/TSMC.2022.3166397
- [13] N. F. Zulkurnain, Y. A. Ahmad, and N. A. M. Nazri, "Web-based safety eyewear detection system in workplace using machine learning," in *Proc. 2023 9th International Conference on Computer and Communication Engineering (ICCCCE)*, 2023, pp. 289–293. doi: 10.1109/ICCCCE58854.2023.10246087
- [14] B. T. Nguyen-Tat, M. Q. Bui, and V. M. Ngo, "Automating attendance management in human resources: A design science approach using computer vision and facial recognition," *International Journal of Information Management Data Insights*, vol. 4, no. 2, 100253, 2024. doi: 10.1016/j.jjime.2024.100253
- [15] H. B. Khoirullah, N. Yudistira, and F. A. Bachtiar, "Facial expression recognition using convolutional neural network with attention module," *JOIV: International Journal on Informatics Visualization*, vol. 6, no. 4, pp. 897–903, 2022. doi: 10.30630/joiv.6.4.963
- [16] C. H. Boe, K. W. Ng, S. C. Haw, P. Naveen, and E. A. Anaam, "An automated face detection and recognition for class attendance," *JOIV: International Journal on Informatics Visualization*, vol. 8, no. 3, pp. 1146–1153, 2024. doi: 10.62527/joiv.8.3.2967
- [17] D. F. Abdulkader and M. F. Ghanim, "Design and analysis of face recognition system based on VGGFace-16 with various classifiers," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 2, pp. 1499–1510, 2024. doi: 10.11591/ijai.v13.i2.pp1499-1510

- [18] V. Adewopo, N. Elsayed, Z. Elsayed *et al.*, “Big data and deep learning in smart cities: A comprehensive dataset for AI-driven traffic accident detection and computer vision systems,” in *Proc. 2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, 2024, pp. 1–6. doi: 10.1109/ICMI60790.2024.10586073
- [19] J. E. Estrada, L. A. Veja, and M. Devaraj, “Modelling proper and improper sitting posture of computer users using machine vision for a human-computer intelligent interactive system during COVID-19,” *Applied Sciences*, vol. 13, no. 9, p. 5402, 2023. doi: 10.3390/app13095402
- [20] S. Zhao and Y. Su, “Sitting posture recognition based on the computer’s camera,” in *Proc. 2024 2nd Asia Conference on Computer Vision, Image Processing and Pattern Recognition*, 2024, pp. 1–5. doi: 10.1145/3663976.3664014
- [21] A. Sabo, N. Mittal, A. Deshpande, H. Clarke, and B. Taati, “Automated, vision-based goniometry and range of motion calculation in individuals with suspected Ehlers-Danlos syndromes/generalized hypermobility spectrum disorders: A comparison of pose-estimation libraries to goniometric measurements,” *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 12, pp. 140–150, 2024. doi: 10.1109/JTEHM.2023.3327691
- [22] P. C. Lin, Y. J. Chen, W. S. Chen, and Y. J. Lee, “Automatic real-time occupational posture evaluation and select corresponding ergonomic assessments,” *Sci. Rep.*, vol. 12, no. 1, p. 2139, 2022. doi: 10.1038/s41598-022-05812-9
- [23] S. Reid, S. Coleman, D. Kerr, P. Vance, and S. O’Neill, “Keypoint changes for fast human activity recognition,” *SN Computer Science*, vol. 4, no. 5, p. 621, 2023. doi: 10.1007/s42979-023-02063-x
- [24] I. Akhter, A. Jalal, and K. Kim, “Pose estimation and detection for event recognition using sense-aware features and Adaboost classifier,” in *Proc. 2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*, 2021, pp. 500–505. doi: 10.1109/IBCAST51254.2021.9393293
- [25] J. Gao, C. Ma, D. Wu, X. Xu, S. Wang, and J. Yao, “Recognition of human motion intentions based on Bayesian-optimized XGBOOST algorithm,” *Journal of Sensors*, vol. 2022, no. 1, 3015645, 2022. doi: 10.1155/2022/3015645
- [26] B. V. Breugel, N. Seedat, F. Imrie, and M. V. D. Schaar, “Can you rely on your model evaluation? Improving model evaluation with synthetic test data,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 1889–1904, 2023.
- [27] D. Podell, Z. English, K. Lacey *et al.*, “SDXL: Improving latent diffusion models for high-resolution image synthesis,” arXiv Print, arXiv:2307.01952, 2023. doi: 10.48550/arXiv.2307.01952
- [28] B. F. Labs, S. Batifol, A. Blattmann *et al.*, “FLUX.1 context: Flow matching for in-context image generation and editing in latent space,” arXiv Print, arXiv.2506.15742, 2025. doi: 10.48550/arXiv.2506.15742
- [29] GBRX. (2025). GongaLomo XL/Flux/Pony - v5.0 FluxL DMD. Civitai. [Online]. Available: <https://civitai.com/models/1513492/gonzalomo-xlfluxpony>
- [30] RunDiffusion. (2025). Juggernaut XL—ragnarok_by_rundiffusion. [Online]. Available: <https://civitai.com/models/133005/juggernaut-xl>
- [31] Sinatra. (2025). Real Dream—SDXL 7. [Online]. Available: <https://civitai.com/models/153568/real-dream>
- [32] O. O. E. Peter, M. M. Rahman, and F. Khalifa, “Advancing AI-powered medical image synthesis: Insights from MedVQA-GI challenge using CLIP, fine-tuned stable diffusion, and dream-Booth + LoRA,” arXiv Print, arXiv:2502.20667, 2025. doi: 10.48550/arXiv.2502.20667
- [33] S. Lin, A. Wang, and X. Yang, “SDXL-Lightning: Progressive adversarial diffusion distillation,” arXiv Print, arXiv:2402.13929, 2024. doi: 10.48550/arXiv.2402.13929
- [34] FLUX.1 Krea [dev]. Black Forest Labs. 2025. [Online]. Available: <https://huggingface.co/black-forest-labs/FLUX.1-Krea-dev>
- [35] M. Li, Y. Lin, Z. Zhang *et al.*, “SVDQuant: Absorbing outliers by low-rank components for 4-bit diffusion models,” arXiv Print, arXiv:2411.05007, 2025. doi: 10.48550/arXiv.2411.05007
- [36] FLUX.1 Turbo Alpha. Alimama Creative. 2024. [Online]. Available: <https://huggingface.co/alimama-creative/FLUX.1-Turbo-Alpha>
- [37] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proc. IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847. doi: 10.48550/arXiv.2302.05543
- [38] X. Zhang, J. Fan, T. Peng, P. Zheng, X. Zhang, and R. Tang, “Multimodal data-based deep learning model for sitting posture recognition toward office workers’ health promotion,” *Sensors and Actuators A: Physical*, vol. 350, 114150, 2023. doi: 10.1016/j.sna.2022.114150
- [39] M. R. Cardoso, A. K. Cardenas, and W. J. Albert, “A biomechanical analysis of active vs static office chair designs,” *Applied Ergonomics*, vol. 96, 103481, 2021. doi: 10.1016/j.apergo.2021.103481
- [40] E. Weston, P. Le, and W. S. Marras, “A biomechanical and physiological study of office seat and tablet device interaction,” *Applied Ergonomics*, vol. 62, pp. 83–93, 2017. doi: 10.1016/j.apergo.2017.02.013
- [41] R. Moreira, S. Teixeira, R. Fialho *et al.*, “Validity analysis of monocular human pose estimation models interfaced with a mobile application for assessing upper limb range of motion,” *Sensors*, vol. 24, no. 24, p. 7983, 2024. doi: 10.3390/s24247983
- [42] S. S. R. Kopanaty, S. V. S. Emani, U. K. Cherukuri, and M. Duggirala, “Human pose tracking with MoveNet,” *IJETMS*, vol. 9, no. 2, pp. 466–475, 2025. doi: 10.46647/ijetms.2025.v09i02.058
- [43] F. C. Lin, H. H. Ngo, C. R. Dow *et al.*, “Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection,” *Sensors*, vol. 21, no. 16, p. 5314, 2021. doi: 10.3390/s21165314
- [44] A. Kumar, K. Sharma, and A. Sharma, “MEMoR: A multimodal emotion recognition using affective biomarkers for smart prediction of emotional health for people analytics in smart industries,” *Image and Vision Computing*, vol. 123, 104483, 2022. doi: 10.1016/j.imavis.2022.104483
- [45] R. Gao and Z. Liu, “An improved AdaBoost algorithm for hyperparameter optimization,” *J. Phys.: Conf. Ser.*, vol. 1631, no. 1, 012048, 2020. doi: 10.1088/1742-6596/1631/1/012048
- [46] X. Liu, Y. Wu, and H. Wu, “Machine learning enabled 3D body measurement estimation using hybrid feature selection and Bayesian search,” *Applied Sciences*, vol. 12, no. 14, p. 7253, 2022. doi: 10.3390/app12147253
- [47] H. Altaheri, G. Muhammad, M. Alsulaiman *et al.*, “Deep learning techniques for classification of Electroencephalogram (EEG) Motor Imagery (MI) signals: A review,” *Neural Comput. & Applic.*, vol. 35, no. 20, pp. 14681–14722, 2023. doi: 10.1007/s00521-021-06352-5
- [48] H. Hamdani, A. Septiarni, A. Sunyoto, S. Suyanto, and F. Utaminigrum, “Detection of oil palm leaf disease based on color histogram and supervised classifier,” *Optik*, vol. 245, 167753, 2021. doi: 10.1016/j.ijleo.2021.167753
- [49] F. Shi, “A motion capture framework for table tennis using optimized SVM and AdaBoost algorithms,” *IJCAI*, vol. 49, no. 6, pp. 191–204, 2025. doi: 10.31449/inf.v49i6.6809
- [50] S. Benghazouani, S. Nouh, and A. Zakrani, “Enhancing breast cancer diagnosis: A comparative analysis of feature selection techniques,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 4, p. 4312, 2024. doi: 10.11591/ijai.v13.i4.pp4312-4322
- [51] W. Castro, J. Oblitas, R. Santa-Cruz, and H. Avila-George, “Multilayer perceptron architecture optimization using parallel computing techniques,” *PLoS ONE*, vol. 12, no. 12, e0189369, 2017. doi: 10.1371/journal.pone.0189369
- [52] R. O. Ogundokun, R. Maskeliūnas, and R. Damaševičius, “Human posture detection using image augmentation and hyperparameter-optimized transfer learning algorithms,” *Applied Sciences*, vol. 12, no. 19, 10156, 2022. doi: 10.3390/app121910156
- [53] O. Rainio, J. Teuhon, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Sci. Rep.*, vol. 14, no. 1, p. 6086, 2024. doi: 10.1038/s41598-024-56706-x

Copyright © 2025 by the authors. This is an open access article distributed under the Creative Commons Attribution License ([CC-BY-4.0](https://creativecommons.org/licenses/by/4.0/)), which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.