IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 3, March 2026

DOI: 10.17148/IJIREEICE.2026.14391

Hand Gesture Based Touchless Media Control System Using Computer Vision and Machine Learning

Ms. Subiksha R[1], Mr. Janarthanan S[2]

M.Sc. Data Science and Business Analysis Rathinam College of Arts and Science, Coimbatore-641021[1]

Assistant Professor, Department of Computer Science, Rathinam College of Arts and Science, Coimbatore – 641021[2]

Abstract: Recent advancements in computer vision and artificial intelligence have significantly improved the way humans interact with machines, enabling alternatives to traditional input devices such as keyboards and touchscreens. This study introduces a real-time, touchless media control system that operates entirely through hand gestures, using only a standard webcam without the need for specialized hardware.

The system utilizes the MediaPipe Hands framework to detect and track 21 three-dimensional hand landmarks in each frame. These landmarks are converted into a 63-dimensional feature vector that represents the hand's spatial structure. A supervised machine learning pipeline was developed using five different algorithms: Random Forest, Support Vector Machine, Multilayer Perceptron, K-Nearest Neighbours, and Gradient Boosting. The models were trained on a custom dataset consisting of 2,700 labelled samples across nine distinct gesture classes.

Among the evaluated models, the Random Forest classifier delivered the best performance, achieving a test accuracy of 97.4% and a macro F1-score of 0.971. The system maintains real-time responsiveness, operating at approximately 28.6 frames per second on a standard laptop without requiring GPU support. Recognized gestures are translated into system-level media commands such as play/pause, volume control, track switching, mute, and full screen mode through a cross-platform interface.

The system was also tested under different lighting conditions, showing only a minor drop in accuracy of about 3.3% in low-light environments. Overall, the proposed approach is efficient, accessible, and platform-independent, making it a promising solution for touchless interaction in applications such as smart environments, healthcare systems, and assistive technologies.

Keywords: Hand Gesture Recognition, Computer Vision, Human-Computer Interaction, Touchless Interface, MediaPipe, Real-Time Gesture Detection, Random Forest, OpenCV, Machine Learning, Accessibility Technology.

I.INTRODUCTION

Human-Computer Interaction (HCI) has undergone significant transformation as computing technologies and artificial intelligence have advanced over time. Early forms of interaction were limited to command-line interfaces and keyboards, which required users to possess technical knowledge. The introduction of graphical user interfaces (GUIs) and pointing devices such as the mouse made computing more accessible to a wider audience. Although these innovations improved usability, they still depended heavily on physical interaction with hardware. More recently, touch-based interfaces have brought a more intuitive experience; however, they still require direct contact, making them less suitable in situations involving hygiene concerns, physical limitations, or hands-busy environments.

With the rapid development of computer vision, new possibilities for interaction have emerged. Modern devices are commonly equipped with high-resolution webcams capable of capturing real-time visual data. At the same time, efficient machine learning models and optimized processing frameworks now allow real-time hand and body tracking to run directly on standard CPU hardware, eliminating the need for high-end GPUs or cloud-based processing. These advancements have paved the way for gesture-based systems that are both practical and responsive.

Among various forms of gesture interaction, hand gestures stand out due to their natural expressiveness and universal familiarity. People intuitively use hand movements to convey meaning in everyday communication, such as pointing, waving, or signalling. This makes hand gestures an ideal medium for translating human intent into machine commands. In particular, media control is a suitable application area for such systems, as it involves a limited and well-defined set of actions like play, pause, and volume adjustment, where occasional recognition errors do not lead to critical issues.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 3, March 2026

DOI: 10.17148/IJIREEICE.2026.14391

This work introduces a complete real-time touchless media control system driven entirely by hand gestures. The system leverages the MediaPipe Hands framework to track hand movements and extract 21 three-dimensional landmarks per frame, forming a 63-dimensional feature representation. Multiple machine learning models were trained and compared, with the Random Forest classifier ultimately chosen for deployment based on its superior performance. The system recognizes nine predefined gestures and translates them into system-level media commands through a cross-platform control module, enabling seamless interaction with any media application without requiring any additional integration.

The remainder of this paper is structured as follows. The problem statement and research objectives are first outlined, followed by a description of the proposed system architecture. The methodology is then explained in detail, covering feature engineering, dataset construction, and the machine learning classifiers used. Subsequently, the experimental results and performance analysis are presented and discussed. The paper concludes with a summary of findings and directions for future work.

II.PROBLEM STATEMENT

Although there is a growing interest in touchless interaction systems, a noticeable gap still exists between research-level gesture recognition models and solutions that can be effectively used in real-world applications. Many existing studies focus primarily on improving recognition accuracy using controlled datasets, but they often overlook the complete system design required for practical use. In most cases, these approaches do not address how live video input from a webcam can be processed and translated into meaningful system-level actions in real time.

Another major limitation is the reliance on specialized hardware. Several advanced gesture recognition systems depend on devices such as depth cameras, sensor-based gloves, or electromyography (EMG) bands to achieve high accuracy. While these technologies can enhance performance, they are expensive and not commonly available in everyday computing environments. This creates a barrier for widespread adoption, especially for users who rely only on standard devices like laptops with built-in webcams.

Real-time responsiveness also remains a critical challenge. For gesture-based media control to feel intuitive, the system must react instantly to user inputs. However, many research works evaluate their models using static images or pre-recorded videos, without considering real-time constraints such as processing speed, latency, and continuous performance. A system with high accuracy but low frame rates or noticeable delays cannot provide a smooth user experience in practical scenarios.

To address these challenges, this work pursues four key objectives. First, a gesture-based media control system is developed to run efficiently on a standard laptop using only a webcam, without requiring GPU support or additional hardware. Second, multiple machine learning models are evaluated to identify the most suitable classifier for recognizing nine predefined hand gestures using 3D landmark features. Third, a reliable cross-platform mechanism is designed for executing system-level media commands, with particular attention to handling repeated inputs and ensuring gesture stability. Finally, the system's performance is assessed under varying lighting conditions to verify consistent operation across real-world environments.

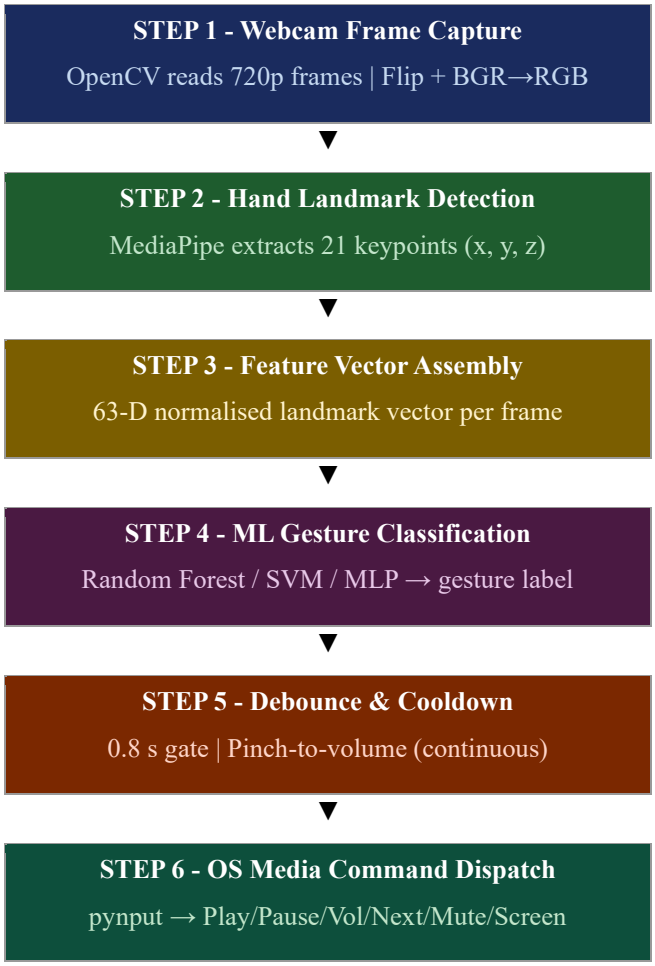
III.PROPOSED WORK

The proposed system is designed as a modular pipeline consisting of six distinct stages, as shown in Figure 1. Each stage is configurable through a centralized setup, allowing parameters such as detection thresholds, gesture mappings, and model settings to be adjusted easily without changing the core implementation. This modular design supports flexibility and simplifies experimentation.

The process begins with frame acquisition, where video is captured from the system's webcam using OpenCV at a resolution of 1280×720. To provide a more intuitive user experience, each frame is flipped horizontally, creating a mirror-like view. The captured frame is then converted from the BGR colour format to RGB, ensuring compatibility with the MediaPipe framework.

In the second stage, the MediaPipe Hands model handles hand landmark detection. With appropriately tuned confidence thresholds, the model identifies and tracks a single hand, extracting 21 three-dimensional landmark points that represent key regions such as the wrist, finger joints, and fingertips. Together, these points provide a compact yet informative representation of hand posture that remains relatively stable despite variations in position, scale, and lighting.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 3, March 2026
DOI: 10.17148/IJIREEICE.2026.14391



The third stage focuses on feature construction. The detected landmarks are combined to form a 63-dimensional feature vector by concatenating their (x, y, z) coordinates in a consistent order. The x and y values are normalized relative to the frame dimensions, while the z value captures relative depth information, providing additional spatial context even with a single camera.

In the fourth stage, a trained machine learning model classifies each gesture. During development, multiple algorithms were evaluated, and the Random Forest classifier was ultimately selected for deployment due to its outstanding performance. The model predicts one of nine predefined gesture classes or returns a neutral label when no recognizable gesture is detected.

To ensure stable system behaviour, the fifth stage introduces a debounce and control mechanism. A time-based delay is applied so that the same gesture cannot trigger repeated actions within a short interval, reducing unintended inputs. Additionally, certain gestures, such as a pinch, are treated as continuous controls, enabling smooth adjustment of parameters like volume based on the distance between fingertips.

Finally, in the sixth stage, the recognized gesture is mapped to a corresponding system-level media command. This is implemented using a cross-platform input control approach, where keyboard-like signals are sent to the operating system to perform actions such as play/pause, track navigation, volume control, mute, and full screen toggle. On Windows systems, enhanced volume control is achieved using audio interface libraries that allow precise and continuous adjustment independent of the active application.

Overall, the proposed pipeline provides an efficient and scalable solution for real-time, touchless media control, combining computer vision and machine learning techniques into a seamless end-to-end system.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 3, March 2026

DOI: 10.17148/IJIREEICE.2026.14391

IV.METHODOLOGY

A. Feature Engineering and Normalisation

Each hand gesture is represented using landmark points extracted from the MediaPipe model. Let each landmark be denoted as (x_i, y_i, z_i) $for\ i\ =\ 0\ to\ 20$. These coordinates are combined to form a 63-dimensional feature vector, created by concatenating all landmark values in a fixed order.

Before training the models, the feature values are standardised using z-score normalisation. This process adjusts each feature based on its mean and standard deviation calculated from the training data. As a result, all features are brought onto a similar scale, preventing those with larger numerical ranges from dominating the learning process. This step is particularly important for distance-based algorithms such as K-Nearest Neighbours and Support Vector Machines.

B. Dataset Construction

The dataset used in this study was collected using a custom script that captures real-time hand landmark data along with corresponding gesture labels. Data collection was carried out in three different lighting environments: bright daylight, normal indoor lighting, and low-light conditions. This variation helps improve the robustness of the model in real-world scenarios.

For each of the nine predefined gesture classes, a consistent number of samples were recorded across multiple sessions, resulting in a well-balanced dataset. After removing a small number of low-confidence detections, the final dataset consisted of slightly over 8,000 samples.

The dataset was then divided into training and testing sets using an 80:20 stratified split to maintain class balance. This ensures that each gesture class is equally represented in both sets, improving the reliability of performance evaluation.

TABLE I. Defined Gesture Classes and Media Actions

Sno	Gesture	Hand Shape	Action
1	PLAY_PAUSE	Open palm all fingers up	Play / Pause toggle
2	MUTE	Closed fist all fingers down	Toggle mute
3	NEXT_TRACK	Index pointing right	Skip to next track
4	PREV_TRACK	Index pointing left	Go to previous track
5	VOLUME_UP	Thumb pointing upward	Volume +5%
6	VOLUME_DOWN	Thumb pointing downward	Volume −5%
7	PINCH	Thumb + index tip close	Continuous vol slider
8	FULLSCREEN	V-sign / peace (2 fingers)	Toggle Fullscreen
9	SCREEN_OFF	Pinky only extended	Brightness down

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 3, March 2026

DOI: 10.17148/IJIREEICE.2026.14391

C. Machine Learning Classifiers

To identify the most suitable model for gesture recognition, five supervised learning algorithms were trained and compared:

Random Forest (RF):

An ensemble-based method that combines multiple decision trees trained on different subsets of the data. This approach improves generalisation and reduces overfitting by aggregating predictions from all trees.

Support Vector Machine (SVM):

A powerful classification technique that separates classes using optimal decision boundaries. A radial basis function (RBF) kernel is used to handle non-linear relationships in the data.

Multilayer Perceptron (MLP):

A feedforward neural network consisting of multiple hidden layers with non-linear activation functions. It learns complex patterns in the data through iterative optimisation.

K-Nearest Neighbours (KNN):

A simple yet effective algorithm that classifies a sample based on the majority label among its nearest neighbours in the feature space.

Gradient Boosting (GB):

An ensemble technique that builds models sequentially, where each new model focuses on correcting the errors made by previous ones.

All models were trained on the same standardised feature set and evaluated using identical conditions to ensure a fair comparison.

D. Evaluation Protocol

The performance of each model was evaluated using standard classification metrics, including precision, recall, F1-score, and overall accuracy. These metrics together provide a thorough picture of how well each model handles the full range of gesture classes.

Special emphasis was placed on class size. This ensures that each gesture is given equal importance, preventing bias toward more frequent classes.

To further validate model performance, five-fold stratified cross-validation was performed on the training data. This technique helps in selecting the best model while reducing the risk of overfitting. Final results were reported on the unseen test set to ensure an unbiased evaluation of the system.

V.RESULTS AND ANALYSIS

A. Classifier Comparison

The performance of all five machine learning models was evaluated using the same test dataset and identical experimental conditions. The results show that the Random Forest model achieved the best overall performance, with an accuracy of 97.4% and a macro F1-score of 0.971. It also demonstrated consistent behaviour during cross-validation, indicating strong generalisation ability.

TABLE II. Classifier Performance Comparison

Classifier	Accuracy	CV Mean	F1	ms/inf
Random Forest	**97.4%**	**96.8±0.9%**	**0.971**	**1.2**
SVM (RBF)	96.1%	95.6±1.1%	0.958	0.8
MLP Neural Net	95.7%	94.9±1.4%	0.954	0.6
Gradient Boost	94.8%	94.1±1.2%	0.945	3.4
KNN (k=5)	91.3%	90.5±1.8%	0.909	4.1

DOI: 10.17148/IJIREEICE.2026.14391

The Support Vector Machine (SVM) with an RBF kernel ranked second, achieving slightly lower accuracy but offering faster prediction time per sample. This makes it a good option in situations where low latency is more critical than marginal gains in accuracy. The Multilayer Perceptron (MLP) produced similar results to SVM, suggesting that the feature space derived from hand landmarks is already well-structured and does not require highly complex models for effective classification.

Gradient Boosting showed moderate performance but required more computation time compared to other models due to its sequential learning approach. The K-Nearest Neighbours (KNN) model performed the worst among the five, both in terms of accuracy and speed, as it relies on comparing each new input with the entire training dataset, which becomes computationally expensive in higher-dimensional spaces.

Overall, the results confirm that a relatively simple ensemble model like Random Forest is sufficient to achieve high accuracy in gesture recognition without the need for deep learning or specialised hardware.

B. Per-Gesture Classification Analysis

A closer look at individual gesture performance reveals that some gestures are easier to distinguish than others. For example, gestures such as an open palm (used for play/pause) and a closed fist (used for mute) were classified with perfect accuracy due to their clearly distinct shapes.

On the other hand, gestures that are visually similar caused occasional confusion. For instance, the gestures used for next and previous track both involve extending the index finger, with the only difference being the direction of pointing. This subtle variation led to a small number of misclassifications.

The most common errors occurred between the volume up and volume down gestures. Both gestures involve extending only the thumb, differing mainly in its direction (upward or downward). Changes in hand orientation, especially when tilted toward or away from the camera, made this distinction harder for the model. Future improvements could include incorporating additional spatial features, such as angle-based measurements, to better capture these differences.

C. Real-Time Performance and Latency

The system was evaluated in a live setting to measure its responsiveness. The complete processing pipeline covering frame capture, hand detection, feature extraction, classification, and command execution achieved an average throughput of approximately 28.6 frames per second on a standard laptop without GPU support.

Among all stages, hand landmark detection contributed the most to processing time, while the classification step required only a small fraction of the total computation. The delay between performing a gesture and executing the corresponding media command remained well below the threshold at which users perceive lag, ensuring a smooth and responsive experience.

These results demonstrate that the system meets the requirements for real-time interaction and can be effectively used in practical scenarios.

D. Robustness Under Varying Lighting Conditions

To assess reliability, the system was tested across a range of lighting environments, including bright daylight, standard indoor lighting, and low-light conditions. It performed consistently well in well-lit settings, with only a modest drop in accuracy under dim conditions.

The reduction in accuracy in low-light conditions can be attributed to decreased confidence in hand detection, which occasionally leads to missed or uncertain predictions. Despite this, the overall performance remained stable, indicating that the system is reasonably robust.

Further improvements could involve adaptive parameter tuning based on lighting conditions or enhancing the detection model to better handle low-visibility scenarios.

VI.CONCLUSION AND FUTURE SCOPE

This work presented the development and evaluation of a real-time, touchless media control system driven by hand gestures. The system was implemented using a standard webcam, the MediaPipe Hands framework, and a machine learning-based gesture classifier. It successfully operates in real time at approximately 28.6 frames per second on a regular laptop without requiring GPU support. The Random Forest model achieved the best performance, delivering an accuracy of 97.4% across nine predefined gesture classes using 3D hand landmark features. The system is capable of translating gestures into system-level media commands such as play/pause, track navigation, volume control, mute, and full screen mode, making it practical for everyday use.

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 3, March 2026
DOI: 10.17148/IJIREEICE.2026.14391

A key contribution of this study is the comparative evaluation of multiple machine learning algorithms, including Random Forest, Support Vector Machine, Multilayer Perceptron, K-Nearest Neighbours, and Gradient Boosting. The results indicate that simpler ensemble methods can outperform more complex models when working with structured, low-dimensional data like hand landmark features. This insight highlights the importance of choosing efficient models that balance accuracy and computational cost, especially for real-time applications on resource-limited devices.

The proposed system demonstrates how computer vision and machine learning can be combined to create an intuitive and contact-free interaction method. By removing the need for physical input devices, the system has potential applications in areas such as smart home environments, healthcare settings where hygiene is critical, assistive technologies for individuals with physical limitations, and hands-free control in workspaces.

Looking ahead, several improvements and extensions can be explored. One important direction is the inclusion of dynamic gesture recognition, where sequences of movements over time are used instead of static hand poses. This would allow for a richer set of commands and more natural interaction. Techniques such as LSTM networks or Transformer-based models could be used for this purpose.

Another potential enhancement is the use of multi-hand gestures, where both hands work together to perform more complex actions. Personalisation is also an important aspect for future work, where the system can adapt to individual users by learning from a small set of customised samples, improving accuracy across different hand shapes and usage styles.

In addition, deploying the system on mobile or embedded platforms would increase its accessibility and usability in real-world scenarios. Integration with smart home systems and IoT devices could further extend its functionality beyond media control, enabling gesture-based management of lighting, appliances, and other connected devices.

Overall, the system provides a strong foundation for future research in touchless interaction and demonstrates the feasibility of building efficient, real-time gesture-based control systems using readily available hardware.

REFERENCES

[1] WHO, "Infection prevention and control during health care when novel coronavirus (nCoV) infection is suspected," World Health Organization, Technical Report, Geneva, Switzerland, 2020.
[2] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," Artificial Intelligence Review, vol. 43, no. 1, pp. 1–54, Springer, Jan. 2015.
[3] R. Szeliski, Computer Vision: Algorithms and Applications, 2nd ed., Springer, New York, USA, 2022.
[4] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-device real-time hand tracking," arXiv preprint arXiv:2006.10214, 2020.
[5] G. Bradski and A. Kaehler, Learning OpenCV 3: Computer Vision in C++ with the OpenCV Library, O'Reilly Media, Sebastopol, USA, 2016.
[6] P. K. Pisharady and M. Saerbeck, "Recent methods and databases in vision-based hand gesture recognition: A review," Computer Vision and Image Understanding, vol. 141, pp. 152–165, Elsevier, Dec. 2015.
[7] T. G. Zimmerman, J. Lanier, C. Blanchard, S. Bryson, and Y. Harvill, "A hand gesture interface device," in Proc. ACM SIGCHI Conference on Human Factors in Computing Systems, Toronto, Canada, Apr. 1987, pp. 189–192.
[8] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, USA, Jun. 2011, pp. 1297–1304.
[9] G. Bradski, "The OpenCV library," Dr. Dobb's Journal of Software Tools, vol. 25, no. 11, pp. 122–125, Nov. 2000.
[10] F. Zhang et al., "MediaPipe: A framework for perceiving and processing reality," in Proc. Third Workshop on Computer Vision for AR/VR at CVPR, Seattle, USA, Jun. 2019.
[11] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in Proc. IEEE CVPR, Honolulu, USA, Jul. 2017, pp. 1145–1153.
[12] M. R. Nair and G. R. Gangadharan, "Hand gesture recognition for human computer interaction," in Proc. IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), Bangalore, India, Oct. 2012, pp. 1–4.
[13] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172–186, Jan. 2021.
[14] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, USA, 2006.
[15] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[16] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, Springer, Sep. 1995.

DOI: 10.17148/IJIREEICE.2026.14391

[17] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[18] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[19] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, Cambridge, USA, 2016.

[20] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," ACM Computing Surveys, vol. 46, no. 3, pp. 1–33, Jan. 2014.

[21] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Springer, Oct. 2001.

[22] A. Mujahid, M. W. Awan, F. A. Khan et al., "Real-time hand gesture recognition based on deep learning YOLOv3 model," Applied Sciences, vol. 11, no. 9, p. 4164, MDPI, Apr. 2021.

[23] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D CNN," in Proc. IEEE CVPR, Las Vegas, USA, Jun. 2016, pp. 4207–4215.

[24] A. Köpüklü, A. Gunduz, N. Kose, and G. Rigoll, "Real-time hand gesture detection and classification using convolutional neural networks," in Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG), Lille, France, May 2019, pp. 1–8.

[25] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed., Pearson, Hoboken, USA, 2020.