Desktop Application for a Virtual Chemistry Lab with Computer Vision-Based Interaction for High School Students

Marlon Llaguento-de-la-Cruz
Universidad Peruana de Ciencias Aplicadas (UPC)
Lima, Peru
u20201b055@upc.edu.pe

Juan Osorio-Quiroz
Universidad Peruana de Ciencias Aplicadas (UPC)
Lima, Peru
u201713107@upc.edu.pe

Willy Ugarte
Universidad Peruana de Ciencias Aplicadas (UPC)
Lima, Peru
willy.ugarte@upc.pe

***Abstract*—The integration of computer vision in educational environments opens new opportunities for interactive and immersive learning. This work proposes a desktop application for a virtual chemistry laboratory aimed at high school students, utilizing hand gesture recognition to simulate real laboratory interactions. Through the development of a proprietary dataset with multi-action gestures and the training of a convolutional neural network model, the system enables students to conduct virtual experiments safely, reinforcing theoretical concepts through practical experiences. This solution promotes active, accessible, and personalized learning in science education, especially in contexts with limited resources.**

***Index Terms*—Object Detection, Convolutional Neural Networks, Gestures, Pose Estimation.**

I. Introduction

The Peruvian educational system faces serious challenges in science education, particularly in the area of chemistry, where over 50% of students perform below the minimum proficiency level according to PISA tests [1].

This issue is further exacerbated by the precariousness of educational infrastructure: as of December 2024, the infrastructure gap reached $42,355 million, and 16.2% of public schools were in critical condition, on the verge of collapse [2].

Furthermore, the Comptroller General of the Republic warned that 70% of supervised educational institutions did not receive educational materials at the beginning of the 2024 school year, and over 50% presented severe structural deficiencies, such as roofs, walls, and floors in poor condition [3].

In this context, emerging technologies—such as artificial intelligence, augmented reality, and computer vision—open new opportunities to strengthen science education.

These tools enable the development of immersive and interactive environments that facilitate virtual experimentation, eliminating the need to handle dangerous or costly materials [4].

Particularly, the use of gesture recognition via computer vision has proven to be an effective strategy to foster student participation and improve learning in educational settings [5][6].

To achieve seamless interaction in the virtual laboratory, a gesture detection model based on YOLOv8 was developed, trained with deep learning on Roboflow.

This system uses real-time video from a standard camera to recognize gestures such as ”grab,” ”pour,” and ”stir,” without requiring physical peripherals.

Libraries like OpenCV, PyTorch, and Pygame were integrated to capture video, process gestures, and display animations. This allows for a more realistic and immersive experience. The model improves precision and naturalness in the simulation of chemistry practices.

A key aspect of this work involved the creation of a robust, proprietary dataset capturing diverse hand gestures in real-world educational contexts.

The entire work was executed through a structured development process, encompassing fundamental research, system architecture design, intensive data collection, model training, and rigorous application evaluation.

This work proposes the development of a desktop application that integrates computer vision to detect hand gestures and simulate interactions within a virtual chemistry laboratory.

This solution aims to enrich the learning experience in schools, increase student motivation, and improve academic performance in a key discipline such as science.

Our main contributions are as follows:

- We have developed a desktop application for a virtual chemistry laboratory that enables active learning through a computer vision-based hand gesture recognition system.
- We created a robust, proprietary dataset of over 1,000 frames featuring 11 specific laboratory gestures, collected in a real educational environment with significant variations in capture conditions.
- We successfully trained a YOLOv8 Pose model that achieves high accuracy (99.50%) and mAP50 (0.941) in real-time gesture detection (10.1ms inference latency), validating our approach for natural and intuitive interaction without external peripherals.

This paper is organized as follows: Section II reviews related work on gesture detection and virtual laboratory applications. Section III describes our main contribution, including the system's architecture and relevant computer vision concepts. Section IV details the experimental protocol, covering dataset construction, model training, and performance evaluation. Finally, Section V presents the conclusions, discusses limitations, and outlines future work

II. Related Works (RW)

In [7], the authors use Mediapipe and propose techniques such as normalization, introducing the FingerComb block, improving the ResNet architecture to create the FingerNet model, and creating a proprietary dataset, all with the aim of enhancing the accuracy of recorded individual gestures. We, in contrast, will collect a dataset with multi-action gestures to train a neural network that will translate into movements within the virtual environment.

In [8], the authors reframe the gesture detection problem using a transformer and YOLO-based model, trained with a loss function focused on samples with gesture centers, utilizing only the hand skeleton as input. This approach was validated solely on the SHERC'22 and IPN Hands datasets. Instead, a convolutional model will be employed alongside a proprietary dataset, which will be expanded using data augmentation techniques and combined with other datasets to strengthen the training process.

In [9], the authors construct a highly variable dataset using RGB-D sensors and varying factors such as camera angle and distance, individual participant characteristics, and environmental conditions. Twenty participants were involved, and a dictionary of 10 gestures was defined. In contrast, we will collect data through short videos taken of students, considering aspects such as lighting, position, occlusion, and human characteristics. Additionally, data augmentation techniques will be applied to increase the variability of the dataset and improve the model's generalization.

In [10], the authors utilize the IPN dataset, specifically designed for hand gesture recognition, which includes 13 gestures classified into two functional categories. Videos were collected from 50 participants across 28 distinct scenarios. However, this dataset does not align with the objectives of our model, which focuses on gestures performed in laboratory environments. Therefore, data will be collected in educational institutions, taking into account the human characteristics of students in real learning contexts.

In [11], the authors highlight that the VRChemEd application represents a significant advancement in secondary-level chemistry education by offering an interactive and safe environment that overcomes the limitations of traditional methods. This application facilitates the understanding of chemical concepts through the use of simulations and interactive content, including didactic quizzes. Similarly, our proposal aims to promote accessibility and a better assimilation of content, encouraging experimentation in an immersive virtual environment that fosters active and safe learning.

III. Main Contribution

In this section, the main concepts of our work are presented.

A. Context/Background/Overview

Our objective is to develop a desktop application for a virtual chemistry laboratory that utilizes a computer vision model to enable experimentation and active learning.

Definition 1 (Gamification [12]:] Gamification is understood as the use of game design elements in non-game contexts, with the aim of increasing student motivation and participation, particularly in digitized educational environments.

Example 1: In the virtual laboratory, gamification will stimulate students to correctly complete chemical mixtures, encouraging repetition and exploration of experiments.

Definition 2 (Pose Estimation [13]:] Pose Estimation is the process of detecting and localizing key anatomical points (such as joints or body parts) in images or videos within a two-dimensional space.

In our virtual chemistry laboratory, this technology enables the tracking of hand gestures to simulate interactions like pouring liquids or manipulating equipment.

Example 2: During the virtual practice, the system recognizes the student's hand position and movement towards the instruments, which activates the animation without the need for a mouse or keyboard.

Definition 3 (Object Detection [14]:] Object Detection [14] is a computer vision technique that localizes and classifies objects within an image using bounding boxes.

Example 3: The palm detector processes RGB images (see Figure 1) and generates square bounding boxes around the hands, ignoring non-rigid aspects like articulated fingers. This allows for isolating regions for subsequent gesture analysis.

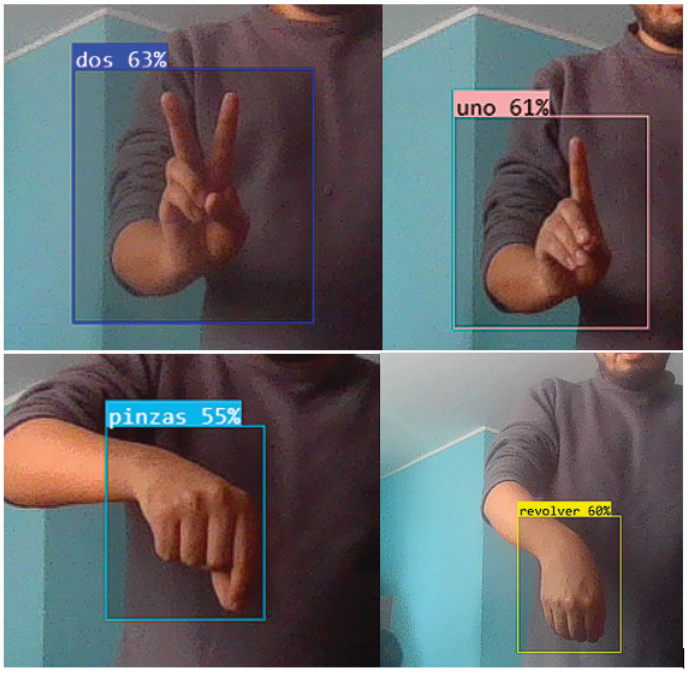

Fig. 1. Hand Detection

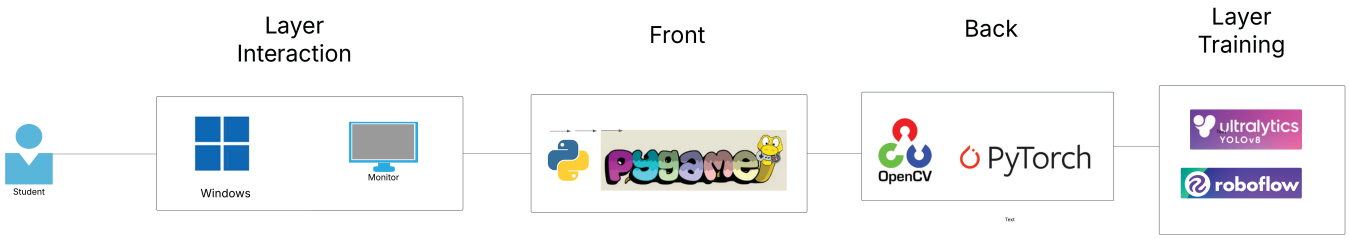

Fig. 2. Architecture Diagram

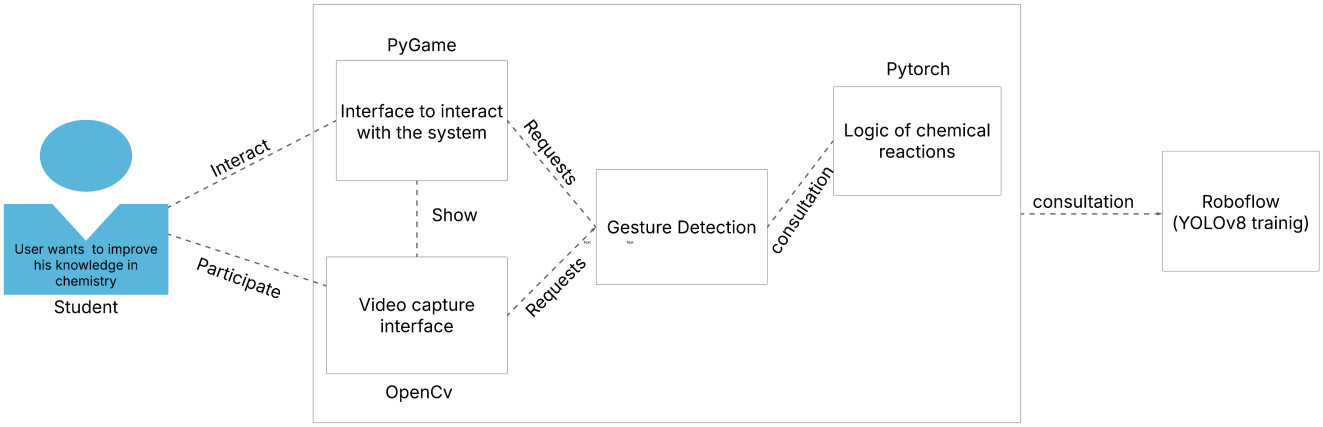

Fig. 3. C4 Model

Fig. 4. Land Marks

Definition 4 (Gesture Detection [14]:] Gesture Detection [14] is the process of interpreting hand movements or postures by means of key points (landmarks).

Example 4: Fig. 4 illustrates how finger flexion (accumulated angles between landmarks) translates into gestures such as 'pinch' or 'x'.

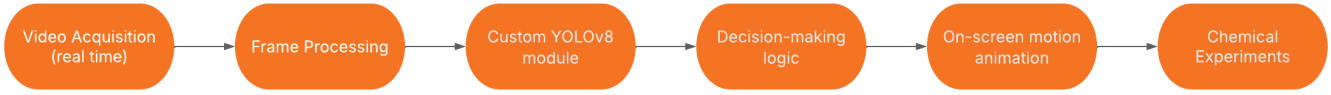

Fig. 5. flow chart

TABLE I. PERFORMANCE COMPARISON BETWEEN YOLO VERSIONS

Models	Authors	Datasets	#Samples	Accuracy(%)
CNN	O. Yulius et al. [9]	ASL datasets	30526	96.30
CNN	N. HERBAZ et al. [10]	Personal datasets	1000	98.70
YOLO-v3	A. Mujahid et al. [7]	Personal datasets	216(train)/ 15(test)	97.68
YOLO-v4	Zhuowen Zheng [15]	self-built datasets	2450	97.80
YOLO-v5	**Proposed approach**	Personal datasets [14]	1500	**97.99**
YOLO-v6	**Proposed approach**	Personal datasets [14]	1500	**98.40**
YOLO-v8	**Proposed approach**	Personal datasets [14]	1500	**99.50**

B. Method

The virtual Chemistry laboratory is a desktop application based on the following design choices. For the front-end, we opted to use Pygame to display results and control the virtual environment. The back-end is based on PyTorch and OpenCV for model execution and video capture. The YOLOv8 model is trained in Roboflow, as illustrated in Fig. 2.

Fig. 3 presents the architecture of the Virtual Chemistry Laboratory, composed of the following main components: Graphical Interface, Gesture Detector, Video Capture, Game Logic, and Roboflow.

In this system, Pygame is responsible for displaying the interactive environment with which the user can interact. Specifically, beyond rendering the environment, Pygame also serves as the direct interface with the student. On the screen, the user interacts with a virtual laboratory bench displaying flasks, test tubes, and other instruments. Each recognized gesture is immediately translated into a real-time animation—for example, a "grasp" gesture triggers the action of picking up a flask, while a "pour" gesture simulates liquid transfer between containers. This design provides an immersive experience through instant visual feedback, without the need for Augmented Reality (AR) or Virtual Reality (VR) headsets. Instead, the system prioritizes a 2D accessible interface, combining gamification elements (points, achievements, and levels) with the direct manipulation of virtual objects, ensuring usability in educational contexts with limited resources. The prototype was implemented as a desktop application for Windows, given its widespread availability in Peruvian educational institutions and the stable support of Pygame, PyTorch, and OpenCV in this environment. This choice also ensures offline functionality—crucial in areas with limited Internet access—while maintaining portability for future adaptations to Linux, macOS, or web-based platforms.

The prototype was developed as a desktop application for Windows due to its widespread availability in Peruvian educational institutions, where most computer laboratories and student devices rely on this operating system. Additionally, Pygame, PyTorch, and OpenCV offer stable support in Windows environments, ensuring ease of deployment and reproducibility. Choosing Windows also allows the system to function offline, a critical factor in contexts with limited Internet access. Nevertheless, the Python-based architecture remains portable, allowing future adaptations for Linux, macOS, or even web-based platforms.

OpenCV captures real-time video from the camera and provides the frames to the detection model. Subsequently, PyTorch executes the YOLO-trained gesture detection model, which analyzes each frame to identify specific gestures. Based on these detections, the Game Logic interprets the user's actions and generates responses within the virtual environment.

Although the system was designed to operate under limited computational resources, the prototype was tested on a machine equipped with an AMD Ryzen 5 4600H processor, 16 GB of RAM, an NVIDIA GTX 1650 GPU, 1 TB of storage, and an HP TrueVision HD Camera (1280 × 720 resolution at 30 fps). These specifications ensured stable real-time execution. However, further research is required to establish the minimum hardware requirements, such as the lowest supported CPU/RAM configurations and the minimum camera resolution necessary to maintain accurate gesture detection.

It is worth noting that Roboflow was used solely for model training and preparation; therefore, it is not part of the system's real-time execution.

As shown in Fig. 5, the system begins with real-time Video Acquisition, where images are captured from the user's camera.

These images are then sent to the Frame Processing module, which is responsible for adjusting the visual data to be compatible with the detection model.

Subsequently, the frames are analyzed by a customized YOLOv8 module, specifically trained to recognize relevant postures and gestures within the educational context.

The detections produced are interpreted by the Decision-Making Logic, which determines the corresponding action based on the identified gesture.

This decision translates into On-Screen Motion Animation, providing immediate visual feedback to the user.

Finally, these interactions allow for the execution of simulated Chemical Experiments, where the user can manipulate virtual reagents, materials, or instruments through gestures, creating an immersive, safe, and intuitive educational experience.

Table II shows the movements designated for model retraining in Roboflow.

The listed gestures are essential for guiding the retraining process. This focused approach ensures targeted improvements in gesture detection accuracy.

TABLE II. GESTURE LIST

Gesture
grab glass
grab spoon
two
trigger
palms
pinch
stir
release spoon
release glass
one
x

IV. EXPERIMENTS

In this section, the choice of the YOLOv8 model is justified, primarily considering precision metrics such as accuracy and mean average precision (mAP).

YOLOv8 demonstrated superior performance, achieving an accuracy of 99.50% when trained with a proprietary dataset, which supports its suitability for object detection tasks in this work [15] (see Table **??**).

As observed in Table I, the mAP (mean Average Precision) metric, provided by the official YOLO website for the different variants of the v8 model, is key for evaluating performance in object detection tasks.

This metric reflects both the precision and recall capabilities of the model and allows for an objective comparison of improvements across versions.

A notable advancement is evident in version 8 compared to its predecessors (v7, v6, and v5), consolidating it as a more precise and efficient option for computer vision applications[15].

A proprietary dataset was developed through video collection at the YMCA Rímac school branch, where hand movements of high school students from 1st to 5th grade, including both male and female students, were captured.

During the recording process—represented in Figure 6 factors such as lighting and camera angles were carefully considered to ensure the diversity and representativeness of the dataset.

Furthermore, in compliance with Law No. 29733 – Personal Data Protection Law (Peru), the collection and processing of data strictly adhered to ethical and legal standards. All recorded information, including images and biometric data related to students' gestures, was anonymized and safeguarded to protect privacy.

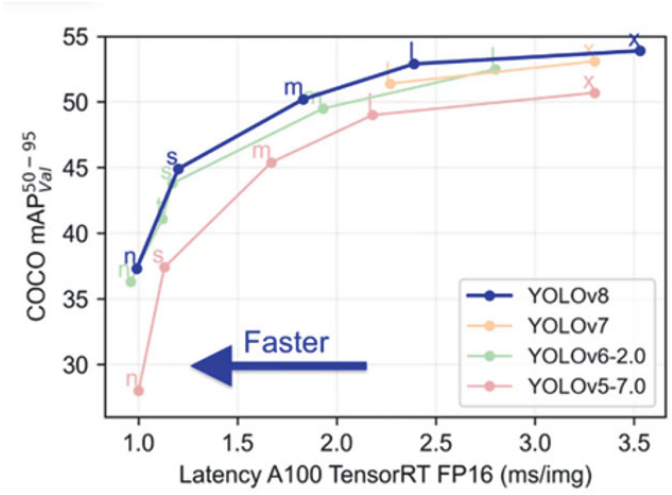

Fig. 6. YOLOv8 performance chart vs. previous versions

This variety allows for training more robust models adaptable to different environmental conditions.

This step is fundamental to ensure that the model accurately learns the characteristics of the gestures, as each frame is manually marked with its corresponding class, which allows for generating consistent and high-quality annotations.

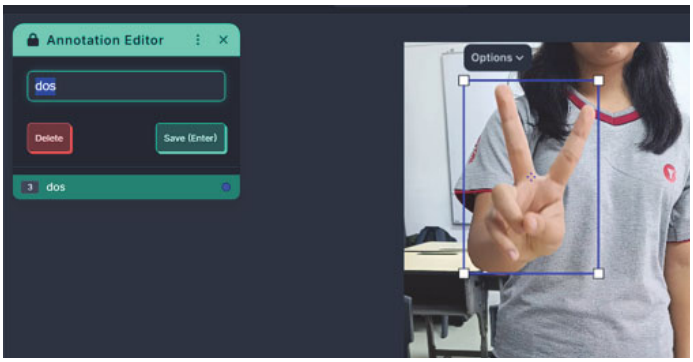

Fig. 7. Manual labeling process in Roboflow for each video frame

Once the meticulous labeling process was completed, a specific distribution of frames was obtained for each of the 11 gestures defined in our dataset.

Table III illustrates this distribution, detailing the number of frames associated with each laboratory gesture. This disaggregation is essential for targeted model training, ensuring that every crucial movement in the virtual environment has adequate representation in the dataset.

The model training was conducted in a Colab Pro environment, leveraging its computational resources to optimize the process.

The detailed performance results of the YOLOv8 model, evaluated on a total of 1237 images and 1215 instances, are summarized below in Table IV.

TABLE III. DISTRIBUTION OF FRAMES PERGESTURE IN THE DATASET

Gesture	No. Frames
grab glass	576
grab spoon	673
two	515
trigger	538
palms	434
pinch	678
stir	658
release spoon	268
release glass	424
one	470
x	986

Speed: 0.2ms preprocess, 10.1ms inference, 0.0ms loss, 1.8ms postprocess per image

These detailed results confirm the model's robustness, with an overall mAP50 of 0.941 and an mAP50-95 of 0.684.

Particularly high performance is observed for gestures such as 'palms' (mAP50: 0.995) and 'x' (mAP50: 0.988), indicating an excellent detection capability for these categories.

The low processing latency, with 10.1ms of inference per image, ensures a real-time response crucial for the virtual laboratory's interactive experience.

For a comprehensive evaluation of the YOLOv8 model's performance at a class level, a confusion matrix was employed, which is presented in Figure 8.

This graphical tool is fundamental for visualizing the effectiveness of a classifier, breaking down the number of correct and incorrect predictions for each gesture category.

Values along the main diagonal indicate instances that were correctly classified (true positives), while values outside this diagonal denote classification errors, where an instance of one class was incorrectly assigned to another, thus precisely identifying the confusions the model experiences between specific gestures.

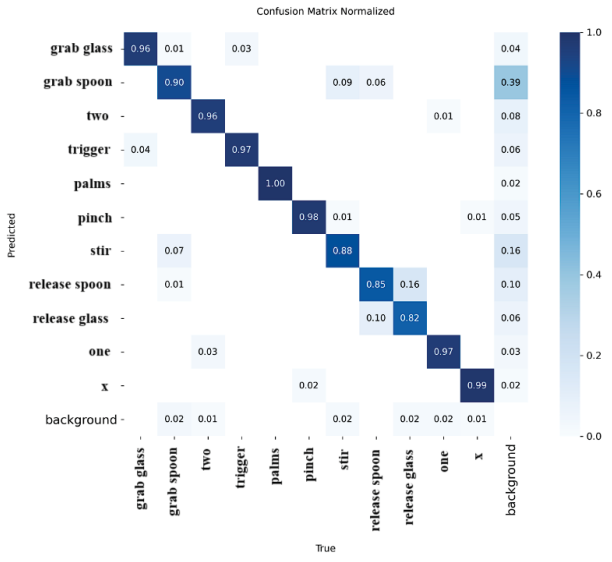

Fig. 8. Confusion Matrix

V. CONCLUSIONS AND PERSPECTIVES

Our main contribution lies in the creation of a robust, proprietary dataset, collected in a real educational environment, encompassing over 1,000 frames with significant variations in capture conditions, and the successful training of a YOLOv8 Pose model.

This model has demonstrated remarkable accuracy of 99.50% and an overall mAP50 of 0.941 in detecting 11 specific laboratory gestures, operating with a low inference latency (10.1ms per image) that ensures a fluid, real-time user experience. These results validate the effectiveness of our approach for natural and intuitive interaction, eliminating the need for additional peripherals.

The implementation of this virtual laboratory represents a significant advancement towards a more active, immersive, and accessible chemistry learning experience.

By offering a safe environment for experimentation, our system not only reinforces theoretical concepts through simulated practice but also fosters student motivation and interest in sciences, contributing to close existing educational gaps, especially in contexts with limited resources.

While the obtained results are highly promising, we acknowledge certain limitations that pave the way for future lines of research.

Evaluating the pedagogical impact in a long-term study with real student control groups is an essential step to quantify the effect on academic performance and knowledge retention.

Furthermore, expanding the dataset to include a greater diversity of users and environmental conditions, as well as integrating more complex or combined gestures, could further enhance the model's robustness.

VI. BIBLIOGRAPHY

[1] Minedu. (2022). Resultados Resultados PISA 2022. Ministerio de Educacion del Peru. http://umc.minedu.gob.pe/resultadospisa2022/

[2] ComexPeru. (2023, mayo 30). Cerrar la brecha en infraestructura educativa costarıa S/ 158,832 millones. Attps://www.comexperu.org.pe/articulo/cerrar-la-brecha-en-infraestructura-educativa-costaria-s-158832-millones

[3] Contraloría General de la Republica. (2024, febrero 20). Contraloría alerta que 70% de colegios supervisados no recibio´ material educativo. https://www.gob.pe/institucion/contraloria/noticias/918845-contraloria-alerta-que-70-de-colegios-supervisados-no-recibio-material-educativo

[4] Zouhri, A., & Mallahi, A. (2024). Improving Teaching Using Artificial Intelligence and Augmented Reality. Journal of Automation, Mobile Robotics & Intelligent Systems. https://doi.org/10.14313/JAMRIS/2-2024/13

[5] Balaji, P., & Ranjan Prusty, M. (2024). Multi- modal fusion hierarchical self-attention network for dynamic hand gesture recognition. Journal of Vi-sual Communication and Image Representation, 98. https://doi.org/10.1016/j.jvcir.2023.104019

[6] Mujahid, A., Awan, M. J., Yasin, A., Mohammed, M. A., Damaˇseviˇcius, R., Maskeli¯unas, R., & Abdulkareem, K. H. (2021). Real-time hand gesture recognition based on deep learning YOLOv3 model. Applied Sciences (Switzerland), 11(9). https://doi.org/10.3390/app11094164

TABLE IV. YOLOv8 MODEL PERFORMANCE METRICS PER CLASS

Class	Images	Instances	Box(P)	R	mAP50	mAP50-95
all	1237	1215	.889	.920	.941	.684
grab glass	112	112	.948	.970	.986	.796
grab spoon	169	169	.866	.787	.917	.615
two	102	102	.938	.971	.977	.642
trigger	105	105	.907	.981	.986	.658
palms	87	87	.979	1.000	.995	.801
pinch	133	133	.949	.975	.967	.824
stir	101	101	.727	.845	.835	.536
release spoon	72	72	.809	.847	.874	.552
release glass	44	44	.737	.773	.835	.554
one	95	95	.951	.979	.990	.680
x	195	195	.974	.990	.988	.867

[7] Meng, Y., Jiang, H., Duan, N., & Wen, H. (2024). Real-Time Hand Gesture Monitoring Model Based on MediaPipe's Registerable System. Sensors, 24(19), 6262. https://doi.org/10.3390/s24196262

[8] Dozdor, Z., Kalafatic, Z., Ban, Z., & Hrkac, T. (2023). TY-Net: Transforming YOLO for Hand Ges-ture Recognition. IEEE Access, 11, 140382–140394. https://doi.org/10.1109/ACCESS.2023.3341702

[9] Hubert, C., Odic, N., Noel, M., Gharib, S., Zargarbashi, S. H. H., & S´eoud, L. (2025). MuViH: Multi-View Hand gesture dataset and recognition pipeline for human-robot interaction in a collaborative robotic finishing platform. https://doi.org/10.5683/SP3/JZJ

[10] Nguyen, T. T., Nguyen, N. C., Ngo, D. K., Phan, V. L., Pham, M. H., Nguyen, D. A., Doan, M. H., & Le, T. L. (2022). A Continuous Real-time Hand Gesture Recognition Method based on Skeleton. 2022 11th International Conference on Control, Automa-tion and Information Sciences, ICCAIS 2022, 273–278. https://doi.org/10.1109/ICCAIS56082.2022.9990122

[11] Chew, C. S., Zafarin, M. N. S., Zain, N. H. M., Aminuddin, R., Tan, T. G., & Chin, K. O. (2024). Trans-forming Secondary School Chemistry Learning with Vir-tual Reality. 2024 IEEE 22nd Student Conference on Research and Development, SCOReD 2024, 482–487. https://doi.org/10.1109/SCOReD64708.2024.10872679

[12] Machado, A., Ten´orio, K., Santos, M. M., Barros, A. P., Rodrigues, L., Mello, R. F., Paiva, R., & Dermeval, D. (2025). Workload perception in educational resource recommendation supported by artificial intelligence: A controlled experiment with teachers. Smart Learning En-vironments, 12(20). https://doi.org/10.1186/s40561-025- 00373-6

[13] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2021). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1), 172- 186. https://doi.org/10.1109/TPAMI.2019.2929257

[14] Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L., & Grundmann, M.(2020). MediaPipe Hands: On-device real-time hand tracking. arXiv preprint arXiv:2006.10214. https://arxiv.org/abs/2006.10214

[15] Ultralytics. (n.d.). Modelos YOLOv8. Ultralytics Docs. Recuperado el 30 de junio de 2025, de https://docs.ultralytics.com/es/models/yolov8/

[16] Herbaz, N., El Idrissi, H., & Badri, A. (2023). Deep Learning Empowered Hand Gesture Recognition: using YOLO Techniques. 2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA), 1-6. https://doi.org/10.1109/SITA60746.2023.10373734