Hand Landmark Detection using YOLO26n for Gesture-Based Smart Wheelchair Navigation

Rakib Ahammed Diptho, Safiul Haque Chowdhury, Md Abdullah Al Mamun, Md. Shakhawat Hosen, Md. Shamsur Rahman, Sarnali Basak and Md Abul Kalam Azad
Department of CSE, Jahangirnagar University, Dhaka-1342
Correspondence *e*-mail: 3788diptho@juniv.eduw

***Abstract*—Assistive mobility technologies enhance independence for individuals with physical disabilities, yet conventional joystick-based interfaces remain difficult to operate for users with limited motor coordination. Gesture-based interaction has emerged as an effective contactless control approach, enabling natural human–computer interaction through computer vision. To address these challenges, this study proposes a smart wheelchair system based on the YOLO26n-pose deep learning model for accurate hand-keypoints detection. The model is trained on the hand-keypoints dataset, which contains 26,768 annotated images with 21 hand landmarks, enabling robust hand landmark detection that performs better than many other existing approaches. Using the detected keypoints, a distance-based gesture recognition method analyzes finger movements and generates wheelchair navigation commands. Experimental results demonstrate strong detection performance, achieving a mean average precision of 0.992 with an inference time of 2.6ms, indicating that the proposed approach is well-suited to real-time wheelchair control. The fast and precise processing capability of the proposed approach outperforms the contemporary methods in terms of speed, efficiency, and reliability.**

***Index Terms*—Assistive Technology, Deep Learning, Gesture-Controlled Wheelchair, Hand Landmark, YOLO26n.**

I. Introduction

Physical disabilities create barriers to independence by limiting an individual's ability to move about freely in their environment. Additionally, millions of individuals worldwide rely on wheelchairs as their primary means of transportation to accomplish daily tasks and interact socially [1]. While manual wheelchairs provide mobility for physically disabled persons, they are generally cumbersome to operate requiring great amounts of strength from the user's upper body. Electric wheelchairs have greatly reduced this burden; however, joystick based interfaces are still common today. Users who have impaired fine motor coordination find it difficult to use these joysticks. The limitations described above demonstrate the necessity for wheelchair control mechanisms that are easier to access, smarter, and more intuitive.

A. Research Gap and Motivation

Motivated by these challenges, recent advances in Artificial Intelligence (AI) and Computer Vision (CV) have enabled the development of gesture-based Human–Computer Interaction (HCI) systems. By utilizing AI-driven vision techniques, wheelchair control can be achieved through natural hand gestures without requiring direct physical interaction. As a result, gesture recognition technologies have gained increasing attention in robotics, healthcare monitoring, and assistive systems. In particular, Hand Landmark (HL) detection and Deep Learning (DL)-based approaches have shown promising potential for real-time gesture recognition.

Despite these advancements, several important limitations remain in existing gesture-based wheelchair control systems:

- **Latency Issues**: Existing MediaPipe-based systems, such as those presented by Khaksar *et al.* [2] and Huda *et al.* [3], typically require 10–15ms inference time per frame. Although effective for general gesture recognition, such latency may reduce responsiveness in real-time wheelchair navigation, where immediate control feedback is essential for safe operation.
- **Limited Customization**: Most MediaPipe-based frameworks employ relatively fixed architectures with limited support for task-specific optimization and domain-focused retraining. Consequently, performance may degrade under challenging real-world conditions, including illumination changes, hand occlusions, and background complexity [2].
- **Incomplete Safety Integration**: In addition to gesture recognition, intelligent wheelchair systems require reliable safety mechanisms for practical deployment. However, studies by Basak *et al.* [4] and Chamalsha *et al.* [5] indicate that many existing systems lack integrated obstacle detection and health monitoring capabilities, potentially limiting operational safety.
- **Insufficient Comparative Analysis**: Furthermore, limited research has systematically evaluated hand pose estimation architectures specifically for wheelchair control applications. Therefore, it remains unclear which framework provides the most effective balance between speed, accuracy, robustness, and deployment flexibility.

B. Proposed Contribution

To address the aforementioned limitations, this work introduces a gesture-based smart wheelchair control framework using the YOLO26n architecture. The major contributions of this research are summarized as follows:

1) **Real-Time Hand Landmark Detection**: The proposed YOLO26n-based framework achieves an inference time of 2.6ms, which is approximately 4 to 6 times faster than existing MediaPipe-based approaches while maintaining high detection accuracy (mAP@50=0.992). This improvement is expected to enable highly responsive real-time wheelchair control with minimal latency.
2) **Customizable Task-Specific Training**: Unlike fixed MediaPipe architectures, the proposed framework supports task-specific fine-tuning using the Hand Keypoints dataset. This is expected to improve robustness under varying illumination conditions, hand occlusions, and complex real-world environments relevant to wheelchair navigation.
3) **Integrated Safety-Critical System**: In addition to gesture recognition, the proposed system is designed to incorporate obstacle detection and health monitoring functionalities to improve operational safety and enhance the practicality of AI-assisted smart wheelchair systems.

Collectively, these contributions highlight the potential of AI-driven gesture recognition to support faster, safer, and more adaptive wheelchair control systems for physically disabled individuals.

II. Literature Review

In recent years, gesture recognition within the domain of Human–Computer Interaction (HCI) has emerged as an important research area with applications in robotics and medical systems, enabling intuitive and contactless interaction between users and technological systems. Among the various approaches used for gesture recognition, Computer Vision (CV) and Deep Learning (DL) have become widely adopted techniques. For example, Zhao *et al.* [6] used hand landmark extraction to create new DL architectures, while Khaksar *et al.* [2] examined how well MediaPipe-based hand landmark detection worked along with its existing limitations.

In addition to advancements in sensing modalities, researchers have started exploring novel sensing modalities such as thermal imaging with meta-surface optics. Chen *et al.* [7] demonstrated that using binocular thermal meta-lens systems combined with DL could be used to perform robust gesture recognition in low-contrast and partially occluded conditions. As a complementary method to those above, radar-based sensing has recently emerged as a privacy-preserving alternative to traditional CV and machine learning-based methods. Luo *et al.* [8] presented a Multi-Task Learning Framework that used miniature radar sensors to perform both gesture recognition and person identification utilizing micro-Doppler and range-Doppler signatures. Researchers have also been exploring feature fusion methods to increase recognition accuracy by fusing different features extracted from images of hand gestures. Alotaibi *et al.* [9] developed a feature fusion-based hand gesture recognition system using ConvNeXt Base, VGG16, and EfficientNet-V2 along with deep belief networks and tornado optimization for sign language accessibility.

The object detection framework is a widely used architecture for gesture recognition due to its fast detection speed and low computational cost. Feng *et al.* [10] proposed a lightweight YOLO-based gesture recognition model for edge devices. Huang *et al.* [11] and Wang *et al.* [12] proposed YOLO-based frameworks for dynamic gesture recognition and detecting small objects. The hybrid approach of combining landmark detection with DL classifiers has also been studied. Padhi *et al.* [13] implemented MediaPipe-based hand detection with Convolutional Neural Networks. Ye *et al.* [14] proposed a spatiotemporal YOLO-based architecture for motion-based gesture recognition.

Recently, there have been several studies on optimizing YOLO models for embedded deployment. Song *et al.* [15] and Dang *et al.* [16] proposed lightweight optimization techniques to decrease computational complexity while maintaining detection performance.

Gesture recognition technology has been incorporated into assistive mobility technologies. Huda *et al.* [3] developed a MediaPipe hand landmark-based gesture-controlled wheelchair system. Basak *et al.* [4] and Chamalsha *et al.* [5] integrated gesture recognition with additional sensing technologies to enhance system safety.

Although there have been numerous advances in the field of gesture recognition, many existing systems rely heavily on predefined landmark detection frameworks and may not include built-in safety and health monitoring components, encouraging the creation of smarter and more flexible assistive mobility solutions.

III. Methodology

A. System Overview

The proposed system consists of three main modules:

- Gesture-based wheelchair control
- Obstacle detection and emergency stop
- Health monitoring system

The overall workflow of the system is illustrated in Fig.1.

B. Dataset Description

To train a Hand Pose Estimation Model, we used the *Hand-Keypoints Dataset* with 26,768 images of human hands that have been manually labeled [17]. It was built to be able to perform both hand-pose estimation and hand-gesture recognition and has detailed labels of hand joints. Images were taken under many environmental conditions; as background images, light images, images of the hand in many angles (as well as) as hand-skin color. As such, it would provide the ability to generalize across environments, so our model will hopefully detect hand gestures effectively in a variety of real world settings. Annotating data for the Hand-

Keypoints dataset was done using the Google MediaPipe API; therefore, each image should be accurately and consistently annotated for all key points of the hand.

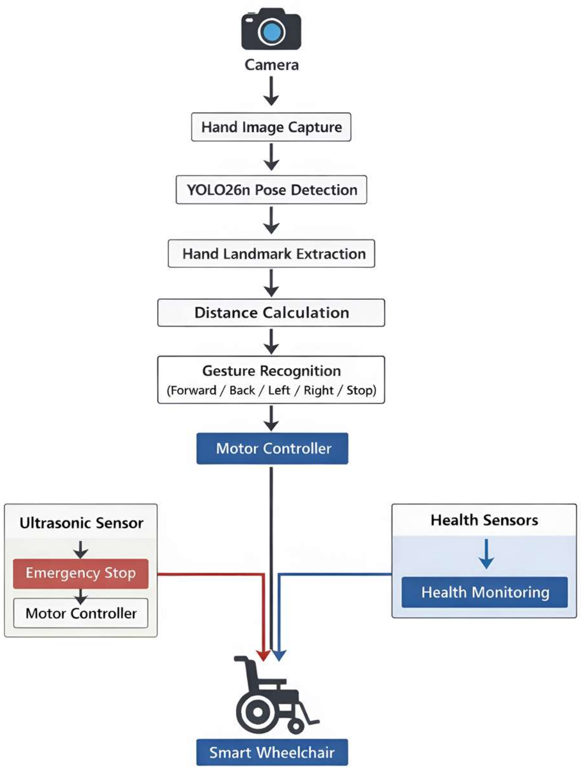

Fig. 1. Proposed methodology of the gesture-based smart wheelchair system using YOLO26n.

We split up the Hand-Keypoints dataset into two groups: one group had 18,776 images for training, while the other group had 7,992 images for validation. In addition to having annotations for these images, each image also contained information on 21 hand keypoints that represent anatomically correct joints found within the human hand (including the wrist joint). These keypoints are an accurate representation of how the hand's movement occurs, enabling the proposed wheelchair control system to track finger movements precisely for recognizing hand gestures.

C. Hand Landmark Detection

The human hand is represented using 21 landmark points that describe the spatial structure of the hand, beginning from the wrist and extending through each finger, as shown in Fig.2. The wrist serves as the base reference point for measuring finger distances and tracking hand movements. Each finger is represented by four keypoints corresponding to its main joints and fingertip, including the Carpometacarpal (CMC) joint, Metacarpophalangeal (MCP) joint, Proximal Interphalangeal (PIP) joint, Distal Interphalangeal (DIP) joint, and the fingertip (the thumb uses an Interphalangeal (IP) joint instead of PIP and DIP). These landmarks enable accurate tracking of finger bending, extension, and spacing. By analyzing the spatial relationships between these keypoints, the system can interpret hand gestures used for wheelchair control.

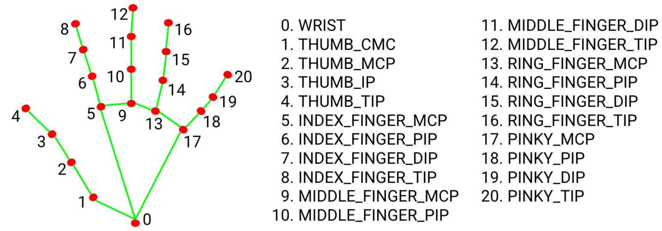

Fig. 2. Hand landmark structure with 21 keypoints for pose estimation [17].

D. Hand Movement Detection

After detecting hand landmarks using the YOLO26n pose estimation model, the next step is to analyze the hand configuration to recognize gestures used for wheelchair control. The YOLO26n model detects 21 hand keypoints representing the wrist and finger joints, providing spatial information that enables real-time tracking of finger positions and movements.

For gesture recognition, the system mainly utilizes the wrist and fingertip landmarks. In the hand landmark indexing scheme, landmark 0 represents the wrist, while landmarks 4, 8, 12, 16, and 20 correspond to the thumb, index, middle, ring, and little fingertips, respectively. These landmarks are used to determine the configuration of the hand and identify the performed gesture. As illustrated in Fig.3, the displacement of the index fingertip is used to determine the movement direction of the wheelchair.

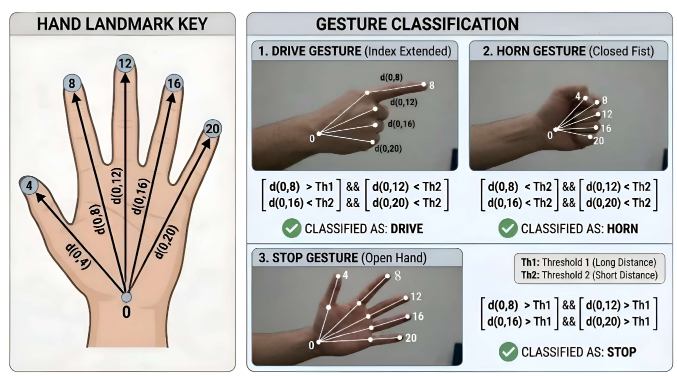

Fig. 3. Hand landmark keypoints and distance-based gesture classification (Drive, Horn, and Stop).

To measure the spatial relationship between two landmarks, the distance between them [18] is calculated as:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

where (x_1, y_1) and (x_2, y_2) represent the Cartesian coordinates of two hand landmarks.

To ensure consistent gesture recognition regardless of hand size or distance from the camera, the distances are normalized using the hand width. The hand width is defined as the distance between keypoints 5 and 17, corresponding to the base joints of the index and little fingers. The normalized distance is calculated as:

$$d_{norm} = \frac{d(a,b)}{d(5,17)} \quad (2)$$

This normalization reduces scale variations and improves the robustness of the gesture recognition process.

Based on the normalized distances between the wrist landmark (0) and fingertip landmarks, three primary gesture classes are defined for wheelchair control.

In the **Drive gesture**, the index finger remains extended while the other fingers remain folded, resulting in a larger distance $d(0,8)$ compared to other fingertip distances, while in the **Stop gesture**, all fingers are extended and the hand remains fully open, producing large distances between the wrist and all fingertip landmarks. On the other hand, the **Horn gesture** corresponds to a closed hand where all fingertips remain close to the wrist, producing smaller distances.

The conditions used to classify these gestures are summarized in Table I, where Th_1 and Th_2 represent experimentally determined thresholds for extended and folded fingers. Because the distances are normalized by the hand width d (5, 17) using *Eq.*2, both the thresholds are dimensionless and they remain stable across users and camera distances. Based on empirical tuning over the validation set, the values were fixed at Th_1 = 1.5 and Th_2 = 1.0 (in normalized hand-width units). A finger is therefore treated as extended when its normalized wrist-to-fingertip distance exceeds Th_1 and as folded when it falls below Th_2, while the intermediate band ($Th_2 \leq d_{norm} \leq Th_1$) is left as a deliberate margin to suppress ambiguous or transitional hand poses.

TABLE I
CONDITIONS FOR BASIC GESTURE RECOGNITION

Gesture	Distance Conditions	Description
Drive	$d(0,8) > T_{h1}$, and $d(0,12), d(0,16), d(0,20) < T_{h2}$	Index finger extended
Stop	$d(0,8), d(0,12), d(0,16), d(0,20) > T_{h1}$	Hand fully open
Horn	$d(0,8), d(0,12), d(0,16), d(0,20) < T_{h2}$	Hand closed

E. Directional Movement Detection

Once the Drive gesture is detected, the direction of wheelchair movement is determined by tracking the index fingertip (keypoint 8). The initial fingertip position is stored as a reference point (x_{init}, y_{init}). During operation, the current fingertip position is detected as (x_{cur}, y_{cur}). The fingertip displacement relative to the reference position is computed as:

$$dx = x_{cur} - x_{init} \quad (3)$$

$$dy = y_{cur} - y_{init} \quad (4)$$

These displacement values are compared with predefined thresholds to determine the direction of wheelchair movement. Horizontal displacement controls left and right motion, while vertical displacement controls forward and backward movement. Following the image-coordinate convention in which the horizontal axis increases from left to right and the vertical axis increases from top to bottom, a rightward fingertip movement produces a positive dx and a leftward movement a negative dx, whereas an upward movement produces a negative dy and a downward movement a positive dy. The directional thresholds are therefore applied symmetrically about zero, so that a movement command is issued only when the displacement along the dominant axis exceeds $\pm Th_3$. The directional drive gesture conditions are summarized in Table II.

TABLE II
DIFFERENT DRIVE GESTURE CONDITION

Gesture	Condition
Drive Right	$dx > Th_3$
Drive Left	$dx < -Th_3$
Drive Forward	$dy < -Th_3$
Drive Backward	$dy > Th_3$

The threshold Th_3 is used for directional movement detection by comparing the displacement of the index fingertip to determine the wheelchair navigation direction. Because absolute pixel displacements depend on the camera resolution, the directional threshold is defined relative to the frame width and set to approximately 6% of the frame width, corresponding to Th_3 = 40 pixels for the 640 × 480 input used in this study. The purpose of this threshold is to distinguish intentional directional hand movements from minor involuntary hand jitter and tracking noise inherent in vision-based pointing systems. Previous studies on vision-based pointing and cursor control have reported that low-amplitude hand jitter can degrade pointing accuracy and that such motion is commonly mitigated using thresholding or filtering techniques [19], [20]. Accordingly, the selected threshold is intended to remain above the typical jitter amplitudes reported in the literature while preserving responsiveness to deliberate hand movements. Although no universally accepted pixel threshold has been established for gesture-controlled wheelchair navigation, the value adopted in this study is used as a practical design parameter. Its optimal value will be investigated through sensitivity analysis and hardware validation in future work.

Based on these conditions, the gesture-based wheelchair control procedure is defined as described in Algorithm 1.

Algorithm 1: Gesture-Based Wheelchair Control

Initialize camera
Load YOLO26n pose estimation model
Capture initial hand landmarks
Store initial fingertip position (x_{init}, y_{init})
while system is running **do**
    Capture current frame
    Detect hand landmarks
    Obtain index fingertip position (x_{cur}, y_{cur})
    $dx \leftarrow x_{cur} - x_{init}$
    $dy \leftarrow y_{cur} - y_{init}$
    if Stop gesture detected **then**
        Stop wheelchair motors
    else if Horn gesture detected **then**
        Activate horn signal

Algorithm 1: Gesture-Based Wheelchair Control

    else if $dx > Th_3$ **then**
        Drive wheelchair right
    else if $dx < -Th_3$ **then**
        Drive wheelchair left
    else if $dy < -Th_3$ **then**
        Drive wheelchair forward
    else if $dy > Th_3$ **then**
        Drive wheelchair backward
    else
        Stop wheelchair movement
    end if
end while

F. *Proposed Hand Landmark Detection Model (YOLO26n Architecture)*

The proposed hand gesture recognition framework employs the YOLO26n architecture, a lightweight and computationally efficient deep learning model specifically designed for real-time hand detection and landmark estimation. As illustrated in Fig.4, the architecture is composed of three primary components: the Backbone, the Neck/Head based on the Feature Pyramid Network and Path Aggregation Network (FPN-PAN), and the multi-scale Detection Heads. The network is configured with a depth scaling factor of 0.50, width scaling factor of 0.25, and a maximum channel capacity of 1024. Overall, the model contains approximately 260 layers, 2.57 million trainable parameters, and 6.1 GFLOPs, making it highly suitable for deployment in resource-constrained embedded assistive systems such as AI-based smart wheelchairs [17].

The model receives an RGB input image of size 640 × 640 × 3, which is first processed through the Backbone network. The Backbone consists of multiple 3 × 3 convolutional layers with stride 2 that progressively down sample the feature maps while increasing semantic representation capability. The spatial dimensions are gradually reduced from 640 × 640 to 320 × 320, 160 × 160, 80 × 80, 40 × 40, and finally 20 × 20. This hierarchical feature extraction process enables the network to learn both low-level spatial structures and high-level semantic patterns associated with hand gestures and landmark positions.

To improve feature learning efficiency, several C3k2 modules are integrated at different stages of the Backbone. These modules are optimized residual bottleneck structures designed to strengthen gradient propagation, improve feature reuse, and reduce computational complexity. The use of C3k2 blocks allows the network to preserve important spatial information while maintaining a lightweight architecture suitable for real-time inference. This is particularly important for gesture-controlled wheelchair systems, where accurate and rapid hand landmark detection is essential for responsive navigation.

At the deeper stages of the Backbone, a Spatial Pyramid Pooling Fast (SPPF) module is incorporated to aggregate contextual information from multiple receptive fields. By combining features at different scales, the SPPF module improves robustness against variations in hand size, viewing angle, and camera distance. This capability is important in practical wheelchair environments, where users may perform gestures under different spatial conditions.

Following the SPPF layer, the architecture integrates a C2PSA attention mechanism to enhance discriminative feature representation. The attention module emphasizes informative hand regions while suppressing irrelevant background information and noise. As a result, the model becomes more robust under challenging real-world conditions such as dynamic illumination changes, cluttered backgrounds, motion blur, and partial hand occlusions [17].

The extracted feature maps are subsequently forwarded to the Neck/Head component, which adopts an FPN-PAN structure for multi-scale feature fusion. The Feature Pyramid Network (FPN) performs top-down feature propagation through upsampling operations and feature concatenation. Specifically, deep semantic feature maps from the 20 × 20 resolution are upsampled and fused with intermediate 40 × 40 and 80 × 80 feature maps. This process preserves fine-grained spatial details required for accurate hand landmark localization.

In parallel, the Path Aggregation Network (PAN) strengthens bottom-up information flow through downsampling and feature aggregation operations. The PAN structure enhances localization capability by combining shallow spatial information with deeper semantic representations. The combination of FPN and PAN structures enables the network to maintain strong feature representations across multiple spatial resolutions, improving detection consistency for both small and large hand regions.

The final stage of the architecture consists of three Detection Heads operating at P3/8, P4/16, and P5/32 scales. These detection branches generate landmark predictions at different spatial resolutions, enabling robust multi-scale hand detection. The P3 branch focuses on smaller hand structures and fine-grained landmark details, while the P4 and P5 branches capture medium- and large-scale contextual information. Multi-scale prediction significantly improves robustness against variations in hand size, user orientation, and camera distance.

Furthermore, the YOLO26n architecture adopts an NMS-free (Non-Maximum Suppression-free) prediction strategy, which simplifies the detection pipeline and tries to improve the inference speed by eliminating redundant post-processing operations [17]. During deployment, the model can additionally be quantized to FP16 or INT8 precision to reduce memory usage and computational cost while maintaining detection accuracy. Finally, the predicted hand landmark points are forwarded to the gesture recognition module, where they are translated into wheelchair navigation commands for real-time assistive control.

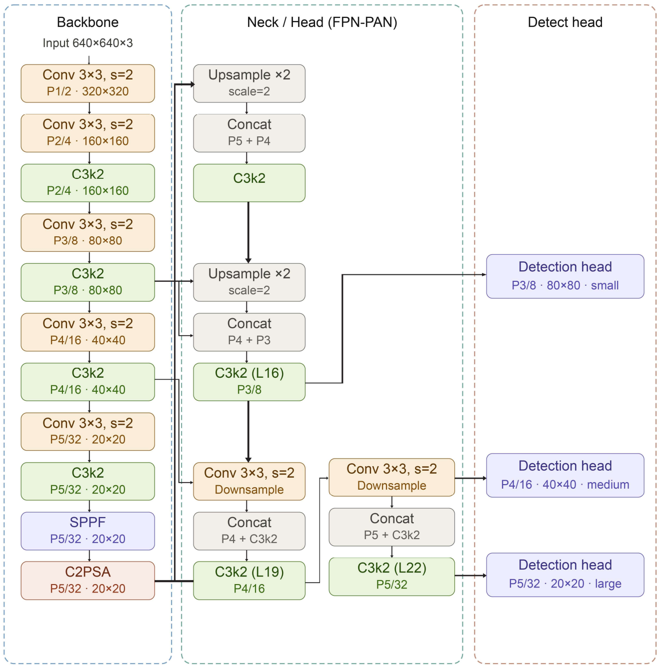

Fig. 4. Architecture of the proposed YOLO26n model.

G. *Architectural Distinctions from Existing YOLO-Pose Models*

Although the proposed framework inherits the single-stage detection philosophy of the YOLO family, several architectural and training-level modifications differentiate YOLO26n from earlier YOLO-pose variants such as YOLOv8-pose and YOLOv11-pose. First, YOLO26n employs an NMS-free, end-to-end prediction head; unlike conventional YOLO-pose models that rely on anchor matching and Non-Maximum Suppression as a post-processing step, this design removes a manually tuned hyperparameter, yields a deterministic and lower inference latency, and simplifies embedded deployment. Secondly, the backbone of YOLOv8 uses optimized C3k2 residual modules to replace the older C2f/C3 blocks of previous YOLO versions *(i.e.,* YOLOv7), to improve both feature reuse and gradient flow while improving performance at lower computational costs. Thirdly, a C2PSA Position-Sensitive Attention Block is added post the SPPF Module. This Attention Stage is not present in previous YOLO Pose Detectors, and it enhances focus on the most informative parts of hands and thus provides improved robustness against partial occlusion, cluttered background and illumination variations. Finally, the model is trained with the MuSGD optimizer, a recent hybrid of SGD and the Muon optimizer, rather than the standard SGD or Adam schedules typically used for prior YOLO-pose models, which improves convergence stability for the keypoint-regression objective.

Beyond these structural differences, the proposed pipeline also differs from existing YOLO-based and MediaPipe-based gesture-recognition frameworks in how detection is used for control. Prior YOLO-gesture works, such as the lightweight edge model of Feng *et al*. [10], the dynamic and small-object detectors of Huang *et al*. [11] and Wang *et al*. [12], and the spatiotemporal architecture of Ye *et al*. [14], predict discrete gesture categories through bounding-box classification. In contrast, the proposed framework performs fine-grained 21-keypoint hand-pose estimation that feeds a lightweight geometric (distance-based) controller, enabling continuous directional navigation rather than a fixed set of gesture labels. Compared with the fixed two-stage MediaPipe pipeline used by Khaksar *et al*. [2] and Huda et al. [3], which cannot be retrained for task-specific conditions, YOLO26n is fully retrainable on the target hand distribution while remaining compact (2.57M parameters, 6.1 GFLOPs) and quantization-ready (FP16/INT8). Collectively, these properties make YOLO26n better suited to real-time, resource-constrained wheelchair control than existing pose-estimation backbones.

H. *Performance Metrics*

The detection performance of the proposed model is evaluated using standard object detection and pose estimation metrics, defined as follows:

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$IoU = \frac{Area_{overlap}}{Area_{union}} \tag{7}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} A\,P_i \tag{8}$$

where TP denotes True Positives, FP denotes False Positives, and FN denotes False Negatives. IoU represents the Intersection over Union between the predicted bounding box and the ground truth box, AP represents Average Precision for a class, and N denotes the total number of classes.

I. *Experimental Setup*

The YOLO26n-pose model was trained using the Hand-Keypoints dataset, containing 26,768 annotated images. Training was done for 50 epochs with an input image resolution of 640 × 640 pixels.

The training was performed using a Kaggle Notebook environment equipped with an NVIDIA Tesla P100 GPU. The model was trained using the MuSGD optimizer, a new hybrid optimizer that combines SGD with Muon. The training was conducted using hyperparameters summarized in Table III.

TABLE III
TRAINING HYPERPARAMETERS

Hyperparameter	Value
Optimizer	MuSGD
Learning Rate	0.01
Momentum	0.90
Weight Decay	0.0005
Epochs	50

J. *Proposed Deployment and System Configuration*

The proposed framework is designed to operate on an assistive wheelchair platform on which the gesture-recognition module would process a live video stream in real time. In the intended configuration, hand gestures would be captured using a commodity RGB webcam operating at a resolution of 640×480 pixels and 30 frames per second. The camera would be mounted on an adjustable arm positioned approximately 30–40 cm in front of the user, near armrest/chest height, and angled slightly downward so that the dominant hand remains within the field of view during normal seating posture. A practical implementation would run the vision model on an embedded single-board computer with GPU acceleration, which forwards the decoded navigation commands to a microcontroller-based motor-driver unit over a serial interface.

Based on the measured model inference time (2.6 ms) and a typical 30 FPS camera pipeline, the proposed system is expected to support low-latency real-time operation. The exact end-to-end latency will be experimentally quantified after implementation on a physical prototype. Regarding power, the additional electrical load introduced by the proposed vision and sensing stages is expected to be modest relative to the drive motors, which dominate the power budget of any powered wheelchair. Embedded vision platforms suitable for this task typically operate on the order of a few watts to a few tens of watts depending on the selected module, while the microcontroller and low-power sensing subsystems are expected to draw only a few watts. The actual power consumption will depend on the selected embedded hardware and will be evaluated experimentally in future prototype implementations.

IV. Results and Discussion

A. Hand Landmark Detection Performance

The qualitative performance results of the model are shown in Table IV. The model achieved a precision of 0.979, a recall of 0.978, and thus it was able to accurately detect hand landmarks while making very few incorrect predictions. The model was also able to achieve an mAP@50 of 0.992 as well as an mAP@50–95 of 0.910; this indicates that the model is capable of detecting hands at a variety of different IoU values. The pose mAP@50 value of 0.925 provides additional evidence of how accurately the model predicts the location of keypoints for the detected hands.

TABLE IV
PERFORMANCE EVALUATION RESULTS

Metric	Value
Precision	0.979
Recall	0.978
mAP@50	0.992
mAP@50–95	0.910
Pose mAP@50	0.925
Inference Time	2.6 ms

Fig.5 shows the training and validation performance metrics of the YOLO26n pose estimation model during the learning process. As the number of epochs increases, the precision, recall, and mean Average Precision (mAP) values gradually improve and stabilize. This trend indicates that the model is effectively learning to estimate hand landmarks and is converging toward stable performance with good generalization capability.

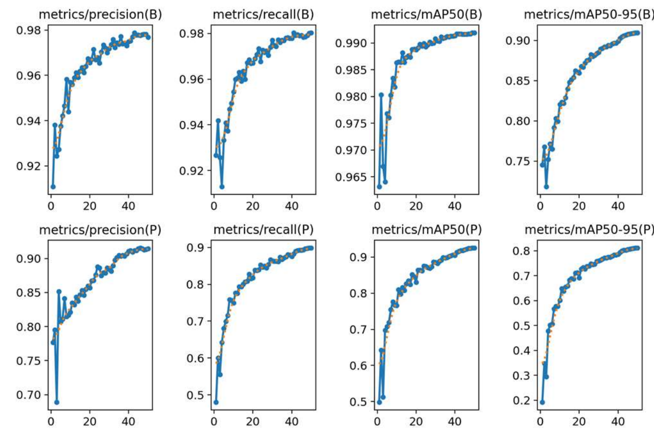

Fig. 5. Learning curves for detection and pose estimation metrics. B represents bounding box detection, while P represents pose (keypoint) detection.

The experimental results clearly demonstrate that the proposed YOLO26n pose estimation model achieves both high detection accuracy and fast processing times. The trained model achieved an average inference time of 2.6 ms per frame, enabling fast real-time gesture recognition. With this inference speed, the proposed model is capable of performing real-time hand landmark detection and is therefore suitable for use with gesture-based wheelchair controls in assistive mobility applications.

B. Comparison: YOLO26n vs. MediaPipe

TABLE V
COMPARATIVE PERFORMANCE: YOLO26N VS. MEDIAPIPE (LITERATURE-BASED ANALYSIS)

Metric	YOLO26n (This Work)	MediaPipe (from [2, 3])	Improvement
Accuracy (mAP@50)	0.992	0.98*	+1.2%
Inference Time (ms)	2.6	12–15	78% faster
Real-time Capability	384 FPS	67–83 FPS	4.6 to 5.7× faster
Customizability	Full task-specific training	Fixed architecture	Fully customizable
Hardware Flexibility	GPU/CPU quantizable	Requires specific config.	More flexible
Robustness to Occlusion	Hand-dataset optimized	Generic training	Hand-optimized design
Gesture Inference Latency	< 3 ms	12–15 ms	Responsive control

*MediaPipe accuracy estimated from Khaksar et al. [2] on similar tasks. Table V presents the key advantages of YOLO26n over MediaPipe:

- **Inference Speed**: YOLO26n is 4.6 to 5.7 times faster (2.6 ms vs. 12–15 ms), enabling responsive wheelchair control without perceptible latency.

- **Task-Specific Training**: Unlike fixed MediaPipe models, YOLO26n can be fine-tuned on hand-specific datasets, improving robustness.
- **Hardware Flexibility**: YOLO26n supports quantization (FP16, INT8) for embedded deployment on resource-constrained wheelchair platforms.
- **System Integration**: YOLO26n permits architectural modifications that could support the future integration of obstacle detection and health monitoring

These results indicate that YOLO26n achieves higher detection accuracy with substantially lower inference time compared with MediaPipe, making it more effective for real-time wheelchair control applications. The higher FPS and lower latency enable faster and more responsive gesture recognition, which is critical for safe navigation and user interaction. Furthermore, the support for task-specific training, quantization, and architectural customization is expected to support improved robustness and deployment flexibility in practical assistive healthcare environments.

C. Comparison with Recent State-of-the-Art Approaches

While the quantitative comparison above focuses on MediaPipe-based pipelines, it is equally important to position the proposed framework relative to the most recent (2025–2026) gesture-recognition research, which increasingly explores alternative sensing modalities and heavier network designs. Chen *et al*. [7] achieve robust recognition in low-contrast and partially occluded conditions using a binocular thermal metalens system; however, this approach depends on specialized thermal-optic hardware, whereas the proposed method attains real-time performance using only a commodity RGB camera, making it considerably cheaper and easier to retrofit onto existing wheelchairs. Luo *et al*. [8] adopt miniature radar sensors for privacy-preserving gesture recognition and person identification; although radar operates in darkness and preserves privacy, it provides coarser spatial resolution than vision and requires a dedicated radar module and signal-processing pipeline, while the proposed framework recovers fine-grained 21-keypoint hand geometry that is directly exploitable for continuous directional control. Alotaibi *et al*. [9] report high accuracy for sign-language recognition using a multi-backbone feature-fusion ensemble (ConvNeXt, VGG16, and EfficientNet-V2 with deep belief networks); such ensembles, however, incur a large computational footprint that is difficult to sustain on embedded hardware, whereas YOLO26n operates as a single lightweight model (2.57M parameters, 6.1 GFLOPs) at 384 FPS.

In summary, recent state-of-the-art methods primarily improve recognition under specific challenging conditions (e.g., darkness, occlusion, or large-vocabulary static signs) at the expense of specialized hardware or heavier computation. The proposed framework instead advances the speed–accuracy–deploying trade-off that is most critical for real-time assistive wheelchair control, delivering high keypoints accuracy and millisecond-level inference latency while remaining deployable on inexpensive, widely available hardware. At the same time, the complementary strengths of thermal and radar sensing motivate a promising future direction, namely fusing RGB-based pose estimation with such modalities to maintain reliable operation under extreme illumination or privacy-sensitive conditions.

D. Safety Module and Health Monitoring System

The framework also proposes an ultrasonic based obstacle detection component for the purpose of detecting nearby obstacles; an emergency stop feature which will automatically halt operation of the vehicle upon the occurrence of an object entering into a defined “safe zone” around the wheelchair; and a health monitoring component to enhance the safety of the users. In this proposed design, the ultrasonic sensors will detect the distance between the wheelchair and nearby objects. When it detects that there is an object in the designated area (defined by the caregiver or user) within a predetermined range from the wheelchair, the system will be halted. In addition to detecting obstacles, the proposed health monitoring component will also capture and process various physiological parameters including heart rate, SpO2, ECG signals and body temperature. This data will allow caregivers to assess whether the user has experienced any abnormalities during use of the wheelchair. It should be emphasized that these obstacle-detection, emergency-stop, and health-monitoring modules are presented as proposed components of the framework; they have not yet been physically implemented or experimentally validated in this study, and their hardware integration and evaluation are planned as future work.

E. Discussion on Proposed System

The YOLO26n pose estimation model provides an accuracy that is sufficient for reliable gesture determination. The proposed system is designed to combine gesture recognition, safety features, and health monitoring within a single integrated framework. The detected results are used by the distance-based gesture recognition method to determine the gestures made by the user.

Our proposed wheelchair control system requires only minimal finger movement and is therefore expected to reduce the physical effort for users with limited physical mobility. Additionally, the integration of obstacle detection and health monitoring is expected to improve the overall safety and usability of the wheelchair system.

In a nutshell, the proposed framework supports the feasibility of using DL-based hand pose estimation together with mathematical gesture recognition as an efficient basis for smart wheelchair control.

V. Conclusion and Future Work

The proposed method provides an efficient framework for extracting 21 hand keypoints and performing real-time gesture recognition for smart wheelchair navigation. The experimental results demonstrate that the YOLO26n model achieves high detection accuracy and fast inference time, supporting its suitability as the vision component of the proposed gesture-based smart wheelchair framework.

Distance-based methods were used to detect hand landmarks and recognize gestures based on those landmarks to generate commands for wheelchair movements. In addition to gesture-based controls, the system is designed to include an ultrasonic sensor-based obstacle detection module as well as a health monitoring module to provide enhanced safety to users and improve the overall functionality of the system.

The proposed system is designed to combine highly accurate hand landmark–based gesture recognition with the intended safety and health-monitoring features to address several limitations of other approaches for controlling wheelchairs by hand-gestures. Within this study, only the YOLO26n hand landmark detection model has been experimentally trained and evaluated; the remaining components are shown here as the supplementary elements of proposed framework and are deferred them to the future research work outlined below.

A. Future Work

The present study experimentally validates only the YOLO26n hand landmark detection model on the Hand-Keypoints dataset; the broader smart-wheelchair framework, including its safety, deployment, and clinical aspects, remains to be implemented and evaluated. Several directions are therefore planned as future work.

A comprehensive robustness evaluation of the detection model is planned, comprising quantitative assessment under different lighting conditions, a motion-blur sensitivity analysis, hand-occlusion experiments, and evaluation under varying background complexity. The effects of multiple camera distances and different camera viewing angles will also be characterized, together with the model's generalization across different user age groups and a range of skin tones, in order to establish the practical operating envelope of the system.

Embedded deployment of the model will subsequently be investigated on resource-constrained platforms such as the Raspberry Pi and NVIDIA Jetson families, accompanied by direct hardware latency measurement and a detailed power-consumption analysis to confirm real-time feasibility on portable hardware.

A complete physical wheelchair will then be implemented to enable end-to-end validation of the full framework. This includes experimental validation of the proposed obstacle-avoidance module and health-monitoring module, which have not yet been built, followed by a user study with disabled participants and a subsequent long-term usability study under realistic daily-use conditions.

Finally, several algorithmic and system-level extensions are envisaged, including IoT-based remote health monitoring, adaptive gesture recognition that personalizes to individual users, and multi-modal sensor fusion that combines RGB vision with complementary depth, radar, or thermal sensing to maintain reliable operation under extreme illumination, occlusion, or privacy-sensitive conditions.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude for the financial support provided under the project titled *Developing a Gesture-based Wheelchair for Assisting the Physically Challenged People in Bangladesh* through the ICSETEP-D-67 project, funded by the Asian Development Bank (ADB) and the University Grants Commission (UGC). Their support significantly contributed to the successful completion of this research.

REFERENCES

[1] L. Yang, W. Huang, X. Liu, and W. Yan, "Wheelchair accessibility to public facilities via transits and analysis of delay factors—A case study of Shanghai, China," Computers, Environment and Urban Systems, vol. 121, p. 102331, 2025.

[2] S. Khaksar, L. Checker, B. Borazjan, and I. Murray, "Design and evaluation of an alternative control for a quad-rotor drone using hand-gesture recognition," Sensors, vol. 23, no. 12, p. 5462, 2023.

[3] M. R. Huda, M. L. Ali, and M. S. Sadi, "Real-time hand-gesture recognition for the control of wheelchair," in Proc. 2022 12th Int. Conf. on Electrical and Computer Engineering (ICECE), 2022, pp. 384–387.

[4] S. Basak, F. F. Nandiny, S. M. M. H. Chowdhury, and A. A. Biswas, "Gesture-based smart wheelchair for assisting physically challenged people," in Proc. 2021 Int. Conf. on Computer Communication and Informatics (ICCCI), 2021, pp. 1–6.

[5] A. C. Chamalsha, R. A. K. Nadeeshan, H. M. K. K. M. B. Herath, S. L. P. Yasakethu, and U. Oruthota, "User-centric design of contactless smart wheelchair system for elderly people," in Proc. 2024 6th Int. Conf. on Advancements in Computing (ICAC), 2024, pp. 456–461.

[6] J. Zhao, X. H. Li, J. C. Dela Cruz, M. S. Verdadero, J. C. Centeno, and J. M. Novelero, "Hand gesture recognition based on deep learning," in Proc. 2023 Int. Conf. on Digital Applications, Transformation & Economy (ICDATE), 2023, pp. 250–254.

[7] B.-J. Chen, Y.-C. Chen, C.-Y. Yu, Q.-C. Zeng, C.-F. Huang, C.-C. Tan, W.-L. Hsu, N. Y.-C. Liu, and C.-M. Wang, "Binocular thermal metalens for gesture recognition," Results in Physics, p. 108592, 2026.

[8] F. Luo, A. Li, Z. Huang, K. Wu, B. Jiang, Z. Sun, and L. Wang, "MTxLSTM: Multi-task learning for gesture recognition and person identification using a miniature radar sensor," IEEE Transactions on Mobile Computing, 2026.

[9] N. Alotaibi, R. Al-Dayil, N. O. Aljehane, and M. Rizwanullah, "Enhanced feature fusion with hand gesture recognition system for sign language accessibility to aid hearing and speech impaired individuals," Scientific Reports, 2026.

[10] Y. Feng, J. Qian, W. Zuo, Y. Tan, L. Li, and Z. Shui, "2MCD-YOLO11: A lightweight algorithm based on cross-scale feature fusion for gesture recognition," The Journal of Supercomputing, vol. 82, no. 3, p. 153, 2026.

[11] X. Huang, H. Zhang, J. Ge, and W. Tang, "YOLO-B2DV: An integrated real-time framework for detection, tracking and recognition of catenary pole number plates in railway video streams," IEEE Access, 2026.

[12] E. Wang, Q. Wang, S. Xu, H. Zhao, J. Yang, B. Shao, P. Qu, and L. Na, "BGR-YOLO: An obstacle detection algorithm based on multi-scale feature fusion for UAVs in low-altitude complex environments," Measurement Science and Technology, vol. 37, no. 3, p. 035406, 2026.

[13] P. Padhi and M. Das, "Hand gesture recognition using DenseNet201-MediaPipe hybrid modelling," in Proc. 2022 Int. Conf. on Automation, Computing and Renewable Systems (ICACRS), 2022, pp. 995–999.

[14] Q. Ye, Y. Q. Niu, A. Y. C. Nee, and S. K. Ong, "Spatiotemporal perception and motion similarity learning for real-time assembly verification," Robotics and Computer-Integrated Manufacturing, vol. 100, p. 103258, 2026.

[15] Z. Song, X. Zhang, and P. Tan, “YOLO-Fast: A lightweight object detection model for edge devices,” The Journal of Supercomputing, vol. 81, no. 5, p. 724, 2025.
[16] L. Dang, Z. Li, S. Li, B. Qiao, and L. Zhou, “Effective plug-and-play lightweight modules for YOLO series models,” The Journal of Supercomputing, vol. 81, no. 3, p. 493, 2025.
[17] Ultralytics, “Hand Keypoints Dataset – Dataset Structure,” 2026. [Online]. Available: https://docs.ultralytics.com/datasets/pose/hand-keypoints/. Accessed: Feb. 24, 2026.
[18] J. C. Gower, “Properties of Euclidean and non-Euclidean distance matrices,” Linear Algebra and its Applications, vol. 67, pp. 81–97, 1985.
[19] A. Pavlovych and W. Stuerzlinger, “The tradeoff between spatial jitter and latency in pointing tasks,” in Proc. 1st ACM SIGCHI Symp. on Engineering Interactive Computing Systems (EICS ’09), 2009, pp. 187–196.
[20] G. Casiez, N. Roussel, and D. Vogel, “1 euro filter: A simple speed-based low-pass filter for noisy input in interactive systems,” in Proc. SIGCHI Conf. on Human Factors in Computing Systems (CHI ’12), 2012, pp. 2527–2530.

Biography

Rakib Ahammed Diptho completed his Higher Secondary Certificate (HSC) from Dhaka Residential Model College, Mohammadpur, Dhaka. After completing his secondary education, he earned a B.Sc. in Computer Science and Engineering from BRAC University with excellent academic performance. Following his undergraduate studies, he completed an M.Sc. in Computer Science and Engineering from Jahangirnagar University. He is an Artificial Intelligence (AI) researcher and is currently serving as a full-time Research Assistant in the Department of Computer Science and Engineering at Jahangirnagar University (ADB-supported ICSETEP). His research interests span artificial intelligence, machine learning, computer vision, explainable artificial intelligence (XAI), predictive modeling, and intelligent systems, with applications in healthcare, agriculture, transportation, and environmental sustainability. He has authored numerous peer-reviewed publications in internationally recognized journals and conference proceedings and serves as a reviewer for several international scientific journals.

Safiul Haque Chowdhury is an M.Sc. student in Computer Science and Engineering at Jahangirnagar University and an accomplished Bangladeshi researcher specializing in machine learning, deep learning, and Explainable AI (XAI) for predictive healthcare informatics. Having earned his Bachelor of Science in Computer Science and Engineering from Bangladesh University, his academic and professional efforts focus on developing advanced predictive models for conditions such as diabetes, stroke, and liver disease. Alongside his studies, Safiul serves as a Full-Time Research Assistant at Jahangirnagar University under the ADB-supported ICSETEP initiative, has valuable industry experience as a former Machine Learning Engineer and Team Lead at Orion Informatics, and actively supports the global scientific ecosystem as an editorial board member of the Journal of Artificial Intelligence and Digital Health as well as a peer reviewer for prominent international journals.

Md. Abdullah Al Mamun completed his Higher Secondary Certificate (HSC) from Nilphamari Government College, Bangladesh. He received his B.Sc. degree in Computer Science and Engineering from Jahangirnagar University, Bangladesh, and is currently pursuing his M.Sc. degree in Computer Science and Engineering at the same institution. He has conducted research in the areas of the Internet of Things (IoT) and Blockchain Technology and is currently engaged in research focusing on Artificial Intelligence, Machine Learning, Robotics, and IoT. His research interests include Artificial Intelligence, Machine Learning, Robotics, the Internet of Things (IoT), Blockchain Technology, and Computer Vision.

Md. Shakhawat Hosen completed his Higher Secondary Certificate (HSC) from Comilla Government College, Cumilla, Bangladesh. He received his B.Sc. degree in Computer Science and Engineering from Jahangirnagar University, Dhaka, Bangladesh, and is currently pursuing an M.Sc. degree in Computer Science and Engineering at the same institution. His research interests include artificial intelligence, machine learning, computer vision, deep learning, and the Internet of Things (IoT), with a particular focus on intelligent vision systems and robust hand tracking under occlusion. In addition to his research, he has professional experience in backend software development and has solved over 800 algorithmic problems on competitive programming platforms.

Dr. Md. Shamsur Rahman is a Bangladeshi physician who serves as the Chief Medical Officer (Acting) and Director of the Medical Centre at Jahangirnagar University in Savar, Dhaka. He received his medical training at Mymensingh Medical College, one of the oldest and most respected medical institutions in Bangladesh. At Jahangirnagar University, Dr. Rahman leads the Medical Centre, which delivers healthcare to the university's students, faculty, and staff through its nursing service, pharmacy (medicine supply), and laboratory (pathological testing) services. As the institution's senior medical officer, he oversees day-to-day clinical care and campus health responses, and he has been cited in national media in his official capacity on matters of student and staff wellbeing, including campus medical emergencies and seasonal health concerns.

Sarnali Basak is an Associate Professor in the Department of Computer Science and Engineering at Jahangirnagar University with over 14 years of academic and research experience. She earned her B.Sc. (2005–2009) and M.Sc. (2010–2012) in Computer Science and Engineering from Jahangirnagar University, and later completed a second M.Sc. in Cognitive Science at the University of Edinburgh (2015–2016) as a Commonwealth Scholar. Specializing in artificial intelligence, cognitive science, and human-computer interaction, she has a robust publication record featuring research on machine learning applications, cyberbullying detection, and predictive modeling in healthcare. She notably serves as a Co-Principal Investigator (Co-PI) under the ADB-supported ICSETEP initiative, where she leads impactful, technology-driven societal projects, including the development of a gesture-based intelligent wheelchair to assist physically challenged individuals in Bangladesh.

Dr. Md. Abul Kalam Azad is a Professor and currently serving as the Chairman of Department of Computer Science and Engineering at Jahangirnagar University, bringing over 26 years of extensive teaching, leadership, and research experience to the institution. Additionally, he is leading the prestigious ICSETEP-D-67 project under the finance of ADB as a Principal Investigator. He began his academic journey by earning his B.Sc. in Electronics and Computer Science from Jahangirnagar University (1993-1996), where he secured the 1st Class 1st position and was awarded the prestigious President Gold Medal. He subsequently completed his M.Sc. in Information Technology with distinction marks at the Royal Institute of Technology (KTH), Sweden (2006-2008). Demonstrating a deep commitment to advanced research, Dr. Azad holds two Ph.D. degrees: one in Cognitive Radio Networks from Jahangirnagar University (2021) and another in Wireless Sensor Networks from the University of Ulsan, South Korea (2022), supported by the ICT Fellowship and Hyundai Heavy Industries. His research expertise focuses on wireless communications, machine learning, and computer security, resulting in a strong publication record of over 32 articles in internationally recognized journals. Throughout his tenure at Jahangirnagar University, where he was promoted to Full Professor in 2017, he has guided the department's academic vision, pioneered transmission protocols, improved network infrastructures, and mentored the next generation of engineers and researchers.