



Universidad Nacional de Rosario
Facultad de Ciencias Médicas
Carrera de Posgrado de Especialización en Anestesiología

**Más allá del Mallampati: inteligencia artificial multimodal para la
predicción de vía aérea dificultosa**

Alumno: Márquez Vallari León Sebastián¹

Tutor: Llorens Antonio²

Cotutor: Ogusuku Diego²

CENTRO FORMADOR: Hospital Provincial del Centenario

AÑO 2026

¹ Médico. Alumno de la Carrera de Posgrado de Especialización en Anestesiología, Facultad de Ciencias Médicas, Universidad Nacional de Rosario.

² Especialista en Anestesiología. Docente Estable de la Carrera de Posgrado de Especialización en Anestesiología de la Facultad de Ciencias Médicas de la Universidad Nacional de Rosario.

RESUMEN

Introducción: Anticipar una laringoscopia dificultosa continúa siendo un desafío relevante en anestesiología debido a la limitada sensibilidad de las pruebas clínicas tradicionales y a las posibles complicaciones derivadas de una dificultad no anticipada. El desarrollo de modelos de inteligencia artificial con capacidad para analizar imágenes permite explorar nuevas herramientas complementarias para la evaluación de la vía aérea. **Objetivo:** Evaluar de manera exploratoria el desempeño de ChatGPT con capacidad multimodal para identificar dificultad laringoscópica a partir de fotografías faciales preoperatorias estandarizadas. **Materiales y métodos:** Se realizó un estudio prospectivo, observacional y exploratorio en 100 pacientes adultos sometidos a cirugía electiva bajo anestesia general en el Hospital Provincial del Centenario. A cada participante se le efectuó una evaluación clínica preoperatoria de la vía aérea y se le tomaron fotografías faciales estandarizadas de frente, perfil y con la boca abierta. Las imágenes fueron analizadas mediante ChatGPT, que emitió una clasificación binaria de vía aérea fácil o difícil y un nivel cualitativo de confianza. Como criterio de referencia se utilizó la dificultad observada durante la laringoscopia, definida por una clasificación de Cormack-Lehane III o IV. De manera complementaria, se desarrolló una red neuronal convolucional mediante *Teachable Machine* como experiencia preliminar de factibilidad. **Resultados:** La edad media fue de 44 años, el 69% de los participantes fue de sexo femenino y el índice de masa corporal medio fue de 26,1 kg/m². Se registraron 8 casos de dificultad laringoscópica. ChatGPT identificó 6 de los 8 casos, con una sensibilidad del 75%, una especificidad del 52,2%, un valor predictivo positivo del 12%, un valor predictivo negativo del 96% y una exactitud global del 54%. La herramienta clasificó como difíciles a 50 pacientes, por lo que mostró una tendencia a sobreestimar la dificultad. Las respuestas con confianza alta presentaron una mayor proporción de clasificaciones correctas que aquellas con confianza media. La red neuronal convolucional mostró elevada especificidad aparente, aunque sus resultados correspondieron al conjunto utilizado para su entrenamiento y no constituyeron una validación independiente. **Conclusión:** El análisis multimodal mediante ChatGPT fue factible para la evaluación preoperatoria de imágenes faciales y mostró un perfil caracterizado por mayor sensibilidad que especificidad. Estos resultados son exploratorios y respaldan la realización de estudios posteriores con muestras más amplias, criterios estandarizados y validación independiente.

PALABRAS CLAVE

INTELIGENCIA ARTIFICIAL - VÍA AÉREA- ANALISIS DE IMAGENES-
APRENDIZAJE AUTOMÁTICO - CHATGPT - VÍA AÉREA DIFÍCIL
ARTIFICIAL INTELLIGENCE – AIRWAY - IMAGE ANALYSIS - MACHINE
LEARNING - DIFFICULT AIRWAY - DIFFICULT LARYNGOSCOPY

ÍNDICE	
<hr/>	
RESUMEN	1
PALABRAS CLAVE	1
INTRODUCCIÓN	3
OBJETIVOS	5
MATERIAL Y MÉTODOS	6
RESULTADOS	10
DISCUSIÓN	17
CONCLUSIÓN	18
REFERENCIAS BIBLIOGRÁFICAS	19
ANEXO	21

INTRODUCCIÓN

La evaluación preoperatoria de la vía aérea constituye un componente fundamental de la práctica anestesiológica. Su finalidad es reconocer, antes de la inducción anestésica, aquellos pacientes en quienes la ventilación, la laringoscopia o la intubación traqueal podrían resultar dificultosas. Una identificación adecuada permite anticipar recursos, seleccionar dispositivos alternativos, solicitar asistencia y establecer una estrategia de rescate. Por el contrario, una dificultad no prevista puede asociarse con hipoxia, traumatismo de la vía aérea, aspiración, daño neurológico, necesidad de acceso quirúrgico de emergencia e incluso muerte (1-3).

A pesar de su importancia, la predicción de dificultad continúa siendo imprecisa. La evaluación convencional combina antecedentes clínicos con pruebas físicas como la clasificación de Mallampati (MAL) que valora la relación entre el tamaño de la lengua y el espacio orofaríngeo visible, y cuyas clases III y IV se asocian con una mayor probabilidad de dificultad laringoscópica. La distancia tiromentoniana (DTM) que estima el espacio disponible para el desplazamiento de la lengua durante la laringoscopia, considerándose disminuida menor a 6 cm. La apertura bucal (AB) que evalúa el espacio disponible para introducir y manipular el laringoscopio, limitada cuando es inferior a 3 cm. La movilidad cervical (MC) necesaria para lograr una posición adecuada durante la laringoscopia. La presencia de retrognatía (RTG) que indica una posición posterior de la mandíbula y el test de mordida del labio superior (ULBT) que permite estimar la movilidad mandibular y la relación entre los incisivos. Sin embargo, ninguna de estas pruebas, utilizada de manera aislada, presenta suficiente capacidad discriminativa. En el metaanálisis de Shiga y colaboradores, los métodos clínicos mencionados mostraron sensibilidades bajas o moderadas, mientras que su combinación produjo solamente una mejora incremental de la capacidad predictiva (4). Esta limitación implica que una proporción importante de las dificultades continúa sin ser reconocida durante la valoración preoperatoria.

La dificultad también depende de la forma en que el profesional integra los hallazgos disponibles. En la práctica cotidiana, la evaluación no se basa únicamente en mediciones anatómicas, sino también en la apreciación global de la cara, el cuello, la mandíbula, la apertura oral y las relaciones entre dichas estructuras. Esta integración presenta un componente subjetivo y puede variar según la experiencia del observador. En una cohorte de 188.064 pacientes, Nørskov y colaboradores observaron que el 93% de las intubaciones difíciles no habían sido anticipadas, mientras que sólo una cuarta parte de los casos considerados difíciles durante el preoperatorio presentó posteriormente una dificultad real (5). Estos hallazgos muestran que el problema no consiste sólo en detectar factores de riesgo, sino en interpretarlos y combinarlos de manera adecuada.

La imposibilidad de anticipar una dificultad puede generar además un incremento abrupto de la carga de trabajo y de la demanda cognitiva durante el manejo de la vía aérea. En estas circunstancias, incluso los equipos experimentados pueden presentar dificultades para reorganizar el plan, mantener una comunicación efectiva y seleccionar oportunamente una estrategia alternativa. El análisis de eventos adversos ha señalado que una evaluación incompleta, una planificación insuficiente y la repetición de técnicas que ya habían fracasado contribuyeron a desenlaces desfavorables (2,3,6). Por este motivo, el desarrollo de herramientas complementarias que ayuden a sistematizar la evaluación preoperatoria podría contribuir a una planificación más segura.

En los últimos años, la inteligencia artificial se ha incorporado progresivamente al análisis de información clínica e imágenes médicas. El aprendizaje automático permite reconocer patrones complejos a partir de datos previamente clasificados, mientras que el aprendizaje profundo utiliza redes neuronales con múltiples capas capaces de extraer características relevantes sin que estas deban ser definidas manualmente (7). Dentro de este campo, las redes neuronales convolucionales han mostrado especial utilidad en tareas de clasificación de imágenes, debido a su capacidad para identificar bordes, formas, texturas y relaciones espaciales.

Algunos estudios comenzaron a aplicar estos modelos al análisis de imágenes faciales para predecir laringoscopia e intubación difícil. Sus resultados iniciales sugirieron que determinadas características morfológicas podrían ser reconocidas mediante algoritmos de visión artificial (8-10). Sin embargo, gran parte de estas experiencias se desarrollaron

en muestras limitadas, con metodologías heterogéneas y sin validaciones independientes suficientes. Por ello, aunque los resultados fueron promisorios, su aplicación clínica continúa siendo exploratoria (11).

Paralelamente, los modelos extensos de lenguaje evolucionaron desde sistemas diseñados exclusivamente para procesar texto hacia arquitecturas multimodales capaces de interpretar también imágenes. ChatGPT (*Generative Pre-trained Transformer*) representa uno de los principales exponentes de esta tecnología. A diferencia de una red convolucional entrenada para ejecutar una clasificación específica, un modelo multimodal puede observar una imagen, reconocer rasgos anatómicos, integrarlos con conocimientos generales y generar una respuesta acompañada de una explicación en lenguaje natural (12).

La capacidad recién mencionada resulta particularmente atractiva para la evaluación de la vía aérea. A partir de fotografías faciales de frente, perfil y con apertura bucal, el modelo puede considerar elementos como la proporción entre mandíbula y maxilar, la presencia aparente de retrognatía, la apertura oral, la visualización de estructuras orofaríngeas, el volumen cervical y otros rasgos morfológicos. Además de emitir una clasificación, puede explicitar qué características influyen en su decisión. Esta posibilidad aporta cierto grado de interpretabilidad, aunque no elimina el riesgo de errores, sesgos o razonamientos aparentemente convincentes sin suficiente fundamento clínico.

ChatGPT no fue desarrollado ni validado originalmente como dispositivo diagnóstico para la vía aérea. Por lo tanto, su empleo no debe entenderse como reemplazo de la valoración realizada por el anestesiólogo, sino como una herramienta experimental de apoyo. Su utilidad potencial reside en ofrecer una segunda lectura estandarizada de las imágenes, señalar rasgos que podrían pasar inadvertidos y generar una estimación complementaria. Al mismo tiempo, su funcionamiento debe ser evaluado críticamente, dado que una elevada sensibilidad acompañada de baja especificidad podría conducir a una sobreestimación del riesgo y a una cantidad excesiva de falsos positivos.

En este trabajo se priorizó el análisis exploratorio mediante ChatGPT, un modelo multimodal generalista que no fue entrenado específicamente con las imágenes del estudio. En una segunda línea se exploró el uso de Teachable Machine, una plataforma

web que permite entrenar clasificadores de imágenes basados en redes neuronales convolucionales. Si bien este modelo aprendió directamente de las imágenes obtenidas durante la investigación, no se dispuso de un conjunto independiente de pacientes que permitiera validar formalmente su rendimiento diagnóstico, por lo que sus resultados fueron considerados únicamente exploratorios. El propósito principal del estudio fue evaluar la capacidad de ChatGPT para identificar dificultad laringoscópica a partir de fotografías faciales preoperatorias estandarizadas y comparar su desempeño con el de los parámetros clínicos habitualmente utilizados, tomando como referencia el resultado observado durante la laringoscopia. A diferencia de los trabajos previos, centrados principalmente en el desarrollo y validación de redes neuronales entrenadas específicamente para esta tarea, este estudio aporta una evaluación prospectiva de un modelo multimodal generalista, aplicado en un entorno hospitalario local, y brinda evidencia preliminar sobre su rendimiento y sus limitaciones como posible herramienta complementaria de la evaluación preoperatoria.

OBJETIVOS

Objetivo Primario

Evaluar el desempeño del modelo multimodal de ChatGPT para identificar dificultad laringoscópica a partir de fotografías faciales preoperatorias estandarizadas.

Objetivos secundarios

1. Determinar la sensibilidad, especificidad, valores predictivos y exactitud global de ChatGPT.
2. Analizar la relación entre el nivel de confianza informado por ChatGPT y la corrección de sus predicciones.
3. Comparar de manera exploratoria el desempeño de ChatGPT con los principales predictores clínicos preoperatorios.
4. Desarrollar y describir preliminarmente una red neuronal convolucional mediante una plataforma accesible como *Teachable Machine*, dejando su validación en una cohorte independiente como línea de investigación futura.

MATERIAL Y MÉTODOS

Se realizó un estudio prospectivo, observacional y exploratorio en el Hospital Provincial del Centenario de la ciudad de Rosario, provincia de Santa Fe, Argentina, durante el período comprendido entre enero y julio de 2026. El protocolo fue aprobado por el Comité de Ética en Investigación de la institución. Todos los participantes firmaron el consentimiento informado antes de su inclusión.

Se incluyó una muestra consecutiva de 100 pacientes adultos sometidos a procedimientos quirúrgicos electivos bajo anestesia general con intubación orotraqueal.

Los criterios de inclusión fueron pacientes clasificados como ASA I, II o III con disponibilidad para realizar la evaluación preoperatoria y obtener las fotografías faciales estandarizadas.

Se excluyeron pacientes con antecedentes de cirugía facial o cervical, deformidades craneofaciales, tumores mandibulares o cervicales, traumatismos recientes cráneo o cervicales, traqueostomía, limitaciones graves de la movilidad cervical o temporomandibular.

Antes del procedimiento anestésico se registraron edad, sexo, índice de masa corporal y clasificación ASA de cada paciente. La evaluación clínica de la vía aérea incluyó clasificación de Mallampati, distancia tiromentoniana, apertura bucal, movilidad cervical, presencia de retrognatia y test de mordida del labio superior.

Durante la evaluación preoperatoria y antes de la inducción anestésica, el investigador principal obtuvo tres fotografías faciales estandarizadas de cada participante:

1. vista frontal con la boca cerrada;
2. vista frontal con máxima apertura bucal y lengua protruida;
3. vista lateral con la boca cerrada.

Las imágenes se obtuvieron con un teléfono iPhone 15, manteniendo condiciones controladas de iluminación, distancia y fondo (Figura 1). Para proteger la confidencialidad, se eliminaron los metadatos, se asignó un código alfanumérico y se almacenaron por separado las fotografías y los datos clínicos. Solo el investigador principal tuvo acceso al código de correspondencia. La participación fue voluntaria, no

modificó la conducta anestésica y todas las personas firmaron el consentimiento informado.

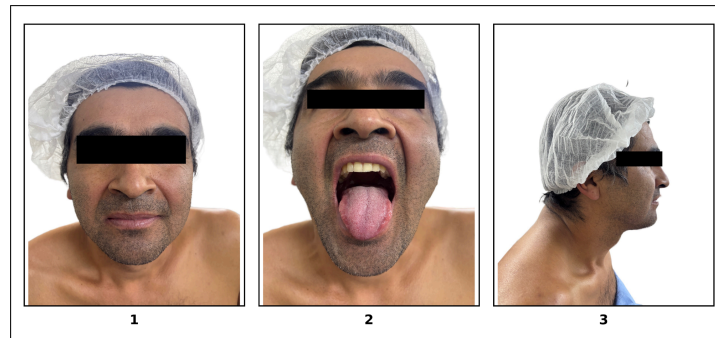


Figura 1. Protocolo fotográfico.

Luego de la toma de imágenes se procedió con la inducción anestésica mediante un protocolo estandarizado con Propofol 1,5 mg/kg; Remifentanilo 0,5 mcg/kg/min y Vecuronio 0,1 mg/kg. La laringoscopia fue realizada por residentes de 4to año de la carrera de Anestesiología, 3 minutos después de la administración del bloqueante neuromuscular. Todos los pacientes fueron posicionados de manera uniforme en rampa. Durante la intubación se registraron el dispositivo utilizado, el número de intentos y la clasificación de Cormack-Lehane.

Se definió dificultad laringoscópica como la presencia de una clasificación de Cormack-Lehane III o IV. Los grados I y II fueron considerados laringoscopias no dificultosas.

Además, se registraron el uso de videolaringoscopia, la necesidad de dispositivos auxiliares y el número de intentos de intubación.

Posteriormente las tres imágenes de cada paciente fueron analizadas mediante ChatGPT (OpenAI) entre enero y julio 2026, mediante un modelo multimodal de la familia GPT-5 disponible en la interfaz durante ese periodo (12). Los casos se procesaron dentro de un contexto conversacional común previamente configurado con el mismo *prompt* y el mismo formato de respuesta. El modelo no recibió datos clínicos ni el resultado de la

laringoscopia antes de emitir su predicción. El prompt completo se reproduce en el Anexo. Se le solicita:

- una predicción binaria: vía aérea fácil o difícil;
- un nivel cualitativo de confianza: alto, medio o bajo;
- una justificación basada en los rasgos anatómicos observados.

El nivel cualitativo de confianza se clasificó de acuerdo con el lenguaje utilizado por el modelo. Se consideró alto cuando empleó expresiones afirmativas o determinantes; medio cuando utilizó términos probabilísticos o sugestivos; y bajo cuando manifestó incertidumbre o falta de elementos suficientes (Figura 2).

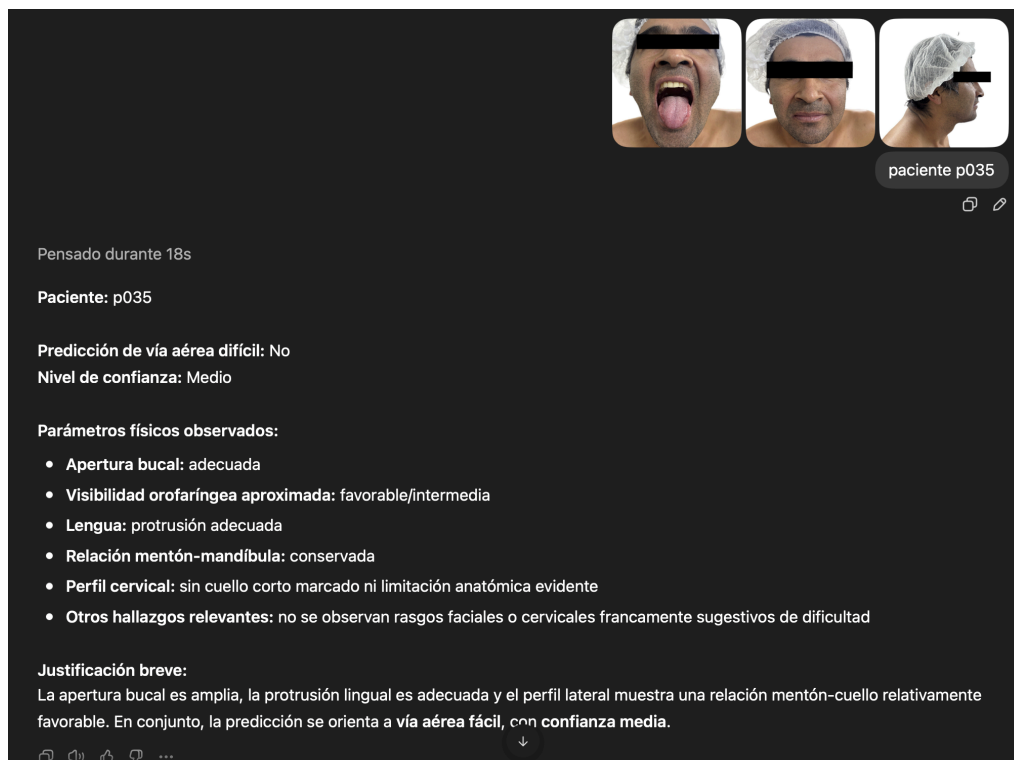


Figura 2. Ejemplo de respuesta multimodal

Como análisis complementario se desarrolló, en la versión web de *Teachable Machine* disponible en 2026, un modelo de clasificación de imágenes basado en una CNN. Se dispuso de 300 fotografías correspondientes a los 100 participantes, etiquetadas como vía aérea fácil o difícil según el desenlace laringoscópico.

El modelo generó una predicción binaria para cada caso (Figura 3). Debido a que la mayor parte de los pacientes fue utilizada durante el proceso de entrenamiento, los

resultados obtenidos se consideraron preliminares y no se interpretaron como una validación clínica independiente.

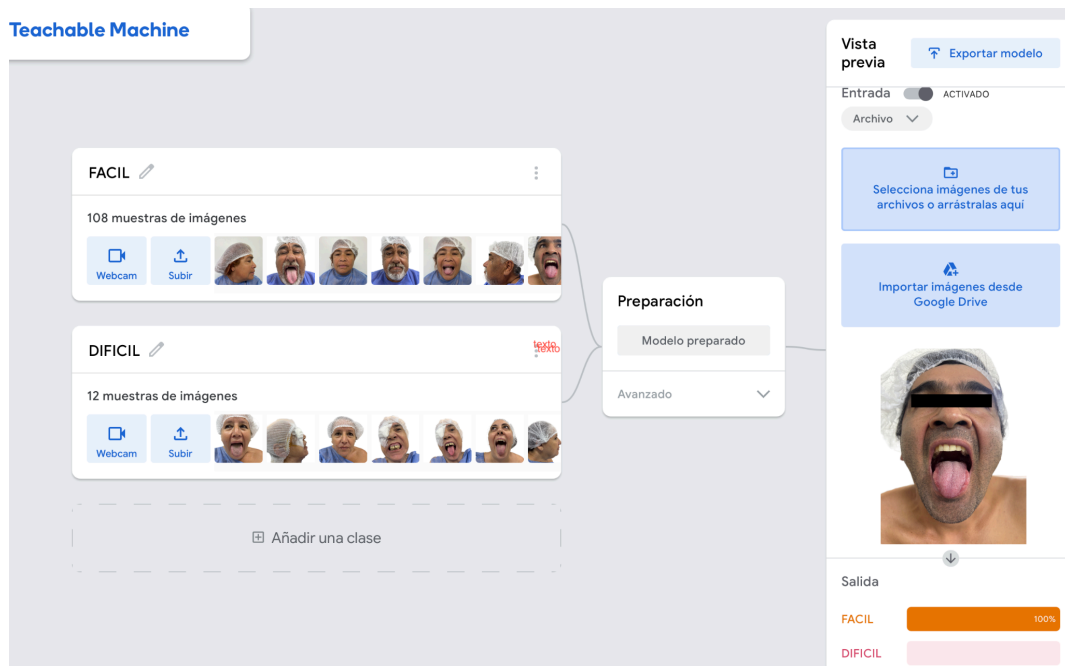


Figura 3. Interfaz de *Teachable Machine*

La variable principal fue la predicción de dificultad laringoscópica emitida por ChatGPT.

Como variables secundarias se analizaron:

- nivel cualitativo de confianza de ChatGPT (alto, medio o bajo)
- predicción aparente de la CNN (fácil o difícil)
- edad (años), sexo, índice de masa corporal (kg/m²) y clasificación ASA (I, II o III)
- MAL (I-IV; para el análisis diagnóstico, I-II frente a III-IV)
- DTM(< 6 cm o ≥ 6 cm)
- AB (< 3 cm o ≥ 3 cm)
- MC (conservada o limitada)
- RTG (presente o ausente)
- ULBT (positivo o negativo)
- Cormack-Lehane (I-IV; desenlace difícil: III-IV)
- número de intentos de intubación (1, 2 o 3)

- dispositivo utilizado (laringoscopio directo o videolaringoscopio)

Análisis estadístico

Las variables cuantitativas se expresaron mediante media, desvío estándar (DE) y rango. Las variables cualitativas se presentaron como frecuencias absolutas y porcentajes. Para evaluar el desempeño de ChatGPT se construyó una tabla de contingencia de 2×2 , utilizando la dificultad laringoscópica observada como criterio de referencia. Se calcularon sensibilidad, especificidad, valor predictivo positivo (VPP), valor predictivo negativo (VPN), falsos positivos (FP), verdaderos positivos (VP), falsos negativos (FN), verdaderos negativos (VN) razones de verosimilitud (LR), positiva (LR+), negativa (LR-) y proporción de clasificaciones correctas. Los respectivos intervalos de confianza del 95% fueron calculados según el método exacto de Clopper-Pearson a través de la función `epi.tests()`. De manera exploratoria, se comparó el desempeño de ChatGPT con los principales predictores clínicos y con la red neuronal convolucional. La diferencia entre las clasificaciones de ChatGPT y la red convolucional se evaluó mediante la prueba de McNemar. También se analizó la relación entre el nivel de confianza informado por ChatGPT y la corrección de la predicción. El procesamiento estadístico se realizó con R Core Team (2026) (13). Se adoptó un nivel de significación de 0,05.

RESULTADOS

Se analizaron 100 pacientes adultos sometidos a cirugía electiva bajo anestesia general. La edad promedio fue de 44,0 años, con un DE de 15,93 años y un rango de 18 a 84 años. El 69% de los participantes fue de sexo femenino. El IMC promedio fue de 26,1 kg/m², con un rango comprendido entre 18,0 y 47,6 kg/m² (Tabla 1).

Tabla 1. Características generales de la población estudiada

Variable	n (%) o media (DE)
Edad, años	44,0 (15,93)
Rango de edad, (años	18-84
Sexo femenino	69 (69,0)
IMC, kg/m ²	26,1 (4,83)
Rango de IMC, kg/m ²	18,0-47,6
ASA I	38 (38,0)
ASA II	58 (58,0)
ASA III	4 (4,0)

La evaluación preoperatoria mostró que 30 pacientes presentaron Mallampati III o IV. Los restantes predictores anatómicos fueron poco frecuentes, con excepción de la distancia tiromentoniana < 6 cm, observada en 22 pacientes (Tabla 2).

Tabla 2. Evaluación preoperatoria de la vía aérea

Evaluación preoperatoria	n (%)
Mallampati I	37 (37,0)
Mallampati II	33 (33,0)
Mallampati III	24 (24,0)
Mallampati IV	6 (6,0)
Distancia tiromentoniana < 6 cm	22 (22,0)
Apertura bucal < 3 cm	10 (10,0)
Movilidad cervical limitada	1 (1,0)
Retrognatía	5 (5,0)
ULBT restringido	2 (2,0)

Se observaron 8 casos de dificultad laringoscópica (8%), definidos como Cormack-Lehane III o IV. La mayoría de las intubaciones se resolvió en el primer intento y mediante laringoscopia directa (Tabla 3).

Tabla 3. Variables relacionadas durante la laringoscopia e intubación

Hallazgo	n (%)
Cormack-Lehane I	39 (39,0)
Cormack-Lehane II	53 (53,0)
Cormack-Lehane III	4 (4,0)
Cormack-Lehane IV	4 (4,0)
1 intento	83 (83,0)
2 intentos	16 (16,0)
3 intentos	1 (1,0)
Laringoscopia directa	99 (99,0)
Videolaringoscopia	1 (1,0)

ChatGPT expresó confianza alta en 33 respuestas y media en 67; no se registró confianza baja. Las predicciones de confianza alta fueron correctas con mayor frecuencia que las de confianza media (72,7% frente a 44,8%; prueba exacta de Fisher, $p = 0,011$; OR = 3,29) (Tabla 4).

Tabla 4. Relación entre el nivel de confianza de ChatGPT y la corrección de la predicción

Confianza de ChatGPT	Predicción correcta	Predicción incorrecta	Total
Alta	24 (72,7%)	9 (27,3%)	33 (100,0%)
Media	30 (44,8%)	37 (55,2%)	67 (100,0%)
Total	54 (54,0%)	46 (46,0%)	100 (100%)

ChatGPT clasificó como potencialmente difíciles a 50 pacientes. Identificó 6 de los 8 casos con dificultad y produjo 44 falsos positivos, con una exactitud global del 54%. Su perfil combinó una sensibilidad del 75% con una especificidad del 52,2% y un VPP del 12%, lo que evidenció una marcada sobreestimación del riesgo (Tablas 5 y 6).

La CNN clasificó 4 participantes como difíciles y 96 como fáciles. Los cuatro positivos correspondieron a casos con dificultad, mientras que otros cuatro casos no fueron identificados (Tabla 5).

Estas métricas describen rendimiento aparente sobre el conjunto utilizado para el desarrollo y no constituyen una validación independiente; por ello, se presentan únicamente como evidencia de factibilidad técnica.

Tabla 5. Dificultad laringoscópica observada y predicción de los modelos

Variable	n (%)
Dificultad laringoscópica observada	8 (8,0)
Predicción de dificultad por CNN*	4 (4,0)
Predicción de dificultad por ChatGPT	50 (50,0)
Confianza alta de ChatGPT	33 (33,0)
Confianza media de ChatGPT	67 (67,0)

El rendimiento diagnóstico de ChatGPT y de los predictores clínicos se resume en la Tabla 6. La CNN se incluye sólo como análisis descriptivo de entrenamiento y se identifica con un asterisco.

Entre los predictores clínicos, Mallampati III-IV mostró la mayor sensibilidad después de ChatGPT, mientras que la apertura bucal < 3 cm presentó la mayor razón de verosimilitud positiva. Los hallazgos de baja frecuencia generaron intervalos de confianza amplios (Tabla 6).

ChatGPT y la CNN coincidieron en 50 clasificaciones y discreparon en las 50 restantes. La prueba de McNemar indicó un patrón clasificatorio diferente ($p < 0,0001$), pero esta comparación no demuestra superioridad diagnóstica y está condicionada por la ausencia de validación independiente de la CNN.

Tabla 6. Comparativa de rendimiento diagnóstico de predictores

Evaluación frente al estándar de referencia. Muestra total N = 100 (Eventos verdaderos = 8, Prevalencia real = 8.0%).

Predictor	Matriz de Confusión				Métricas diagnósticas e Intervalos de confianza (95%)						
	VP	FP	VN	FN	Sensibilidad (%)	Especificidad (%)	VPP (%)	VPN (%)	LR+	LR-	Exactitud (%)
GPT	6	44	48	2	75.0% (35.0 - 97.0)	52.2% (41.6 - 62.6)	12.0% (4.5 - 24.3)	96.0% (86.3 - 99.5)	1.57 (1.00 - 2.47)	0.48 (0.14 - 1.62)	54.0% (43.7 - 64.0)
CNN	4	0	92	4	50.0% (15.7 - 84.3)	100.0% (96.1 - 100.0)	100.0% (39.8 - 100.0)	95.8% (89.7 - 98.9)	-	0.50 (0.25 - 1.00)	96.0% (90.1 - 98.9)
DTM	3	19	73	5	37.5% (8.5 - 75.5)	79.3% (69.6 - 87.1)	13.6% (2.9 - 34.9)	93.6% (85.7 - 97.9)	1.82 (0.68 - 4.84)	0.79 (0.46 - 1.36)	76.0% (66.4 - 84.0)
AB	4	6	86	4	50.0% (15.7 - 84.3)	93.5% (86.3 - 97.6)	40.0% (12.2 - 73.8)	95.6% (88.9 - 98.8)	7.67 (2.71 - 21.66)	0.53 (0.27 - 1.07)	90.0% (82.4 - 95.1)
MC	1	0	92	7	12.5% (0.3 - 52.7)	100.0% (96.1 - 100.0)	100.0% (2.5 - 100.0)	92.9% (85.8 - 97.1)	-	0.88 (0.67 - 1.14)	93.0% (86.1 - 97.1)
RTG	1	4	88	7	12.5% (0.3 - 52.7)	95.7% (89.2 - 98.8)	20.0% (0.5 - 71.6)	92.6% (85.4 - 96.8)	2.88 (0.36 - 22.76)	0.91 (0.70 - 1.19)	89.0% (81.2 - 94.4)
ULBT	2	0	92	6	25.0% (3.2 - 65.1)	100.0% (96.1 - 100.0)	100.0% (15.8 - 100.0)	93.9% (87.1 - 97.7)	-	0.75 (0.50 - 1.12)	94.0% (87.4 - 97.8)
MAL	5	25	67	3	62.5% (24.5 - 91.5)	72.8% (62.6 - 81.6)	16.7% (5.6 - 34.7)	95.7% (88.0 - 99.1)	2.30 (1.22 - 4.33)	0.51 (0.21 - 1.27)	72.0% (62.1 - 80.5)

DISCUSIÓN

Este estudio evaluó de manera exploratoria el desempeño de un modelo multimodal de ChatGPT para identificar dificultad laringoscópica a partir de fotografías faciales. El modelo detectó 6 de los 8 casos difíciles, pero clasificó como difíciles a 44 pacientes sin el desenlace. Por lo tanto, su sensibilidad moderada se obtuvo a costa de una baja especificidad y de un VPP del 12%, rendimiento insuficiente para una aplicación clínica autónoma.

La dificultad para anticipar estos eventos también se observa con las pruebas convencionales. Shiga y colaboradores informaron sensibilidades aproximadas del 20% al 62% y especificidades del 82% al 97% para los principales métodos de cabecera (4). En la presente muestra, Mallampati III-IV alcanzó una sensibilidad del 62,5%, mientras que la apertura bucal < 3 cm mostró la mejor razón de verosimilitud positiva. Estos resultados respaldan que ninguna medición aislada permite descartar o confirmar con seguridad una laringoscopia dificultosa.

La predicción clínica global tampoco es completamente confiable: Nørskov y colaboradores observaron que la mayoría de las intubaciones difíciles no había sido anticipada en la evaluación preoperatoria (5). Frente a esa limitación, el aporte potencial de la inteligencia artificial consiste en integrar rasgos visuales de forma reproducible, pero su utilidad depende del entrenamiento, la calibración y la validación externa.

Los estudios con redes específicamente entrenadas sobre imágenes faciales han comunicado desempeños superiores. Cuendet y colaboradores demostraron la factibilidad del análisis facial automatizado (9), y Kim y colaboradores informaron áreas bajo la curva de 0,81 a 0,88 utilizando miles de imágenes y particiones de entrenamiento y validación (10,11). La comparación con este trabajo debe ser cautelosa: ChatGPT es un modelo generalista, no fue entrenado para este desenlace y se evaluó en sólo 100 participantes.

La sobreestimación observada sugiere un umbral implícito orientado a evitar omisiones, pero una tasa alta de falsos positivos podría incrementar intervenciones, recursos y alertas sin beneficio demostrado. Las respuestas de confianza alta tuvieron más aciertos

que las de confianza media; sin embargo, esa confianza fue una categoría cualitativa derivada del lenguaje del modelo y no una probabilidad calibrada.

La CNN desarrollada con *Teachable Machine* produjo un rendimiento aparente elevado, pero no contó con un conjunto independiente de prueba. Esta ausencia impide distinguir capacidad de generalización de sobreajuste y hace que sus métricas no sean directamente comparables con las de ChatGPT ni con las pruebas clínicas. Su resultado debe interpretarse exclusivamente como demostración de factibilidad tecnológica.

Las principales limitaciones fueron el tamaño muestral, los ocho eventos observados, el diseño unicéntrico, y el uso de Cormack-Lehane III-IV como sustituto de dificultad laringoscópica. Además, la plataforma de ChatGPT pudo actualizar el modelo durante la recolección, los casos se procesaron en un contexto conversacional compartido y no se evaluó la estabilidad de las respuestas ante repeticiones. Estas condiciones reducen la precisión y la generalización de las estimaciones.

Futuros estudios deberán emplear muestras multicéntricas, un identificador fijo de modelo, conversaciones independientes por participante, repetición de consultas y comparación ciega con anesthesiólogos. Para las CNN será imprescindible separar por paciente los conjuntos de entrenamiento, validación y prueba. La literatura reciente considera prometedor el uso de IA en la evaluación de la vía aérea, pero subraya que la validación externa y la integración responsable con la valoración clínica son requisitos previos a cualquier implementación (11).

CONCLUSIÓN

En esta muestra de 100 pacientes, el modelo multimodal de ChatGPT identificó 6 de los 8 casos de dificultad laringoscópica, pero presentó baja especificidad y un valor predictivo positivo del 12%, con marcada sobreestimación del riesgo. En consecuencia, los resultados demuestran factibilidad como prueba de concepto, pero no respaldan su utilización clínica aislada ni permiten afirmar un papel complementario sin validación adicional. La CNN desarrollada mediante *Teachable Machine* mostró factibilidad técnica, aunque su rendimiento no fue evaluado en un conjunto independiente. Se requieren estudios prospectivos multicéntricos, versiones fijas de los modelos y cohortes externas antes de considerar una aplicación asistencial.

REFERENCIAS BIBLIOGRÁFICAS

1. American Society of Anesthesiologists Task Force on Management of the Difficult Airway. Practice guidelines for management of the difficult airway: an updated report. *Anesthesiology*. 2022;136(1):31-81.
2. Cook TM, MacDougall-Davis SR. Complications and failure of airway management. *Br J Anaesth*. 2012;109 Suppl 1:i68–i85.
3. Cook TM, Woodall N, Frerk C; Fourth National Audit Project. Major complications of airway management in the UK: results of the Fourth National Audit Project of the Royal College of Anaesthetists and the Difficult Airway Society. *Br J Anaesth*. 2011;106(5):617–31.
4. Shiga T, Wajima Z, Inoue T, Sakamoto A. Predicting difficult intubation in apparently normal patients: a meta-analysis of bedside screening test performance. *Anesthesiology*. 2005;103(2):429–37.
5. Nørskov AK, Rosenstock CV, Wetterslev J, Astrup G, Afshari A, Lundstrøm LH. Diagnostic accuracy of anaesthesiologists' prediction of difficult airway management in daily clinical practice: a cohort study of 188,064 patients registered in the Danish Anaesthesia Database. *Anaesthesia*. 2015;70(3):272–281.
6. Huitink JM. Airway triage: a novel application-based method for airway assessment and risk stratification. *Br J Anaesth*. 2024;132(5):987–90.
7. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–44.
8. Hayasaka T, Kawano K, Kurihara K, Suzuki H, Nakane M, Kawamae K. Creation of an artificial intelligence model for intubation difficulty classification by deep learning using face images: an observational study. *J Intensive Care*. 2021;9:38.
9. Cuendet GL, Schoettker P, Yüce A, Sorci M, Gao H, Perruchoud C, Thiran JP. Facial image analysis for fully automatic prediction of difficult endotracheal intubation. *IEEE Trans Biomed Eng*. 2016;63(2):328-339.

10. Kim JH, Jung HS, Lee SE, Hou JU, Kwon YS. Improving difficult direct laryngoscopy prediction using deep learning and minimal image analysis: a single-center prospective study. *Sci Rep*. 2024;14(1):14209.
11. Wilk M, Pikiewicz W, Florczak K, Jakóbczak D. Use of artificial intelligence in difficult airway assessment: the current state of knowledge. *J Clin Med*. 2025;14(5):1602.
12. OpenAI. Models: GPT-5 and multimodal capabilities. OpenAI Developer Documentation; 2026 [consultado el 11 de agosto de 2026]. Disponible en: <https://developers.openai.com/api/docs/models>
13. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2026. Disponible en: <https://www.R-project.org/>

ANEXO.

PROMPT ESTANDARIZADO PARA EL ANÁLISIS MEDIANTE CHATGPT

La siguiente consigna fue aplicada de manera estandarizada a las tres fotografías correspondientes a cada participante. El modelo no recibió información clínica adicional ni conoció el resultado de la laringoscopia al momento de emitir su predicción.

“Analizá las tres fotografías faciales preoperatorias correspondientes al mismo paciente: una imagen frontal con la boca cerrada, una imagen frontal con máxima apertura bucal y la lengua protruida, y una imagen lateral con la boca cerrada. A partir exclusivamente de los rasgos anatómicos visibles en las imágenes, evaluá si el paciente presenta características asociadas con una posible vía aérea dificultosa durante la intubación orotraqueal. Considerá, cuando puedan observarse, los siguientes elementos: apertura bucal, visualización de las estructuras orofaríngeas, relación entre la lengua y la cavidad oral, proporción mandibular, presencia aparente de retrognatia, perfil facial, características del mentón, volumen y longitud aparente del cuello y otros rasgos morfológicos que pudieran influir en el manejo de la vía aérea. No utilices información clínica que no esté presente en las imágenes y no infieras antecedentes del paciente. Respondé utilizando exactamente el siguiente formato:”

“Predicción: vía aérea fácil o vía aérea difícil.

Confianza: alta, media o baja.

Justificación: descripción breve de los principales rasgos anatómicos observados que fundamentan la predicción.”