



INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE  
MONTERREY

Escuela de Ingeniería y Ciencias

Ingeniería en Ciencia de Datos y Matemáticas

**Class Snap: Toma de asistencia y participación con visión  
computacional**

**Reporte Final**

INTELIGENCIA ARTIFICIAL AVANZADA PARA LA CIENCIA DE  
DATOS II

*Federico Medina García Corral A01721441*

*Michelle Yareni Morales Ramón A01552627*

*Paola Sofia Reyes Mancheno A00831314*

*Maria Fernanda Torres Alcubilla A01285041*

**Supervisado por:**

Iván Mauricio Amaya Contreras, Ph.D.

Edgar Covantes Osuna, Ph.D.

Hugo Terashima Marìn, Ph.D.

Monterrey, Nuevo León. Fecha, 28 de noviembre de 2023

## **Resumen**

Este proyecto presenta un sistema integral de reconocimiento facial y seguimiento de gestos diseñado para optimizar la toma de asistencia y evaluar la participación en entornos académicos. Integrando las tecnologías face\_recognition y MediaPipe, nuestro sistema ha demostrado eficacia en tiempo real, proporcionando una solución automatizada para monitorear la asistencia y la interacción de lxs estudiantes durante las sesiones de clase.

La implementación exitosa de la tecnología de reconocimiento facial simplifica los procesos administrativos y establece un método cuantificable para evaluar la participación activa de lxs estudiantes. Aunque hemos logrado avances notables, reconocemos áreas de mejora, centrándonos en la optimización de la detección de gestos a través de MediaPipe y la identificación de las herramientas más adecuadas para la implementación práctica.

Visualizamos nuestro proyecto como una contribución valiosa al ámbito educativo superior, mejorando la eficiencia administrativa y proporcionando datos esenciales para la mejora continua de la calidad educativa. Este trabajo sienta las bases para una integración más amplia de la tecnología de reconocimiento facial y seguimiento de gestos en entornos académicos, destacando nuestro compromiso con la colaboración con personas expertas en tecnología educativa y visión computacional. **Palabras clave:** MediaPipe, face\_recognition, automatización, innovación académica

## **1. Introducción**

A lo largo del tiempo, en los salones de clase lxs docentes toman un tiempo al inicio o fin de la clase para tomar la asistencia de sus estudiantes, una tarea relativamente sencilla pero que quita valioso tiempo y volviéndose tedioso e inexacto, tomando en cuenta que lxs estudiantes pueden llegar a diferentes horas. Sumado a esto, la participación de lxs estudiantes en el salón es muy importante para una fructífera sesión de apren-

dizaje; y, al no ser contabilizada comúnmente, no permite tener una retroalimentación sobre el interés de lxs estudiantes o sobre la efectividad de la planeación de lx maestrx.

Para esta situación, se ha planteado desde la industria de la Inteligencia Artificial y la visión computacional posibles soluciones para generar una toma automática de asistencia y participación. Estos aportes de investigación a la educación se comenzaron a ver aproximadamente desde el 2020, donde la educación en línea, a raíz de la pandemia, mostró la necesidad de evaluar la atención de lxs estudiantes durante las clases. Esto se puede ver evidenciado en el artículo *The Attention of Students in Online Education: Using Head Pose Techniques to Detect Attention in Videoconferencing Platforms: A Case Study* [2] realizado en 2022, donde se aplicó Detección de rostro para verificar la atención en clases en línea. Un año antes, el artículo *Comparison between attendance system implemented through haar cascade classifier and face recognition library* [1], donde se presenta una comparación entre dos algoritmos de detección de rostro para la toma de asistencia, ahora de forma presencial.

A pesar de tener varias aportaciones durante los últimos años con respecto a este tema, las limitaciones o problemas asociados a la toma automática de asistencia y participación son bastante complejos, por lo que muy pocas de estas aportaciones han sido realmente aplicadas en la vida real. Por un lado, se encuentra la dificultad de detectar una participación normal, pues dependiendo de lxs estudiantes o de la clase, una participación puede no verse acompañada de una mano alzada, o incluso una mano alzada no puede significar una participación. Esto significa que la identificación de participación necesitaría modelos adicionales al de identificación de pose, lo que resultaría en algoritmos todavía más pesados automáticamente. Complicación, que lleva al tema de los costos; la inversión que una institución educativa debe hacer para implementar estas soluciones es importante, pues son sistemas que utilizan big data y por ende todos sus com-

ponentes estarían hosteados en la nube.

Estas limitaciones llevan a la reflexión de la importancia de un sistema que, al mismo tiempo de ser relativamente ligero a nivel computacional, funcione correctamente, y además que se trabaje desde el diseño hasta su uso junto con profesores; tomando en cuenta la necesidad de que maestrxs puedan editar la información recopilada automáticamente a través de la solución planteada. De esta manera pudiendo mitigar los errores que pueden ocurrir al aplicar un modelo de visión computacional.

En este sentido, este proyecto tiene como finalidad desarrollar un sistema de reconocimiento facial que permita tomar asistencia de manera automatizada y medir la participación del alumnado en tiempo real dentro de un salón de clases. Esto conectado con una plataforma de acceso para profesores y estudiantes con funcionalidades específicas para cada una de estas credenciales, donde se muestre la información actualizada al final de las sesiones de clase en forma de dashboards. Pudiendo contribuir con un programa que una estos dos requerimientos en un salón de clases, los cuales en investigaciones previas se han tratado de forma separada. Asimismo, al plantear una solución integral se busca brindar una guía para implementar estas soluciones en una institución educativa.

A continuación el reporte está dividido en siete secciones, contando la introducción. La segunda sección está concentrada en los fundamentos, donde se encuentra el marco teórico utilizado para la solución. La Sección 3, trata sobre la explicación y justificación del método de solución propuesto. La metodología y resultados del proyecto conforman la Sección 4. Y finalmente se encuentra la conclusión, referencias y anexos como las Secciones 5, 6 y 7 respectivamente.

## 2. Fundamentos

Desde sus inicios, los modelos de reconocimiento facial han evolucionado de manera extraordinaria. En las primeras etapas, las técnicas se basaban en características geométricas simples,

pero con el advenimiento del aprendizaje profundo en la década de 2010, se produjo un cambio significativo. La capacidad de las redes neuronales para aprender patrones complejos permitió un reconocimiento facial más preciso y robusto, superando obstáculos como variaciones en la iluminación y expresiones faciales.

El desarrollo de bibliotecas como OpenCV marcó hitos clave. Face\_recognition, en particular, ha simplificado enormemente la implementación, ofreciendo una interfaz amigable para la detección y comparación de rostros. Gracias a estos avances, la tecnología de reconocimiento facial se ha vuelto más accesible y eficiente, abriendo puertas a aplicaciones innovadoras en diversos campos.

Es por esto que, para el reconocimiento de cara, se utilizó el modelo de face\_recognition, el cual utiliza una arquitectura basada en la red neuronal profunda ResNET y el algoritmo Max-Margin Object Detection (MMOD). MMOD utiliza una variante de SVM para la detección de objetos mediante la maximización de márgenes entre las instancias positivas y negativas en el espacio de características [3]. A continuación se muestra una imagen en donde se observa visualmente cómo detecta las características faciales el modelo.

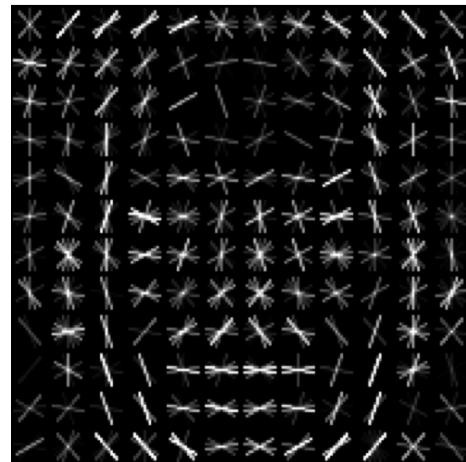


Figura 1: Extracción de características de un rostro utilizando face\_recognition

Por otro lado, para la detección de pose se eligió el modelo de MediaPipe, específicamente BlazePalm. Este modelo está basado en la arquitectura de MobileNetV1 con la diferencia de que para detectar las palmas utiliza una red de regresión para la estimación precisa de las posiciones de los puntos clave de la mano [4]. Esta se eligió para no tener que entrenar al modelo para hacer detección de poses específicas, sino que cuando el modelo detecte una mano alzada por encima del hombro lo contabilice como una participación nueva de lx alumnx.

### 3. Método propuesto

Para crear el modelo se utilizaron dos modelos de aprendizaje profundo combinados. Estos modelos son face\_recognition para la toma de asistencia y MediaPipe, específicamente BlazePalm para la toma de participaciones. Donde la asistencia se toma durante toda la clase y se clasifica mediante niveles: *inasistencia (0-5 %)*, *baja (5-20 %)*, *media (20-60 %)* y *alta(60-100 %)*, con el fin de indicar el tiempo de presencia en la clase de lxs estudiantes. Con respecto a la participación, esta se toma siempre que una persona levante su mano arriba de su hombro, tomando en cuenta que diferentes personas tienen diferentes formas de levantar su mano, y tomar una participación con una posición de la mano o el brazo específica puede perjudicar a ciertxs alumnxs.

Ahora bien, las razones por las que se decidió tomar este camino con respecto a las librerías y modelos escogidos, fueron dos puntos principales:

- No hay muchas opciones en el mercado que cumplan con las especificaciones que el Socio Formador solicitó en el proyecto.
- Siguiendo al punto anterior, se solicitó que en el proyecto el modelo pueda recopilar las asistencias de los alumnos al igual que sus participaciones cuando levanten alguna de sus manos bajo ciertas restricciones que cada equipo puede decidir poner.

Igualmente, es importante recalcar que para la creación de este, se probaron muchos otros algoritmos y modelos que al final fueron descartados por diferentes cuestiones tales como el Haar Cascade, YOLOv8, PoseNET, etc. Finalmente elegimos los dos mencionados anteriormente gracias a las especificaciones que nos pedía el socio formador al igual que por detalles computacionales y cuestiones de gustos personales del equipo. Originalmente, la idea del modelo era que detectara la pose de lx usuarix, sin embargo, encontramos que el utilizar un modelo que sea integrado únicamente por la mano de lxs estudiantes (MediaPipe con BlazePalm) facilita computacionalmente el modelo al igual que limita los errores que se pueden tomar por colores de vestimenta (el cual era un problema que teníamos con YOLOv8) y por poses que puedan ser consideradas como alguien estirandose o rascándose. Por otro lado, para elegir el face\_recognition fue simplemente porque consideramos que fue el modelo que mejor sirvió ya que fue comparado con Haar Cascade, lo cual significa que estábamos comparando un modelo de aprendizaje profundo (face\_recognition) contra un algoritmo antiguo (Haar Cascade) y obsoleto.

Para el desarrollo de la interfaz se trabajó con Streamlit, ya que representa una opción que cumple con los requerimientos del proyecto y proporciona una aplicación web sencilla de realizar basada en Python, donde cualquier persona puede acceder a esta, con el único requisito de contar con una cuenta, misma que se puede crear directamente en la interfaz con un correo institucional. Además, la interfaz se encuentra conectada con Firestore y Google Storage para la obtención y edición de registros de clase, asignación de clases y almacenamiento de credenciales, esta conectividad nos permite aterrizar nuestro proyecto a un escenario lo más cercano a tiempo real y que tanto docentes como estudiantes la información esté disponible lo más rápido posible.

La interfaz y lo relacionado a esta cuenta con 3 roles, el primero es el administrativo, el cual tiene permiso a las bases de datos, tanto de Google Storage como de Firestore, las acciones que

se pueden realizar son:

- Firestore:

- Crear, eliminar e inhabilitar cuentas.
- Modificar contraseñas de las cuentas.
- Crear, modificar o eliminar cursos.
- Modificar cursos asignados a lxs docentes y estudiantes.
- Otorgar permisos a otras personas (se cuentan con diversos permisos según los privilegios)

- Google Storage:

- Eliminar o modificar fotos para el modelo.
- Borrar, modificar o agregar reportes de los cursos.
- Otorgar permisos a otras personas (se cuentan con diversos permisos según los privilegios)

Por otro lado, se tienen los roles de estudiantes y docentes, donde al crear una cuenta se asigna un rol automáticamente correspondiente a tu ID de la institución. En cuanto a la plataforma de docentes, se tienen las siguientes características:

- Filtrado por curso, fechas y estudiantes
- A partir de un curso y una fecha en específico, se puede editar el reporte de asistencia y participación y actualizarlo en la nube.
- Descarga de los reportes generados de tus cursos asignados.
- Visualizaciones de asistencia y participaciones.

Asimismo, la plataforma de estudiantes es similar a la anterior, con la única diferencia que la persona registrada puede solamente visualizar sus estadísticas y desempeño individuales en los cursos asignados, sin la posibilidad de edición.

Por último, al momento de transmitir la información de las identidades de lxs usuarixs, se priorizó su integridad y confidencialidad, es por esto que, además de hacer uso de la seguridad que nos brindan las herramientas, se realizó un proceso de anonimización. En este a cada usuarix se le otorga un LocalID generado aleatoriamente, el cual es una cadena de caracteres que de manera pública no se puede relacionar a información sensible de las personas.

Para validar el modelo se utilizó un dataset creado por nosotrxs mismxs debido a que el modelo creado es un modelo completamente nuevo y por lo mismo no se cuentan métricas de evaluación previas. Para esto se creó un folder llamado “dataset”, en donde subimos fotos de lxs integrantes del equipo y de compañeros de clase con las cuales se probó y evalúo el modelo. Para realizar las métricas de evaluación se tomaron distintos videos en donde el modelo se ejecutaba mientras que al menos cuatro personas se encontraban en el frame participando. Una vez tomado el video se contaban los frames en donde el modelo contaba participaciones incorrectas o hacía detección de caras incorrectas o hasta nulas. Este proceso se repitió 5 veces y se encontró un promedio de lo mencionado anteriormente, con lo cual obtuvimos las siguientes métricas:

- Accuracy: 89 %
- Precision: 87 %
- Recall: 67 %
- F-Score: 73 %

Cabe recalcar que las métricas anteriores son para todo el modelo general, no se tomaron mediciones para face\_recognition y MediaPipe por separado.

## 4. Metodología

Para medir y comparar nuestros modelos, implementamos un benchmark con tres pruebas clave:

- Rendimiento del modelo con registros completos de personas y rostros.
- Rendimiento cuando se tienen registros de personas pero solo se muestra parte del rostro.
- Comportamiento del modelo con personas no registradas, para evaluar su respuesta ante rostros desconocidos.

Se inició usando el modelo de reconocimiento facial Haar Cascade, un clasificador preentrenado para patrones visuales. A diferencia de las técnicas modernas de Machine Learning, Haar Cascade utiliza un enfoque más tradicional, analizando imágenes mediante patrones de claros y oscuros y clasificándolos en conjuntos de entrenamiento positivos y negativos.

Aunque este modelo podía reconocer rostros, dependía excesivamente de características específicas, como el uso de accesorios, lo cual limitaba su eficacia. Además, por ser una tecnología más antigua, mostraba un rendimiento lento y poco fluido, lo que no cumplía con los objetivos del proyecto.

Para evaluar el benchmark, se realizaron pruebas en variados escenarios. Se utilizó la imagen de “Federico” en la Figura 2 y las de otros integrantes del equipo. El modelo, aplicado en la Figura 3, mostró una predicción correcta. No obstante, al emplearlo en la Figura 4 con diferente iluminación, aunque reconoció el rostro, identificó erróneamente un rostro en la prenda, evidenciando limitaciones en la precisión bajo condiciones cambiantes.

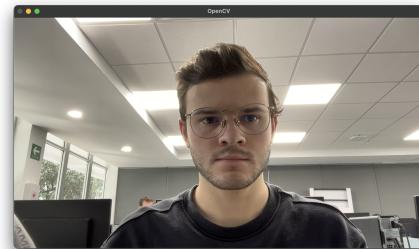


Figura 2: Foto utilizada para entrenar modelo de Haar Cascade

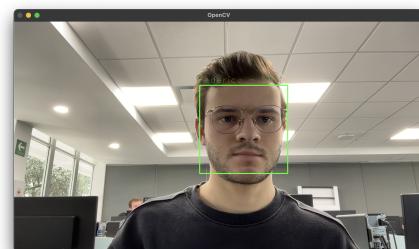


Figura 3: Prueba Haar Cascade

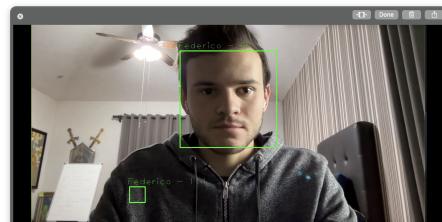


Figura 4: Reconocimiento de rostro Erróneo con el modelo Haar Cascade

La prueba de rostro parcialmente cubierto se observa en la Figura 5, en esta el modelo logra tanto la detección como reconocimiento correcto.



Figura 5: Predicción correcta con rostro parcialmente cubierto

Por último, en la imagen 6, se observa la predicción del modelo entrenado con fotos del resto del equipo (excluyendo a “Federico”), podemos observar que el modelo fuerza la predicción a tener un resultado conocido, por lo que en este modelo no existe el caso de “unknown”.

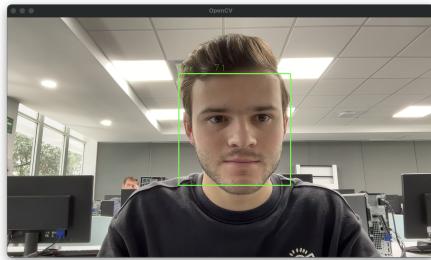


Figura 6: Reconocimiento erróneo de rostro con el modelo Haar Cascade

Tanto el error de identificación observado en la Figura 4 y el etiquetado obligado en los rostros no registrados observado en la Figura 6 no se alinean con los objetivos y contexto de la situación planteada, por lo que se decide que este modelo no es viable.

La solución escogida fue el modelo de reconocimiento facial de dlib, destacando por su alta precisión del 99.38 % en el benchmark “Labeled Faces in the Wild”. Para entrenarlo con nuestro dataset, se necesita solo una foto por persona. Para la comparativa, se emplearán los mismos datos de prueba que en el modelo Haar Casca-

de, garantizando una evaluación uniforme entre ambos modelos.

El modelo permite ajustar el parámetro de *tolerancia*, que determina la rigurosidad en la comparación de rostros. Con una tolerancia alta de 0.8, el modelo identifica correctamente rostros claros como en la Figura 3, pero puede confundir rostros parciales con personas similares. Además, si un rostro no está registrado, tiende a asociarlo con el rostro más similar, reduciendo la probabilidad de identificarlo como “unknown”, como se muestra en la Figura 8 donde ambos son clasificados como “Federico”.

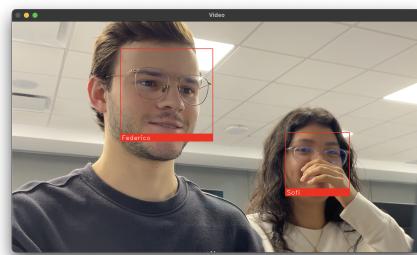


Figura 7: Reconocimiento rostros registrados

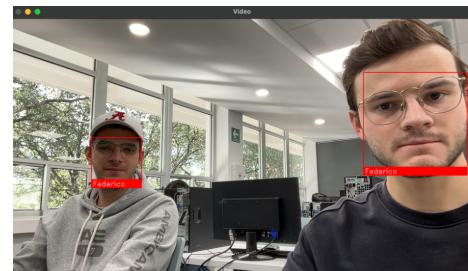


Figura 8: Error en el modelo face\_recognition con parámetro de *tolerance* = 0.8

Se redujo el parámetro de *tolerancia* a 0.6. En la Figura 9, el modelo reconoce correctamente rostros despejados y clasifica rostros parciales como “unknown”, mejorando la toma de asistencia. En el escenario de rostros no registrados, como en la Figura 11, el modelo también clasifica correctamente como “unknown”, a diferencia del

anterior.

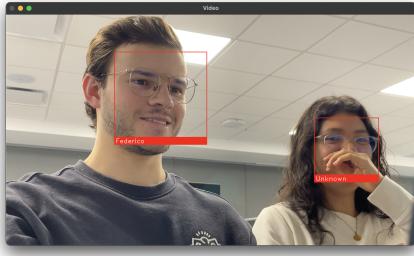


Figura 9: Reconocimiento rostros registrados

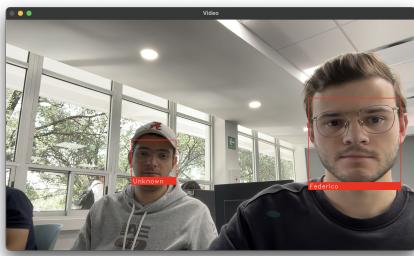


Figura 10: Modelo face\_recognition con parámetro de *tolerance* = 0.6 con obstáculo

Finalmente, se probó con una tolerancia de 0.4 y se observó que incluso con rostros claros, el modelo no predice correctamente.

Al comparar ambos modelos, el de *face\_recognition* superó al Haar Cascade en reconocimiento facial, gracias a su capacidad de aprendizaje profundo y el valor óptimo de *tolerance* fue 0.6, equilibrando la identificación correcta de personas registradas y la clasificación de no registrados como “unknown”.

Para las pruebas de detección de poses, se evaluaron tres modelos distintos. El primero fue MoveNet, un modelo de visión por computadora especializado en la detección de poses humanas en tiempo real. Aunque MoveNet es eficaz en el seguimiento de puntos clave del cuerpo humano usando deep learning, resultó no ser el más adecuado para nuestro contexto. Su enfoque en la

detección de todo el cuerpo presentó limitaciones en situaciones donde solo el torso está visible, como es este caso donde los alumnos están sentados, dificultando así la precisión en la detección de poses.

Posteriormente se probó el modelo de YOLOv7 y YOLOv8, en donde el seguimiento de objetos obtuvo resultados muy positivos. Estos modelos fueron muy buenos para la detección de las poses que los alumnos podían hacer a la hora de participar, sin embargo, se decidió deshabilitar dicho modelo debido a que muchos alumnos participan en clases haciendo señas de distintas maneras, por lo que se tenían que hacer pruebas con múltiples poses, las cuales por cuestiones de tiempo y logística no dieron resultados.

Finalmente, optamos por utilizar MediaPipe, enfocándonos en la detección exclusiva de manos. Esta decisión resultó en una mejora significativa en el rendimiento del modelo, que previamente enfrentaba problemas de estabilidad y eficiencia debido a la complejidad de los modelos utilizados. La especialización en manos con MediaPipe ofreció un equilibrio ideal entre precisión y fluidez en el procesamiento.

Finalmente se fusionaron ambos modelos, lo que permitió que trabajaran en conjunto. El reconocimiento facial asegura una identificación constante y precisa de los individuos, mientras que la detección de manos se enfoca en captar los gestos de participación. Esta combinación resultó en un sistema más completo, capaz de ofrecer un seguimiento detallado de la interacción en el aula.

Para evaluar la efectividad de nuestro sistema, se realizaron grabaciones de video de 30 segundos en las cuales se monitoreó y contabilizó la frecuencia con la que los participantes levantaban las manos. Paralelamente, el sistema también se encargó de reconocer y registrar las caras de los participantes. Las métricas resultantes, detalladas a continuación, reflejan la precisión del sistema en la identificación correcta de las personas y el seguimiento de su participación. Estos resultados son cruciales para entender la eficacia del sistema en un entorno dinámico y propor-

cionan insights valiosos sobre su rendimiento en situaciones reales de interacción grupal.

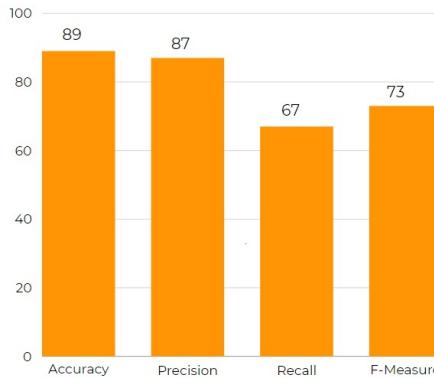


Figura 11: Score del modelo final

#### 4.1. Resultados

Finalmente después de hacer las comparaciones debidas con los primeros modelos que se elaboraron al igual que las comparaciones entre el propio modelo de face\_recognition pero utilizando nuevos parámetros, obtenemos el mejor modelo que pudimos lograr elaborar en estas diez semanas. El modelo final cuenta con las métricas mostradas en la Sección 3 y, en base a las pruebas realizadas, fue el mejor modelo y el más cercano a lo que consideramos que el Socio Formador esperaba como modelo “ideal”. En base a lo que se presentó en el día 27 de noviembre del 2023, el equipo notó al Socio Formador satisfecho tanto con el modelo como con la interfaz, con la única sugerencia de que se mejoren los gráficos en donde se plasma la información, lo cual por cuestiones de tiempo es imposible cambiar en solo un día.

### 5. Conclusión

En el transcurso de este proyecto, hemos alcanzado una satisfacción significativa al desarrollar un sistema integral de reconocimiento facial

y seguimiento de gestos, empleando las tecnologías face\_recognition y MediaPipe al igual que plasmarlas en una interfaz de StreamLit, utilizando datos guardados en la nube. Este sistema se ha demostrado efectivo para la toma de asistencia y la evaluación de la participación en tiempo real en entornos académicos.

Nuestro enfoque ha permitido abordar la necesidad de automatizar y mejorar los procesos de seguimiento en el ámbito educativo. La implementación exitosa del reconocimiento facial y el seguimiento de manos no solo simplifica la toma de asistencia, sino que también proporciona un método cuantificable y objetivo para evaluar la participación activa de los estudiantes durante las sesiones de clase.

Aunque hemos logrado avances significativos, reconocemos la existencia de áreas de mejora al igual que áreas de oportunidad. Específicamente, la optimización de la detección de gestos a través de MediaPipe y la determinación de las herramientas ideales para su implementación práctica son aspectos clave para el perfeccionamiento continuo de nuestro sistema, la cual sin duda alguna mejorará con la incrementación de presupuesto al igual que al abarcar mayor tiempo para el desarrollo tanto de la interfaz como del modelo.

Mirando hacia el futuro, visualizamos nuestro proyecto como una herramienta invaluable para la educación superior. La capacidad de monitorear la asistencia y la participación de manera automatizada no solo mejora la eficiencia administrativa, sino que también proporciona datos valiosos para la mejora continua de la calidad educativa.

En resumen, este proyecto sienta las bases para una integración más amplia de la tecnología de reconocimiento facial y seguimiento de gestos en entornos educativos. Sabemos bien que al estar colaborando con expertxs en tecnología educativa y visión por computadora se logrará un impacto significativo en la experiencia de aprendizaje y evaluación en el ámbito académico utilizando como base el proyecto presentado.

## Referencias

- [1] Samiksha Malhotra et a. «Comparison between attendance system implemented through haar cascade classifier and face recognition library». En: *IOP Conference Series: Materials Science and Engineering* (2021). DOI: : 10 . 1088 / 1757 - 899X / 1022 / 1 / 012045.
- [2] Luis Rodrigo Barba-Guaman y Priscila Valdiviezo-Diaz. «The attention of students in online education». En: *International Journal of Emerging Technologies in Learning (iJET)* 17.22 (2022), págs. 144-155. DOI: 10.3991/ijet.v17i22.34241.
- [3] Subhransu Maji y Jitendra Malik. «Object detection using a Max-margin Hough Transform». En: *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009). DOI: 10.1109/cvpr.2009.5206693.
- [4] Hyeonsoo Yoo, Igor Goncharenko y Yanlei Gu. «Real-time Dynamic Sign Language Recognition using LSTM based on Me-

diaPipe hand data». En: *2023 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)* (2023). DOI: 10.1109/icce-taiwan58799.2023.10226687.

## 6. Anexos

Github de la creación del modelo utilizando: [https://github.com/fedemedina72/Reto\\_IA2\\_Equipo8/blob/main/CreacionModelo/Modelo.ipynb](https://github.com/fedemedina72/Reto_IA2_Equipo8/blob/main/CreacionModelo/Modelo.ipynb)

Enlace a la interfaz: <https://class-snap.streamlit.app>

Credencial docente:

- Correo: profel@tec.mx
- Contraseña: 123456

Credencial estudiante:

- Correo: a0fer@tec.mx
- Contraseña: 123456