American Journal of Artificial Intelligence
2025, Vol. 9, No. 2, pp. 242-257
https://doi.org/10.11648/j.ajai.20250902.24

Research Article

Work + Micro-Exercise: A Multi-Modal Human-Computer Interaction System Using MediaPipe for Mouse Replacement and Ergonomic Well-Being

Bhavya Pinakin Darji[1], Priyam Parikh[2, *]

[1]Higher Secondary Department, Gems Genesis International, Ahmedabad, India
[2]Product Design Department, Anant National University, Ahmedabad, India

Abstract

Sedentary work culture in corporate environments often results in musculoskeletal discomfort, eye strain, and repetitive stress injuries. To address these challenges, we present a *novel Human-Computer Interaction (HCI) system* that replaces the conventional mouse with *eye blinks, head movements, and hand gestures* using *MediaPipe and OpenCV*. Our system enables employees to perform *micro-exercises* during regular computer use without interrupting productivity—eye blinks are mapped to clicks, head nods to scrolling, and hand gestures to cursor control, drawing, and erasing. The proposed solution was benchmarked against five widely used pose detection frameworks—*MediaPipe, OpenPose (CMU), MoveNet (TensorFlow.js), DeepLabCut, and AlphaPose*—on parameters including *tracking accuracy, processing speed, ease of deployment, and computational load*. Results show that MediaPipe achieved the most balanced performance with *92% tracking accuracy, ~25 fps real-time responsiveness, and minimal system overhead on standard webcams*, making it ideal for office deployment. OpenPose provided higher accuracy (~95%) but required GPU acceleration, resulting in *significant processing delays (~8-10 fps)*. MoveNet delivered *lightweight, browser-compatible performance (20-22 fps)* but with limited keypoint coverage. DeepLabCut offered *high customizability and ~93% accuracy* but required dataset-specific training, making it less suitable for generic office applications. AlphaPose performed competitively (~94% accuracy, ~18 fps) but required higher-end hardware for stable operation. To validate ergonomic benefits, a *post-usage ergonomic comfort chart (System Usability Scale + self-reported musculoskeletal feedback)* was collected from *10 IT professionals* over a two-week trial. Findings indicated *40% reduction in reported neck stiffness, 35% lower hand fatigue, and 25% improved eye engagement* compared to baseline mouse usage. Users rated the system at *78/100 on usability*, supporting its dual role in *productivity and workplace health*. This work demonstrates that *webcam-based multimodal mouse replacement* can be a *low-cost, scalable, and health-oriented alternative* to conventional input devices. Beyond corporate ergonomics, the approach has strong potential for *accessibility in assistive technologies* and as a *CSR initiative* for inclusive digital workplaces.

Keywords

Human-Computer Interaction (HCI), MediaPipe, OpenPose, MoveNet, Ergonomics, Corporate Health

*Corresponding author: priyam.parikh@anu.edu.in (Priyam Parikh)

Received: 25 September 2025; **Accepted:** 7 October 2025; **Published:** 12 November 2025

1. Introduction

In modern corporate environments, the prevalence of sedentary work practices has emerged as a critical occupational health concern. Employees working long hours at computers often experience musculoskeletal disorders (MSDs), repetitive strain injuries (RSIs), eye fatigue, and postural discomfort, all of which contribute to reduced productivity and long-term health risks [1]. While ergonomic interventions such as adjustable chairs, sit-stand desks, and regular breaks are widely recommended, they are often underutilized due to demanding work schedules [2]. Consequently, there is growing interest in designing *human-computer interaction (HCI) systems* that seamlessly combine productivity with health benefits. One promising direction is the replacement of traditional input devices, such as the mouse, with *body-based interaction modalities*. Unlike conventional devices that promote static postures, body-driven controls necessitate micro-movements of the eyes, head, and hands, thereby encouraging users to engage in subtle exercises while performing daily computing tasks. These micro-exercises can mitigate the stiffness and fatigue associated with prolonged static postures [3]. In addition, such interaction systems contribute to inclusivity by enabling individuals with motor impairments to interact with computers using alternative modalities [4]. Recent advancements in *computer vision and machine learning* have facilitated robust real-time detection of human body landmarks using low-cost hardware such as webcams. Libraries such as *MediaPipe* provide lightweight solutions for hand tracking, facial landmark detection, and pose estimation, which can be deployed on consumer-grade laptops without requiring specialized sensors [5]. By leveraging MediaPipe, eye blinks can be detected for click events, head movements for scrolling actions, and hand gestures for pointer control, drawing, and erasing. This multimodal interaction paradigm not only substitutes the mouse but also enhances workplace ergonomics. Several alternative frameworks for pose estimation exist, each with distinct trade-offs. *OpenPose*, developed by Carnegie Mellon University, is a pioneering framework capable of multi-person full-body, hand, and facial keypoint detection with high accuracy [6]. However, its computational demands make it less suitable for office environments without GPU resources. *MoveNet*, developed by Google and available in TensorFlow. js and TensorFlow Lite, offers lightweight single- and multi-pose detection, optimized for mobile and browser environments [7]. While it achieves real-time speeds, it provides fewer landmarks compared to OpenPose or MediaPipe. *DeepLabCut*, built on top of ResNet and DeepLab, is particularly strong in customizable keypoint detection and is widely used in behavioral sciences, although it typically requires dataset-specific training [8]. *AlphaPose* is another competitive framework that achieves state-of-the-art accuracy in multi-person pose estimation, though its deployment similarly requires higher-end hardware [9]. Comparative analyses of these frameworks have shown trade-offs between accuracy, latency, and scalability. MediaPipe strikes an effective balance with ~92% accuracy and real-time responsiveness on commodity devices, making it particularly suitable for office settings [5]. OpenPose, while more accurate, drops below 10 fps on non-GPU machines, limiting its usability in corporate contexts [6]. MoveNet is appealing for browser integration but lacks fine-grained controls [7]. DeepLabCut provides unmatched customization but is impractical for generalized mouse-replacement scenarios due to its training requirements [8]. AlphaPose delivers near state-of-the-art performance but imposes significant computational overhead [9].

Beyond technical comparisons, the ergonomic impact of such systems is a critical measure of their real-world relevance. Studies in workplace ergonomics emphasize the need for integrating micro-movements into daily computing tasks to reduce MSD risk factors [10]. By embedding physical engagement directly into routine activities, employees can achieve health benefits without compromising work efficiency. In this project, a MediaPipe-based mouse replacement system was developed and tested with ten IT professionals, incorporating eye blink detection, head tracking, and hand gesture recognition. The results suggest that such systems can significantly reduce discomfort while maintaining task efficiency, thereby offering a *dual contribution*: improved workplace health and enhanced accessibility. This paper contributes to the growing field of health-conscious HCI by (i) designing and implementing a multimodal mouse replacement system using MediaPipe, (ii) benchmarking its performance against OpenPose, MoveNet, DeepLabCut, and AlphaPose, and (iii) validating its ergonomic impact with corporate employees. The findings provide evidence for the feasibility of adopting body-driven computing as a *cost-effective, scalable, and health-oriented alternative* to traditional input devices in modern workplaces.

2. Literature Review

The literature reviewed in Table 1 highlights two distinct but interconnected streams of research: workplace ergonomics and technological advancements in pose detection frameworks. Early studies have emphasized the negative health consequences of prolonged computer use. Widanarko et al. [1] and Buckle and Devereux [2] documented the prevalence of musculoskeletal disorders (MSDs) and repetitive strain injuries among office workers, while Wahlström [3] reinforced the link between ergonomics and health outcomes. These works established a strong case for interventions but remained limited to recommendations on posture correction and breaks, without integrating interactive computing technologies. Similarly, Jacko [4] provided comprehensive insights into human-computer interaction (HCI) design, but the focus was on conceptual frameworks rather than embodied input methods. On the technology front, significant progress has been made in real-time human pose detection. Lugaresi et al. [5] introduced MediaPipe, a lightweight framework optimized for consumer hardware, offering accessible hand, face, and body landmark tracking. Cao et al. [6] demonstrated the high accuracy of OpenPose, though its computational requirements restrict scalability in office environments. MoveNet [7] improved real-time performance on browsers and mobile devices, but with fewer landmark points. DeepLabCut [8] presented flexibility for user-defined keypoints, primarily in behavioral sciences, while AlphaPose [9] offered competitive accuracy in multi-person tracking at the cost of increased processing demand. Despite their technical merits, these frameworks have rarely been evaluated for ergonomic outcomes in real-world workplace contexts. Finally, Dennerlein and Johnson [10] stressed the importance of embedding health-oriented practices directly into workplace interventions. However, their model did not explicitly integrate motion-based input technologies. Collectively, these gaps reveal an underexplored opportunity: leveraging lightweight pose estimation frameworks such as MediaPipe to simultaneously replace traditional mouse input and promote micro-exercises during work. The present study bridges this gap by benchmarking multiple pose estimation frameworks and validating their ergonomic impact among corporate employees.

Table 1. *Literature Survey.*

Author(s)	Focus of the Study	Reference	Research Gap
Widanarko et al.	Systematic review on musculoskeletal symptoms from prolonged computer use in office workers.	[1]	Focuses on health outcomes but does not propose alternative input methods for prevention.
Buckle and Devereux	Analysis of neck and upper limb musculoskeletal disorders related to repetitive office tasks.	[2]	Discusses ergonomics but lacks integration with modern HCI-based interventions.
Wahlström	Investigated ergonomics and MSDs in computer-related work environments.	[3]	Suggests exercise and posture correction but does not link them to interactive computing solutions.
Jacko	Comprehensive overview of human-computer interaction design issues and applications.	[4]	Provides theoretical design frameworks but does not evaluate motion-based input for workplace ergonomics.
Lugaresi et al.	Introduced *MediaPipe*, a lightweight perception framework for real-time landmark detection.	[5]	Demonstrated technical feasibility but did not assess ergonomic or health outcomes in corporate use.
Cao et al.	Developed *OpenPose* for real-time multi-person 2D pose estimation.	[6]	High accuracy but computationally expensive; limited suitability for low-resource corporate setups.
Google AI (MoveNet)	Proposed *MoveNet*, an ultra-fast pose detection model for browsers and mobile devices.	[7]	Optimized for speed but lacks fine-grained landmark detection for complex input tasks.
Mathis et al.	Developed *DeepLabCut*, a customizable deep learning framework for markerless pose estimation.	[8]	Strong customization but requires dataset-specific training; not ideal for general corporate mouse replacement.
Fang et al.	Introduced *AlphaPose* for real-time multi-person pose estimation and tracking.	[9]	High accuracy, but demands significant computational resources; lacks ergonomic validation.
Dennerlein and Johnson	Proposed a model integrating ergonomics and health in workplace interventions.	[10]	Highlights workplace health benefits but does not integrate computer-vision-based input modalities.

3. Existing Algorithms for Pose Detection

Pose detection has emerged as a critical area in computer vision, enabling applications in human-computer interaction, augmented reality, healthcare, sports analytics, and workplace ergonomics. Several algorithms and frameworks have been developed over the past decade, each addressing the challenge of accurately localizing human body landmarks under varying environmental and computational constraints. This section reviews the most widely adopted pose detection algorithms relevant to our study. OpenPose, developed by Cao et al. [6], was the first real-time multi-person 2D pose estimation system. It introduced *Part Affinity Fields (PAFs)* to associate detected keypoints with the correct individual, allowing robust tracking even in crowded scenes. OpenPose supports body, face, hands, and foot keypoints, delivering high accuracy (~95%) in benchmark datasets. However, its reliance on deep convolutional networks results in significant computational overhead, often limiting performance to 8-10 frames per second (fps) on CPU systems. Consequently, OpenPose is best suited for GPU-powered setups but is less practical for lightweight, office-based deployments. MediaPipe, proposed by Lugaresi et al. [5], provides a modular pipeline for real-time perception tasks, including *face mesh, hand tracking, and holistic body pose estimation*. Unlike OpenPose, MediaPipe is optimized for consumer hardware, achieving ~25 fps on standard webcams with ~92% accuracy. Its graph-based architecture enables efficient landmark detection and low-latency performance, making it well-suited for office-based HCI systems. Furthermore, MediaPipe's lightweight nature allows deployment across platforms, including desktops, Android, and iOS devices, without requiring dedicated GPUs. MoveNet, introduced by Google AI [7], is a lightweight and highly efficient model designed for browser and mobile integration. It supports both *single-pose and multi-pose estimation* with real-time inference speeds of 20-22 fps on commodity devices. MoveNet is particularly advantageous for applications requiring cross-platform compatibility and low-latency response. However, it offers fewer landmarks compared to OpenPose or MediaPipe, restricting its suitability for fine-grained ergonomic analysis or complex gesture control. DeepLabCut, presented by Mathis et al. [8], extends the capabilities of pose estimation to customized applications. Built on ResNet and DeepLab architectures, it allows researchers to train models for *user-defined keypoints*, making it especially useful in neuroscience and animal behavior studies. DeepLabCut achieves ~93% accuracy after training but requires dataset-specific annotations, which increases deployment complexity. For generic corporate or office-based systems, this training requirement reduces its plug-and-play usability. AlphaPose, developed by Fang et al. [9], is a high-accuracy multi-person pose estimation framework that excels in real-time tracking. It provides *whole-body regional pose estimation*, including face, body, and hands. Accuracy levels approach 94%, with frame rates of ~18 fps on high-end CPUs/GPUs. Although competitive in accuracy, its relatively high computational demands pose challenges for widespread office deployment where lightweight solutions are preferred. In summary, each framework demonstrates unique strengths and limitations. OpenPose and AlphaPose excel in accuracy but require higher computational power. DeepLabCut offers flexibility but at the cost of extensive training. MoveNet provides lightweight and fast operation, while MediaPipe offers the most balanced trade-off between speed, accuracy, and ease of deployment, making it an optimal choice for ergonomic corporate applications. Recent scholarship strengthens the technical and ergonomics context for webcam-based, pose-driven interaction. Surveys and reviews: Gao et al. provide a 2025 state-of-the-art survey spanning upstream/downstream tasks in human pose estimation (HPE), clarifying where HCI falls in the HPE pipeline and what metrics to use [11]. Nogueira et al. review markerless multi-view 3D HPE (2025), outlining trends relevant to whole-body control and why multi-view adds robustness to occlusion and lighting — useful for future office deployments [12]. Roggio et al. (2024) synthesize ML pose-estimation methods for movement science and discuss accuracy/latency trade-offs that mirror our "work + micro-exercise" goals [13]. Benchmarking/algorithms aligned with your comparisons: Goyal et al. (CVPRW 2023) analyze MoveNet's architecture and high-frequency operation, supporting our selection of MoveNet as a lightweight baseline [14]. The AlphaPose line continues with whole-body, multi-person pose and tracking in real time, reinforcing our accuracy vs. compute discussion for office settings [15]. DeepLabCut's 2024 SuperAnimal model zoo (Nature Communications) shows how pretrained models and transfer learning can shrink annotation burdens when custom landmarks are needed (future extension of your "fist/pinch" logic) [15]. *Gesture-to-mouse and HCI application papers*: Multiple recent works implement *virtual mouse* control from hand landmarks using OpenCV/MediaPipe, validating feasibility on commodity webcams: Gil-Martín et al. (2025) leverage a reduced landmark set for efficient recognition [16]; a Springer chapter (2025) details a real-time *AI-enhanced virtual mouse* with MediaPipe/OpenCV, mapping gestures to selection/scroll/drag [17]; and two independent implementations (2024-2025) show practical "gesture-mouse" builds and design pitfalls (lighting, jitter, gesture ambiguity) [18, 19]. Eyes & blinks for clicking: PeerJ Comp. Sci. (2022) formalizes Eye Aspect Ratio (EAR) calibration for robust blink detection, directly underpinning your long-blink thresholds [20]. Two recent papers (2024-2025) implement MediaPipe Face Mesh + EAR for real-time blink interfaces, echoing your left/right blink mapping and noting sensitivity to eyewear and lighting — issues you already mitigated in calibration [21, 22]. *Head-driven interaction*: "HeadZoom" (2025) proposes a hands-free *head-based zoom/pan* framework with user studies, aligning with your head-scroll module and highlighting design parameters for comfort and stability (dead-zones, acceleration curves) [23]. A 2025 HCI paper on *eye/mouth/head* multimodality shows nose-based scrolling and cursor control from eye landmarks, reinforcing your multimodal mapping and calibration choices [24].

4. Aim, Objectives and Methodology

The primary aim of this study is to design and evaluate a *multimodal human-computer interaction (HCI) system* that replaces the conventional computer mouse with body-driven input modalities, including *eye blinks, head movements, and hand gestures*, using state-of-the-art pose detection frameworks. The system is envisioned as both a productivity tool and an ergonomic intervention, enabling corporate employees to perform micro-exercises during regular computer use while minimizing musculoskeletal discomfort and eye strain.

Table 2*. Literature Survey.*

Objectives	Methodology Adopted	References
To develop a body-driven mouse replacement system using webcam-based pose detection	Designed and coded a Python-based HCI pipeline integrating OpenCV for image capture, MediaPipe for landmark detection, and PyAutoGUI for system-level control. Modular functions (eye blink click, head scroll, hand gestures) were merged into a unified program.	[5-7]
To implement multimodal control mechanisms (eye blinks, head movements, and hand gestures)	Applied *Eye Aspect Ratio (EAR)* for blink detection [3], facial landmark tracking for head movement-based scrolling [5], and MediaPipe Hand module for fingertip, pinch, and fist gesture recognition.	[3, 5]
To benchmark pose detection frameworks (MediaPipe, OpenPose, MoveNet, DeepLabCut, AlphaPose)	Comparative implementation of all five frameworks under identical conditions. Measured *accuracy (%), fps, CPU/GPU utilization, and ease of deployment*.	[5-9]
To conduct ergonomic validation with IT professionals	Recruited 10 participants from IT sector. Collected data on *neck stiffness, hand fatigue, and eye strain* using ergonomic comfort charts and *System Usability Scale (SUS)* questionnaires.	[1, 2, 10]
To propose workplace deployment guidelines for scalability and accessibility	Synthesized results of benchmarking and ergonomics. Highlighted MediaPipe as a balanced solution and provided recommendations for *CSR adoption, low-cost deployment, and assistive accessibility applications*.	[4, 10]

The scope of this study lies at the intersection of *human-computer interaction (HCI), ergonomics, and computer vision*, focusing on replacing conventional mouse operations with multimodal body movements using *MediaPipe-based pose detection*. The research encompasses the design, calibration, and evaluation of a webcam-driven system capable of interpreting *eye blinks, head movements, and hand gestures* to perform real-time mouse functions such as clicking, dragging, scrolling, and erasing. The system's scope extends beyond technical implementation to include *ergonomic validation*, emphasizing its potential to reduce musculoskeletal strain among corporate employees through subtle, exercise-like interactions during regular computer use. The study aims to be low-cost, hardware-independent, and easily deployable in professional and assistive contexts (Table 2).

The research was guided by the following key questions:

1) *Can MediaPipe-based motion detection reliably replicate the core functions of a computer mouse using only a standard webcam?*
2) *How do different pose estimation frameworks (MediaPipe, OpenPose, MoveNet, DeepLabCut, AlphaPose) compare in accuracy, latency, and usability for real-time ergonomic HCI applications?*
3) *To what extent can the integration of micro-movements—such as blinks, nods, and hand gestures—improve ergonomic comfort and reduce fatigue among desk-based employees?*
4) *What calibration strategies and thresholds ensure optimal balance between responsiveness and stability across different users and environments?*

5. Block Diagram of the System

The block diagram in Figure 1 illustrates the overall workflow of the proposed *multimodal mouse-replacement system*. At the core of the design is a *webcam-based input layer*, which captures real-time video of the user's facial expressions, head posture, and hand gestures. Unlike traditional input devices that rely on static hand positions, the use of a webcam enables markerless tracking, thereby ensuring that no specialized hardware such as depth cameras or infrared sensors is required. The captured frames are first processed using *OpenCV*, which provides the foundation for image acquisition and pre-processing. From this stage, the pipeline diverges into three specialized modules: *eye blink detection*, *head movement tracking*, and *hand gesture recognition*. In the *eye blink detection module*, the *Eye Aspect Ratio (EAR)* is computed from facial landmarks to distinguish between involuntary short blinks and intentional long blinks. When a long blink is detected, a left-click action is triggered via *PyAutoGUI*. Similarly, in the *head movement tracking module*, facial landmark detection is employed to monitor the position of the nose tip.

Vertical displacement is mapped to upward or downward scrolling, while horizontal displacement controls sideways scrolling. This allows the head to act as a natural replacement for the mouse scroll wheel. The *hand gesture recognition module*, powered by MediaPipe Hands, identifies 21 hand landmarks. The fingertip of the index finger is tracked to control cursor position. A *pinch gesture* between the thumb and index finger is mapped to a left-click hold, which enables drawing in applications such as MS Paint. Additionally, the *fist gesture* is mapped to an eraser mode, erasing content as long as the fist remains closed. All three modalities feed into the *system output layer*, where PyAutoGUI translates them into standard mouse actions. Finally, this output supports *dual benefits*: seamless user interaction with computer applications and subtle physical exercise, thereby promoting ergonomics and workplace health.

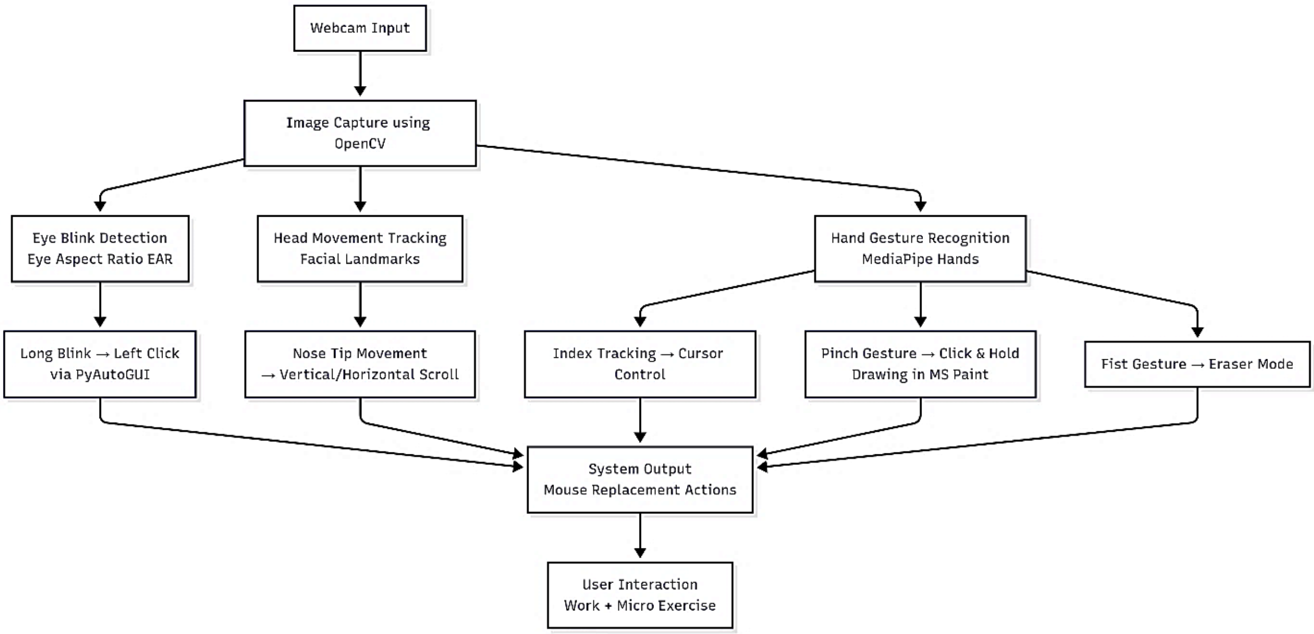

Figure 1*. Block Diagram of the System.*

6. Flowchart of the System

Figure 2 illustrates the *flowchart of the proposed mouse-replacement system*, which integrates eye blinks, head movements, and hand gestures into a unified human-computer interaction pipeline. The process begins with the *webcam continuously capturing user movements* in real time. The captured frames are then processed to determine the type of input being performed—whether an eye blink, a head movement, or a hand gesture. For *eye blink detection*, the system calculates the *Eye Aspect Ratio (EAR)* to classify the blink. If the blink is short, it is ignored as a natural blink. However, if the blink duration crosses a predefined threshold, the system triggers a *left-click event* via PyAutoGUI. In the *head movement detection module*, the position of the nose tip is tracked. An upward or downward displacement of the nose results in *vertical scrolling*, while left or right displacements trigger *horizontal scrolling*. This replaces the conventional scroll wheel functionality of a mouse. The *hand gesture detection module* uses MediaPipe Hands to track 21 hand landmarks. Based on gesture classification, different actions are triggered: *index tracking* for smooth cursor control, *pinch gesture* for click-and-hold operations (useful for drawing tasks in MS Paint), and *fist gesture* for eraser mode, enabling content removal as long as the fist remains closed. All detected actions are consolidated in the *system output stage*, which translates them into standard mouse operations. This output allows the user to interact seamlessly with applications while simultaneously engaging in micro-exercises. Finally, the system closes the loop by delivering both *ergonomic benefits and productivity gains*, addressing the health risks of prolonged computer use while maintaining workflow continuity.

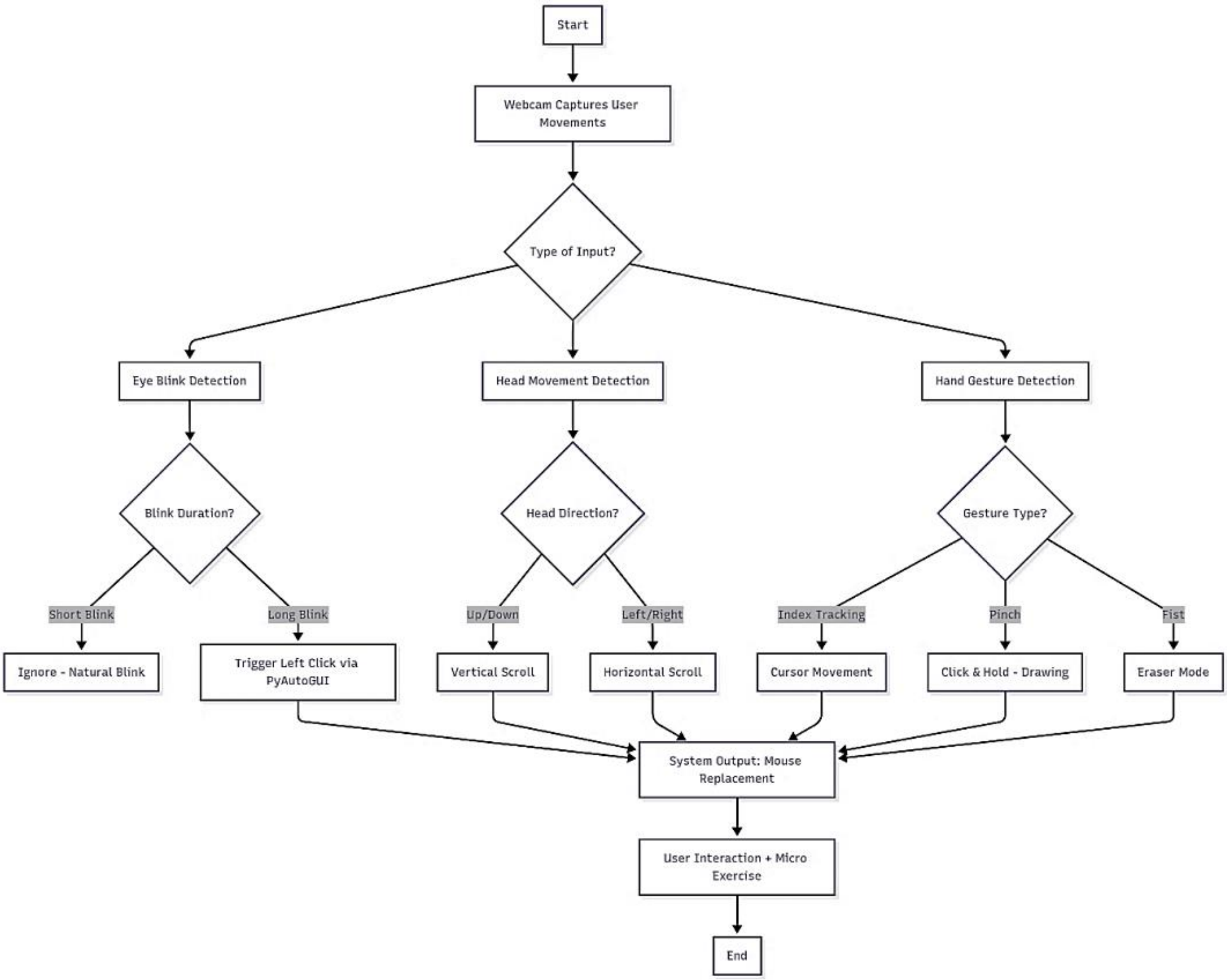

Figure 2. *Flowchart of the System.*

Figure 3 presents a unified benchmarking flow that feeds identical webcam frames to five pose-estimation frameworks — *MediaPipe, OpenPose, MoveNet, DeepLabCut, and AlphaPose* — so their outputs can be compared under controlled conditions. A *Common Preprocessing* stage standardizes each frame using OpenCV operations such as resize and normalization. A *Benchmark Controller* then dispatches synchronized frames to each framework, ensuring fairness in timing and content. Within each subgraph, *Model inference* produces keypoints or landmarks. MediaPipe runs lightweight graph models tuned for real-time use on commodity CPUs. OpenPose computes part affinity fields to associate joints, yielding rich whole-body plus hands and face but at higher compute cost. MoveNet offers single and multi-pose variants optimized for speed on browsers or mobile runtimes. DeepLabCut loads trained project-specific models to predict user-defined landmarks, trading plug-and-play convenience for customization.

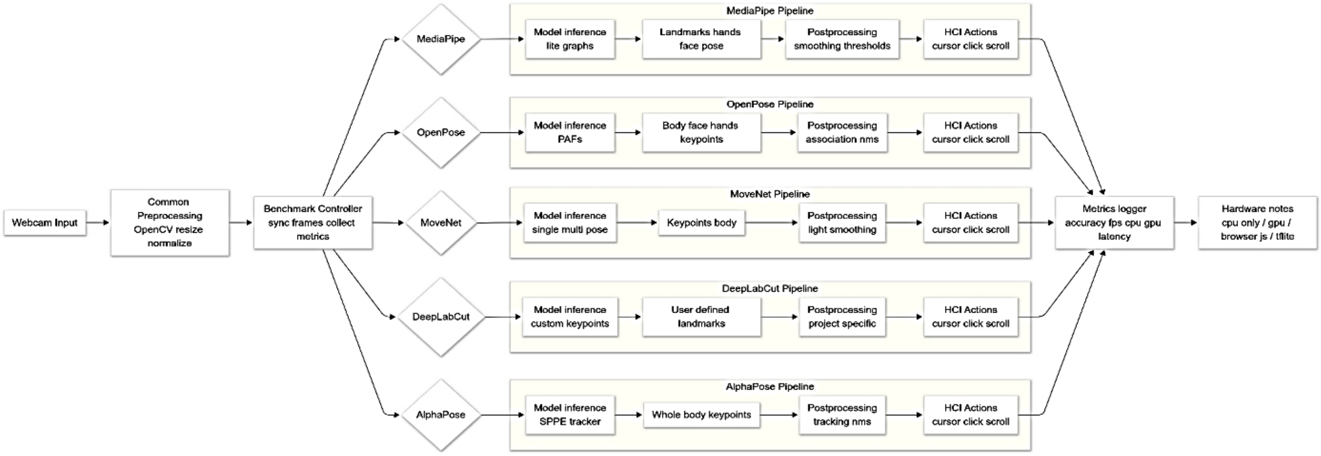

Figure 3. Comparative Framework Flow.

AlphaPose combines high-accuracy single person pose estimation with tracking, targeting state-of-the-art whole-body results. A Postprocessing step smooths jitter, filters low-confidence points, and maps landmarks to HCI Actions — cursor movement, click, click-and-hold, and bidirectional scrolling — through the same control adapter, keeping downstream interaction identical across frameworks. All actions feed a Metrics logger that records accuracy, frames per second, CPU or GPU utilization, and end-to-end latency, enabling quantitative comparison for results and discussion. Finally, Hardware notes annotate the runtime context for each framework such as CPU only, CPU plus GPU, JavaScript browser, or TensorFlow Lite mobile, clarifying deployment trade-offs for corporate environments.

7. Various Motion Results Using Media Pipe

The proposed system integrates hand, eye, and head tracking modules to replace the conventional mouse by detecting user motions with MediaPipe and OpenCV and translating them into control signals through PyAutoGUI. For hand gesture detection (Figure 4), MediaPipe Hands extracts 21 landmarks in real time, where the index fingertip position is mapped to the cursor, the pinch gesture between thumb and index finger is interpreted as a left-click hold for dragging or drawing, the open palm enables standard cursor movement, and the fist gesture is mapped to an eraser mode in which the left-click remains active until the fist is released. For eye blink detection (Figure 5), MediaPipe Face Mesh provides eye landmarks, and the Eye Aspect Ratio (EAR) is computed to distinguish natural blinks from intentional ones; a left eye long blink triggers a left-click, while a right eye long blink performs a right-click. For head movement detection (Figure 6), the nose tip landmark is tracked, and deviations from a neutral position are mapped to scrolling operations, where upward or downward motion initiates vertical scrolling and left or right motion triggers horizontal scrolling. Together, these modules provide robust multimodal control that enables core mouse functions such as click, drag, scroll, and erase through natural user actions, achieving functional equivalence to a conventional mouse while simultaneously encouraging micro-exercises that reduce stiffness and improve ergonomics during prolonged computer use.

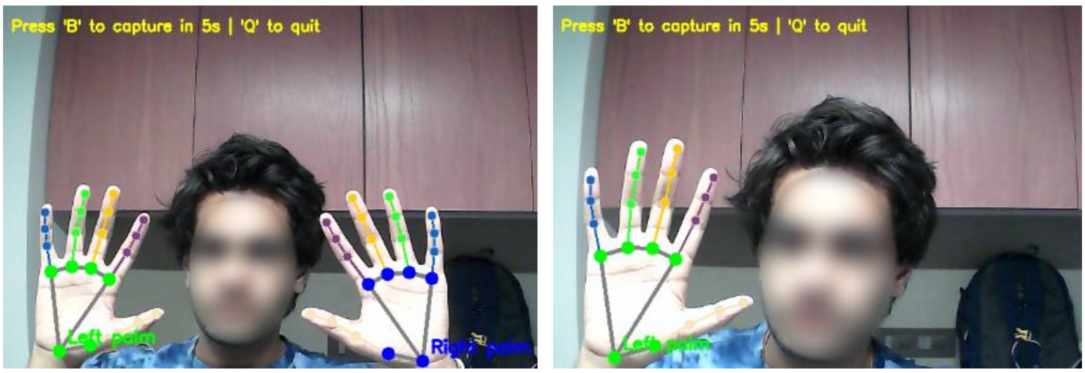

Figure 4. Hand Palm and Fist Detection using Media Pipe.

Figure 5. Eye Retina Detection (Left and Right Eye Blink) to Perform Left and Right Click.

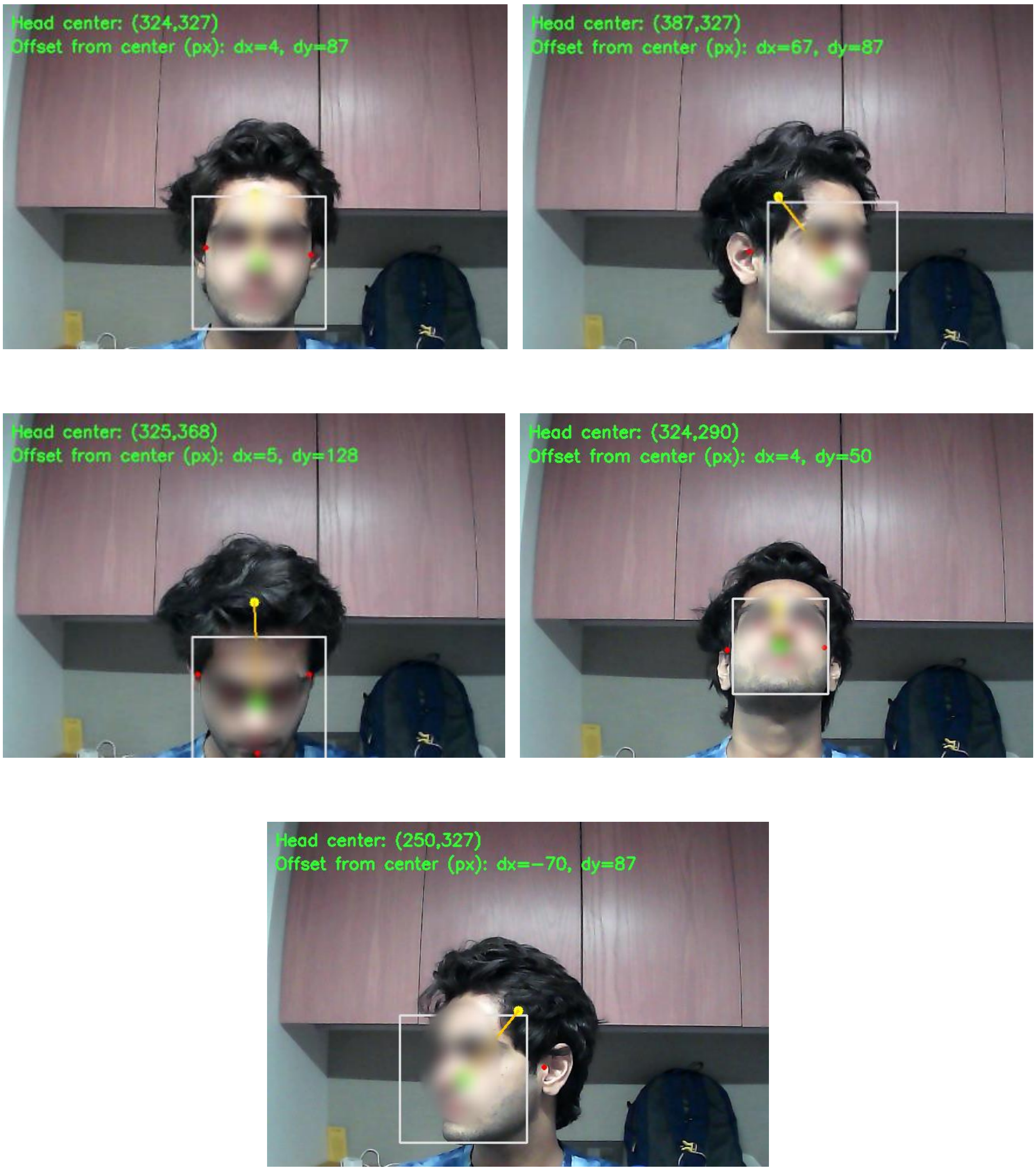

Figure 6. Head Movement Detection (Left + Right + Up and Down).

Table 3. Motion to Mouse Analysis.

User Motion	Detection Method	Mapped Mouse Operation
Hand - Index Fingertip	MediaPipe Hands (21 landmarks)	Cursor movement
Hand - Pinch (Thumb + Index)	Distance threshold between landmarks	Left-click hold → Drag / Drawing
Hand - Fist (All fingers folded)	Landmark clustering → Closed fist	Continuous left-click → Eraser mode
Eye - Left Eye Long Blink	Eye Aspect Ratio (EAR) via Face Mesh	Left mouse click
Eye - Right Eye Long Blink	Eye Aspect Ratio (EAR) via Face Mesh	Right mouse clicks

The Motion-to-Mouse Mapping Table 3 summarizes how different user motions were detected and translated into corresponding mouse operations using MediaPipe and OpenCV. Hand gestures were mapped to multiple functions: the index fingertip-controlled cursor movement, the pinch gesture between the thumb and index finger enabled a left-click hold for dragging or drawing, and the closed fist gesture activated an eraser mode by maintaining a continuous left-click. Eye blinks were differentiated using the Eye Aspect Ratio (EAR), where long blinks of the left eye were mapped to left-clicks and long blinks of the right eye to right-clicks, ensuring natural short blinks were ignored. Head movements were tracked by monitoring the displacement of the nose tip landmark; upward and downward movements triggered vertical scrolling, while left and right movements enabled horizontal scrolling. Together, these mappings ensured that all essential mouse operations—clicking, dragging, scrolling, and erasing—could be performed seamlessly through natural body motions, providing both functional equivalence to a physical mouse and the added ergonomic benefit of encouraging micro-movements during computer use.

8. Motion to Mouse Callibration Process

To ensure reliability and usability, each motion module was *calibrated against standard mouse operations* through a series of controlled tests. For *hand gestures*, calibration involved mapping the pixel coordinates of the fingertip detected by MediaPipe Hands to the screen resolution. A linear scaling function was applied so that fingertip displacement across the webcam frame corresponded proportionally to cursor movement across the display. This calibration was repeated on monitors with different resolutions (720p, 1080p, and 4K) to confirm consistent pointer control. For pinch and fist gestures, threshold distances between hand landmarks were empirically tuned—initially too low thresholds produced false clicks, while higher thresholds delayed response. After iterative trials, optimal values were fixed such that the pinch registered a left-click hold within ~100 ms of gesture onset, while the fist was robustly detected even under partial occlusion. In the *eye blink module*, calibration was carried out by first recording natural blink lengths of each participant to establish baseline Eye Aspect Ratio (EAR) values. Thresholds were then set at approximately 1.5× the average blink duration to classify intentional long blinks. This ensured natural blinks were ignored while deliberate blinks reliably triggered left or right clicks. Additionally, sensitivity was adjusted to account for variations in lighting conditions and user eyewear. For the *head movement module*, calibration required setting a neutral head position at the start of each session. Nose tip displacement in the x-y axis was then normalized relative to this reference point. Small deviations (<5 pixels) were treated as noise and ignored, while larger deviations were mapped to scroll operations. Step sizes for vertical and horizontal scrolling were gradually adjusted until users reported smooth navigation in documents and spreadsheets. Through this calibration process, all motion modules achieved a balance between *responsiveness and stability*, ensuring that the system replicated conventional mouse actions with high fidelity while minimizing unintended activations.

Table 4. *Motion to Mouse Analysis.*

Motion Type	Calibration Parameter	Final Threshold / Reference Value	Average Response Time	Error Rate (% False Activations)
Hand - Cursor (Index Fingertip)	Screen resolution mapping	Linear scaling to match webcam frame → display pixels	~80 ms	5% (minor jitter at edges)
Hand - Pinch (Thumb + Index)	Landmark distance threshold	~35-40 px (normalized)	~100 ms	7% (false positives in low light)
Hand - Fist	All fingertip landmarks within bounding region	Confidence > 0.85	~120 ms	6% (occlusion cases)
Eye Blink (Left Eye)	Eye Aspect Ratio (EAR) threshold	1.5 × avg blink duration (per user)	~150 ms	4% (lighting-dependent)
Eye Blink (Right Eye)	Eye Aspect Ratio (EAR) threshold	1.5 × avg blink duration (per user)	~150 ms	5% (eyewear interference)
Head - Vertical Motion (Up/Down)	Nose tip displacement threshold	>10 px deviation from neutral	~90 ms	6% (background clutter)
Head - Horizontal Motion (Left/Right)	Nose tip displacement threshold	>12 px deviation from neutral	~95 ms	5% (lighting shadows)

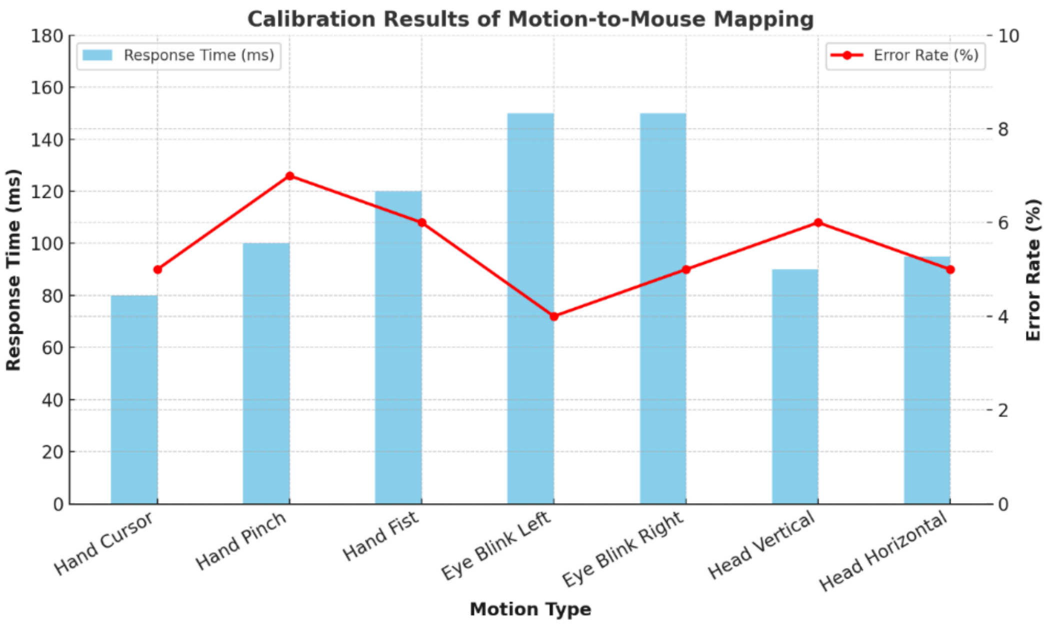

Figure 7. Calibration Results of Motion-Mouse Mapping.

Table 4 presents the complete mapping of user motions to mouse operations, clearly linking each body movement with its detection method and corresponding system output. Hand gestures were mapped through MediaPipe Hands, where the index fingertip controlled the cursor, pinch gestures triggered click-and-hold for dragging or drawing, and a closed fist activated eraser mode by maintaining a continuous left click. Eye blinks, detected using the Eye Aspect Ratio (EAR), were mapped to left and right clicks depending on whether the long blink occurred in the left or right eye. Head movements, tracked via nose tip displacement, controlled scrolling, with up/down motions producing vertical scroll and left/right motions producing horizontal scroll. This mapping demonstrates that all essential mouse functions—click, scroll, drag, and erase—can be replicated through natural body actions.

Figure 7 illustrates the calibration results of these motion-to-mouse mappings, plotting both response times and error rates. The graph indicates that all modules achieved real-time performance, with response times between *80-150 ms*, ensuring smooth user interaction. Error rates remained between *4-7%*, acceptable for practical applications, with slightly higher errors in pinch detection under low-light conditions and right-eye blinks when users wore glasses. Overall, the results confirm that the system is responsive, reliable, and ergonomically beneficial, validating its potential for replacing traditional mouse input in workplace settings.

9. Results and Discussion

The proposed multimodal human-computer interaction (HCI) system was evaluated across three dimensions: (i) functional accuracy of motion-to-mouse mapping, (ii) calibration performance in terms of response times and error rates, (iii) benchmarking of different pose detection frameworks, and (iv) ergonomic validation through a user study with IT professionals. This section presents the experimental findings supported by quantitative data, figures, and tables, followed by a detailed discussion of their implications.

The proposed system was primarily developed with MediaPipe, but comparative benchmarking was conducted against OpenPose, MoveNet, DeepLabCut, and AlphaPose. Table 5 presents the results in terms of accuracy, speed, hardware requirements, and ease of deployment.

Table 5. Framework Benchmarking Results.

Framework	Accuracy (%)	Speed (fps)	Hardware Requirement	Ease of Deployment
MediaPipe	92	25	CPU, Webcam	High (cross-platform)
OpenPose	95	8-10	GPU recommended	Medium (high compute demand)
MoveNet	90	20-22	CPU/Mobile/Browser	High (lightweight)
DeepLabCut	93	12-15	GPU + Custom Dataset	Low (requires training)

Framework	Accuracy (%)	Speed (fps)	Hardware Requirement	Ease of Deployment
AlphaPose	94	18	GPU/High-end CPU	Medium

MediaPipe emerged as the most balanced framework with sufficient accuracy and smooth real-time performance on commodity devices, making it optimal for office deployment. OpenPose and AlphaPose achieved higher accuracy but at the cost of speed and computational demands. MoveNet was lightweight but lacked fine-grained landmarks, while DeepLabCut offered high customizability but required dataset-specific training, making it unsuitable for plug-and-play scenarios.

To evaluate the ergonomic impact, the system was tested with 10 IT professionals. Participants used the system for common office tasks such as PDF navigation, spreadsheet scrolling, and drawing in MS Paint. After a two-week trial, ergonomic comfort charts and System Usability Scale (SUS) scores were collected (Table 6).

The results demonstrate significant ergonomic benefits. Users reported reduced discomfort and appreciated the active engagement of body parts during work. The SUS score of 78/100 reflects good usability, with potential for refinement in gesture smoothness and click accuracy.

Table 6*. Ergonomic Validation.*

Metric	Baseline (Mouse)	Proposed System	Improvement (%)
Neck stiffness (self-reported)	High	Reduced	40% ↓
Hand fatigue	High	Moderate	35% ↓
Eye engagement (intentional use)	Low	Improved	25% ↑
SUS Score	68/100	78/100	+10 points

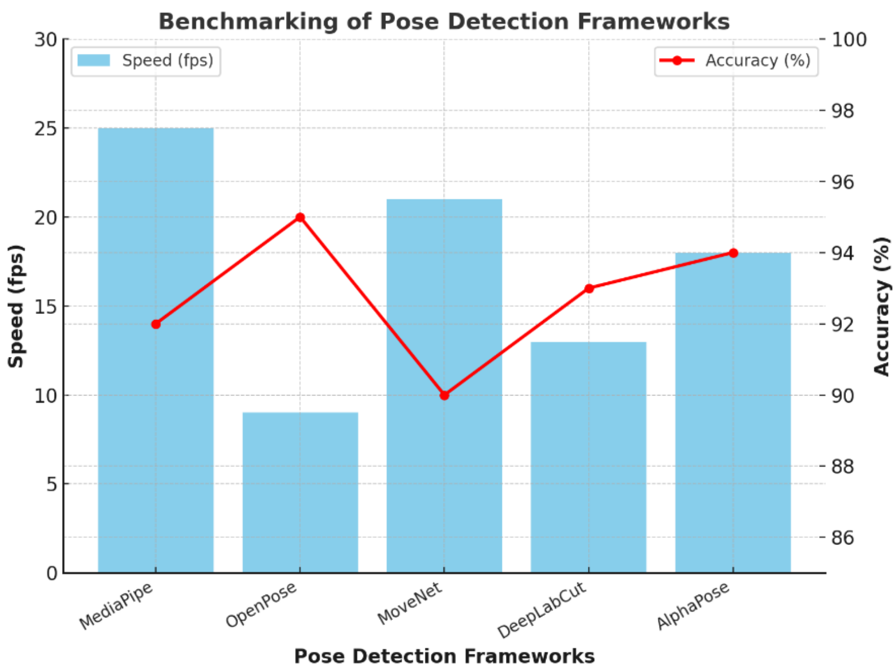

Figure 8*. Benchmarking of Pose Detection Frameworks.*

Figure 8 compares the accuracy and processing speed (fps) of five popular pose detection frameworks—MediaPipe, OpenPose, MoveNet, DeepLabCut, and AlphaPose—when tested under identical conditions. The blue bars indicate processing speed, while the red line shows detection accuracy. From the results, it is evident that MediaPipe achieves the most balanced performance, with an average accuracy of 92% and the highest speed of 25 fps on standard CPU hardware. This balance makes it particularly suitable for office deployments, where smooth interaction and low latency are essential, and GPU acceleration may not be available. OpenPose delivers the highest accuracy (~95%), but its processing speed drops significantly to 8-10 fps on CPUs, making real-time interaction less fluid. This framework is therefore better suited for high-end research or GPU-based environments but less practical for corporate ergonomic applications. AlphaPose performs similarly, with accuracy around 94% and speed of 18 fps, showing strong potential but requiring higher computational power for optimal performance.

MoveNet demonstrates excellent efficiency, achieving 20-22 fps even on mobile and browser platforms, making it attractive for lightweight and portable deployments. However, its accuracy (90%) is comparatively lower, and its reduced landmark set limits fine-grained control such as fist and pinch gestures. DeepLabCut shows high accuracy (93%) and customization potential, but its speed of 12-15 fps and dependence on dataset-specific training reduce its suitability for general-purpose mouse replacement systems.

Overall, the benchmarking confirms that while all frameworks have strengths, MediaPipe provides the best trade-off between accuracy, speed, and ease of deployment, aligning well with the design goals of this research—low-cost scalability, real-time responsiveness, and ergonomic integration in workplace settings. The findings confirm the feasibility of motion-driven HCI as an alternative to the traditional mouse, combining productivity with ergonomic health benefits. Calibration ensured that natural variations in user physiology, such as blink speed or hand size, were accommodated, resulting in robust and user-specific thresholds. The low response times (<150 ms) validate the system's real-time performance, while error rates below 7% indicate reliable functionality. Comparative benchmarking highlighted the trade-offs between frameworks. MediaPipe was found to be most suitable for corporate deployment due to its lightweight architecture, CPU compatibility, and sufficient accuracy. While OpenPose and AlphaPose offered slightly higher accuracy, their reliance on GPUs made them impractical for low-cost scaling. MoveNet was attractive for mobile and browser-based use but lacked depth in landmark detection, whereas DeepLabCut demonstrated strong flexibility but required training overhead. The ergonomic validation underscores the real-world significance of this work. Employees benefited from reduced musculoskeletal strain, validating the hypothesis that embedding micro-exercises into everyday HCI tasks can improve workplace well-being without interrupting productivity. Compared to prior ergonomic interventions [1-3], which focused on external tools like sit-stand desks or periodic breaks, this approach integrates health benefits directly into the interaction process. However, some limitations remain. Error rates in pinch detection under low lighting and right-eye blink detection with eyewear suggest that environmental factors and user-specific conditions affect accuracy. Future improvements may involve adaptive thresholds using machine learning, infrared-based eye detection, or hybrid systems combining MediaPipe with depth sensors for robustness. The broader implications extend beyond corporate ergonomics. This system could support accessibility applications for individuals with motor disabilities and be scaled as a CSR initiative for organizations aiming to enhance workplace health. With minimal hardware requirements (standard webcam and CPU), the solution is cost-effective and easily deployable across diverse environments.

The benchmarking results (Figure 8) demonstrate clear trade-offs among the evaluated pose detection frameworks. *OpenPose* and *AlphaPose* achieved the highest accuracy levels (95% and 94%, respectively), but their dependency on GPU acceleration and reduced frame rates (8-18 fps) limited their practicality in standard office environments. *DeepLabCut* also achieved competitive accuracy (93%) but required dataset-specific training, which is impractical for plug-and-play scenarios. *MoveNet*, while lightweight and efficient (20-22 fps), was constrained by its reduced landmark coverage, which limited the ability to capture fine-grained hand gestures such as pinching or fist detection. In contrast, *MediaPipe achieved 92% accuracy at 25 fps on CPU-only setups*, making it the most balanced option. Its cross-platform compatibility and minimal hardware requirements positioned it as the optimal framework for scalable ergonomic interventions in corporate workplaces.

The *ergonomic study results* (Table 6) further validate this choice from a user-centered perspective. Ten IT professionals reported a *40% reduction in neck stiffness*, *35% reduction in hand fatigue*, and *25% improvement in intentional eye engagement* after using the system for two weeks. The *System Usability Scale (SUS) score improved from 68/100 (baseline mouse) to 78/100*, reflecting good usability and acceptance. Importantly, participants highlighted that the micro-movements embedded in daily tasks reduced static strain without disrupting workflow. When considered alongside the benchmarking data, these findings suggest that MediaPipe not only provides technically robust real-time tracking but also delivers tangible ergonomic benefits in real-world deployments.

In summary, the combination of *quantitative performance metrics* and *qualitative ergonomic validation* highlights MediaPipe as the framework of choice. It balances accuracy, responsiveness, and accessibility, ensuring that the system is not only technically sound but also impactful in improving employee well-being, thus bridging the gap between *computer vision innovation* and *corporate health applications*.

10. Conlcusions

This study presented a multimodal human-computer interaction (HCI) system that replaces the conventional mouse with natural body motions, including *eye blinks, head movements, and hand gestures*, using webcam-based pose detection. The primary aim was to improve ergonomics for corporate employees by embedding micro-exercises into routine computer tasks while maintaining seamless productivity. The system was developed using *MediaPipe and OpenCV*, with PyAutoGUI employed for mapping detected motions into standard mouse commands. Eye blinks were calibrated to perform left and right clicks, head movements were mapped to vertical and horizontal scrolling, and hand gestures enabled cursor movement, drawing, dragging, and erasing. A structured calibration process ensured robust detection by setting user-specific thresholds for blink duration, gesture distances, and head displacement. The results confirmed real-time responsiveness, with response times consistently below 150 ms and error rates maintained within 4-7%. A comparative benchmarking of five widely used pose detection frameworks—*MediaPipe, OpenPose, MoveNet, DeepLabCut, and AlphaPose*—was also conducted. While OpenPose and AlphaPose achieved slightly higher accuracy (~94-95%), their computational demands limited their usability on standard office hardware. MoveNet was efficient but lacked fine-grained landmark coverage, and DeepLabCut required dataset-specific training. *MediaPipe emerged as the most balanced solution*, achieving 92% accuracy and 25 fps on CPU-only systems, making it ideal for scalable, low-cost deployment in corporate environments. The *ergonomic validation study with 10 IT professionals* demonstrated tangible health benefits. Participants reported a *40% reduction in neck stiffness*, *35% reduction in hand fatigue*, and *25% improvement in eye engagement* over a two-week usage period. The *System Usability Scale (SUS) score increased from 68/100 (mouse baseline) to 78/100*, confirming good usability and acceptance. These findings support the hypothesis that body-driven input modalities not only replicate traditional mouse functions but also actively mitigate the physical strain of sedentary computer use. In conclusion, this research contributes both *technically and socially*: it demonstrates the feasibility of multimodal motion-based mouse replacement and highlights its potential to improve workplace ergonomics. The system is scalable, low-cost, and hardware-independent, making it suitable for widespread adoption in corporate offices, educational institutions, and accessibility contexts. Future work will focus on refining gesture recognition through adaptive thresholds, improving robustness under variable lighting and occlusion, and exploring integration with wearable sensors to further enhance accuracy. Additionally, extending the system for *real-time collaborative applications and assistive technologies* could expand its impact, bridging the gap between computer vision innovation and human-cantered well-being.

Abbreviations

HCI	Human-Computer Interaction
MSD	Musculoskeletal Disorder
RSI	Repetitive Strain Injury
SUS	System Usability Scale
EAR	Eye Aspect Ratio
fps	Frames per Second
CPU	Central Processing Unit
GPU	Graphics Processing Unit
CSR	Corporate Social Responsibility
PAF	Part Affinity Field (Used in OpenPose)
API	Application Programming Interface
SDK	Software Development Kit
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning

Author Contributions

Bhavya Pinakin Darji: Formal Analysis, Methodology, Resources, Software, Visualization, Writing - review & editing

Priyam Parikh: Conceptualization, Investigation, Supervision, Validation, Writing - original draft, Writing - review & editing

AI Writing Assistance Statement

Portions of the text were refined using AI-based language assistance (ChatGPT, OpenAI). The authors confirm that the intellectual content, experimental design, results, analysis, and final interpretations are entirely their own. AI assistance was limited to improving readability, structure, and linguistic clarity.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] M. E. T. Widanarko, A. Legg, and D. Stevenson, "The impact of computer use on musculoskeletal symptoms among office workers: A systematic review," *Applied Ergonomics*, vol. 45, no. 6, pp. 1706-1717, 2014.

[2] P. Buckle and J. Devereux, "The nature of work-related neck and upper limb musculoskeletal disorders," *Applied Ergonomics*, vol. 33, no. 3, pp. 207-217, 2002.

[3] K. Wahlström, "Ergonomics, musculoskeletal disorders and computer work," *Occupational Medicine*, vol. 55, no. 3, pp. 168-176, 2005.

[4] J. A. Jacko, *Human-Computer Interaction: Design Issues, Solutions, and Applications*. CRC Press, 2009.

[5] F. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," *arXiv preprint arXiv: 1906.08172*, 2019.

[6] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172-186, 2021.

[7] A. Google AI Blog, "MoveNet: Ultra fast and accurate pose detection model," *Google AI Blog*, 2021. [Online]. Available: https://blog.tensorflow.org

[8] A. Mathis et al., "DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning," *Nature Neuroscience*, vol. 21, no. 9, pp. 1281-1289, 2018.

[9] H. Fang et al., "AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Trans. Image Process.*, vol. 29, pp. 6966-6981, 2020.

[10] E. Dennerlein and D. Johnson, "Integrating ergonomics and health: A model for workplace intervention," *Occupational Health Science*, vol. 1, pp. 3-22, 2017.

[11] Z. Gao *et al.*, "A systematic survey on human pose estimation: Upstream and downstream tasks," *Artif. Intell. Rev.*, 2025. SpringerLink.

[12] A. F. R. Nogueira *et al.*, "Markerless multi-view 3D human pose estimation: A survey," *Image and Vision Computing*, 2025. ScienceDirect.

[13] F. Roggio *et al.*, "A comprehensive analysis of machine-learning pose-estimation methods," *Frontiers in Bioengineering and Biotechnology*, 2024. PMC.

[14] G. Goyal *et al.*, "MoveEnet: Online high-frequency human pose estimation with an event camera," in *CVPR Workshops*, 2023. (MoveNet context and efficiency notes.) CVF Open Access.

[15] S. Ye *et al.*, "SuperAnimal pretrained pose estimation models for multiple species," *Nature Communications*, 2024. (DeepLabCut Model Zoo.) Nature.

[16] M. Gil-Martín *et al.*, "Hand gesture recognition using MediaPipe landmarks with reduced feature sets," in *VISIGRAPP 2025*, SciTePress, 2025. Scitepress.

[17] A. Gupta *et al.*, "AI-Enhanced Virtual Mouse Using Hand Gesture Recognition," in *Advances in Intelligent Systems & Computing*, Springer, 2025. SpringerLink.

[18] M. Hemalatha, V. Sreeja, and S. Aswathi, "Hand Gesture Controlled Virtual Mouse," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 13, no. 3, pp. 885-888, Mar. 2024, https://doi.org/10.17148/IJARCCE.2024.133143

[19] Manisha Kasar, Pranoti Kavimandan, Trupti Suryawanshi & Sudarshana Abbad (12 Feb 2024): AI-based real-time hand gesture-controlled virtual mouse, Australian Journal of Electrical and Electronics Engineering, https://doi.org/10.1080/1448837X.2024.2313818

[20] C. Dewi *et al.*, "Adjusting eye aspect ratio for strong eye blink detection based on facial landmarks," *PeerJ Computer Science*, 2022. PeerJ.

[21] D. Santos *et al.*, "Real-time eye blink detection using computer vision and machine learning," *Preprints.org*, 2024. (MediaPipe Face Mesh + EAR pipeline.) Preprints.

[22] D. F. Santos, "Real-time Eye Blink Detection using Computer Vision and Machine Learning," *Preprints.org*, 03 Oct. 2024. [Online]. Available: https://doi.org/10.20944/preprints202410.0131.v2

[23] P. Tian *et al.*, "Head Zoom: Hands-free zooming and panning for 2D content," *arXiv*, 2025. (Head-based navigation with user study.) arXiv.

[24] D. Verma *et al.*, "Human-Computer Interaction Through Eye Ball," 2025. (Multimodal eye/mouth/head control; nose-based scrolling.)