Recent Advances in Vision-Based Hand Gesture Recognition for AI Virtual Mouse Systems: A Review

1st Manali Sapkal
Department of Computer Engineering (SPPU)
G. H. Raisoni College of Engineering and management Pune

2nd Dr. Geeta Atkar
Department of Computer Engineering (SPPU)
G. H. Raisoni College of Engineering and management Pune

3rd Dr. Vidya Dhamdhere
Department of Computer Engineering (SPPU)
G. H. Raisoni College of Engineering & Management Pune

4th Dr. Sarita Patil
Department of Computer Engineering (SPPU)
G. H. Raisoni College of Engineering & Management Pune

5th Jaydeep Shinde
Department of Computer Engineering (SPPU)
G. H. Raisoni College of Engineering & Management Pune

6th Yuvraj Bargaje
Department of Computer Engineering (SPPU)
G. H. Raisoni College of Engineering & Management Pune

Abstract— **Hand gesture recognition has emerged as a promising approach for enabling intuitive and touchless human–computer interaction (HCI). Conventional input devices such as the mouse and keyboard limit natural interaction, particularly in environments where contactless operation is preferred. This paper presents the design and performance evaluation of a real-time AI-based virtual mouse system using vision-based hand gesture recognition. The proposed framework utilizes a lightweight Convolutional Neural Network (CNN) to classify hand gestures captured through a standard RGB webcam.**

A custom dataset consisting of approximately 6000 gesture images was collected under varying lighting conditions, backgrounds, and hand orientations to improve model generalization. The system supports four primary gestures corresponding to cursor movement, scrolling, left click, and right click operations. To enhance usability in real-time environments, cursor smoothing and click debouncing mechanisms are incorporated to reduce jitter and prevent unintended actions.

Experimental evaluation demonstrates that the proposed CNN model achieves a classification accuracy of 97.9%, outperforming traditional machine learning algorithms including Support Vector Machine, K-Nearest Neighbors, Decision Tree, and Random Forest. The system operates at approximately 28–30 frames per second on standard CPU hardware, demonstrating its practicality for real-time human–computer interaction applications. The results indicate that the proposed framework provides a reliable and cost-effective solution for gesture-based touchless interfaces in assistive technology, public systems, and immersive computing environments.

Keywords— Hand Gesture Recognition, Virtual Mouse, Convolutional Neural Network, Human–Computer Interaction, Computer Vision, Touchless Interface.

I. INTRODUCTION

Human–Computer Interaction (HCI) has experienced significant transformation with the rapid development of artificial intelligence and computer vision technologies. Traditional interaction devices such as keyboards and physical mice remain widely used; however, they require direct physical contact and limit the natural interaction between users and computing systems. In recent years, gesture-based interaction has emerged as an effective alternative that enables intuitive, touchless communication between humans and digital devices. These systems are particularly valuable in environments where contactless operation is desirable, such as public kiosks, healthcare settings, and immersive computing applications [1], [2].

Hand gesture recognition has therefore become a major research topic within computer vision and artificial intelligence. Recent advances in deep learning have significantly improved the reliability of gesture recognition systems by enabling automatic extraction of spatial features from image data. Several studies have demonstrated the potential of vision-based gesture recognition systems for controlling user interfaces, robotic systems, and interactive displays [3], [4]. These approaches provide more natural interaction mechanisms compared to conventional input devices.

Early gesture recognition systems relied on traditional computer vision techniques such as skin-color segmentation, contour detection, and handcrafted feature extraction methods. Although these approaches were computationally efficient, they often suffered from limitations when operating under varying lighting conditions, background clutter, and hand orientation changes [14]. To address these challenges, researchers introduced machine learning techniques such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forest classifiers to improve recognition performance [9].

With the advancement of deep learning, Convolutional Neural Networks (CNNs) have become the dominant approach for image-based gesture recognition. CNN models can automatically learn hierarchical spatial features from raw image data, eliminating the need for manual feature engineering. Several recent works have demonstrated that CNN-based architectures significantly outperform classical machine learning models in gesture classification tasks [6], [7]. In addition, hybrid deep learning frameworks incorporating temporal modeling techniques such as Long Short-Term Memory (LSTM) networks have further improved recognition accuracy for dynamic gestures [8].

Despite these advancements, many existing gesture-based virtual mouse systems focus primarily on classification accuracy while overlooking practical usability and real-time interaction stability. Some systems rely on specialized hardware such as depth cameras or sensor-based devices, which increases system cost and limits widespread adoption [5]. Furthermore, complex deep learning architectures often require high-performance GPUs, making them unsuitable for deployment on standard consumer hardware.

To address these limitations, this paper proposes a **lightweight CNN-based real-time virtual mouse system** capable of operating using a standard RGB webcam without requiring specialized hardware. The proposed system recognizes four fundamental hand gestures corresponding to common mouse operations: cursor movement, scrolling, left click, and right click. To improve user experience and interaction reliability, additional stabilization techniques such as cursor smoothing and click debouncing are incorporated.

The major contributions of this work can be summarized as follows:

1. Development of a lightweight CNN architecture optimized for real-time hand gesture classification.
2. Creation of a diverse gesture dataset consisting of approximately 6000 images captured under varying lighting conditions and backgrounds.
3. Implementation of gesture-to-action mapping for cursor control, scrolling, and clicking operations.
4. Comparative performance evaluation with classical machine learning algorithms including SVM, KNN, Decision Tree, and Random Forest.
5. Demonstration of real-time performance achieving approximately 28–30 frames per second on standard CPU hardware.

The remainder of this paper is organized as follows. Section II presents a detailed review of related research in gesture-based virtual mouse systems. Section III describes the proposed system architecture and methodology. Section IV explains the CNN model design and training procedure. Section V discusses the experimental results and performance evaluation. Finally, Section VI concludes the paper and outlines potential directions for future research.

II. Research Motivation and Contributions

Many existing gesture-based mouse systems emphasize classification accuracy but overlook deployment feasibility and interaction stability. Complex deep architectures often increase computational cost, reducing real-time performance on standard hardware.

The primary objective of this work is to design a lightweight yet reliable virtual mouse framework optimized for CPU-based real-time execution.

The main contributions are:

1. Development of a computationally efficient CNN architecture for four-class gesture recognition.
2. Creation of a diverse 6000-image gesture dataset.
3. Integration of smoothing and debouncing mechanisms to enhance interaction stability.
4. Comparative evaluation against classical machine learning classifiers.
5. Real-time performance validation on consumer-grade hardware.

III. Literature Review

A. *Chained Literature Review*

1. **A. Ślesicka and A. Kawalec, "Real-time hand gesture recognition for IoT devices using FMCW mmWave radar and continuous wavelet transform" [1] (2026, *Electronics*)**

   Ślesicka and Kawalec propose a real-time gesture recognition framework that utilizes FMCW mmWave radar signals combined with continuous wavelet transform for gesture classification. The approach focuses on radar-based sensing, which enables gesture recognition without relying on optical cameras. Their system demonstrates strong robustness in environments with poor lighting conditions and offers potential applications for IoT-based human–machine interfaces. **Open issues:** Although radar-based gesture recognition provides robustness to lighting variations, the requirement of specialized radar hardware increases system complexity and cost. This limitation restricts its accessibility for general consumer applications. These challenges motivate the exploration of vision-based systems using standard cameras that can achieve comparable performance without specialized sensing hardware.

2. **D. Kalaivani et al., "Enhancing accessibility through gesture-based human–computer interaction: A virtual mouse approach" [2] (2026, Springer ICDSMLA)**

   Kalaivani et al. present a gesture-based virtual mouse system designed to improve accessibility in human–computer interaction. Their approach enables users to control cursor movement and mouse operations through hand gestures detected via computer vision techniques.

The study highlights the usefulness of gesture-controlled interfaces in assistive technologies, particularly for users with limited physical mobility. **Open issues:** While the system demonstrates the potential of gesture-based interfaces for accessibility applications, the study primarily focuses on system usability rather than improving gesture recognition accuracy. Furthermore, robustness under varying environmental conditions and complex hand orientations is not deeply evaluated. This suggests the need for more reliable gesture classification models capable of operating consistently in real-world environments.

3. **C. Cui, M. S. Sunar, and G. Eg Su, "Deep vision-based real-time hand gesture recognition: A review" [3] (2025, *PeerJ Computer Science*)**

Cui et al. provide a comprehensive review of deep learning approaches for vision-based hand gesture recognition. The authors analyze various CNN-based and hybrid deep learning architectures that have been proposed for recognizing static and dynamic gestures in real-time environments. Their study demonstrates that deep learning models significantly outperform traditional computer vision techniques by automatically learning spatial features from image data. **Open issues:** Despite the improvements offered by deep learning models, many systems require large computational resources and complex network architectures. Such requirements may limit their practical deployment on consumer-grade hardware. This highlights the importance of designing lightweight CNN models capable of achieving high accuracy while maintaining real-time performance.

4. **Y. Yaseen et al., "Evaluation of benchmark datasets and deep learning models with pre-trained weights for vision-based dynamic hand gesture recognition" [4] (2025, *Applied Sciences*)**

Yaseen et al. evaluate several deep learning architectures using benchmark gesture datasets to analyze the performance of pre-trained models for dynamic gesture recognition. Their research demonstrates that transfer learning and pre-trained CNN models can significantly improve recognition accuracy and reduce training time. The study emphasizes the importance of dataset diversity and model generalization in gesture recognition tasks. **Open issues:** Although pre-trained deep learning models improve classification accuracy, they often introduce increased computational complexity and require substantial memory resources. This can negatively impact real-time performance, particularly on systems without dedicated GPUs. Therefore, lightweight architectures remain necessary for practical human–computer interaction systems.

5. **J. Wang et al., "Hand gesture recognition for user-defined textual inputs and gestures" [5] (2025, *Universal Access in the Information Society*)**

Wang et al. present a gesture recognition framework designed to interpret user-defined gestures for interactive computing applications. The system enables users to perform custom gestures that can be mapped to textual inputs and control commands. Their research highlights the flexibility of gesture-based interfaces in modern human–computer interaction systems. **Limitation addressed by next paper:** Although the system provides flexible gesture interpretation capabilities, it primarily focuses on gesture-to-command mapping rather than continuous cursor control required in virtual mouse systems. Additionally, the study does not extensively analyze the real-time interaction stability necessary for smooth cursor movement and click operations.

Research Gap Leading to the Proposed Work

From the above literature, it can be observed that existing gesture recognition systems either rely on specialized hardware, computationally intensive deep learning models, or lack robust real-time interaction mechanisms. Therefore, there remains a need for a lightweight and reliable gesture recognition framework that can operate using standard camera hardware while maintaining high accuracy and real-time performance.

To address these limitations, the present work proposes a **CNN-based real-time AI virtual mouse system** that focuses on efficient gesture classification, interaction stability, and practical deployment on consumer-grade hardware.

IV. Research Gap and Motivation

Although significant research has been conducted in the domain of hand gesture recognition and gesture-based human–computer interaction, several limitations still exist in current systems. Earlier works primarily focused on traditional computer vision techniques for gesture detection. For instance, the work of M. Oudah and N. A. Abuhasan presented a computer-vision-based gesture recognition approach that demonstrated the feasibility of gesture-based interaction but lacked robustness under varying environmental conditions [14].

Subsequent studies introduced deep learning models to improve recognition performance. Research by J. P. Sahoo utilized fine-tuned convolutional neural networks for real-time gesture recognition, showing improved accuracy compared with classical approaches [10]. Similarly, S. Waichal proposed a CNN-based virtual mouse system capable of performing cursor movements and click operations using hand gestures [11]. However, these systems still face challenges related to real-time responsiveness, gesture variability, and computational efficiency.

Recent advancements in deep learning have explored hybrid architectures combining CNN models with additional frameworks such as LSTM and MediaPipe. For example, Y. Yaseen and colleagues proposed a dynamic gesture recognition model integrating MediaPipe with Inception-V3 and LSTM networks to enhance gesture interpretation accuracy [8]. Although such methods achieve higher recognition accuracy, they often require increased computational resources, which may limit real-time performance on standard consumer devices.

Furthermore, recent studies have also explored alternative sensing technologies such as radar-based gesture recognition. The work by A. Ślesicka and A. Kawalec demonstrated gesture recognition using FMCW mmWave radar signals and wavelet transforms [1]. While this approach improves robustness against lighting variations, it requires specialized hardware, making it less practical for widespread deployment.

Despite these advancements, several research gaps remain:

- Many gesture recognition systems rely heavily on specialized sensors or high-computational deep learning models.
- Some approaches lack robustness in real-time environments with varying lighting conditions or complex hand orientations.
- Real-time cursor control using simple and accessible hardware remains an important challenge.

To address these limitations, this work proposes a **CNN-based AI Virtual Mouse system** that utilizes real-time computer vision and deep learning techniques to recognize hand gestures for cursor control and user interaction. The proposed approach aims to provide **high accuracy, improved gesture recognition performance, and real-time usability using standard webcam hardware**, making it suitable for practical human–computer interaction applications.

V. Proposed System

The proposed system introduces an **AI-driven virtual mouse framework** that enables users to control computer cursor movements and mouse functions using hand gestures captured through a standard webcam. The system integrates computer vision techniques with deep learning-based gesture recognition to interpret hand movements in real time.

The architecture of the proposed system consists of the following main components:

1. **Image Acquisition Module**
   A webcam continuously captures video frames containing hand movements. These frames serve as the input to the gesture recognition pipeline.
2. **Hand Detection and Landmark Extraction**
   Computer vision algorithms detect the presence of a hand in each frame and extract important hand landmarks such as fingertips and palm positions.
3. **Feature Processing and Gesture Classification**
   Extracted features are passed to a Convolutional Neural Network (CNN) model that classifies the detected hand gesture into predefined commands such as cursor movement, click, scroll, or drag operations.
4. **Cursor Control Module**
   The recognized gesture is mapped to corresponding mouse operations, allowing the user to control the cursor without using a physical mouse.

The proposed framework aims to achieve **efficient real-time performance while maintaining high recognition accuracy**, making it suitable for applications such as touch-free interfaces, accessibility systems, and gesture-based computing environments.

A. Gesture Definitions

The system supports four primary gestures:

TABLE I. Gesture Wise Accuracy

Gesture	**Mouse Operation**
Index finger only	Cursor movement
Index + middle finger	Scroll
Open palm	Left click
Thumb + index pinch	Right click

B. Dataset Collection

We collected around 6000 images of hand gestures covering the four gesture classes above. The dataset includes variation in lighting (indoor daylight, artificial light), backgrounds (plain wall, cluttered room), hand orientations (rotated ±30°, scaling, different users/hands), and skin tones. A portion (~20%) was set aside for validation/testing.

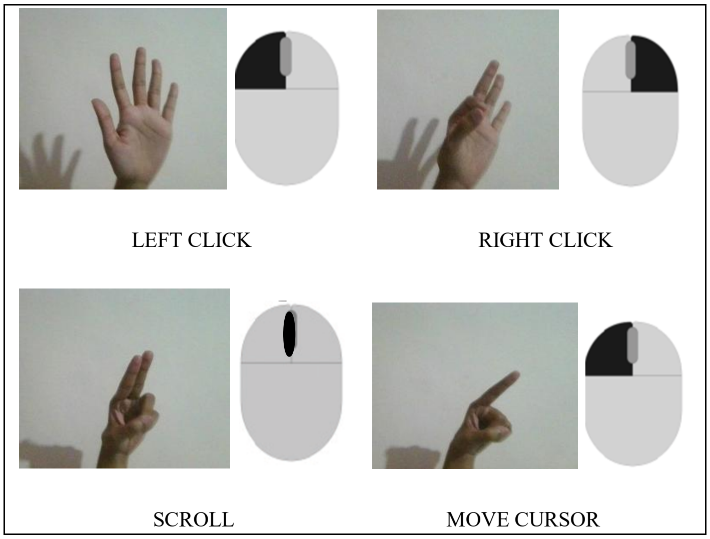

Fig 1. Postures and their Corresponding Mouse Actions

VI. Data Flow Diagram

A. Level 0 DFD (Context Diagram)

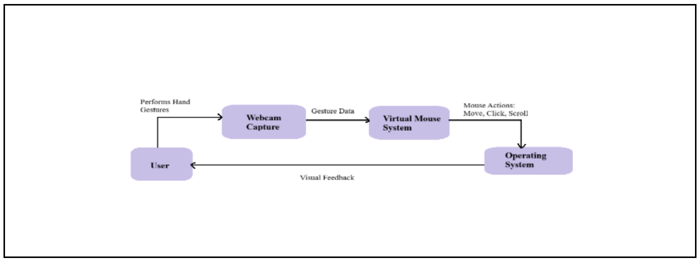

Fig 2. Data Flow Diagram (Level 0)

1. The User performs various hand gestures (input).
2. The Webcam captures these gestures as video frames and sends them to the Virtual Mouse System.
3. The system processes the input to interpret the intended action (Move, Scroll, Left-Click, Right-Click).
4. Recognized gestures are transmitted to the Operating System, which executes corresponding cursor operations.
5. The display screen provides visual feedback back to the user, completing the interactive loop.

Inference:
The Level-0 diagram confirms that the system functions as a closed-loop HCI model where gesture data continuously flows from the user to the OS and feedback is visually returned.

B. Level 1 DFD (Main Functional Modules)

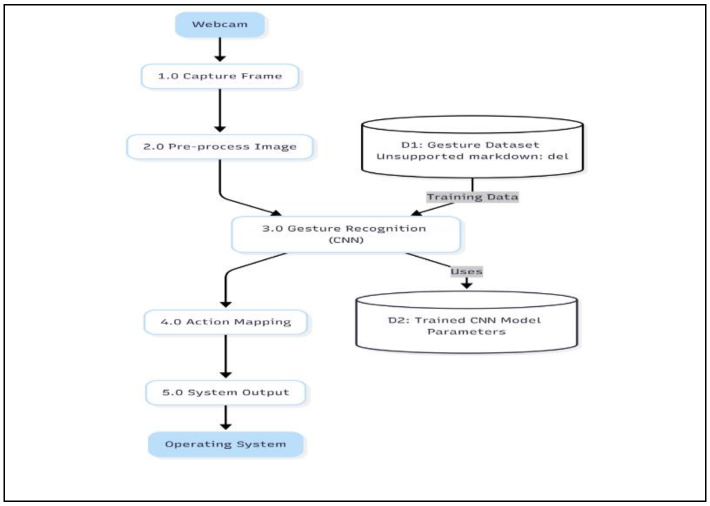

Fig 3. Data Flow Diagram(Level 1)

1. Capture Frame – The webcam captures raw video frames and sends them to pre-processing.
2. Pre-process Image – Removes background noise, enhances hand region, converts colour spaces, and normalizes image size.
3. Gesture Recognition (CNN) – Processes pre-processed images using a trained Convolutional Neural Network to classify gestures.
4. Action Mapping – Maps the recognized gesture into a corresponding mouse command using APIs (e.g., PyAutoGUI).
5. System Output – Sends the control signal to the Operating System, which performs the corresponding operation on the user's screen.

Data Stores:

- D1: Gesture Dataset (~6000 Images) – Used for training and validation of the CNN model.
- D2: Trained CNN Model Parameters – Contains model weights utilized during runtime for real-time classification.

Inference:
Level-1 DFD shows the complete data pipeline, from raw input acquisition to actionable output, confirming modular design and data independence between components.

C. Level 2 DFD (Gesture Recognition Details)

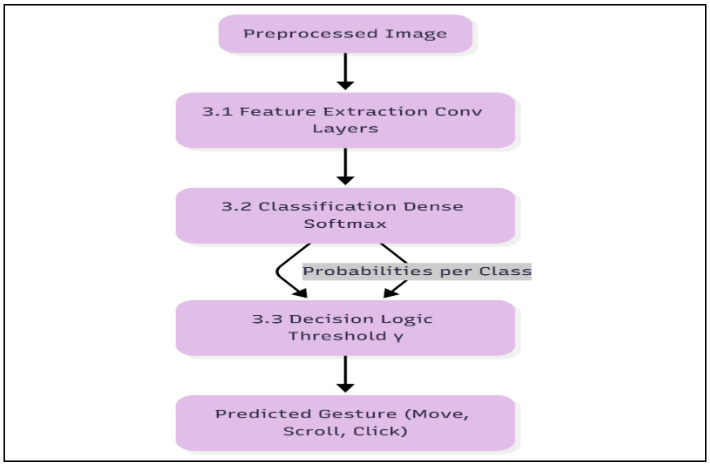
Preprocessed Image
3.1 Feature Extraction Conv Layers
3.2 Classification Dense Softmax
Probabilities per Class
3.3 Decision Logic Threshold γ
Predicted Gesture (Move, Scroll, Click)

Fig 4. Data Flow Diagram(Level 2)

1. **Feature Extraction (Convolutional Layers)** – Extracts spatial and edge features from pre-processed hand images using convolution and pooling operations.
2. **Classification (Fully Connected + Softmax Layers)** – Converts extracted features into class probability distributions representing each gesture type.
3. **Decision Logic** – Applies a confidence threshold γ\gammaγ (e.g., 0.6–0.8) to select the most probable gesture and filters uncertain predictions.

Data Flow Explanation:

- Pre-processed Image → Feature Extraction → Feature Maps → Classification → Probability Scores → Decision Logic → Predicted Gesture (Move / Scroll / Click / Right-Click**).**

Inference:
Level-2 DFD confirms how real-time frame data flows through the CNN model and decision logic before being mapped to user actions.
It highlights the machine learning intelligence layer responsible for accurate, marker-free gesture recognition.

VII. System Architecture

The proposed AI-driven virtual mouse system is designed to enable natural human–computer interaction using real-time hand gesture recognition. The architecture integrates computer vision techniques with deep learning models to detect hand gestures and convert them into mouse control commands. The system processes video input captured from a webcam, identifies hand landmarks, classifies gestures using a convolutional neural network (CNN), and maps the detected gestures to corresponding cursor actions.

The overall architecture consists of four primary modules: image acquisition, hand detection, gesture classification, and cursor control. These modules work sequentially to ensure efficient real-time processing.

The image acquisition module continuously captures frames from the webcam. Each frame is then processed by the hand detection module, which identifies the hand region and extracts key landmarks such as fingertip positions. These landmarks represent spatial features that are used for gesture interpretation.

The extracted features are then passed to the gesture classification module, where a CNN-based model analyzes spatial relationships between hand landmarks to determine the intended gesture. Deep learning techniques enable the system to recognize complex hand gestures more accurately compared to traditional rule-based approaches.

Finally, the cursor control module converts the recognized gestures into system commands such as cursor movement, click operations, and scrolling actions. These commands allow users to interact with the computer without requiring a physical mouse device.

The proposed architecture aims to provide high accuracy, real-time performance, and minimal hardware requirements, making it suitable for practical applications such as touchless interaction systems, accessibility technologies, and smart computing environments.

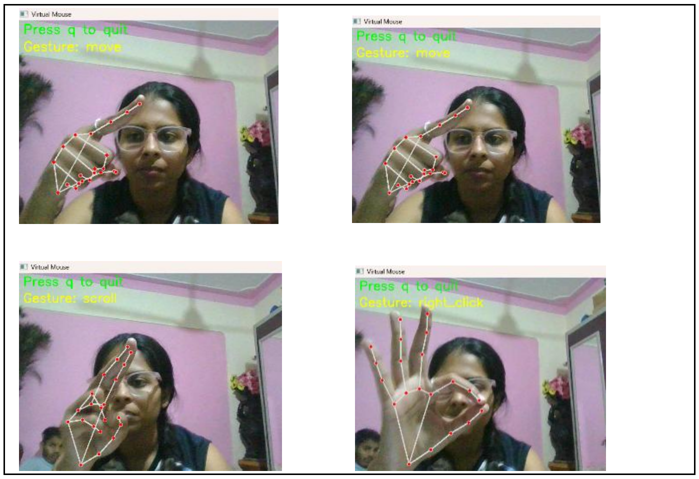

Fig 5. Real-Time Gesture Mapping

VIII.. EXPRIMENTAL RESULTS

A. Accuracy Comparison

TABLE II. Accuracy Comparison of Algorithm

Algorithm	**Accuracy (%)**
SVM	87.4
KNN	90.6
Decision Tree	85.3
Random Forest	91.5
Proposed CNN	**97.9**

The proposed AI virtual mouse system was evaluated using multiple machine learning and deep learning algorithms to analyze performance differences in gesture recognition accuracy. The experiments were conducted using a dataset consisting of various hand gestures captured under different environmental conditions.

The performance of classical machine learning algorithms such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, and Random Forest was compared with the proposed CNN-based approach. The evaluation results demonstrate that deep learning models outperform traditional algorithms in gesture recognition tasks due to their ability to learn complex spatial features.

The comparison results are presented in Table I, which shows the accuracy achieved by each algorithm.

IX. APPLICATIONS

The system can be applied in:

- Assistive technology
- Public kiosks
- Hospital environments
- AR/VR systems
- Touchless industrial interfaces

X. CONCLUSION

This paper presented an AI-driven virtual mouse system based on real-time hand gesture recognition using convolutional neural networks. The proposed framework enables users to control cursor movements and mouse operations through natural hand gestures captured by a webcam.

Experimental results demonstrate that the proposed CNN model achieves higher accuracy compared to traditional machine learning algorithms. The system provides an efficient and intuitive alternative to conventional input devices, offering potential applications in touchless interfaces, accessibility technologies, and smart computing environments.

Future work will focus on improving gesture recognition robustness under challenging conditions such as complex backgrounds and varying lighting environments. Additionally, integrating advanced deep learning architectures and expanding the gesture dataset may further enhance system performance and usability.

XI. REFERENCES

[1] A. Ślesicka and A. Kawalec, "Real-time hand gesture recognition for IoT devices using FMCW mmWave radar and continuous wavelet transform," Electronics, vol. 15, no. 2, Art. no. 250, Jan. 2026, doi: 10.3390/electronics15020250.

[2] D. Kalaivani, S. J. Mahure, A. Rajput, and P. B. Mane, "Enhancing accessibility through gesture-based human–computer interaction: A virtual mouse approach," in Proc. 6th Int. Conf. Data Science, Machine Learning and Applications (ICDSMLA), Lecture Notes in Electrical Engineering, vol. 1528. Singapore: Springer, 2026, pp. 75–87.

[3] C. Cui, M. S. Sunar, and G. Eg Su, "Deep vision-based real-time hand gesture recognition: A review," PeerJ Computer Science, vol. 11, Art. no. e2921, Jun. 2025, doi: 10.7717/peerj-cs.2921.

[4] Y. Yaseen, O.-J. Kwon, J. Kim, J. Lee, and F. Ullah, "Evaluation of benchmark datasets and deep learning models with pre-trained weights for vision-based dynamic hand gesture recognition," Applied Sciences, vol. 15, no. 11, Art. no. 6045, May 2025.

[5] J. Wang, I. Ivrissimtzis, Z. Li, and S. Zhang, "Hand gesture recognition for user-defined textual inputs and gestures," Universal Access in the Information Society, vol. 24, pp. 1315–1329, Jun. 2025.

[6] S. Saranya, "An ensembled real-time hand-gesture recognition using CNN," in Proc. 15th Int. Conf. Computing Communication and Networking Technologies (ICCCNT), 2025, pp. 1–5.

[7] I. El Magrouni, A. Ettaoufik, S. Aouad, and A. Maizate, "Hand gesture recognition for virtual mouse control," International Journal of Interactive Mobile Technologies, vol. 19, no. 2, pp. 53–64, Jan. 2025.

[8] Y. Yaseen, O.-J. Kwon, J. Kim, and J. Lee, "Next-generation dynamic hand gesture recognition using MediaPipe, Inception-V3 and LSTM based deep learning model," Electronics, vol. 13, no. 16, Art. no. 3233, Aug. 2024.

[9] G. Datta, A. S. Vadana, A. V. Akhil, K. M. S. Chandana, and V. V. Padyala, "Improved method for hand gesture recognition using CNN algorithm with OpenCV datasets," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 3, pp. 3621–3629, 2024.

[10] J. P. Sahoo, "Real-time hand gesture recognition using fine-tuned convolutional neural networks," Sensors, vol. 22, no. 3, 2022.

[11] S. Waichal, "Hand gesture recognition based virtual mouse using CNN," International Journal of Computer Applications, vol. 184, no. 20, pp. 1–6, 2022.

[12] W. Zhang, J. C. Wang, and F. P. Lan, "Dynamic hand gesture recognition based on short-term features and deep network," IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 2, pp. 1–11, 2021.

[13] S. Shriram, V. B. Suraj, and R. S. Kumar, "Deep learning-based real-time AI virtual mouse system using computer vision to avoid COVID-19 spread," 2021.

[14] M. Oudah and N. A. Abuhasan, "Hand gesture recognition based on computer vision," Journal of Imaging, vol. 6, no. 8, Art. no. 73, 2020.

[15] O. Köpüklü, Y. Rong, and G. Rigoll, "Talking with your hands: Scaling hand gestures and recognition with CNNs," arXiv preprint arXiv:1904.08722, 2019.

[16] P. Xu, "A real-time hand gesture recognition and human–computer interaction system," arXiv preprint arXiv:1704.07296, 2017.