Eescuela Superior Politécnica del Litoral

Facultad de Ingeniería en Electricidad y Computación

Desarrollo de un modelo de Aprendizaje Profundo para la detección de movimientos y ademanes en presentaciones del sistema RAP de ESPOL

TECH-361

Proyecto Integrador

Previo la obtención del Título de:

Ingeniero/a en Ciencias de la Computación

Presentado por:

Kevin Isaac Chévez Coronel

Pamela Nayeli Rugel Díaz

Guayaquil - Ecuador

Año: 2024

Dedicatoria

Kevin Chévez

Dedico este trabajo a Dios, por ser mi guía y fortaleza en todo momento. A mi familia cercana, mi mamá Mariuxi, mi papá Luis y mi hermano Luigi, por su amor incondicional, apoyo y comprensión. Sin ustedes, este logro no habría sido posible. Gracias por estar siempre a mi lado.

Pamela Rugel

Dedico este trabajo a mis padres, Mariuxi y Carlos, y a mis hermanos, Anggie y Jean Carlos, por su constante apoyo y motivación, que me dieron la fuerza para alcanzar mis metas. También agradezco a mis amigos de la carrera, cuya compañía hizo más llevadero el camino académico, siempre impulsándome a mejorar y explorar nuevas oportunidades.

Agradecimientos

Kevin Chévez

Agradezco profundamente a los profesores del CTI y colaboradores de ESPOL de quienes siempre aprendí algo nuevo. Su guía y apoyo fueron fundamentales para alcanzar nuestros objetivos. Su dedicación y compromiso han sido esenciales en mi formación académica y personal.

Pamela Rugel

Agradezco profundamente a los profesores de la universidad por su invaluable dedicación y sabiduría, que han sido clave en mi formación. También extiendo mi gratitud a las personas que conocí en los diferentes trabajos, cuya colaboración y compañerismo han sido esenciales en mi crecimiento personal y profesional. Gracias a todos por su constante apoyo.

Declaración Expresa

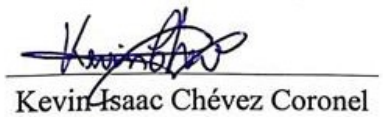
Kevin Isaac Chévez Coronel y Pamela Nayeli Rugel Díaz acordamos y reconocemos que:

La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por nosotros durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que nos corresponda de los beneficios económicos que la ESPOL reciba por la explotación de nuestra innovación, de ser el caso.

En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique los autores que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Guayaquil, 21 de Mayo del 2024.



Kevin Isaac Chévez Coronel

Pamela Nayeli Rugel Díaz

Evaluadores

Firmado electrónicamente por:
RONALD RAUL CRIOLLO BONILLA

Ing. Ronald Raúl Criollo Bonilla

PROFESOR DE LA MATERIA

Firmado electrónicamente por:
FEDERICO XAVIER DOMINGUEZ BONINI

Ph.D. Federico Domínguez Bonini

PROFESOR TUTOR

Resumen

Esta investigación desarrolla un módulo de retroalimentación automática para presentaciones orales usando un modelo de detección de ademanes basado en inteligencia artificial. El objetivo es mejorar la precisión y utilidad del sistema RAP mediante este módulo. Se hipotetiza que un modelo entrenado con datos históricos y sintéticos ofrecerá una retroalimentación más precisa. El proyecto se justifica por la necesidad de optimizar la evaluación del RAP.

Se mplearon técnicas de procesamiento de video y análisis de datos con herramientas como Python, OpenCV y MediaPipe. Se aplicaron técnicas de normalización y data augmentation para enriquecer el dataset, y se entrenaron tres tipos de modelos: uno con datos históricos, otro con datos sintéticos y un modelo combinado.

Los resultados obtenidos demostraron que el modelo entrenado con datos generados en un entorno controlado ofreció el mejor desempeño en la identificación de ademanes. La retroalimentación generada incluyó un análisis detallado de la frecuencia de acciones, acompañado de ejemplos visuales y estadísticas precisas. El modelo implementado permite una evaluación efectiva y cuantitativa de las acciones durante las presentaciones y facilita una comprensión clara del desempeño del usuario y mejora la utilidad del sistema RAP.

Palabras Clave: Detección de ademanes, Retroalimentación Automática, Procesamiento de Video, Inteligencia Artificial, Análisis de Datos.

Abstract

This research develops an automatic feedback module for oral presentations using a gesture detection model based on artificial intelligence. The main objective is to improve the accuracy and usefulness of the RAP system through this module. It is hypothesized that a model trained with both historical and synthetic data will provide more precise feedback. The project is justified by the need to optimize the current RAP evaluation.

Video processing and data analysis techniques were employed using tools such as Python, OpenCV, and MediaPipe. Techniques like normalization and data augmentation were applied to enrich the dataset, and three types of models were trained: one with historical data, another with synthetic data, and a combined model.

The results showed that the model trained with data generated in a controlled environment performed the best in gesture identification. The feedback provided included a detailed analysis of action frequency, along with visual examples and precise statistics. The implemented model enables effective and quantitative evaluation of actions during presentations, facilitating a clear understanding of user performance and enhancing the utility of the RAP system.

Keywords: Gesture Detection, Automatic Feedback, Video Processing, Artificial Intelligence, Data Analysis.

Índice General

Abreviaturas

BiLSTM	Long Short-Term Memory Bidireccional
BiLSTM-CNN	Long Short-Term Memory Bidireccional with Convolutional Neural Network
CNN	Convolutional Neural Network
Conv-LSTM	Convolutional Long Short-Term Memory
CPM	Capacidad, Planificación y Monitoreo
CTI	Centro de Tecnología de Información
DB-LSTM	Densely-Connected Bi-directional Long Short-Term Memory
ESPOL	Escuela Superior Politécnica del Litoral
FC	Fully Connected
HAR	Human Action Recognition
LSTM	Long Short-Term Memory
MMT	Multimodal Transformer Model
NTU RGB+D	Dataset RGB+D de la Universidad Tecnológica de Nanyang
RAP	Retroalimentación Automática de Presentaciones
RNN	Recurrent Neural Network
RGB	Red, Green, Blue
Wifi	Wireless Fidelity

Índice de figuras

Capítulo 1

1. Introducción

1.1 Descripción del Problema

En los entornos empresarial, industrial y estudiantil, la oratoria pública destaca como la forma de comunicación más crucial. A través de presentaciones ante audiencias grandes y pequeñas, los oradores buscan persuadir, informar, impresionar y entretener. Sin embargo, la oratoria pública también es propensa a errores, lo que puede llevar a un desempeño deficiente, malentendidos, impacto negativo en la imagen y fallas en eventos.

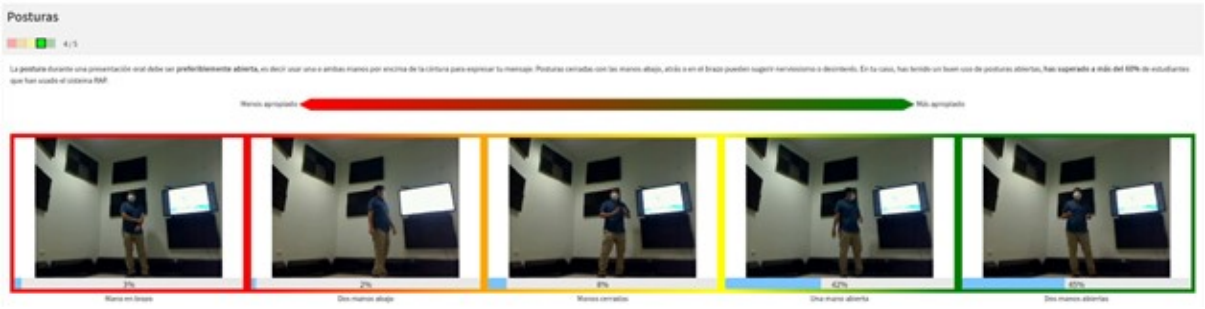
Una exposición se compone de tres fases: preparación (objetivos, análisis del público, estrategias y recopilación de información), presentación (entrega verbal y visual, adaptación a la audiencia) y preservación (manejo de preguntas para conservar el impacto positivo del discurso) Baird, 1981. Para mejorar las habilidades de comunicación del estudiante, es esencial recibir retroalimentación del profesor. No obstante, el proceso puede demorarse debido al tiempo requerido para analizar cada presentación. Por esta razón, se desarrolló el Sistema de Retroalimentación Automática de Presentaciones (RAP) en el Centro de Tecnología de Información (CTI) de la ESPOL. Utilizando una cámara y dos computadoras, RAP evalúa las presentaciones orales y envía retroalimentación automática por correo electrónico, con el objetivo de formar profesionales capaces de comunicar ideas de manera clara y convincente.

El sistema RAP actual analiza la presentación del estudiante, evaluando posturas y miradas, proporcionando un resultado en porcentaje de posturas correctas. También detecta direcciones de la mirada y analiza el audio para retroalimentar sobre volumen, uso de muletillas y pausas. La precisión para detectar comportamientos incorrectos es del 0.80 para

postura, 0.85 para mirada, 0.87 para pausas sonoras y 0.87 para volumen. Se comparó la retroalimentación del sistema con evaluadores humanos, obteniendo un consenso del 65% para postura, 78% para mirada, 75% para pausas sonoras y 71% para volumen Ochoa et al., 2018.

Al presente, el informe completo de la retroalimentación de la exposición es enviado al correo electrónico del estudiante. Incluye la calificación de la postura y el porcentaje de las cinco posturas realizadas, calificación de la calidad de la mirada, calificación del volumen de la voz, el uso de muletillas, el uso de las pausas sonoras y la calificación del contenido de las diapositivas Ochoa et al., 2018 1.

Figura 1
Reporte de retroalimentación del Sistema RAP sobre las posturas.



El lenguaje corporal es crucial en presentaciones, impactando la atención y percepción del público hacia el presentador. Sin embargo, el sistema RAP no analiza el lenguaje corporal completo del estudiante, ya que utiliza frames estáticos del video, lo que limita la evaluación de acciones dinámicas. Por ejemplo, movimientos constantes de manos podrían ser malinterpretados como una postura adecuada. Esta limitación impide al sistema proporcionar retroalimentación detallada sobre las acciones de movimiento durante la presentación, diferenciándolo de la evaluación humana.

1.2 Justificación del Problema

El proyecto pretende ampliar las funcionalidades del sistema RAP enfocadas en ademanes de los estudiantes cuando presentan. Mucho del potencial de retroalimentación del sistema RAP se pierde al estar limitado al audio y a imágenes estáticas Ochoa et al., 2018 del estudiante, especialmente en un aspecto importante como el lenguaje corporal. Un estudiante al utilizar el sistema RAP en su versión actual sin supervisión podría recibir una retroalimentación incompleta y no contemplar la importancia del lenguaje corporal y ademanes durante la presentación, impidiendo un desarrollo completo de esta habilidad. Además, el objetivo del proyecto es analizar videos cortos para ademanes que realizan las personas, por lo cual amplía la utilidad del sistema para poder ser generalizado a espacios diferentes o para vídeos ya grabados de exposiciones de otras personas como referencia. Es conveniente de implementar ya que el sistema ya utiliza una cámara para la detección de la postura del estudiante, por lo que se podría adicionar el análisis de la nueva funcionalidad de ademanes a partir de los videos grabados.

Por consiguiente, se propone agregar una sección adicional en el informe remitido por correo electrónico, con el propósito de visualizar los patrones recurrentes de ademanes inapropiados durante un intervalo de tiempo. Esta sección utilizará el mismo formato del reporte existente, incluyendo la calificación general del estudiante, el puntaje y ejemplos de los momentos del video donde se detectaron los patrones de ademanes. Dichos patrones serán evaluados durante la inspección de los videos para obtener clasificaciones pertinentes del lenguaje corporal. Los patrones más importantes a detectar incluyen el movimiento de las manos al presentar, ya que son los que más influyen en la presentación. Otros ademanes importantes pueden ser el movimiento excesivo del cuerpo y ademanes asociados con la ansiedad o la falta de confianza Rodero, 2022. Esto beneficiaría a los docentes encargados de la retroalimentación de presentaciones, a los encargados de este proyecto interno de la

ESPOL, y principalmente a los estudiantes que utilizan las aulas RAP para mejorar sus exposiciones.

1.3 Objetivos

1.3.1 Objetivo General

Desarrollar un modelo de Aprendizaje Profundo que detecte ademanes de expositores utilizando inteligencia artificial, optimizado y validado mediante vídeos de exposiciones reales del sistema RAP de ESPOL.

1.3.2 Objetivos Específicos

1. Seleccionar y etiquetar ademanes relevantes a partir de videos históricos de las aulas RAP.
2. Generar datos sintéticos para cada ademán seleccionado, con el fin de crear un conjunto de datos idóneo y libre de ruido, adecuado para un buen aprendizaje del modelo.
3. Entrenar un modelo inicial utilizando los datos sintéticos, creando así una base sólida para la posterior clasificación precisa de los ademanes.
4. Automatizar la creación de un dataset utilizando el modelo preentrenado con datos sintéticos, aplicándolo a videos de las aulas RAP de ESPOL.
5. Realizar técnicas de aumento de datos, incluyendo la normalización de puntos, inversión de puntos en el eje X y desplazamiento en 5 frames, para reentrenar el modelo preentrenado y obtener un modelo final robusto y altamente confiable.

1.4 Marco Teórico

1.4.1 CNN

En un estudio Siriak et al., 2019 sobre la detección de acciones de manos por videos se propuso un modelo enfocado en la eficiencia para detectar acciones en tiempo real utilizando dispositivos móviles. El modelo utilizó redes neuronales convolucionales (CNN) para la extracción de características de las poses que luego eran ingresadas a una Máquina de Soporte de Vectores (SVM) para la clasificación de las poses. El modelo se implementó en dispositivos móviles y evidenció una alta precisión (95%) en pruebas de tiempo real. Se concluyó que este método de aprendizaje profundo se puede usar para clasificar acciones directamente de los frames del video y que, gracias a su mayor eficiencia en comparación a otros modelos tradicionales, puede implementarse en dispositivos móviles ejecutándose en tiempo real Siriak et al., 2019.

1.4.2 RNN

En un estudio Luo et al., 2018 sobre detección de poses en videos, se creó un modelo de inteligencia artificial utilizando Redes Neuronales Recurrentes (RNN) con unidades de Memoria a Largo-Corto Plazo (LSTM). Este modelo pretendía sustituir detectores de pose que utilizaban imágenes y por ende capturas estáticas de videos que eran limitados e ineficientes computacionalmente. El modelo tenía que superar problemas derivados de los videos como degradación de la imagen en el tiempo o la coherencia geométrica del cuerpo para poder identificar la correlación temporal de las poses (las poses dependen de frames anteriores). Se concluye que las unidades LSTM pueden ayudar a aprender dependencias de largo alcance y mejorar los resultados en detecciones de videos superando de manera satisfactoria los problemas mencionados. El modelo resultante se probó con un dataset considerable y se concluyó que tenía más precisión y eficiencia comparado con los

detectores de poses por capturas estáticas.

1.4.3 DB-LSTM Y LSTM

En Ullah et al., 2018 se propuso un método de reconocimiento de acciones que combina RNN y Redes Profundas de Memoria Bidireccional a Largo Plazo (DB-LSTM). . CNN se usó para extraer características de fotogramas individuales de un video, mientras que DB-LSTM se usó para aprender las relaciones secuenciales entre estas características. El método propuesto es capaz de aprender dependencias a largo plazo en videos, lo que lo hace muy adecuado para tareas de reconocimiento de acciones. Se utilizó una red LSTM bidireccional profunda para aprender información secuencial a largo plazo, procesar videos largos mediante el análisis de características durante un cierto intervalo de tiempo y se evaluó en tres conjuntos de datos de referencia: UCF-101, YouTube 11 Actions y HMDB51. Logró mejoras significativas en el reconocimiento de acciones en comparación con los métodos más avanzados en estos conjuntos de datos.

En He et al., 2020 se propuso un método para avanzar el reconocimiento inteligente de vehículos autónomos dirigido a la detección de acciones de agentes de tránsitos. Se generó un conjunto de datos de acciones agentes que incluía videos de la policía de tránsito, donde se etiquetaron acciones presentes en cada frame de videos en RGB. Se empleó un método para extraer características creadas por los investigadores denominadas “Longitud Relativa del Hueso” y “Ángulo con la Gravedad” de los datos esqueléticos humanos. Estas características actuaban como un enlace entre la Máquinas de Pose Convolucional (CPM) y la LSTM, mejorando significativamente la precisión del reconocimiento. Se realizó una prueba de reconocimiento en tiempo real (prueba en línea) para validar la disponibilidad del método cuando está conectado a internet. Esta prueba se llevó a cabo en cuatro diferentes lugares y como resultado, se obtuvo un promedio total de precisión del 0.9329.

Se observa que una de las opciones viables es la combinación de las CNN con LSTM

o como se refiere también Conv-LSTM. Existen modelos disponibles para omitir mucho del entrenamiento de la parte convolucional de la red como lo es Detrecton2 de Facebook y la librería de mediapipe de Google, que extraen coordenadas de puntos clave de la persona para mejorar la rapidez del entrenamiento y eficiencia de predicción con la desventaja de desligar el contexto del fondo de la imagen del entrenamiento. En Wang et al., 2023 se comparan estos métodos usando bases de datos de acciones públicas, concluyendo que ambos alcanzan precisiones superiores al 90%, siendo opciones viables para la detección de gestos en diversas aplicaciones.

1.4.4 Transformers

Una investigación Do and Kim, 2022 reciente destaca el interés creciente en el reconocimiento de acciones humanas en interiores (HAR). Se propone un novedoso modelo transformer multimodal (MMT) que integra eficientemente datos de imágenes RGB y secuencias de esqueletos para mejorar el rendimiento del reconocimiento de acciones. Se utilizaron tres conjuntos de datos de referencia públicos para evaluar el rendimiento de su método. El conjunto de datos NTU RGB+D contiene 56.880 secuencias de acciones de 3D, el conjunto de datos NTU RGB+S contiene 112.745 secuencias de acciones y el conjunto de datos Northwestern-UCLA contiene 20.000 secuencias de acciones. Los tres conjuntos de datos incluyen diferentes clases de acciones, desde acciones simples como caminar hasta acciones complejas como jugar al baloncesto. El método propuesto logra un rendimiento competitivo en los tres conjuntos de datos, lo que demuestra su versatilidad.

Los resultados muestran que el MMT supera consistentemente a los métodos unimodales y logra resultados comparables a enfoques de última generación con tan solo ocho cuadros de entrada. Este enfoque demuestra la utilidad del transformador como una herramienta poderosa para mejorar el reconocimiento de acciones humanas al capturar dependencias a largo plazo en datos secuenciales.

Una investigación Ren et al., 2023 propuso un nuevo método para el reconocimiento de acciones de estudiantes en aulas inteligentes se basa en un transformer SlowFast Swin mejorado. Este modelo combina las ventajas de los modelos SlowFast y Swin Transformer para aprender relaciones a largo y corto plazo en videos. Se entrenó en un conjunto de datos de video de aulas inteligentes que contiene videos de estudiantes realizando diferentes acciones como escribir, hablar escuchar, etc. Los resultados muestran que el método es capaz de reconocer con precisión las acciones de los estudiantes, con una precisión del 92,3%. Este método tiene el potencial de ser utilizado para una variedad de aplicaciones en aulas inteligentes, como el seguimiento del aprendizaje, la detección de problemas de comportamiento y la personalización de la instrucción.

1.4.5 Random Forest

En Meenakshisundaram and Ramkumar, 2022 se exponen otro método de aprendizaje supervisado que sirve para la clasificación a partir de datos llamado Random Forest, utilizado en el ámbito médico para detección de cáncer cervical. El concepto del clasificador es tomar diferentes valores para tener varios árboles de decisión en el entrenamiento con datos reales. Luego los resultados de los diferentes árboles se combinan para llegar a una sola clasificación que en el estudio tuvo un alto grado de acierto y de detección. Los resultados muestran que este método de aprendizaje de máquina puede tener mejores resultados con menores tiempos de entrenamiento en clasificación de categorías a partir de un conjunto grande de datos reales.

En Ding et al., 2021 se propone un método novedoso para la clasificación de acciones utilizando varias cámaras en diferentes puntos conectadas en un entorno Wi-Fi. Se basa en el concepto de aprendizaje por transferencia o metaaprendizaje, que es utilizar el conocimiento final adquirido de un modelo como base de otro para ampliar el entrenamiento de una manera eficiente. En el estudio se propone que los modelos aprenden las

características comunes de cada localidad para extraer solo las características distintivas para cada acción y poder realizar una mejor predicción. Se concluye que este sistema tiene altos grados de acierto y que puede ser implementado en diferentes localidades con la ventaja de necesitar pocos pocos datos de ejemplo de las acciones para el entrenamiento.

1.4.6 Multilabel

El etiquetado multilabel implica asignar a un objeto varias etiquetas, un ejemplo es la clasificación de emociones en un texto. En Rajabi et al., 2020 se propuso un modelo, BiLSTM-CNN multicanal, el cual aborda esta tarea capturando información contextual con la capa BiLSTM y detectando patrones locales con la capa CNN en secuencias de texto. Esta combinación proporciona una representación vectorial que logra una precisión de clasificación multietiquetada de emociones en texto informal en español, superando el rendimiento de modelos basados únicamente en RNN o CNN. La arquitectura BiLSTM-CNN multicanal se destaca como una herramienta eficaz para esta tarea.

En Javadi and Lim, 2021 se trabaja con un conjunto diverso de 200 videos etiquetados que capturan expresiones de ira en contextos culturales norteamericanos y persas. Este conjunto de datos, único en su enfoque cultural, aborda la carencia de diversidad en conjuntos de datos emocionales previos. Cada video, anotado con seis categorías de emociones y 13 emojis relacionados con la ira, permite entrenar modelos de reconocimiento de emociones capaces de identificar múltiples emociones en un solo video.

La relevancia radica en la capacidad de estos modelos para detectar combinaciones de emociones, como ira y frustración, lo cual es común en la expresión emocional. Esta habilidad tiene aplicaciones en la interacción humano-computadora, robótica social y diagnóstico de salud mental. En la interacción humano-computadora, los modelos multilabel pueden adaptar el comportamiento del sistema en tiempo real según las emociones detectadas, mejorando la experiencia del usuario. En la robótica social, estos modelos

permiten a los robots ajustar su comportamiento para ser más comprensivos y compasivos en situaciones emocionales. En el ámbito de la salud mental, la detección de signos emocionales múltiples posibilita la identificación temprana de trastornos como la depresión o la ansiedad, facilitando un diagnóstico y tratamiento más oportunos.

Capítulo 2

2. Metodología

En este capítulo se detalla la metodología utilizada para el desarrollo del sistema de detección de ademanes en las aulas RAP. Se presentan los requerimientos funcionales y no funcionales del sistema, así como el análisis y clasificación de los ademanes más comunes observados en los videos históricos de las presentaciones. Además, se describen las técnicas de preprocesamiento de datos y las etapas de prototipado del modelo, seguidas de la arquitectura del modelo de aprendizaje profundo empleado. Finalmente, se abordan las evaluaciones y validaciones realizadas para asegurar la precisión y efectividad del sistema propuesto.

2.1 Análisis

En esta sección, se analizan los diversos aspectos de la solución propuesta para la detección y análisis de ademanes en las aulas RAP de ESPOL. Este análisis abarca desde los requerimientos funcionales y no funcionales, hasta el alcance, limitaciones, riesgos, beneficios, usuarios, prototipado y evaluación de la solución.

2.1.1 Requerimientos

Para desarrollar adecuadamente el sistema de detección y análisis de ademanes en las aulas RAP de ESPOL, es fundamental definir claramente los requerimientos que guiarán el diseño y la implementación. Estos requerimientos se dividen en funcionales y no funcionales, abarcando desde la selección intuitiva de ademanes relevantes hasta la validación del modelo final utilizando datos reales. Los requerimientos no funcionales, por

otro lado, aseguran la eficiencia, precisión, escalabilidad, adaptabilidad y mantenimiento del sistema para garantizar su operatividad continua y efectiva en diversos contextos y condiciones.

2.1.1.1. Requerimientos Funcionales

- **Selección y etiquetado de ademanes relevantes:** se debe permitir a los investigadores seleccionar y etiquetar ademanes relevantes a partir de videos históricos de las aulas RAP. Este proceso debe ser intuitivo y permitir una selección precisa y eficiente.
- **Generación de datos sintéticos:** una vez seleccionados los ademanes, se debe generar datos sintéticos para cada ademán. Estos datos deben ser representativos y de alta calidad para garantizar un buen aprendizaje del modelo.
- **Entrenamiento del modelo inicial:** el sistema debe facilitar el entrenamiento de un modelo inicial utilizando los datos sintéticos generados. Este modelo debe servir como base para la posterior clasificación precisa de los ademanes.
- **Automatización de la creación de datasets:** el sistema debe automatizar la creación de un dataset utilizando el modelo preentrenado con datos sintéticos. Este dataset debe incluir datos obtenidos de las aulas RAP de ESPOL y ser representativo y de alta calidad.
- **Aplicación de técnicas de Data Augmentation:** el sistema debe permitir la aplicación de técnicas de Data Augmentation para reentrenar el modelo preentrenado. Estas técnicas deben incluir normalización, inversión de puntos y desplazamiento de frames.
- **Validación del modelo final:** el sistema debe validar el modelo final utilizando vídeos de exposiciones reales del sistema RAP de ESPOL para asegurar que funcione adecuadamente en escenarios reales.
- **Interfaz gráfica para usuarios finales:** el sistema debe proporcionar una interfaz gráfica que permita a los usuarios finales probar el modelo y visualizar los resultados.

La interfaz debe ser intuitiva y mostrar la frecuencia de los ademanes, recomendaciones basadas en percentiles y los movimientos mayormente realizados.

2.1.1.2. Requerimientos No Funcionales

- **Eficiencia:** el sistema debe ser capaz de procesar y analizar grandes volúmenes de datos de manera eficiente, minimizando el tiempo de espera para los usuarios.
- **Precisión y calidad de los datos:** los datos generados y etiquetados por el sistema deben ser de alta confiabilidad y bajo error para asegurar resultados confiables en el análisis de los ademanes.
- **Escalabilidad:** el sistema debe ser escalable para manejar un aumento en la cantidad de datos y usuarios sin comprometer el rendimiento.
- **Adaptabilidad:** el sistema debe ser capaz de adaptarse a diferentes contextos culturales y estilos de presentación, ya que los ademanes y el lenguaje corporal pueden variar considerablemente.
- **Mantenimiento:** el sistema debe ser fácil de mantener y actualizar, asegurando que tanto el hardware como el software se mantengan en buen estado y operativos.

2.1.2 Alcance y limitaciones de la solución

Las limitaciones y los alcances que se destacan en el desarrollo de la solución son los siguientes:

- La dependencia de la posición y las características de la cámara al grabar al expositor, ya que cualquier modificación puede afectar la precisión en la predicción de los ademanes.
- La necesidad de un conjunto de datos amplio y diverso para entrenar el modelo de manera efectiva, lo que puede requerir tiempo y recursos significativos.
- La capacidad del sistema para adaptarse a diferentes contextos culturales y estilos de presentación, ya que los ademanes y el lenguaje corporal pueden variar

considerablemente.

2.1.3 Riesgos y beneficios de la solución

Riesgos

- **Precisión del Modelo:** existe el riesgo de que el modelo no alcance la precisión deseada en la detección de ademanes. El uso del modelo podría generar retroalimentaciones incorrectas o incompletas debido a que existe la posibilidad que detecte más de un ademán incluido en el movimiento del expositor.
- **Dependencia Tecnológica:** la solución puede depender de hardware específico (cámaras de alta calidad) y software que deben mantenerse actualizados para asegurar su correcto funcionamiento.

Beneficios

- **Mejora en la Retroalimentación:** al proporcionar una evaluación más completa del lenguaje corporal y los ademanes, se espera que los estudiantes reciban retroalimentación más precisa y útil, mejorando sus habilidades de presentación.
- **Automatización del Proceso:** la automatización del análisis de presentaciones permite una retroalimentación rápida y consistente, reduciendo la carga de trabajo de los profesores y permitiendo un seguimiento más continuo del progreso de los estudiantes.
- **Aplicabilidad Generalizada:** la solución puede ser utilizada en diversas áreas y contextos, no solo en entornos académicos sino también en capacitaciones empresariales y otras actividades que requieran presentaciones orales.

2.1.4 Usuarios de la solución

- **Estudiantes:** que utilizan el sistema RAP para mejorar sus habilidades de presentación a través de la retroalimentación automática y detallada sobre su desempeño.
- **Docentes:** que supervisan y evalúan las presentaciones de los estudiantes, beneficiándose de una herramienta que proporciona retroalimentación precisa y reduce su carga de trabajo.
- **Administradores de Sistemas de Aprendizaje:** Encargados de implementar y mantener el sistema RAP, asegurando su correcta integración y funcionamiento.

2.1.5 Prototipado

En esta sección se presenta el pipeline 2 del nuevo módulo de detección de ademanes para las Aulas RAP. Este pipeline describe el proceso completo desde que una persona comienza su exposición en las aulas, pasando por la evaluación automatizada de los ademanes, hasta la retroalimentación final mostrada en el Sistema Web del RAP. También se presenta una interfaz gráfica 4 5 6 7 8 9 10 11 con el objetivo de proporcionar una vista clara y comprensible de cómo la solución analiza y califica las presentaciones.

Figura 2

Pipeline del módulo de detección de ademanes en las Aulas RAP, desde la grabación hasta la retroalimentación por correo.

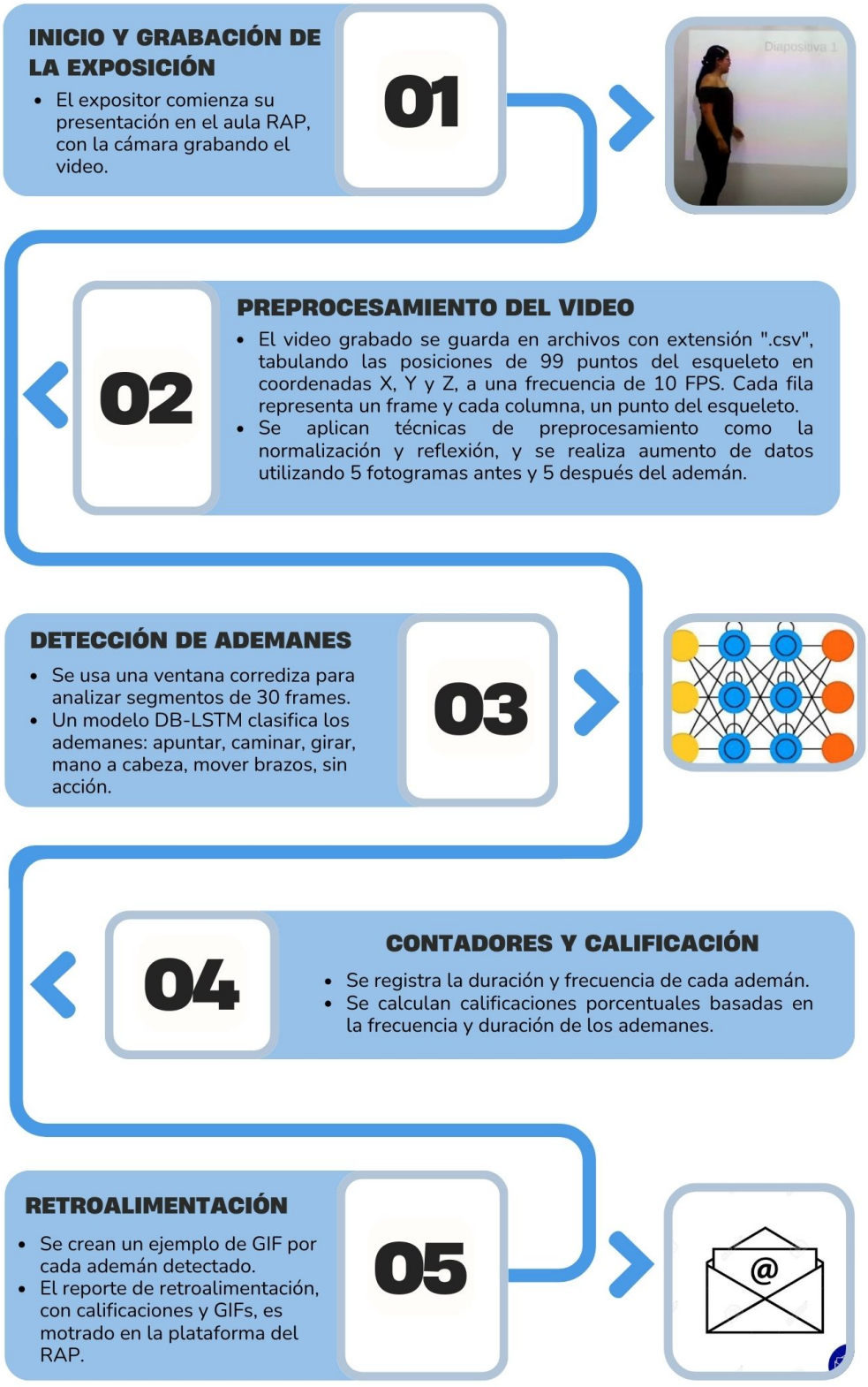


Figura 3
Pantalla de inicio de la interfaz.

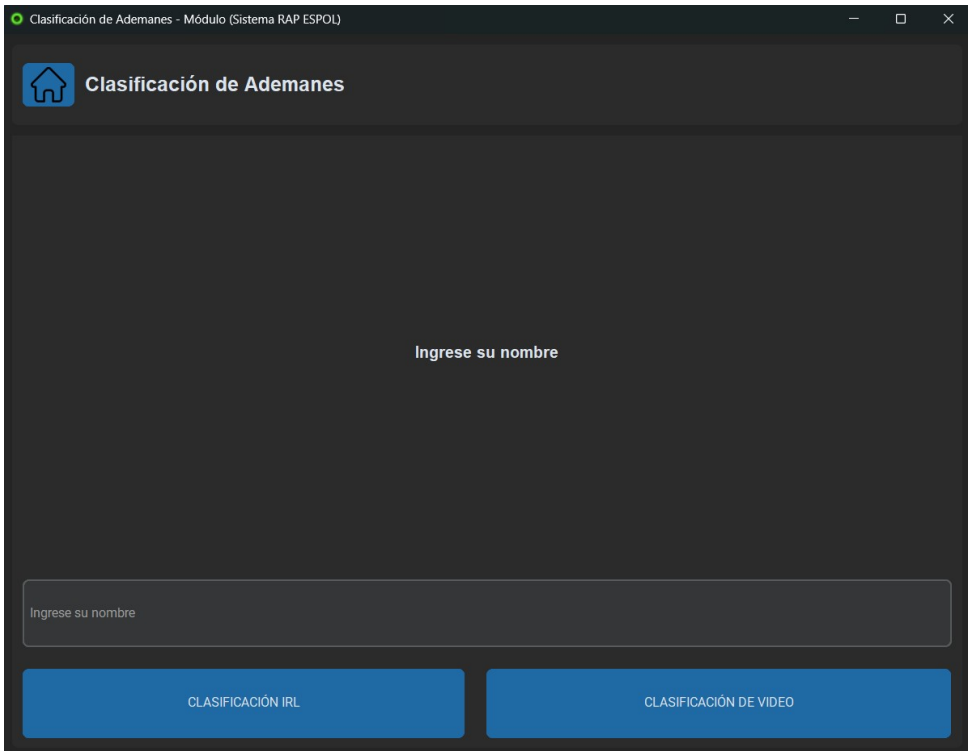


Figura 4
Pantalla de inicio de la interfaz.

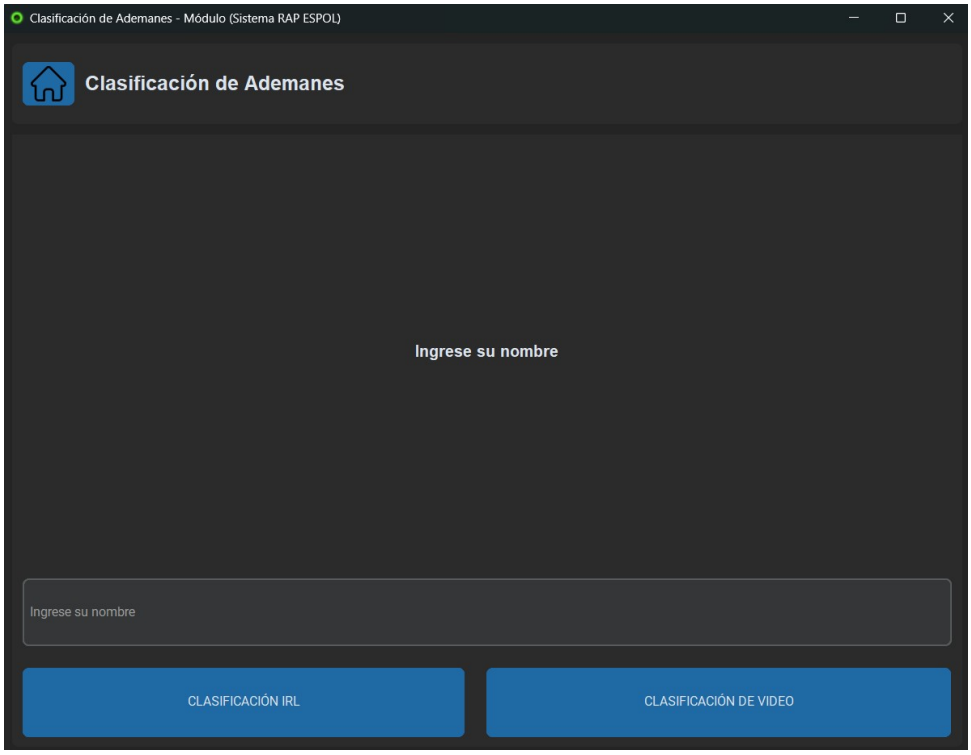


Figura 5
Pantalla en caso de que el usuario desee grabar su video desde la interfaz.

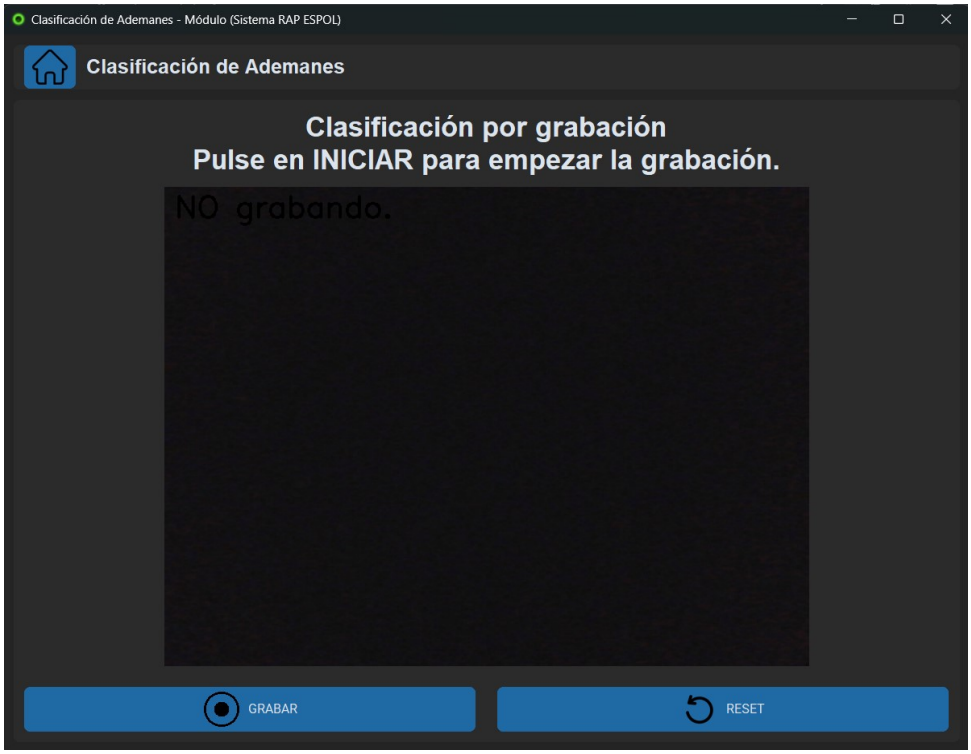


Figura 6
Pantalla mientras se graba el video desde la interfaz.

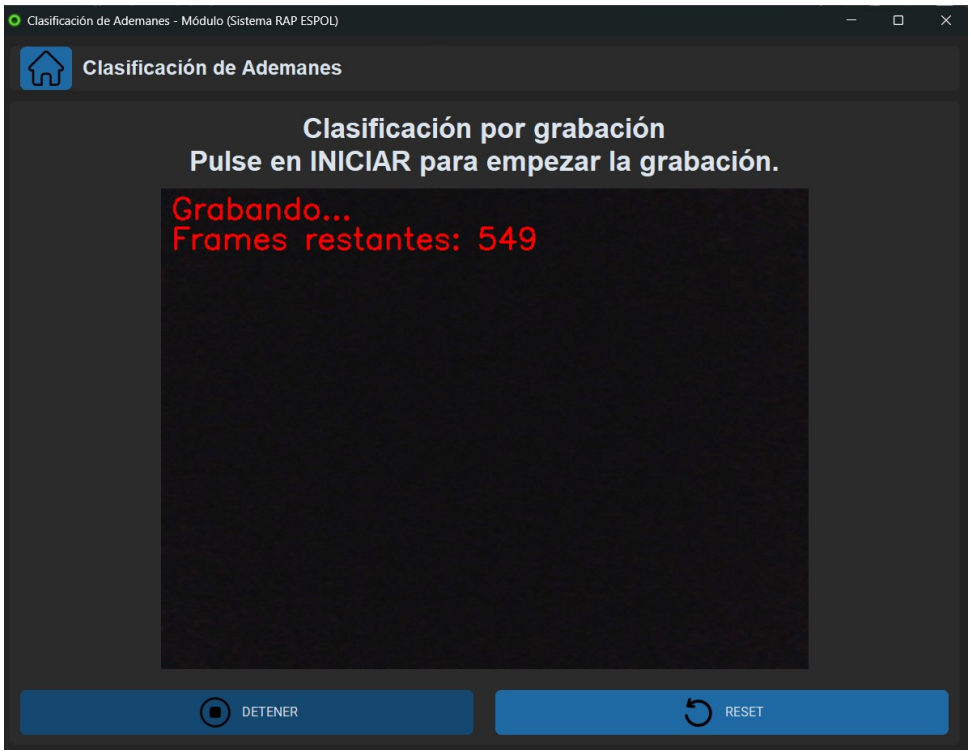


Figura 7
Pantalla en caso de que desee subir un video pregrabado.

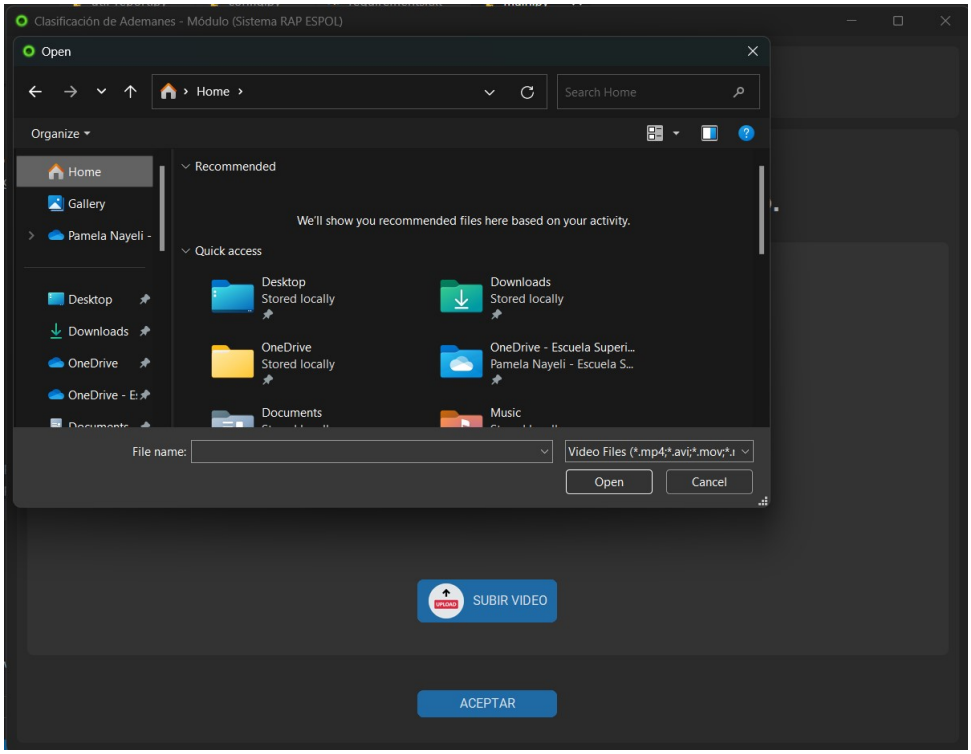


Figura 8
Pantalla mientras el video se analiza.

Figura 9
Pantalla con resultados sección 1 - Top 3 de ademanes más frecuentes durante todo el video.

Figura 10
Pantalla con resultados sección 2 - Porcentaje de duración de cada ademán durante todo el video.

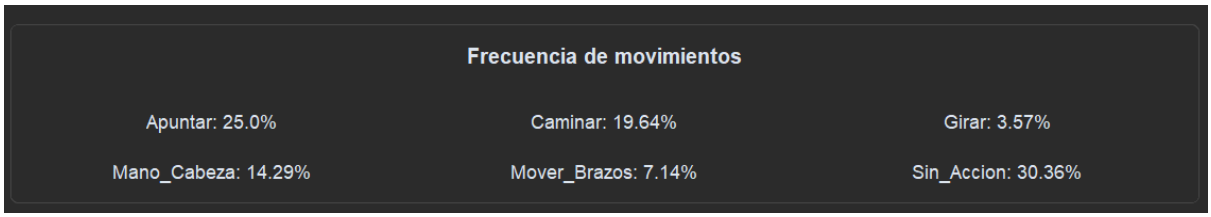
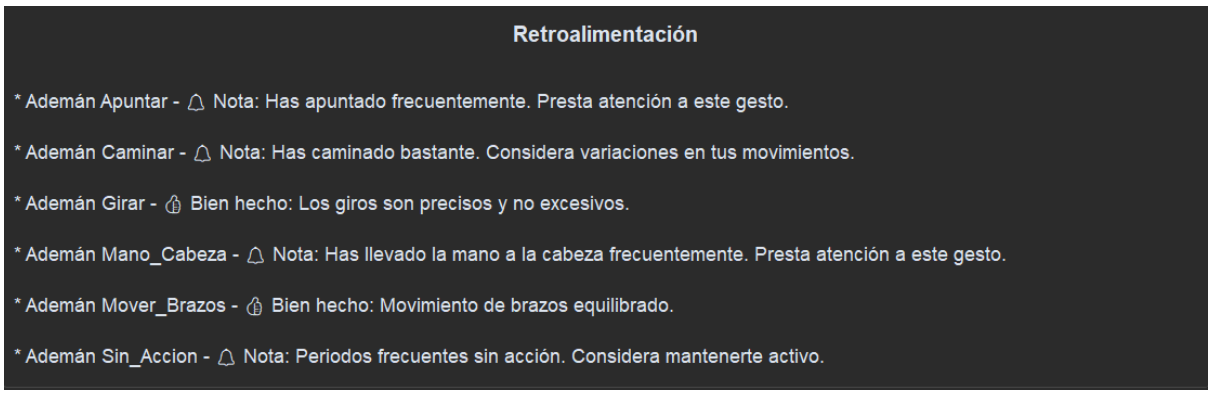


Figura 11
Pantalla con resultados sección 3 - Mensajes de retroalimentación.



2.1.6 Evaluación

La evaluación del prototipo se llevó a cabo con la retroalimentación de nuestro tutor, quien proporcionó varias recomendaciones valiosas para mejorar el sistema.

Nuestro tutor recomendó probar tres tipos de modelos diferentes para evaluar cuál ofrece un mejor desempeño en la detección de ademanes:

- Un modelo entrenado únicamente con datos históricos.
- Un modelo entrenado exclusivamente con datos generados bajo un entorno controlado.
- Un modelo combinado, entrenado con una mezcla de datos históricos y sintéticos.

Además, sugirió que al momento de mostrar los ademanes en la retroalimentación, se incluyan GIFs de ejemplo para cada ademán detectado. Esto permitiría a los usuarios finales entender mejor las acciones que se están evaluando y cómo se representan visualmente.

Estas recomendaciones tienen como objetivo asegurar que el sistema no solo sea preciso y efectivo, sino también intuitivo y fácil de comprender para los usuarios finales.

2.2 Diseño de la solución

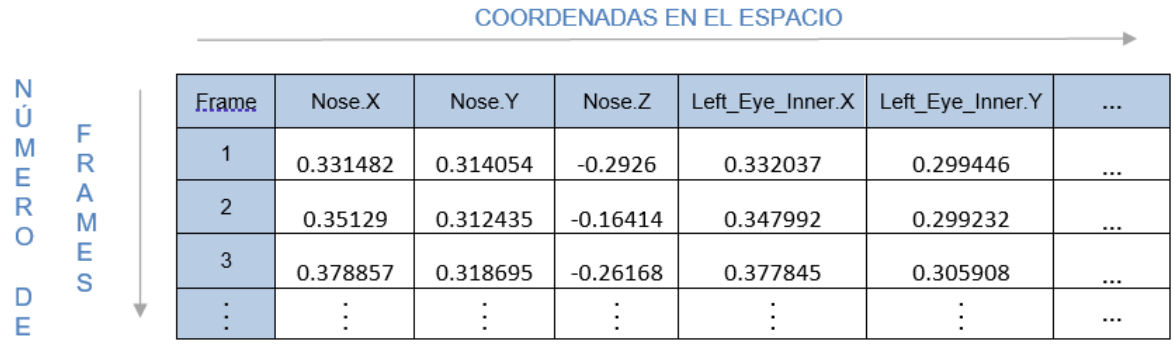
En el diseño de la solución, se llevó a cabo un proceso de procesamiento de datos que implicó el análisis de cada video, seguido por la identificación de los ademanes críticos más recurrentes y el aumento de datos.

2.2.1 Preprocesamiento

Para el preprocesamiento del conjunto de datos, se analizaron 150 videos históricos de las aulas RAP. Los videos fueron grabados con cámaras de video comunes (espacio RGB) y posteriormente almacenados en archivos con extensión ”.csv” 12, se tabularon las posiciones de los puntos del esqueleto pertinentes de cada video, con una frecuencia de 10 frames por segundo. Cada columna representa un punto del esqueleto en coordenadas X, Y

y Z, teniendo un total de 99 puntos, y cada fila representa un frame. Los datos contenidos en la tabulación son números que van desde 0 hasta 1 en las coordenadas X y Y y desde -1 a 0 en la coordenada Z.

Figura 12
Ejemplo de datos tabulados de los archivos ".csv."



Frame	Nose.X	Nose.Y	Nose.Z	Left_Eye_Inner.X	Left_Eye_Inner.Y	...
1	0.331482	0.314054	-0.2926	0.332037	0.299446	...
2	0.35129	0.312435	-0.16414	0.347992	0.299232	...
3	0.378857	0.318695	-0.26168	0.377845	0.305908	...
⋮	⋮	⋮	⋮	⋮	⋮	...

A través de los datos tabulados y con la ayuda de OpenCv, Numpy, Pandas y Mediapipe 13, se generaron videos conectando los puntos y simulando los ademanes que realiza la persona al momento de exponer 14. Se realizó un proceso manual exhaustivo, en el cual se visualizaron los videos históricos de las aulas RAP generados con Mediapipe para identificar ademanes comunes que los estudiantes realizan al exponer. Paralelamente, se extrajeron datos tanto del frame inicial como del frame final para los ademanes identificados en los videos 12.

Dando como resultado los siguientes ademanes detectados en los videos, los cuales serán las clases para el entrenamiento del modelo:

Apuntar

- Se capturó el ademán del expositor al apuntar a la diapositiva que se encuentra detrás de él.
- Se consideró la acción de levantar una o dos manos extendidas hacia las diapositivas.
- Se incluyó el giro cuando la persona está apuntando, siempre y cuando en el movimiento se incluya mínimo el levantamiento de una mano con el objetivo de señalar

Figura 13
Pseudocódigo de la generación de videos

```
función agregar_frame(img, frame):
    landmarks_de_frame = frame.iloc[1:134]
    crear lista_de_landmarks
    para cada índice en el rango de 0 a la longitud de landmarks de frame con salto 3:
        crear un landmark
        guardar la coordenada x del landmarks_de_frame en el índice actual
        guardar la coordenada y del landmarks_de_frame en el índice actual + 1
        guardar la coordenada z del landmarks_de_frame en el índice actual + 2
        agregar el landmark a la lista_de_landmarks
    crear instancia de mp_drawing
    crear instancia de mp_pose
    dibujar landmarks en la imagen usando mp_drawing y lista_de_landmarks

función renderizar_video(datos, nombre_archivo, fps=5, predicciones=None):
    XMAX = 1500
    YMAX = 1000
    obtener los frames de datos
    crear objeto VideoWriter para guardar el video
    para cada índice, frame en los frames:
        crear una imagen en blanco del tamaño (YMAX, XMAX, 3)
        definir fuente para el texto
        crear texto con información del video y el frame actual
        agregar texto a la imagen
        agregar landmarks al frame usando la función agregar_frame
        escribir el frame en el objeto VideoWriter
    devolver el objeto VideoWriter
```

Figura 14
Video generado con las coordenadas conectadas

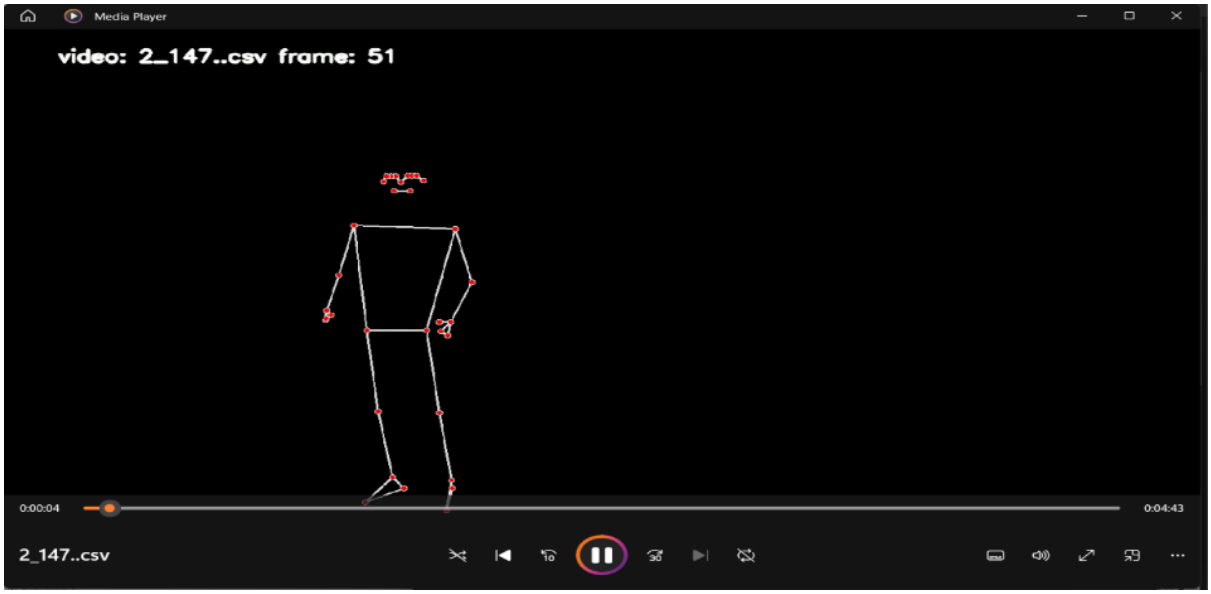


Figura 15
Ejemplo de datos extraídos a partir del análisis de cada video de un ".csv"

Código de video	Rango de frames	Descripción del gesto
2_50	1122-1150	Gesto señalando la diapositiva con la mano derecha
2_50	1880-1890	Gesto señalando la diapositiva con la mano derecha

hacia la diapositiva.

Caminar

- Se capturó el ademán del expositor al caminar a través del escenario de grabación, bien sea de un extremo hacia el otro o hasta la mitad del escenario.
- Se consideró una vez que el cuerpo está girado aproximadamente 90 grados de su posición inicial y está a punto de dar el primer paso.
- Los pasos pueden ser hacia adelante o hacia atrás.

Girar

- Se capturó el ademán del expositor al girar aproximadamente 90 grados hacia la diapositiva que se encuentra detrás de él, .
- Se debe tomar en cuenta que ninguna mano esté extendida para evitar la confusión con el ademán de apuntar.
- El ademán puede ser un giro parcial, es decir, sin incluir el giro que haría a la persona regresar su postura hacia el frente. Si el ademán es rápido, puede incluir el giro de vuelta al frente.

Mano_Cabeza

- Se consideró el ademán del expositor tocándose la cabeza con una o ambas manos. La mano o manos pueden estar en cualquier posición relativa: una sola mano en la cabeza, ambas manos en la cabeza, una mano en cada lado de la cabeza, etc.
- La acción puede incluir movimientos como rascarse la cabeza, colocarse la mano en la frente, en las sienes, o cualquier otra parte de la cabeza.
- Se incluyeron tanto movimientos sutiles como más evidentes, siempre y cuando el contacto con la cabeza sea claro.

Mover_Brazos

- Se consideraron los movimientos de los brazos de manera significativa, ya sea con un solo brazo o con ambos.

- La acción puede incluir movimientos amplios y abiertos, como agitar los brazos, levantar y bajar los brazos, o cualquier movimiento brusco y visible de los brazos.
- Los movimientos pueden ser hacia adelante, hacia los lados, hacia arriba o hacia abajo, siempre y cuando sean claramente perceptibles y significativos.
- Se incluyeron tanto los movimientos simétricos (ambos brazos moviéndose de manera similar) como los asimétricos (cada brazo moviéndose de manera diferente).
- Se consideraron tanto movimientos rápidos como movimientos más lentos y sostenidos, siempre y cuando sean evidentes y distintivos.

Sin_Accion

- Se consideraron los periodos en los que el expositor no realiza movimientos significativos o sus movimientos son mínimos.
- La acción incluye la inactividad de los brazos, tanto en posiciones arriba como abajo, con el cuerpo relativamente estático.
- También se considera sin acción cuando el expositor se mantiene en una postura estática, con movimientos leves y naturales que no constituyen una acción específica.
- Se incluye cualquier pequeña oscilación o ajuste postural que ocurra de forma natural pero que no sea parte de una acción intencionada.
- La acción puede ser identificada durante los momentos en los que el expositor está hablando pero sin realizar ademanes destacables.
- Esta categoría ayuda a establecer una línea base para diferenciar las acciones significativas del expositor, facilitando la identificación de acciones más distintivas.

Cabe destacar que los ademanes Mano_Cabeza, Mover_Brazos y Sin_Accion son considerados independientemente de la orientación del expositor, es decir el expositor puede estar de frente, de lado o de espaldas realizando la acción.

Se observó que el promedio de duración de un ademán es de aproximadamente 3 segundos, lo cual equivale a 30 FPS. Por lo tanto, los archivos ”.csv” de las exposiciones

fueron segmentados por ademanes, con cada segmento conteniendo 30 frames. Se utilizó la técnica de la ”ventana corrediza” para extraer varios archivos de 30 frames para cada ademán identificado. De esta manera, se crearon 10 archivos adicionales por ademán, comenzando desde 5 frames antes y después de lo identificado en el análisis de los videos. Esto permitió generar datos de entrenamiento más completos, evitando la elección arbitraria del inicio del ademán. Se realizaron etapas de preprocesamiento de los datos, incluyendo el escalamiento de los puntos críticos, la reflexión en el eje X y la normalización de los puntos sin afectar el movimiento en dicho eje.

Es importante destacar la definición de normalización en el contexto de los datos utilizados en el proyecto. La normalización de los videos consiste en representar todo el conjunto de puntos en un espacio ”estándar”. El objetivo principal fue eliminar cualquier distintivo en los videos para que el modelo pudiera enfocarse en los ademanes realizados, eliminando dependencias de ademanes por altura y complexión de la persona. De esta forma, se eliminaron muchas variables dependientes de características específicas de cada individuo.

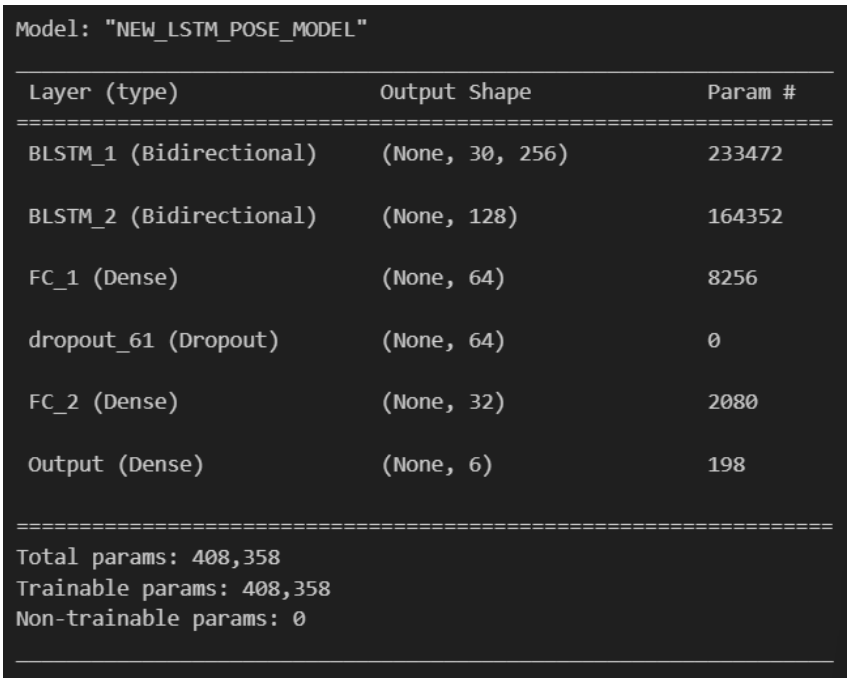
Además, se creó un dataset sintético, con 12 videos de con 5 minutos duración grabados por dos personas con características diferentes, realizando un video por cada ademán. Estos videos pasaron por el mismo preprocesamiento que los archivos históricos.

2.2.2 Arquitectura del modelo

La arquitectura del modelo 16 se creó considerando la cantidad de datos para el entrenamiento como también la cantidad de características a analizar en el modelo.

El modelo se inicia con una capa de entrada (inputs) que toma datos de entrada con la forma 30, 99 y la cantidad de los datos que se van a enviar para el modelo. El valor de 30 se refiere a la cantidad de frames que se van a enviar para cada gesto y 99 representa las

Figura 16
Estructura del Modelo LSTM creado para la clasificación de ademanes



coordenadas para los 33 puntos del cuerpo correspondiente a las columnas de los archivos csv.

A continuación se detallan las capas seleccionadas en el modelo creado junto a los hiperparámetros seleccionados.

2.2.2.1. Capas LSTM Bidireccionales. Se utilizan dos capas LSTM Bidireccionales para procesar la secuencia de frames. La primera capa LSTM tiene 128 unidades (neuronas) con un dropout de 0.3, la segunda capa LSTM tiene 64 unidades con un dropout de 0.5. Se utiliza la configuración Bidireccional por la naturaleza de acciones y de la forma en que se realizan, debido a que se debe dar igual importancia el final del ademán como el principio.

2.2.2.2. Capa Densa (Fully Connected 1 | FC_1). Luego de las capas LSTM Bidireccionales, hay una capa densa con 64 unidades, con una función de activación ReLU. Esta capa busca extraer características más abstractas de las salidas de las capas LSTM.

2.2.2.3. Capa Dropout. Luego de la capa densa, se aplica una de dropout del 0.5 con la intención de reducir el overfitting del modelo.

2.2.2.4. Capa Densa (Fully Connected 2 | FC_2). Luego de la capa Dropout vuelve a haber una capa densa, esta capa tiene 32 unidades, con una función de activación ReLU. En esta capa se reducen las características de manera que se extraen valores más cercanos a la clasificación.

2.2.2.5. Capa de Clasificación. Finalmente, la capa de salida consiste en una capa densa que produce las salidas de clasificación. El número de neuronas en esta capa depende del número de clases a clasificar. Si hay únicamente dos clases la capa tiene una neurona y utiliza una función de activación sigmoide. Si hay más de dos clases, utiliza una función de activación SoftMax. En este caso hay 6 clases, es decir la salida es de 6 unidades con una función de activación softmax la cual predice el top 6 de probabilidades en la clasificación de ademanes.

2.2.2.6. Hyperparámetros y Optimizador. Se utilizó el método de optimización "Adam" con una taza de aprendizaje *(learning_rate)* de 0.001, tamaño de lote *(batch_size)* de 32, épocas *(epoch)* de 100 iteraciones con la posibilidad de tener una finalización temprana en el entrenamiento *(early_stop)* con una paciencia *(patience)* de 5 iteraciones enfocada en el valor de pérdida del conjunto de validación *(val_loss)*. Se utilizó un 20% del conjunto de datos de entrenamiento *(training_set)* para validación *(validation_set)*, y de ese 20% se seleccionó de manera aleatoria el 20% para en conjunto de datos de prueba *(testing_set)*.

Se diseñó una arquitectura para tener en cuenta la secuencialidad de los puntos críticos a lo largo del tiempo, permitiendo que cada dato influya en el modelo utilizando unidades DB-LSTM. Por ello, se procedió a entrenar un primer modelo basado en DB-LSTM, utilizando solo datos sintéticos del aula RAP. Posteriormente, se reentrenó el mismo modelo añadiendo los archivos históricos. Ambos modelos fueron evaluados mediante el envío de un video completo y en tiempo real, buscando la identificación de los ademanes.

Capítulo 3

3. Resultados y análisis

Los resultados del proyecto ofrecen una visión detallada sobre el desempeño del modelo de detección de ademanes, evaluando su capacidad para mejorar la precisión y utilidad del sistema de retroalimentación automática. A continuación, se presentan y analizan los datos obtenidos, proporcionando una evaluación comprensiva del rendimiento del modelo y su efectividad en la retroalimentación durante las presentaciones orales.

3.1 Plan de Implementación

El diagrama de Gantt 17 proporciona una visión detallada y estructurada del cronograma del proyecto, desde su inicio hasta la entrega final. El diagrama ilustra claramente las etapas clave del proyecto, incluyendo la socialización del problema, la identificación de objetivos y requerimientos, el desarrollo y ajuste del modelo, y la integración final al Sistema RAP. Además, se detallan las fases de documentación y pruebas, que son fundamentales para asegurar la calidad y efectividad del proyecto. La secuencia temporal de cada tarea y la interdependencia entre las fases destacan la planificación rigurosa necesaria para cumplir con los objetivos del proyecto y asegurar una implementación exitosa.

3.2 Pruebas

Se analizaron tres enfoques principales: entrenamiento con datos históricos, entrenamiento con datos generados en un entorno controlado, y entrenamiento con una combinación de datos históricos y generados. Cada enfoque fue evaluado tanto en entornos controlados como reales, utilizando matrices de confusión para medir la confiabilidad y

Figura 17
Diagrama de Gantt

NÚMERO EDT	TÍTULO DE LA TAREA	FECHA DE INICIO	FECHA DE ENTREGA	DURACIÓN
1	**Análisis e inicio del proyecto**			
1.1	Socialización del problema	17/05/24	21/05/24	3
1.1.1	Reunión con el tutor	22/05/24	22/05/24	1
1.2	Reunión con el cliente	23/05/24	23/05/24	1
1.3	Uso de las Aulas RAP	24/05/24	24/05/24	1
1.4	Identificación de los objetivos	27/05/24	29/05/24	3
1.5	Identificación de los requerimientos	28/05/24	30/05/24	3
1.6	Estado del arte	30/05/24	31/05/24	2
1,7	Documentación de la definición del Problema	21/05/24	31/05/24	9
2	**Desarrollo del proyecto**			
2,1	Elección de arquitectura del modelo	31/05/24	03/06/24	2
2,2	Selección de los ademanes en base a videos históricos	03/06/24	04/06/24	2
2,3	Grabaciones en las aulas RAP para generar datos sintéticos	03/06/24	07/06/24	5
2,4	Creación de un modelo inical con los datos sintéticos	10/06/24	11/06/24	2
2,5	Automatización en la creación del dataset que contenga datos reales del Aula RAP	05/06/24	05/06/24	1
2,6	Aplicación de ténicas de aumento de datos	13/06/24	14/06/24	2
2,7	Reentrenar el modelo inicial con datos reales	14/06/24	21/06/24	6
2,8	Documentación de la metodología	04/06/24	21/06/24	14
3	**Pruebas del Proyecto**			
3.1	Ajuste de parámetros del modelo	24/06/24	01/07/24	6
3.2	Validación del modelo con videos reales	02/07/24	09/07/24	6
3.3	Pruebas del modelo en tiempo real	10/07/24	12/07/24	3
3.4	Documentación de la solución	24/06/24	12/07/24	15
4	**Ensamblaje del proyecto**			
4.1	Código integrador del modelo al Sistema RAP	15/07/24	19/07/24	5
4.2	Integración al Sistema RAP	22/07/24	26/07/24	5
4.3	Pruebas del modelo ensamblado en el Sistema RAP	29/07/24	02/08/24	5
4.4	Documentación Final	15/07/24	29/07/24	11
4.5	Preparación 5 min Pitch	30/07/24	02/08/24	4

precisión del modelo en la detección de ademanes. Los resultados muestran variaciones significativas en el desempeño del modelo según el tipo de datos y el entorno de prueba, proporcionando una visión integral de sus fortalezas y áreas de mejora.

3.2.1 Pruebas de modelo entrenado únicamente con datos históricos

Para entrenar el modelo, se utilizó un conjunto de datos compuesto exclusivamente por exposiciones grabadas en las aulas RAP. El conjunto de datos históricos incluyó específicamente los ademanes más comunes identificados, proporcionando una representación auténtica de los movimientos y comportamientos de los expositores en el entorno natural del aula. El modelo fue entrenado utilizando solo estos datos históricos para aprender las características y patrones específicos de los ademanes observados en las grabaciones, con el objetivo de asegurar que el sistema pudiera identificar y clasificar los ademanes en contextos similares a los del entorno de origen.

3.2.1.1. Evaluación del modelo con datos de un entorno controlado. A través de la matriz de confusión 18, utilizando datos de validación generados en un entorno controlado, se obtuvo de manera general una **confiabilidad del 50.00%**, además se observó que el modelo tiene un alto desempeño en la predicción de los ademanes *Caminar* y *Mano a Cabeza*. Sin embargo, el ademán *Apuntar* se confunde frecuentemente con *Caminar*. Además, el modelo detectó *Girar* como *Caminar* en un 62% de los casos, y *Mover Brazos* se confundió con *Mano a Cabeza* en un 69% de las ocasiones.

3.2.1.2. Evaluación del modelo con datos de un entorno real. A través de la matriz de confusión 19, utilizando datos de validación obtenidos de videos grabados por usuarios en las aulas RAP, se obtuvo de manera general una **confiabilidad del 44.31%**, además se observó que el modelo detecta correctamente los ademanes *Caminar* y *Mano a Cabeza*. Sin embargo, el ademán *Girar* fue clasificado erróneamente como *Caminar* en un 100% de los casos.

Figura 18

Matriz de confusión con datos de un entorno controlado para la evaluación del modelo con datos históricos

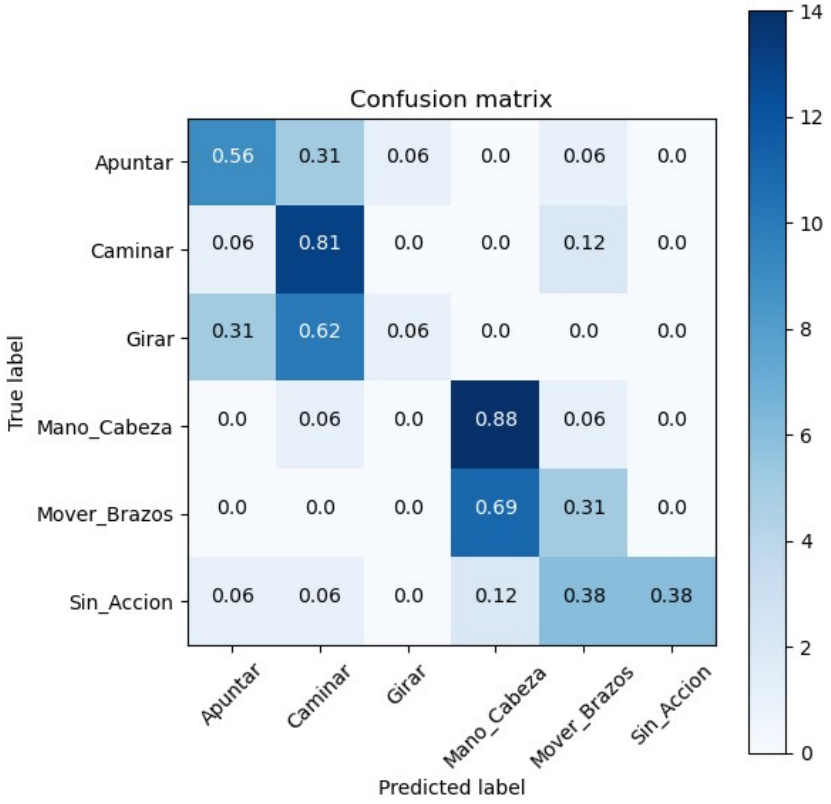
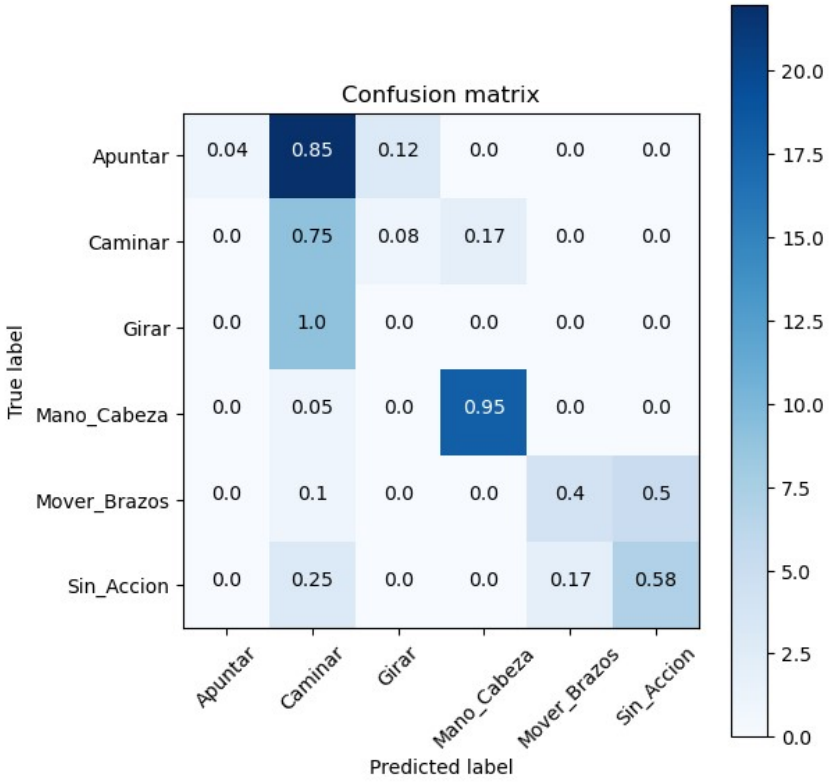


Figura 19

Matriz de confusión con datos de un entorno real para la evaluación del modelo con datos históricos



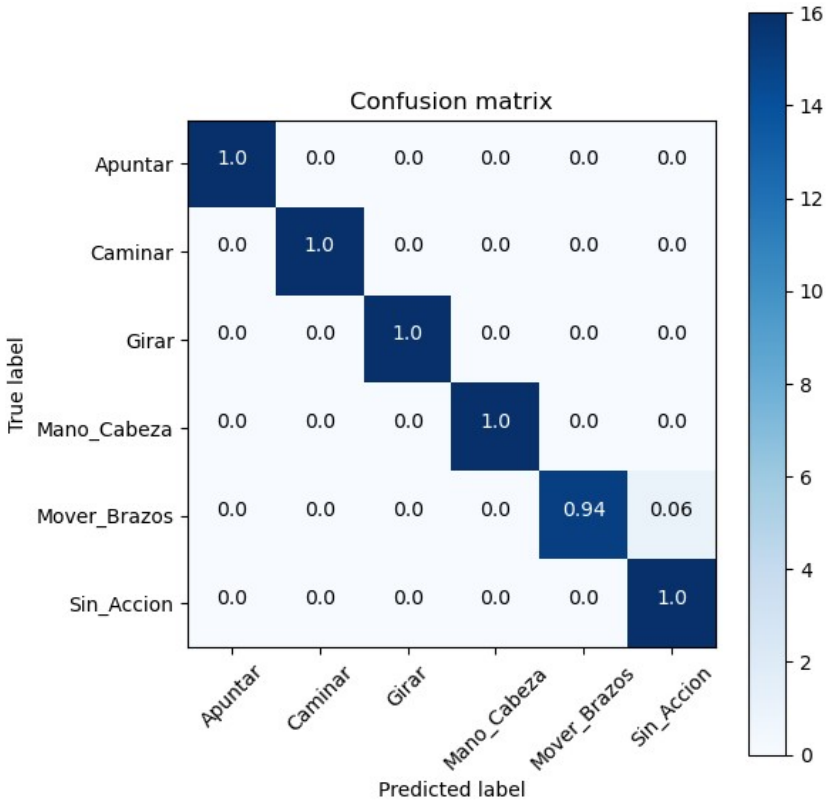
3.2.2 Pruebas de modelo entrenado únicamente con datos generados bajo un entorno controlado

Se realizaron pruebas con un modelo entrenado exclusivamente con datos generados en un entorno controlado. Este enfoque permitió evaluar el desempeño del modelo en condiciones artificialmente optimizadas, garantizando que los datos de entrenamiento fueran consistentes y específicos para el entorno simulado. Las pruebas se centraron en determinar cómo el modelo generaliza a partir de estos datos controlados y en identificar posibles áreas de mejora antes de su aplicación en escenarios más variados y reales.

3.2.2.1. Evaluación del modelo con datos de un entorno controlado. A través de la matriz de confusión 20, utilizando datos de validación generados en un entorno controlado, se obtuvo de manera general una **confiabilidad del 98.95%**, además se observó que el modelo muestra un alto desempeño en la predicción de todos los ademanes, siendo *Mover*

Brazos el menos preciso con un 94% de aciertos.

Figura 20
Matriz de confusión con datos de un entorno real para la evaluación del modelo con datos de un entorno controlado



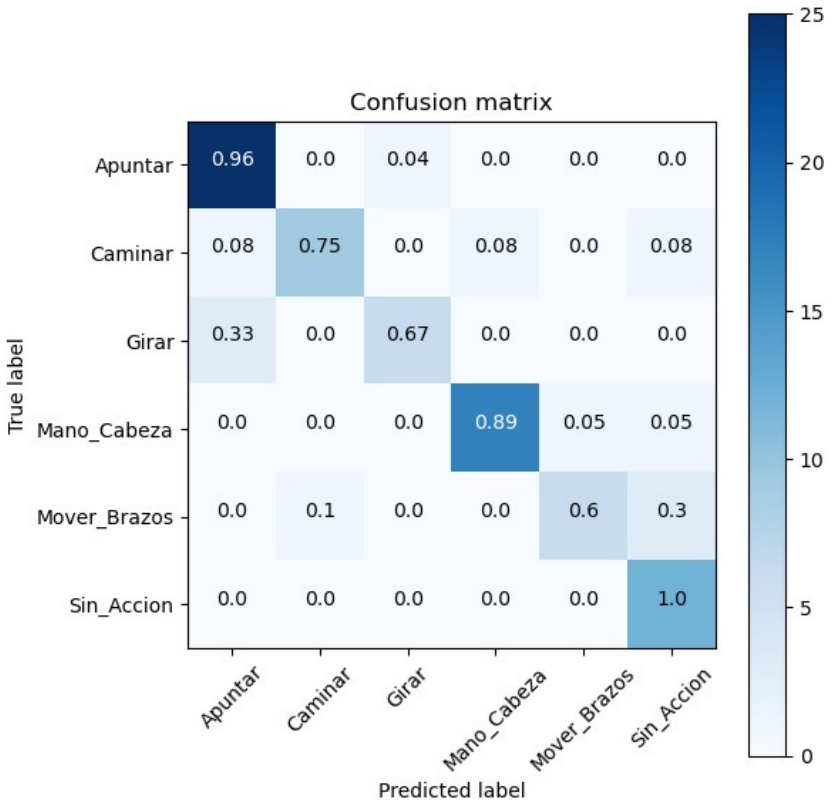
3.2.2.2. Evaluación del modelo con datos de un entorno real. A través de la matriz de confusión 21, utilizando datos de validación obtenidos de videos grabados por usuarios en las aulas RAP, se obtuvo de manera general una **confiabilidad del 85.22%**, además se observó que el modelo predice *Girar* con un 67% de precisión y *Mover Brazos* con un 60%, siendo estas las predicciones con menor acierto. En contraste, el ademán *Sin Acción* fue detectado con un 100% de precisión.

3.2.3 Pruebas de modelo entrenado con datos históricos y datos generados bajo un entorno controlado.

Se llevaron a cabo pruebas con un modelo entrenado utilizando una combinación de datos históricos y datos generados en un entorno controlado. Este enfoque híbrido permitió

Figura 21

Matriz de confusión con datos de un entorno real para la evaluación del modelo con datos de un entorno controlado



evaluar cómo el modelo se desempeña al integrar información proveniente de exposiciones reales en las aulas RAP junto con datos sintéticos creados en condiciones optimizadas. Las pruebas se enfocaron en analizar la capacidad del modelo para generalizar y adaptarse a una variedad de escenarios, combinando la riqueza de datos históricos con la precisión y uniformidad de los datos generados artificialmente.

3.2.3.1. Evaluación del modelo con datos de un entorno controlado. A través de la matriz de confusión 22, utilizando datos de validación generados en un entorno controlado, se obtuvo de manera general una **confiabilidad del 57.29%**, además se observó que el modelo muestra un buen desempeño en la detección de *Caminar* y *Mano a Cabeza*, con precisiones promedio del 81% y 88% respectivamente. Sin embargo, el ademán *Girar* se detecta erróneamente como *Apuntar* en un 62% de los casos, y *Mover Brazos* se confunde con *Mano a Cabeza* en un 56% de las ocasiones. El resto de los ademanes se predice con un promedio del 52%.

3.2.3.2. Evaluación del modelo con datos de un entorno real. A través de la matriz de confusión 23, utilizando datos de validación obtenidos de videos grabados por usuarios en las aulas RAP, se obtuvo de manera general una **confiabilidad del 78.41%**, además se observó que el modelo predice los ademanes con una precisión promedio de alrededor del 70%. Los ademanes *Sin Acción* y *Mano a Cabeza* muestran las tasas de precisión más bajas y más altas respectivamente, con un 50% y un 89%.

3.3 Resultados

- El modelo entrenado únicamente con datos históricos mostró alta precisión en *Caminar* y *Mano a Cabeza* en un entorno controlado. Sin embargo, en un entorno real, *Girar* fue clasificado erróneamente como *Caminar* en un 100% de los casos.
- El modelo entrenado únicamente con datos generados en un entorno controlado mostró un alto desempeño general. En un entorno real, *Girar* tuvo un 67% de precisión y

Figura 22

Matriz de confusión con datos de un entorno real para la evaluación del modelo con datos históricos y de un entorno controlado

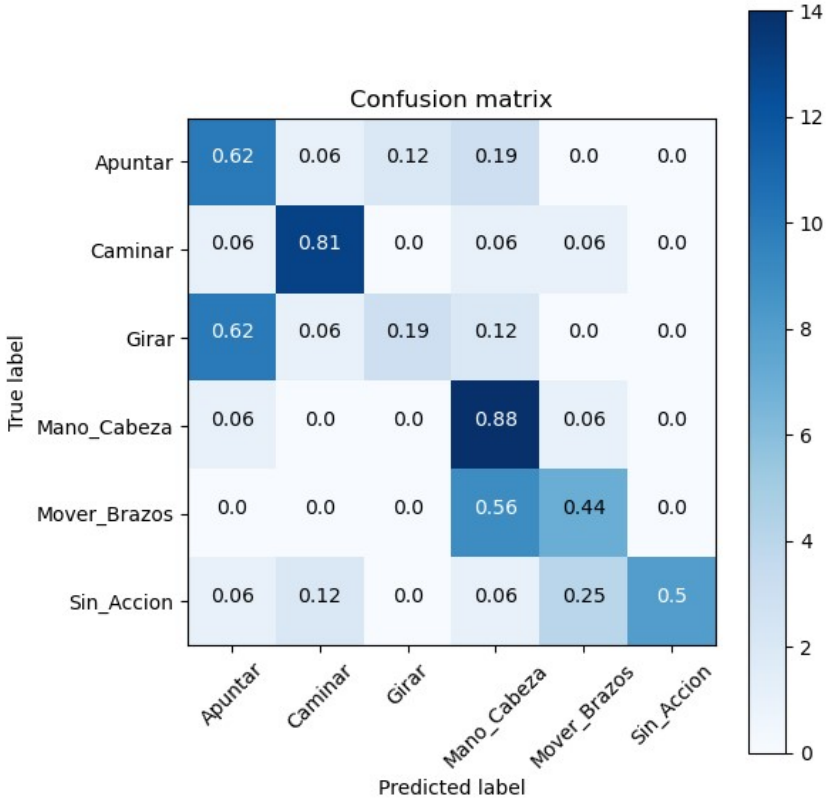
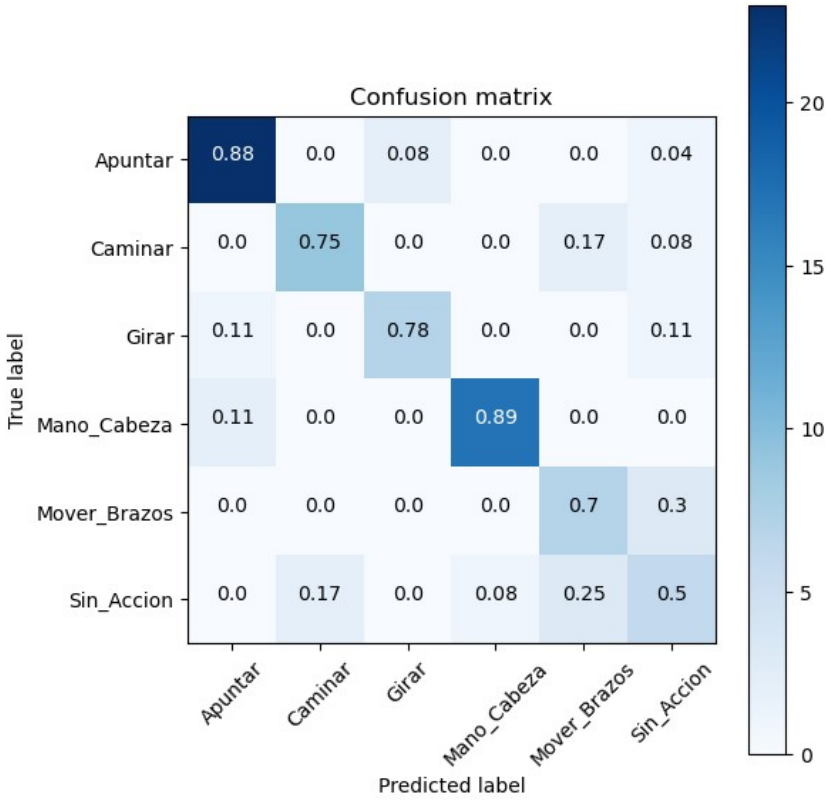


Figura 23

Matriz de confusión con datos de un entorno real para la evaluación del modelo con datos históricos y de un entorno controlado



Mover Brazos un 60%, mientras que *Sin Acción* fue detectado con un 100% de precisión. En promedio, este modelo tuvo el mejor desempeño global.

- El modelo entrenado con una combinación de datos históricos y datos generados en un entorno controlado mostró una precisión promedio del 81% para *Caminar* y del 88% para *Mano a Cabeza* en un entorno controlado. En un entorno real, el modelo logró una precisión promedio del 70%, con *Sin Acción* alcanzando un 50% y *Mano a Cabeza* un 89%.

3.4 Análisis de Costos

En el análisis de costos realizado para este proyecto, se observa que no se incurrieron en gastos adicionales debido a que el proyecto ya había sido previamente ejecutado. La única actividad realizada fue la modificación de un módulo específico, en este caso, el módulo de ademanes. Esta tarea se llevó a cabo utilizando los recursos y herramientas ya disponibles, sin necesidad de adquirir nuevos materiales o servicios externos. Por lo tanto, el proyecto no generó costos adicionales, ya que la intervención se realizó en el contexto de una infraestructura ya establecida y operativa.

Capítulo 4

4. Conclusiones y Recomendaciones

Se han obtenido diversos aprendizajes y hallazgos clave que han permitido evaluar su efectividad y áreas de mejora. Las conclusiones derivadas reflejan el cumplimiento de los objetivos planteados, mientras que las recomendaciones proporcionan un camino claro para futuras optimizaciones y expansiones del trabajo realizado. A continuación, se presentan las principales conclusiones y recomendaciones del proyecto.

4.1 Conclusiones

Como conclusión del trabajo se muestra la importancia de lo realizado en las dos grandes actividades que fueron el preprocesamiento y el diseño del modelo.

- Se destaca que, dentro del entorno real del sistema RAP, el modelo con el mejor desempeño es el que fue entrenado utilizando datos generados en un entorno controlado. Este enfoque ha demostrado ser el más efectivo en la identificación precisa de las acciones en presentaciones reales, gracias a su capacidad para generalizar y adaptarse a las variaciones del entorno.
- A partir de la visualización de los videos, se pudo identificar varias acciones distintivas para el entrenamiento del video y se escogieron las más pertinentes para la creación del modelo. El proceso que se siguió para la identificación de longitud y procesamiento de los datos de acciones podría ser continuado para contar con un dataset más amplio y ayudar a que el modelo detecte aún mejor las acciones detectadas porque no realiza una predicción aún completamente precisa.
- El modelo creado servirá para detectar las acciones y que los resultados puedan ser

interpretados de la manera más conveniente para el reporte final.

- Al proporcionar retroalimentación al estudiante, realizar interpretaciones de las acciones realizadas durante su exposición expresadas en porcentajes. Esto permitirá una evaluación cuantitativa de su desempeño, siendo que la retroalimentación sea más general ya que acciones pueden llegarse a confundir o predecir de manera incorrecta.
- Como parte del proceso de validación sobre los resultados del modelo, se hicieron pruebas con datos normalizados y que nunca habían sido utilizados para el proceso de entrenamiento. De esta forma se evidenció que el modelo está en la capacidad de identificar acciones a partir de nuevas presentaciones del sistema.

4.2 Recomendaciones

- Establecer un plan de mantenimiento y actualización periódica del modelo para incorporar nuevos datos y mejorar continuamente el rendimiento.
- Crear un mecanismo para recopilar y analizar comentarios de los usuarios sobre la precisión del modelo y la utilidad de la retroalimentación para realizar ajustes y mejoras continuas.
- Explorar la inclusión de nuevos tipos de ademanes y acciones para hacer el sistema más versátil y aplicable a una gama más amplia de presentaciones y situaciones.

Referencias

Baird, J. E. (1981). How to overcome errors in public speaking. *IEEE Transactions on Professional Communication*, *PC-24*(2), 94–98. https://doi.org/10.1109/TPC.1981.6447846

Ding, X., Jiang, T., Zhong, Y., Huang, Y., & Li, Z. (2021). Wi-fi-based location-independent human activity recognition via meta learning. *Sensors*, *21*(8). https://doi.org/10.3390/s21082654

Do, J., & Kim, M. (2022). Multi-modal transformer for indoor human action recognition. *2022 22nd International Conference on Control, Automation and Systems (ICCAS)*, 1155–1160. https://doi.org/10.23919/ICCAS55662.2022.10003914

He, J., Zhang, C., He, X., & Dong, R. (2020). Visual recognition of traffic police gestures with convolutional pose machine and handcrafted features. *Neurocomputing*, *390*, 248–259. https://doi.org/https://doi.org/10.1016/j.neucom.2019.07.103

Javadi, R., & Lim, A. (2021). The many faces of anger: A multicultural video dataset of negative emotions in the wild (mfa-wild).

Luo, Y., Ren, J., Wang, Z., Sun, W., Pan, J., Liu, J., Pang, J., & Lin, L. (2018). Lstm pose machines.

Meenakshisundaram, N., & Ramkumar, G. (2022). An efficient and robust model for cervical cancer risk classification based on random forest classifier. *2022 International Conference on Computer, Power and Communications (ICCPC)*, 330–335. https://doi.org/10.1109/ICCPC55978.2022.10072264

Ochoa, X., Domínguez, F., Guamán, B., Maya, R., Falcones, G., & Castells, J. (2018). The rap system: Automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 360–364. https://doi.org/10.1145/3170358.3170406

Rajabi, Z., Shehu, A., & Uzuner, O. (2020). A multi-channel bilstm-cnn model for multilabel emotion classification of informal text. *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, 303–306. https://doi.org/10.1109/ICSC.2020.00060

Ren, Y., Li, C., Bao, W., Chen, X., & Jing, Y. (2023). A study of student action recognition in smart classrooms based on improved slowfast swin transformer. *2023 8th International Conference on Signal and Image Processing (ICSIP)*, 59–63. https://doi.org/10.1109/ICSIP57908.2023.10270896

Rodero, E. (2022). Effectiveness, attractiveness, and emotional response to voice pitch and hand gestures in public speaking. *Frontiers in Communication*, *7*. https://doi.org/10.3389/fcomm.2022.869084

Siriak, R., Skarga-Bandurova, I., & Boltov, Y. (2019). Deep convolutional network with long short-term memory layers for dynamic gesture recognition. *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, *1*, 158–162. https://doi.org/10.1109/IDAACS.2019.8924381

Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., & Baik, S. W. (2018). Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE Access*, *6*, 1155–1166. https://doi.org/10.1109/ACCESS.2017.2778011

Wang, Z., Yang, Y., Liu, Z., & Zheng, Y. (2023). Deep neural networks in video human action recognition: A review.