

ARTÍCULO DE INVESTIGACIÓN

**DESARROLLO DE UN SISTEMA DE RECONOCIMIENTO
DE GESTOS DE LA MANO MEDIANTE VISIÓN POR
COMPUTADORA Y APRENDIZAJE AUTOMÁTICO BASADO
EN RANDOM FOREST**

**DEVELOPMENT OF A HAND GESTURE RECOGNITION
SYSTEM USING COMPUTER VISION AND MACHINE
LEARNING BASED ON RANDOM FOREST**

ARTÍCULO DE INVESTIGACIÓN

Bermúdez González, José Erick

Universidad del Valle de Puebla

im45420@uvp.edu.mx

ORCID: <https://orcid.org/0009-0005-3977-4200>

López Nieto, Sergio Raúl

Universidad del Valle de Puebla

sergio.lopez@uvp.edu.mx

ORCID: <https://orcid.org/0000-0001-9762-8109>

Recibido el 29 de mayo de 2026. Aceptado el 7 de julio de 2026.

Publicado el 31 de agosto de 2026.

Reseña de Autor 1

José Erick Bermúdez González es estudiante de Ingeniería Mecatrónica en la Universidad del Valle de Puebla. Sus principales áreas de interés se enfocan en la programación y la automatización de procesos. Actualmente, desarrolla proyectos enfocados en inteligencia artificial, robótica y procesamiento de datos, integrando herramientas de software y hardware para aplicaciones en tiempo real. Sus intereses académicos incluyen el *machine learning*, sistemas embebidos y control.

Reseña de Autor 2

Sergio Raúl López Nieto es Ingeniero Industrial, Licenciado en Administración y Gestión de PyME, Maestro en Ingeniería Administrativa y Calidad, Doctor en Alta Dirección. Posdoctor en Administración de Negocios y en Alta Dirección y Tecnología. Actualmente cursa un Posdoctorado en educación y una maestría en Marketing Digital.

Ha desarrollado investigación en temas relacionados con productividad, liderazgo, gestión del recurso humano, desarrollo de procesos, calidad en el servicio, finanzas verdes, tecnología y educación y estrategias de mejora en las PyME, de los cuales se han publicado algunos artículos en revistas y se han expuesto en diversos eventos de divulgación.

Ha colaborado con empresas del sector manufacturero, comercial y de servicios en las áreas de Seguridad e Higiene, Mantenimiento, Calidad, Producción, Logística y almacenes, desarrollando proyectos conjuntos con marcas como Sabormex, La Costeña, Grupo Bimbo, Huf de México, Laboratorios Novag, Laboratorios Similares, Veladoras Aramo, Genomma Lab, Truper Herramientas, entre otros.

Resumen

El reconocimiento de gestos representa actualmente una de las áreas de mayor interés y relevancia dentro de la robótica y la visión por computadora, debido a la necesidad de desarrollar sistemas más naturales, intuitivos y accesibles para quienes se interesen por el tema.

En esta investigación se desarrolló un sistema basado en Random Forest con el objetivo de identificar movimientos de la mano en tiempo real. El sistema fue implementado utilizando Python, OpenCV y MediaPipe para la detección de puntos clave conocidos como *landmarks* de la mano permitiendo extraer coordenadas tridimensionales correspondientes a las articulaciones detectadas. A partir de la recopilación de los mismos *landmarks*, se construyó un set de datos compuesto por 4970 muestras distribuidas en seis diferentes gestos: puño, mano abierta, pulgar arriba, uno, dos y tres dedos levantados. Posteriormente, se entrenó un modelo de clasificación basado en el algoritmo Random Forest para reconocer los diferentes gestos previamente definidos, los resultados obtenidos mostraron una precisión del 99.30% en la clasificación de gestos además de tiempos de respuesta adecuados para aplicaciones en tiempo real. Al final de esta investigación se encontró que el uso de técnicas de visión por computadora y aprendizaje automático permite desarrollar interfaces eficientes y accesibles para aplicaciones de interacción humano-máquina.

Palabras clave: robótica, gestos de mano, aprendizaje automático, algoritmo, técnicas de visión por computadora.

Abstract

Gesture recognition is currently one of the most interesting and significant areas in robotics and computer vision, due to the need to develop systems that are more natural, intuitive, and accessible to anyone interested in the subject.

In this research, a Random Forest-based system was developed with the goal of identifying hand movements in real time. The system was implemented using Python, OpenCV, and MediaPipe to detect key points known as “landmarks” on the hand, allowing for the extraction of three-dimensional coordinates corresponding to the detected joints. From the collected landmarks, a dataset was constructed consisting of 4,970 data points distributed across six different gestures: fist, open hand, thumbs-up, and one, two, and three fingers raised. Subsequently, a classification model based on the Random Forest algorithm was trained to recognize the different predefined gestures. The results showed 99.30% accuracy in gesture classification, along with response times suitable for real-time applications. At the conclusion of this research, it was found that the use of computer vision and machine learning techniques enables the development of efficient and accessible interfaces for human-machine interaction applications.

Keywords: robotics, hand gestures, machine learning, algorithm, computer vision techniques.

Introducción

El *machine learning* y la visión por computadora han experimentado un crecimiento significativo durante los últimos años, permitiendo el desarrollo de sistemas capaces de interpretar información visual de manera automática y en tiempo real. Una de las áreas de oportunidad en este contexto es el reconocimiento de gestos de la mano ya

que se ha convertido en una de las aplicaciones más relevantes debido a su potencial en campos como la robótica, automatización, realidad virtual, accesibilidad e interacción humano-máquina.

El reconocimiento de gestos se basa en poder identificar movimientos o posiciones específicas, en este caso de la mano, mediante el procesamiento de imágenes y algoritmos de aprendizaje automático. El objetivo principal de este tipo de tecnologías es mejorar la interacción entre humanos y sistemas computacionales mediante métodos más accesibles y eficientes con el fin de reducir la dependencia de dispositivos tradicionales como teclados, controles o incluso sensores físicos; por esta razón, el reconocimiento de gestos ha adquirido una gran relevancia en aplicaciones relacionadas con sistemas inteligentes, automatización industrial y control de dispositivos en tiempo real. Estos sistemas enfrentan un gran problema, lograr una detección precisa y estable en diferentes condiciones de funcionamiento como lo podrían ser la iluminación, posición de la mano, velocidad del movimiento y la complejidad del entorno, por esta razón, el uso de técnicas de *machine learning* y visión por computadora son una alternativa atractiva para mejorar la precisión y la capacidad de reconocimiento en aplicaciones en tiempo real.

En la actualidad existen diferentes herramientas enfocadas al desarrollo de aplicaciones de visión por computadora, entre las que destaca MediaPipe, esta herramienta es una biblioteca utilizada en Python para la detección de un punto conocido como *landmarks* además de poder identificar coordenadas tridimensionales (X,Y,Z) de las distintas articulaciones que componen la mano utilizando únicamente una cámara convencional. También existe la biblioteca llamada OpenCV que facilita el procesamiento de imágenes en tiempo real mientras que algún tipo de algoritmo de *machine learning* permite clasificar los patrones visuales y reconocer el gesto a partir de un *dataset* previamente creado. Hablando sobre los algoritmos de *machine learning* uno de los más utilizados para tareas de clasificación es *Random Forest* el cual destaca por su fácil implementación,

estabilidad y precisión. Este algoritmo permite analizar múltiples características simultáneamente para generar modelos capaces de reconocer patrones de manera eficiente, convirtiéndose en una alternativa adecuada para sistemas de reconocimiento de gestos en tiempo real.

Planteamiento del problema

El uso de la visión artificial para el control robótico vive una transformación muy importante, gracias al impulso de la inteligencia artificial, ya que lo que antes se consideraba exclusivo de grandes laboratorios o de investigaciones con presupuestos millonarios, hoy en día se ha vuelto mucho más accesible por el abaratamiento del *hardware* y a la disponibilidad de *software* de código abierto, este cambio ha permitido que se puedan realizar implementaciones prácticas en espacios académicos y de investigación aplicada, lo cual abre la posibilidad de experimentar con sistemas que antes parecían inalcanzables. A pesar de todos estos avances todavía existen obstáculos que se deben superar, por ejemplo, lograr un reconocimiento de gestos y movimientos que sea realmente preciso o conseguir algoritmos que puedan ejecutarse de manera eficiente en tiempo real, ya que ambos aspectos son determinantes para que la tecnología se pueda aplicar en la práctica; estos retos representan el potencial teórico que ya se ha demostrado y la realidad de poder usarlo en entornos cotidianos o industriales.

La problemática expuesta encuentra su desarrollo a nivel mundial y tiene presencia tanto en países altamente industrializados como en aquellos que buscan ponerse al día en materia tecnológica, por lo que se pueden encontrar avances significativos en regiones como Estados Unidos, Japón, Corea del Sur y Alemania, donde los proyectos de visión artificial aplicada a la robótica están más consolidados. Sin embargo, también en América Latina comienzan a aparecer iniciativas en universidades y centros de investigación que aprovechan el acceso a plataformas

abiertas y a *hardware* de bajo costo, lo que permite que estudiantes y profesionales puedan realizar pruebas y prototipos sin necesidad de grandes presupuestos. En el caso de México, poco a poco, este tipo de trabajos empieza a tener presencia dentro de las universidades, donde se fomenta la experimentación con brazos robóticos, prótesis y sistemas de interacción humano-máquina, de modo que ya no se trata de un desarrollo limitado a ciertas regiones del mundo, sino de un movimiento global que se adapta a diferentes contextos y recursos.

De manera que es importante atender este tema porque la interacción humano-máquina se está convirtiendo en un eje fundamental dentro del desarrollo de la robótica moderna y no actuar sobre ello significaría quedar rezagados en un ámbito que avanza a gran velocidad. Abordar esta línea de investigación no solo representa un avance desde la perspectiva tecnológica, sino que también tiene una repercusión directa en la solución de problemas reales ya sea el diseño de prótesis más avanzadas, la asistencia a personas con movilidad reducida o la optimización de procesos industriales que requieren precisión y rapidez. Además, mientras más temprano se impulsen estos desarrollos en espacios académicos y de investigación más fácil será que se adapten a las necesidades del futuro lo que permitirá que la sociedad se encuentre preparada para los cambios que traerá la automatización. Derivado de lo anterior, surge la necesidad de desarrollar sistemas capaces de reconocer movimientos de la mano, de manera precisa y en tiempo real, mediante técnicas de visión por computadora e inteligencia artificial.

Por ello, la presente investigación tiene como objetivo desarrollar una interfaz basada en la visión por computadora y el aprendizaje automático que permita reconocer movimientos de la mano en tiempo real, evaluando su desempeño mediante métricas de precisión y tiempo de respuesta.

Revisión bibliográfica

La visión por computadora es un campo que busca que las máquinas puedan interpretar imágenes con un nivel de precisión comparable al del ser humano, para lograrlo se desarrollan métodos capaces de analizar, identificar y extraer información relevante a partir de imágenes digitales, ya que este proceso permite detectar elementos, reconocer formas y comprender características visuales presentes en distintos entornos.

Según Oudah *et al.* (2020), la digitalización de una imagen puede entenderse como el proceso desde el cual se representa una matriz de valores en correspondencia a la intensidad de cada píxel. A partir de esta estructura es posible aplicar diferentes operaciones que permiten transformar la imagen o identificar patrones relevantes, entre estas operaciones se encuentran técnicas para resaltar bordes, detectar regiones específicas o analizar texturas, lo que facilita la interpretación automática del contenido visual. El análisis de imágenes también requiere identificar objetos dentro de la escena, para ello se emplean métodos que detectan regiones conectadas o áreas que comparten ciertas propiedades. Este tipo de procesamiento permite segmentar elementos individuales y obtener mediciones como forma, orientación o tamaño, estas características son fundamentales para tareas de clasificación y reconocimiento, donde el sistema debe distinguir entre diferentes tipos de objetos o estructuras presentes en la imagen.

Las herramientas de visión por computadora han permitido automatizar procesos que antes requerían intervención humana, como la detección de características, el reconocimiento de patrones y la interpretación de escenas. De acuerdo con Trigueiros *et al.* (2015), la visión por computadora está basada en técnicas que procesan imágenes y extraen información de relevancia que pueda aplicarse en tareas de clasificación, segmentación o identificación. Estas herramientas combinan métodos matemáticos con modelos de aprendizaje profundo que permiten analizar datos visuales con un nivel de precisión cada vez mayor.

El aprendizaje automático y la visión por computadora han permitido desarrollar sistemas inteligentes capaces de identificar patrones, interpretar imágenes y reconocer movimientos humanos en tiempo real. Estas tecnologías han sido ampliamente utilizadas en aplicaciones relacionadas con reconocimiento de gestos, lenguaje de señas y clasificación de objetos mediante técnicas de inteligencia artificial. Según Lanzarini *et al.* (2019), la aplicación de modelos de aprendizaje automático y redes neuronales convolucionales implica una alta capacidad para resolver problemas de reconocimiento de patrones en imágenes y video. Esto se percibe, según los mismos autores, en tareas relacionadas con la clasificación de formas de mano, el reconocimiento de la lengua de señas y la detección automática de objetos. Con esto se puede decir que una de las principales ventajas de estos modelos es su capacidad de generalización y adaptación a diferentes conjuntos de datos.

Una de las herramientas más utilizadas es OpenCV que constituye una biblioteca diseñada para realizar operaciones de procesamiento de imágenes y visión por computadora, esta biblioteca integra funciones para la detección de bordes, el filtrado espacial, la segmentación y la extracción de características lo que permite construir sistemas capaces de interpretar contenido visual de forma eficiente, estas operaciones representan la base de muchos sistemas biométricos ya que permiten identificar características distintivas como patrones en la piel, puntos clave faciales o contornos específicos dentro de la imagen.

En los últimos años también se ha utilizado la detección de puntos clave o *keypoints* donde se identifican coordenadas relevantes del cuerpo humano, especialmente de las articulaciones como muñeca, codo y hombro; este enfoque ha permitido describir el gesto mediante estructuras geométricas conocidas como poses (Trigueiros *et al.* 2015).

La extracción de características representa una de las etapas más importantes dentro de los sistemas de reconocimiento de gestos basados en aprendizaje automático debido a que permite transformar la información visual de la mano

en datos útiles para los modelos de clasificación en este tipo de sistemas se emplean coordenadas espaciales obtenidas a partir de articulaciones o puntos clave de la mano, *landmarks*, permitiendo describir posiciones, orientaciones y movimientos mediante representaciones geométricas. Según Andreu Villar *et al.* (2024), las transformaciones aplicadas sobre coordenadas tridimensionales de las articulaciones permiten la extracción de patrones relevantes que guardan una relación con la información posicional y temporal de los gestos. Esto da pie, en seguimiento de los autores, a la generación de representaciones que facilitan el proceso de clasificación automática.

En los sistemas de reconocimiento de gestos, en tiempo real, la comunicación entre los dispositivos y los modelos de *machine learning* representa un elemento fundamental para garantizar una respuesta rápida y eficiente. Este tipo de sistemas requiere transmitir información relacionada con las coordenadas y movimientos de la mano hacia modelos de clasificación capaces de interpretar los gestos realizados por el usuario. Según Andreu Villar *et al.* (2024), resulta viable y eficiente el uso de servidores externos para realizar las inferencias del modelo de aprendizaje automático, lo que permite reducir la carga computacional del dispositivo principal y mejorar la fluidez de la interacción. El uso de protocolos de comunicación de baja latencia para enviar información de las articulaciones de la mano y recibir las predicciones del sistema en tiempo real, favoreciendo una experiencia de usuario más dinámica e intuitiva dentro de aplicaciones basadas en reconocimiento de gestos.

El reconocimiento de gestos y movimientos constituye una de las líneas más relevantes dentro de la visión por computadora ya que permite interpretar acciones humanas a partir de información visual sin necesidad de sensores adicionales. De acuerdo con Oudah *et al.* (2020) este tipo de sistemas se basa en el análisis de características espaciales y temporales presentes en la mano y el brazo con el fin de identificar posturas o desplazamientos que pueden emplearse como comandos dentro de una interfaz hombre-máquina. El reconocimiento de gestos puede

dividirse de manera general en dos categorías principales: por un lado, los gestos estáticos relacionados con posiciones específicas de la mano y, por el otro, los gestos dinámicos que implican trayectorias o movimientos continuos en una secuencia de imágenes, ambos requieren técnicas especializadas que permitan extraer características relevantes incluso bajo variaciones de iluminación, fondo o diferencias físicas entre usuarios, lo que convierte este proceso en un desafío dentro del análisis visual (Oudah *et al.*, 2020).

En los últimos años, el uso de herramientas basadas en visión por computadora ha permitido mejorar significativamente el seguimiento de movimientos de la mano en tiempo real, donde una de las tecnologías más utilizadas es Google MediaPipe Hands, debido a su capacidad para detectar y rastrear puntos clave de la mano mediante coordenadas tridimensionales sin necesidad de sensores físicos adicionales, en el estudio de *Hand tracking for clinical applications: Validation of the Google MediaPipe Hand (GMH) and the depth-enhanced GMH-D framework*.

En sus etapas iniciales se utilizaron métodos basados en segmentación y análisis de contornos para identificar la forma de la mano ya que estos algoritmos permitían extraer la silueta o los bordes para estimar características geométricas como la apertura de los dedos, la orientación del brazo o la posición relativa de la mano dentro de la imagen sin embargo este enfoque presenta limitaciones cuando el fondo no es uniforme o cuando existen variaciones bruscas en la iluminación, lo que puede afectar la calidad de la segmentación y disminuir la precisión del reconocimiento en situaciones reales (Meena, 2011). Pese a ello estos métodos sentaron las bases para el estudio del reconocimiento visual de gestos porque permitieron analizar patrones fundamentales de forma y movimiento.

La selección del algoritmo de clasificación representa una etapa fundamental dentro de los sistemas de aprendizaje automático debido a que cada modelo presenta diferentes niveles de precisión, estabilidad y capacidad de generalización dependiendo del tipo de datos utilizados. En diversos estudios se han realizado comparaciones

entre algoritmos como *Support Vector Machine* (SVM) y *Random Forest* (RF) con el objetivo de evaluar su desempeño en tareas de clasificación automática. De acuerdo con Garzón Barrero *et al.* (2023), a pesar de que el algoritmo SVM presentó mayores niveles de precisión en ciertos conjuntos de datos, *Random Forest* logró destacar por su capacidad para prevenir el sobreajuste y mantener, en consecuencia, un rendimiento estable durante el proceso de clasificación. Se elige *Random Forest* porque utiliza múltiples árboles de decisión entrenados con subconjuntos aleatorios de datos permitiendo compensar errores individuales y mejorar la capacidad de generalización del modelo es por eso que estas características convierten a *Random Forest* en una alternativa eficiente y robusta para aplicaciones relacionadas con visión por computadora y reconocimiento de patrones.

El algoritmo *Random Forest* es una de las técnicas de aprendizaje automático más utilizadas en tareas de clasificación y predicción debido a su alta precisión y capacidad para reducir problemas de sobreajuste. Este método funciona mediante la combinación de múltiples árboles de decisión entrenados con subconjuntos aleatorios de datos permitiendo mejorar la capacidad de generalización del modelo frente a nuevos datos. Según Salman *et al.* (2024), como principal ventaja de *Random Forest* se cuenta su eficiencia para trabajar con grandes volúmenes de información, al igual que su capacidad para manejar variables faltantes y datos desbalanceados. Asimismo, este algoritmo puede usarse tanto en problemas de clasificación como de regresión, siendo ampliamente utilizado en aplicaciones de salud, finanzas, comercio electrónico y sistemas inteligentes y es por ello que *Random Forest* representa una alternativa robusta y confiable para sistemas de reconocimiento y clasificación basados en visión por computadora.

Ahorabien, la calidad de los datos utilizados, durante el entrenamiento de modelos de aprendizaje automático, representa un elemento fundamental para garantizar resultados precisos y confiables en tareas de clasificación y reconocimiento, en las aplicaciones relacionadas con visión por computadora e inteligencia artificial los

conjuntos de datos, mejor conocidos como *datasets*, deben contener información consistente, equilibrada y correctamente estructurada para evitar errores durante el proceso de aprendizaje del modelo. Según Gong *et. al* (2023), uno de los factores más importantes para garantizar un desempeño adecuado en los modelos de aprendizaje automático es la calidad de los *datasets*, debido a que datos incompletos, inconsistentes o desbalanceados afectan la precisión y capacidad de generalización de los algoritmos.

La interacción humano-máquina representa un área fundamental dentro del desarrollo de sistemas inteligentes basados en visión por computadora y *machine learning* debido a que busca establecer formas de comunicación más naturales entre los usuarios y los dispositivos tecnológicos. En aplicaciones relacionadas con reconocimiento de gestos y seguimiento de movimientos.

El reconocimiento de gestos mediante inteligencia artificial y *machine learning* representa una alternativa innovadora para mejorar la interacción entre los usuarios y los sistemas digitales especialmente en aplicaciones relacionadas con realidad virtual, visión por computadora y automatización. Estas tecnologías permiten desarrollar interfaces más naturales e intuitivas mediante el análisis de movimientos realizados frente a una cámara, favoreciendo experiencias inmersivas y dinámicas. Según Cortinez *et al.* (2022), integrar herramientas como MediaPipe, TensorFlow y modelos de aprendizaje automático activa la implementación de sistemas con la capacidad de reconocer y clasificar gestos en tiempo real con resultados eficientes dentro de entornos virtuales interactivos.

Método y Metodología

La presente investigación se desarrolló bajo un diseño no experimental, debido a que las variables analizadas no fueron manipuladas directamente, sino observadas en su entorno de funcionamiento. Asimismo, el estudio presenta un carácter

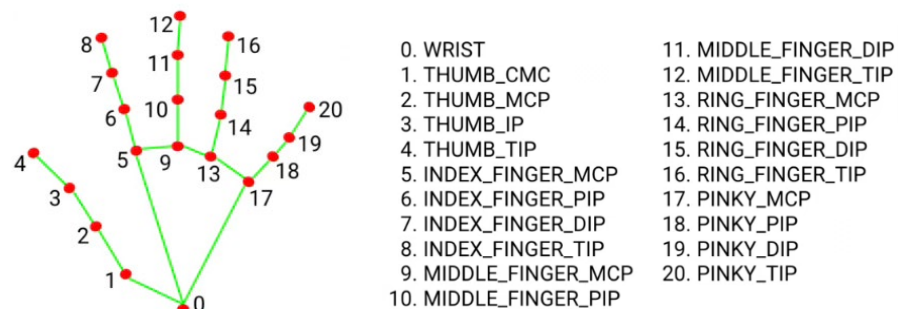
transversal o transeccional, ya que la recopilación y análisis de datos se realizaron en un único periodo de tiempo. De acuerdo con la profundidad de la investigación, el proyecto corresponde a un nivel descriptivo, puesto que tiene como propósito identificar, analizar y describir el comportamiento y desempeño de un sistema basado en visión por computadora y *machine learning* para el reconocimiento de movimientos de la mano en tiempo real.

Por otra parte, la investigación adopta un enfoque cuantitativo, debido a que se fundamenta en la recolección y análisis de datos numéricos relacionados con el desempeño del sistema desarrollado con modelos computacionales y técnicas de reconocimiento visual.

La recolección de información en la presente investigación se realizó mediante la captura y procesamiento de datos visuales obtenidos, en tiempo real, a partir de los movimientos de la mano del usuario utilizando la cámara de un teléfono conectado vía WiFi a una computadora. Las imágenes fueron procesadas mediante técnicas de visión por computadora utilizando la biblioteca MediaPipe, específicamente los modelos MediaPipe Pose y MediaPipe Hands, los cuales permitieron detectar puntos de referencia de mano y dedos como se muestra en la figura 1.

Figura 1

Landmarks de mano



Nota. Recuperado de
https://ai.google.dev/edge/mediapipe/solutions/vision/hand_landmarker?hl=es-419

A partir de los puntos de referencia detectados se realizó la extracción de las coordenadas de cada landmark identificado, estas características fueron utilizadas para crear el dataset compuesto por múltiples registros asociados a diferentes gestos, donde cada muestra incluyó las coordenadas tridimensionales (x, y, z) junto con una etiqueta correspondiente al gesto realizado.

Tabla 1

Ejemplo de estructura del dataset

X_0	Y_0	Z_0	X_2	Y_2	Z_2	X_{20}	Y_{20}	Z_{20}	Gesto
0.34	0.62	-0.01	0.36	0.60	-0.02	0.52	0.40	-0.03	Puño
0.40	0.58	-0.02	0.42	0.56	-0.01	0.60	0.30	-0.01	Mano abierta
0.30	0.58	-0.01	0.33	0.48	-0.01	0.48	0.35	-0.02	Pulgar Arriba

Nota. Elaboración propia.

El conjunto de datos generado se utilizó para entrenar un modelo de aprendizaje automático capaz de identificar diferentes gestos realizados por el usuario, este modelo recibe como entrada las coordenadas espaciales de los puntos de referencia del brazo y la mano, analizando ejemplos previamente etiquetados para aprender patrones que permitan distinguir entre distintos tipos de gestos.

El algoritmo utilizado fue *Random Forest*, seleccionado por su capacidad para manejar datos de alta dimensionalidad y su eficiencia en aplicaciones en tiempo real y una vez finalizado el entrenamiento, el modelo fue integrado al sistema para clasificar nuevos gestos detectados por la cámara en tiempo real.

Finalmente, para el desarrollo de la interfaz y la implementación del sistema se utilizaron diferentes herramientas de software libre y código abierto orientadas al procesamiento de imágenes, extracción de características y reconocimiento de movimientos.

Tabla 2
Herramientas y uso

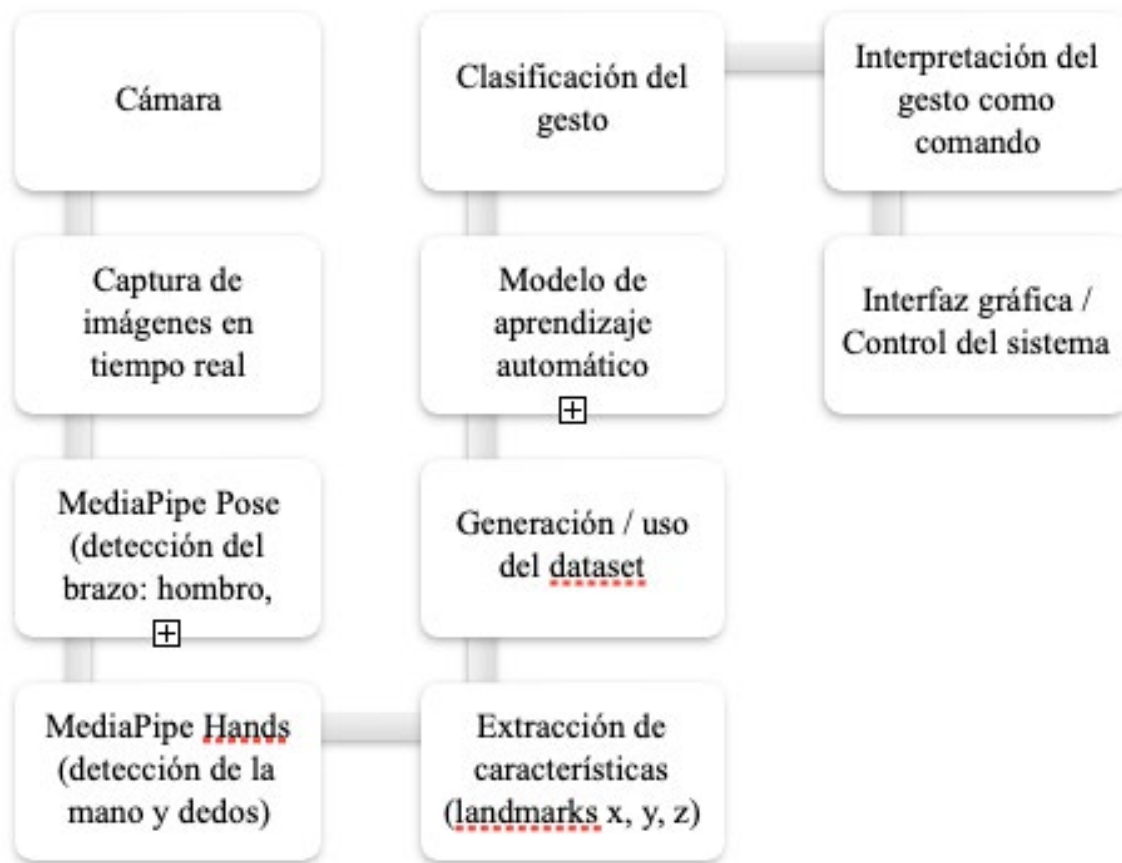
Herramienta	Uso
Visual Studio Code	Entorno de desarrollo utilizado para la programación, edición y depuración del código del sistema.
Python	Lenguaje de programación empleado para la implementación de los algoritmos de visión por computadora, procesamiento de datos y control del sistema.
MediaPipe	Biblioteca utilizada para la detección de puntos de referencia (<i>landmarks</i>) del brazo y la mano mediante visión por computadora.
OpenCV	Biblioteca utilizada para la captura y procesamiento de imágenes en tiempo real desde la cámara.
Numpy	Herramienta para el manejo y procesamiento de datos numéricos, especialmente las coordenadas de los <i>landmarks</i> .
Scikit-Learn	Biblioteca de aprendizaje automático utilizada para entrenar el modelo de clasificación basado en Random Forest.
CustomTkinter	Biblioteca utilizada para el diseño e implementación de la interfaz gráfica del sistema.

Nota. Elaboración propia.

En la Figura 2 se presenta el diagrama de flujo del sistema desarrollado, el cual describe el proceso completo desde la captura de imágenes en tiempo real hasta la clasificación de gestos mediante un modelo de aprendizaje automático. Este flujo integra las etapas de detección del brazo y la mano, extracción de características, generación del conjunto de datos y clasificación del gesto, permitiendo la interpretación de los movimientos del usuario como comandos dentro del sistema.

Figura 2

Landmarks de mano



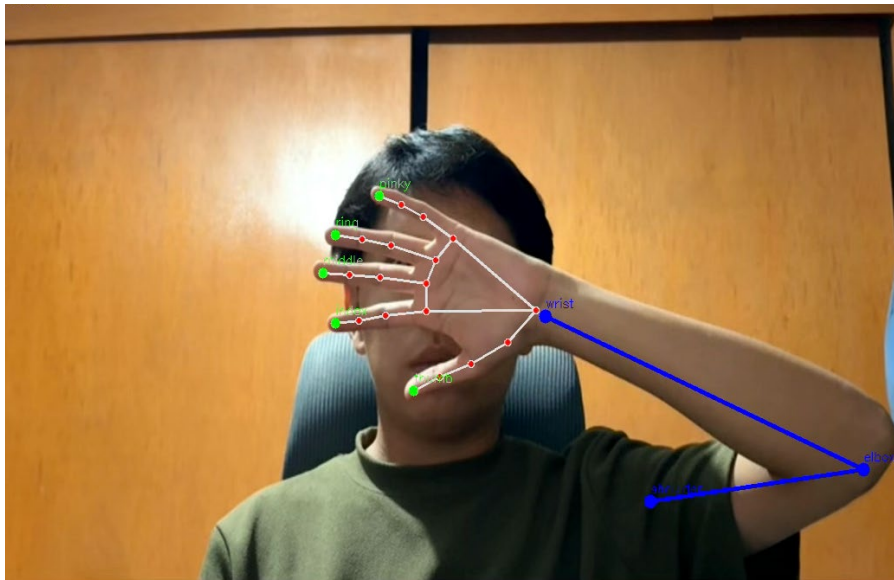
Nota. Elaboración propia.

Resultados

Se describen las distintas etapas del sistema. Primero, a través de la detección de puntos clave *landmarks*, posteriormente, la generación del conjunto de datos, el entrenamiento del modelo de clasificación y finalmente la evaluación del sistema en tiempo real analizando las métricas más relevantes como la precisión del modelo y el tiempo de respuesta con el propósito de determinar el desempeño y la eficacia del sistema.

Figura 3

Extracción de datos de los puntos clave de la mano (63 variables)



Nota. Elaboración propia.

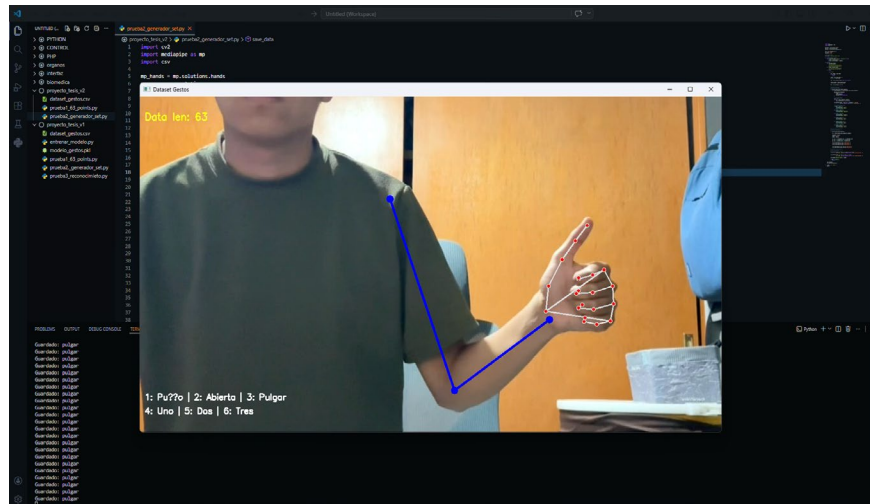
En la Figura 3 se muestra la obtención de un vector de datos de longitud 63, correspondiente a las coordenadas (x, y, z) de los 21 puntos clave de la mano. Este conjunto de datos representa la base para la construcción del *dataset* utilizado en el entrenamiento del modelo de reconocimiento de gestos.

Una vez obtenidas las coordenadas de los puntos clave de la mano, se procedió a la construcción del conjunto de datos, para ello, se implementó un mecanismo de captura mediante el cual se registran múltiples muestras de distintos gestos realizados por el usuario.

Cada muestra se almacena junto con una etiqueta que identifica el gesto correspondiente, tales como puño, mano abierta, pulgar arriba, un dedo, dos dedos y tres dedos. La captura de datos se realiza de forma manual mediante la interacción con el teclado ya que se presiona una tecla específica para registrar cada gesto.

Figura 4

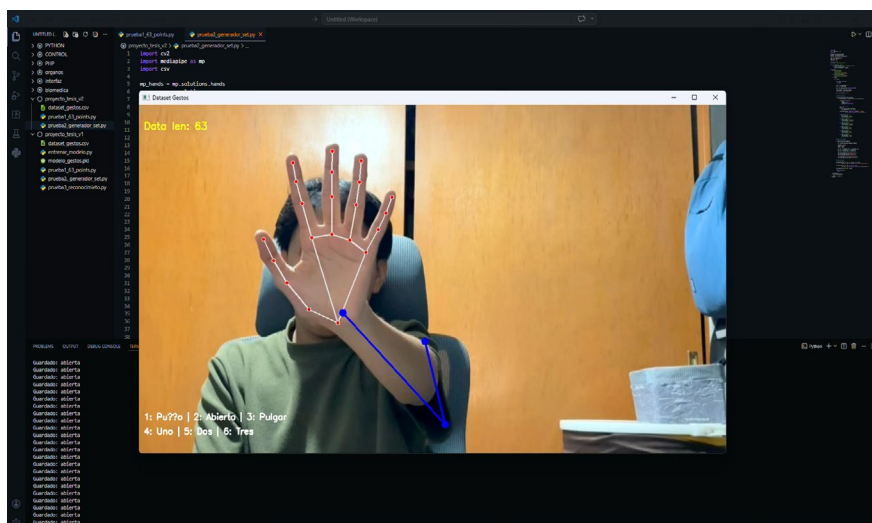
Proceso de captura de datos para la generación del dataset (pulgar)



Nota. Elaboración propia.

Figura 5

Proceso de captura de datos para la generación del dataset (mano abierta)



Nota. Elaboración propia.

En las figuras 4 y 5 se observa el proceso de captura de datos, en tiempo real, donde el sistema detecta los puntos clave de la mano y permite registrar diferentes gestos mediante la presión de teclas. Cada registro generado contiene las coordenadas tridimensionales de los puntos detectados junto con la etiqueta del gesto correspondiente, formando un *dataset* que será utilizado para el entrenamiento del modelo.

Como resultado del proceso de captura, se generó un conjunto de datos estructurado en formato CSV, en el cual cada fila representa una muestra de un gesto específico, mientras que cada columna corresponde a las coordenadas tridimensionales de los puntos clave detectados.

Figura 6

Estructura del conjunto de datos generado para el entrenamiento del modelo

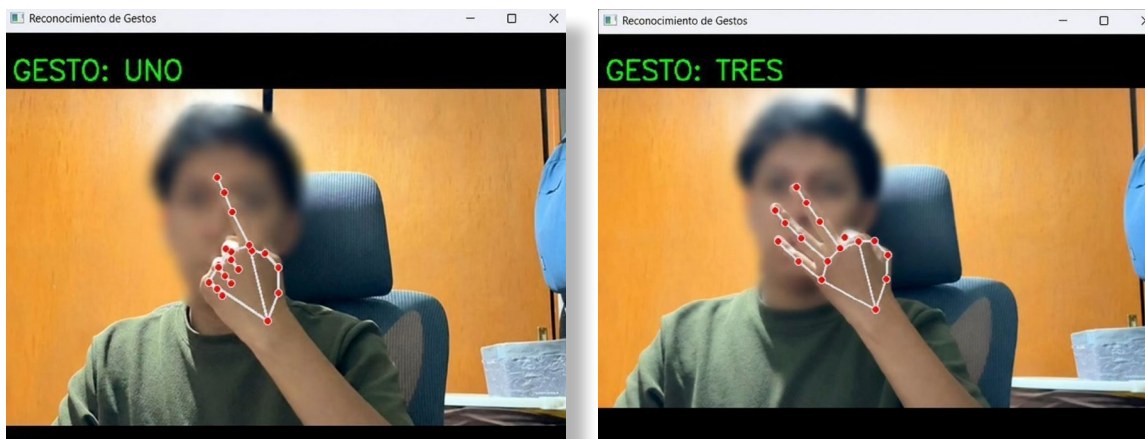
[illegible]

Nota. Elaboración propia.

En la figura 6 se observa un fragmento del *dataset* generado donde se incluyen las coordenadas (x, y, z) de los puntos clave de la mano.

Figura 8

Reconocimiento de gesto en tiempo real mediante el modelo entrenado (Un solo dedo levantado y tres dedos levantados)



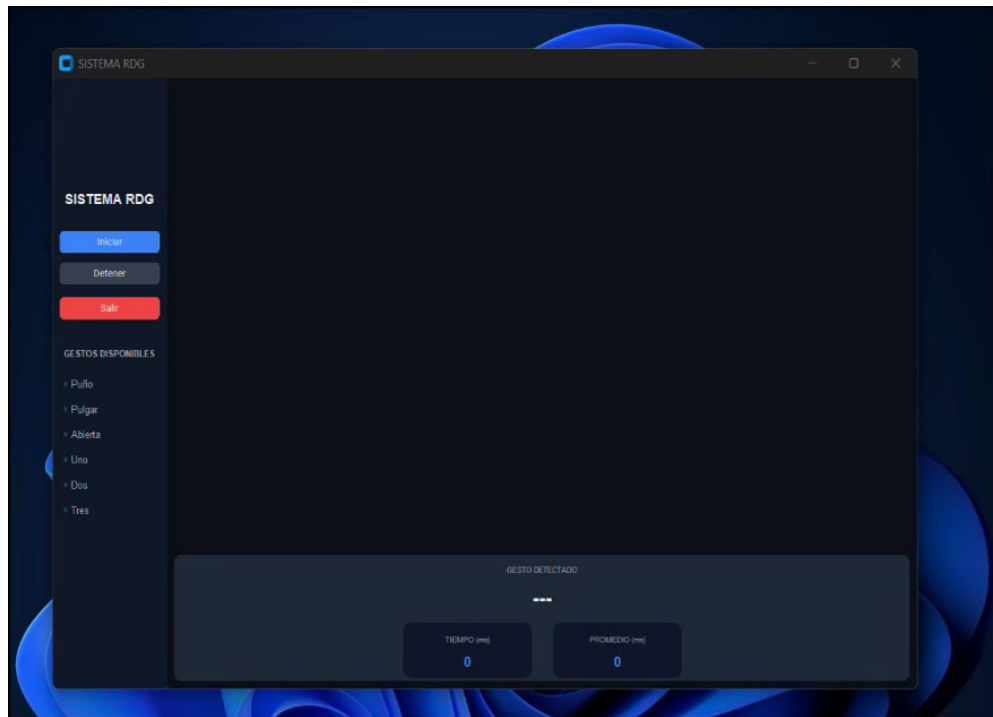
Nota. Elaboración propia.

En la Figura 8 se presentan ejemplos del sistema en funcionamiento donde se observa la correcta clasificación de gestos en tiempo real, tales como uno y tres dedos levantados. El sistema es capaz de identificar la configuración de la mano a partir de los puntos clave detectados y mostrar el resultado de manera inmediata en pantalla.

La interfaz se divide en tres secciones principales: panel lateral de control, área de visualización y panel de resultados. El panel lateral de control está ubicado en el lado izquierdo de la aplicación y contiene los elementos principales para manejar el sistema. En este panel se incluyen tres botones fundamentales: el botón “Iniciar”, que activa la captura de video y el proceso de reconocimiento; el botón “Detener”, que finaliza la captura de la cámara y el botón “Salir”, que cierra completamente la aplicación. La sección central de la interfaz está destinada a la visualización en tiempo real del video capturado por la cámara en esta área se muestra la imagen en vivo del usuario.

Figura 9

Interfaz del sistema en estado inicial

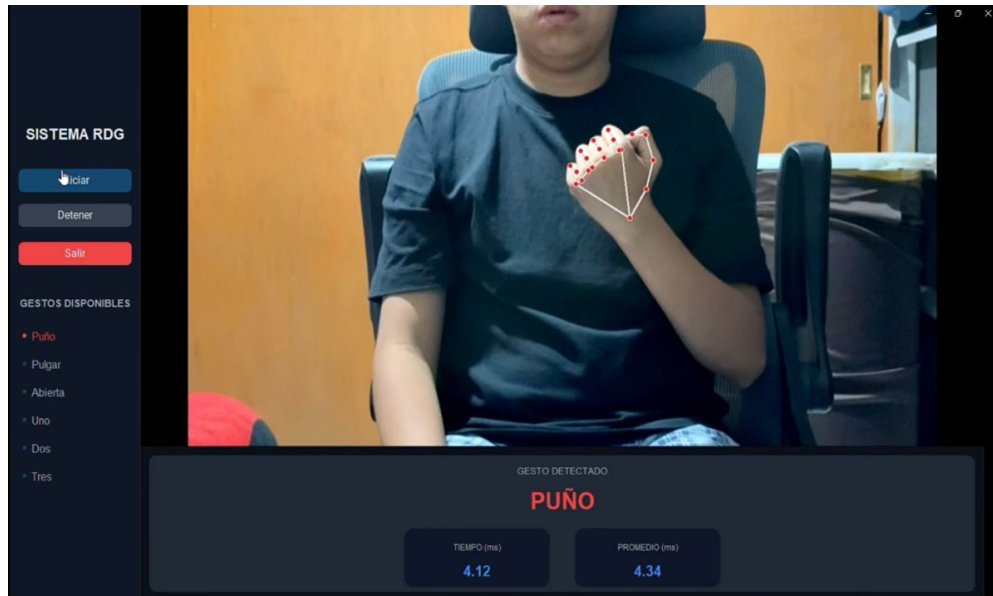


Nota. Elaboración propia.

Como se observa en la figura 9 el sistema se encuentra en espera antes de iniciar la captura de video mostrando únicamente la estructura de la interfaz y los controles disponibles. Una vez que el sistema es activado, se realiza la captura en tiempo real y se muestra la detección de la mano mediante puntos clave (*landmarks*), generados por MediaPipe.

Figura 10

Interfaz del sistema durante la detección de gestos en tiempo real



Nota. Elaboración propia.

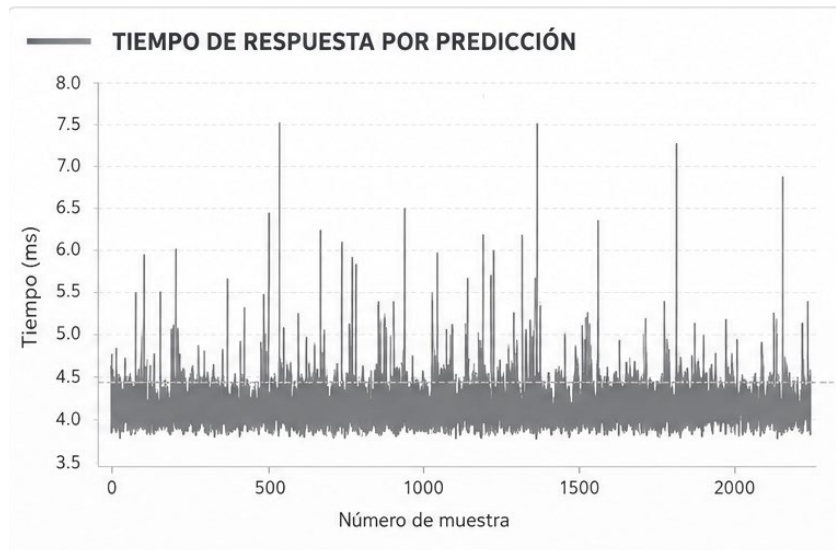
En esta etapa el sistema identifica la posición de la mano y utiliza dicha información para realizar la clasificación del gesto correspondiente.

El panel de resultados se encuentra en la parte inferior de la interfaz y muestra la información generada por el sistema en tiempo real. En esta sección se presenta el gesto detectado mediante texto destacado y con un color dinámico que varía según el tipo de gesto identificado. También se incluye el tiempo de procesamiento, que indica cuánto tarda el modelo en realizar una predicción, así como el promedio de tiempo, que muestra el valor acumulado de los tiempos de respuesta.

Con el objetivo de analizar el desempeño del sistema se realizó un registro del tiempo de respuesta durante múltiples ejecuciones del modelo. A partir de estos datos se lograron generar gráficas para que permitan evaluar la estabilidad del sistema.

Figura 11

Tiempo de respuesta por predicción



Nota. Elaboración propia.

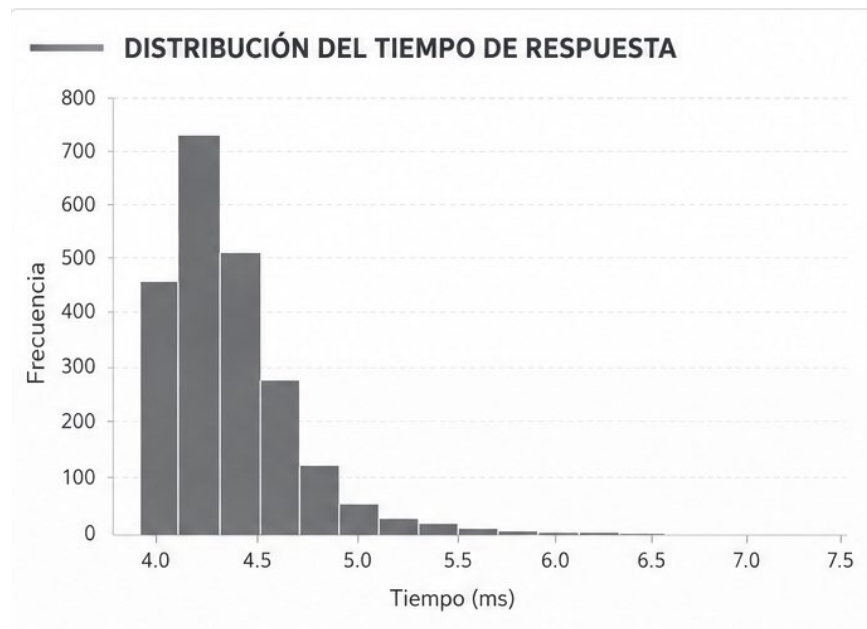
La Figura 11 muestra la variación del tiempo de respuesta por cada predicción realizada durante la ejecución del sistema.

Como se puede observar la mayoría de los valores se mantienen dentro de un rango cercano a los 4 ms, presentando únicamente algunas variaciones puntuales que no afectan significativamente el desempeño general, este comportamiento indica que el sistema mantiene una respuesta estable a lo largo del tiempo.

Adicionalmente se generó un histograma que presenta la distribución de los tiempos de respuesta obtenidos durante la ejecución del sistema.

Figura 12

Distribución del tiempo de respuesta del sistema



Nota. Elaboración propia.

Como se observa en la figura, la mayor concentración de valores se encuentra en el intervalo de aproximadamente 4.0 ms a 4.5 ms lo que indica un comportamiento consistente en el tiempo de procesamiento. Asimismo, se identifican algunos valores aislados que superan los 6 ms los cuales corresponden a variaciones poco frecuentes y no representan una tendencia significativa dentro del sistema, pero en general la distribución evidencia una baja dispersión y una alta estabilidad en los tiempos de respuesta lo cual es adecuado para aplicaciones en tiempo real.

Con el objetivo de evaluar el desempeño del modelo de clasificación, se realizaron pruebas donde se analizaron métricas de desempeño como precisión, exhaustividad (*recall*) y *F1-score*, las cuales permiten medir la capacidad del modelo para reconocer correctamente los diferentes gestos.

Tabla 3

Métricas de clasificación por gesto

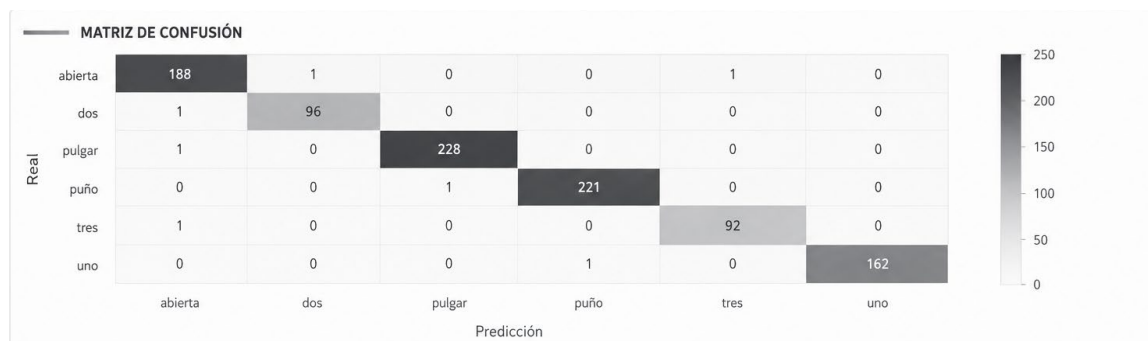
Gesto	Precisión	Recall	F1-Score
Mano abierta	0.98	0.99	0.99
Dos dedos levantados	0.99	0.99	0.99
Pulgar	1.00	1.00	1.00
Puño	1.00	1.00	1.00
Tres dedos levantados	0.99	0.99	0.99
Un dedo levantado	1.00	0.99	1.00

Nota. Elaboración propia.

Los resultados obtenidos muestran un desempeño favorable del modelo de clasificación, alcanzando una precisión general del 99.30%. Los gestos “pulgar” y “puño” presentaron los valores más altos de clasificación, mientras que las confusiones observadas fueron mínimas.

Figura 13

Matriz de confusión



Nota. Elaboración propia.

La matriz de confusión muestra que el modelo clasificó correctamente la mayoría de los gestos evaluados. Los gestos pulgar y puño presentaron un desempeño perfecto en la prueba, mientras que las confusiones fueron mínimas y se presentaron principalmente entre gestos con posiciones similares de la mano. En general, el modelo alcanzó una precisión global del 99.30%, lo que evidencia un desempeño alto en la clasificación de los seis gestos definidos.

Estos resultados evidencian que el modelo entrenado logra generalizar adecuadamente los patrones aprendidos durante la fase de entrenamiento permitiendo reconocer distintos gestos con una respuesta rápida y precisa. No obstante, se pueden presentar ligeras variaciones en la clasificación dependiendo de factores como la posición de la mano, la iluminación y la forma en que el usuario realiza el gesto.

Conclusiones y discusión

A partir de los resultados obtenidos durante el desarrollo y la evaluación del sistema, se puede afirmar que fue posible reconocer gestos de la mano en tiempo real usando técnicas de visión por computadora y aprendizaje automático. El modelo de *Random Forest* utilizado alcanzó una precisión general del 99.30% lo que indica que las coordenadas obtenidas con los *landmarks* efectivamente permiten representar de manera adecuada las diferentes posiciones de la mano para su clasificación.

A partir de la matriz de confusión se observó que la mayoría de los gestos fueron reconocidos correctamente, presentándose únicamente algunos errores mínimos entre los movimientos con características similares, esto demuestra que el modelo logró diferenciar de manera eficientes los gestos definidos en el sistema, pero si se deben considerar factores como la iluminación, la orientación de la mano o la variación en la ejecución del movimiento pueden influir en ciertos casos específicos.

Por otra parte, los tiempos de respuesta obtenidos evidencian que el sistema puede ejecutarse de manera estable en tiempo real, permitiendo una interacción rápida entre el usuario y la interfaz. Esto representa una ventaja importante para aplicaciones relacionadas con interfaces humano-máquina, ya que no requiere dispositivos físicos adicionales para el reconocimiento de movimientos.

Finalmente, aunque el sistema presentó resultados favorables, todavía existen aspectos que pueden mejorarse. Entre ellos se encuentra ampliar el conjunto de datos, incorporar más usuarios y evaluar el comportamiento del sistema en diferentes condiciones de iluminación y captura. Asimismo, como trabajo futuro se considera la posibilidad de integrar este tipo de reconocimiento en aplicaciones robóticas, sistemas inteligentes o dispositivos de asistencia basados en interacción gestual.

Referencias

- Andreu Villar, M., Martínez Hinarejos, C. D., Pons Tomas, P., Soler Domínguez, J. L., Navas Medrano, S., Ortiz Castelló, V. y García Ballesteros, M. (2024). Reconocimiento de Gestos para HoloLens2 con Puntos 3D de las Manos Bajo Limitaciones de Datos. *Interacción. Revista de la Asociación Interacción Persona Ordenador (AIPO)*, 5(2), 18-29.
- Cortinez, F. M., Montenegro, A. y Gabriel, D. (2022). *Sistema de reconocimiento y clasificación de gestos con inteligencia artificial y machine learning*. Grupo Gemis. https://grupogemis.com.ar/wp-content/uploads/2023/01/ExtraCatedra_AVRV_Articulo.pdf
- Garzón Barrero, J., Sánchez Pineda, N. E. y Londoño Pinilla, D. F. (2023). Evaluación comparativa de los algoritmos de aprendizaje automático Support Vector Machine y Random Forest: efectos del tamaño del conjunto de entrenamiento. *Revista Ciencia e Ingeniería Neogranadina*, 33(2), 131–148. <https://doi.org/10.18359/rcin.6996>
- Gong, Y., Liu, G., Xue, Y., Li, R. y Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, 162, 107268. <https://doi.org/10.1016/j.inf-sof.2023.107268>
- Lanzarini, L. C., Estrebou, C.A, Ronchetti, F., Quiroga, F., Camele, G., Rios, G., Cornejo Fandos, U., Rey, B. y Rosete, A. (2019). *Aprendizaje automático y visión por computadora*. Universidad Nacional de La Plata. <https://sedici.unlp.edu.ar/handle/10915/76973>
- Meena, S. (2011). *A Study on Hand Gesture Recognition Technique*. National Institute of Technology, Rourkela. <http://ethesis.nitrkl.ac.in/2887/>
- Oudah, M., Al-Naji, A. y Chahl, J. (2020). Hand Gesture Recognition Based on Computer Vision: A Review of Techniques. *Journal of Imaging*, 6(73), 1-29. <https://doi.org/10.3390/jimaging6080073>
- Salman, H. A., Kalakech, A. y Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, 2024, 69–79. <https://doi.org/10.58496/BJML/2024/007>
- Trigueiros, P., Ribeiro, F. y Reis, L. P. (2015). Hand Gesture Recognition System Based in Computer Vision and Machine Learning. En J. M. R. S. Tavares y R. Natal Jorge (Eds.), *Developements in Medical Image Processing and Computational Vision* (págs. 355-377). Springer. https://link.springer.com/chapter/10.1007/978-3-319-13407-9_21