

Preliminary Performance Assessment of a Customized GPT Model for Geomechanical Classification of Rock Masses Using Images and Input Data

Olger Andrés Cruzado-Araujo¹, Juan Alexander Urquiso-Segura², Julián Díaz-Ruiz³

¹Universidad Privada del Norte (UPN). Cajamarca, Perú., N00257141@upn.pe

²Universidad Privada del Norte (UPN). Cajamarca, Perú., N00261774@upn.pe

³Universidad Privada del Norte (UPN). Cajamarca, Perú., julian.diaz@upn.edu.pe

Abstract– Geomechanical characterization of rock masses is essential for the stability of mining and civil works; however, it is still mostly performed manually, making the process time-consuming and highly dependent on human judgment. This study presents a preliminary evaluation of the performance of a customized GPT model for geomechanical classification using images and input data, comparing its results with field classifications obtained using the Bieniawski RMR system. The research followed a quantitative approach with an applied level, a non-experimental and cross-sectional design, and an exploratory and descriptive scope, analyzing 30 observation points from rock outcrops in the Cajamarca region (Peru). The model evaluated each image in three independent runs and was configured using the technical criteria of the RMR system. The results showed an average accuracy of 83.4%, a simple Kappa of 0.76, and a weighted quadratic Kappa of 0.88, corresponding to substantial to almost perfect agreement, with an overall inter-run reliability of 75.6%. Overall, the customized GPT model demonstrated solid preliminary performance and results consistent with human classification, suggesting potential for future application in the geomechanical characterization of rock masses.

Keywords– geomechanical classification, RMR, artificial intelligence, customized GPT, rock masses.

Evaluación preliminar del desempeño de un modelo GPT personalizado aplicado a la clasificación geomecánica de macizos rocosos mediante imágenes y datos de entrada

Olger Andrés Cruzado-Araujo¹, Juan Alexander Urquiso-Segura², Julián Díaz-Ruiz³

¹Universidad Privada del Norte (UPN). Cajamarca, Perú., N00257141@upn.pe

²Universidad Privada del Norte (UPN). Cajamarca, Perú., N00261774@upn.pe

³Universidad Privada del Norte (UPN). Cajamarca, Perú., julian.diaz@upn.edu.pe

Resumen— La caracterización geomecánica de macizos rocosos es esencial para la estabilidad de obras mineras y civiles, pero aún se realiza mayormente de forma manual, lo que hace que el proceso sea lento y dependiente del criterio humano. En este estudio se evaluó de manera preliminar el desempeño de un modelo GPT personalizado para la clasificación geomecánica a partir de imágenes y datos de entrada, comparando sus resultados con los obtenidos en campo mediante el sistema RMR de Bieniawski. La investigación tuvo un enfoque cuantitativo, nivel aplicado, diseño no experimental y transversal, y un alcance exploratorio y descriptivo. Se trabajó con 30 puntos de observación en afloramientos de la región Cajamarca (Perú). El modelo analizó cada imagen en tres ejecuciones independientes, configurado con los criterios técnicos del sistema RMR. Los resultados mostraron una precisión promedio del 83.4 %, un Kappa simple de 0.76 y un Kappa ponderado cuadrático de 0.88, equivalentes a concordancias sustanciales a casi perfectas. La confiabilidad global entre ejecuciones fue del 75.6 %. En general, el modelo GPT personalizado presentó un desempeño preliminar sólido y resultados coherentes con la clasificación humana, lo que sugiere un potencial de aplicación futura en la caracterización geomecánica de macizos rocosos.

Palabras clave— Clasificación geomecánica, RMR, inteligencia artificial, GPT personalizado, macizos rocosos.

I. INTRODUCCIÓN

La caracterización y clasificación geomecánica de los macizos rocosos desempeñan un papel crucial en el diseño y la seguridad de los proyectos mineros y de infraestructura, incluidas excavaciones subterráneas, taludes, túneles y presas. Por lo tanto, la toma de decisiones vitales relacionadas con el sostenimiento, la estabilidad y la gestión del riesgo geotécnico depende en gran medida de la calidad de esta evaluación. Los procesos de caracterización geomecánica se han basado generalmente en el mapeo geomecánico de campo y sistemas empíricos de clasificación. Estos métodos se basan principalmente en la observación directa y el juicio del especialista, lo que introduce un considerable componente de subjetividad y variabilidad en los resultados [1–3].

A pesar de los avances tecnológicos, la caracterización geomecánica continúa haciéndose mayormente de forma manual. Esta situación va más allá del tiempo que consume: estandarizar los procedimientos se vuelve complicado y

manejar grandes volúmenes de información resulta casi impracticable, más aún cuando la geología es compleja y las condiciones del macizo rocoso varían considerablemente de un punto a otro [4,5]. Los análisis resultan menos eficientes, y su confiabilidad se ve afectada precisamente por esa heterogeneidad tan común en estos ambientes.

En años recientes, la inteligencia artificial (IA) ha comenzado a abrirse paso como herramienta para automatizar y optimizar distintas tareas dentro de la ingeniería geotécnica y minera. Hay trabajos que muestran cómo el aprendizaje automático, las redes neuronales y la visión por computador pueden mejorar la precisión en la clasificación de macizos rocosos y disminuir esa carga de subjetividad que trae consigo el criterio humano, incluyendo la estimación de índices como el GSI y el RMR [6–9]. Sin embargo, la mayoría de estos modelos son bastante especializados y exigen condiciones que no siempre están disponibles: cantidades importantes de datos etiquetados, entrenamientos extensos y recursos computacionales que implican inversiones significativas. Por esta razón, su aplicación tiende a concentrarse en proyectos grandes o en el ámbito de la investigación.

Esta situación se siente particularmente en América Latina, y el Perú no es la excepción. Aunque algunas operaciones mineras de gran envergadura han logrado incorporar modelos digitales y representaciones tridimensionales del macizo rocoso, estas experiencias generalmente dependen de consultorías externas y se enfrentan a limitaciones relacionadas con costos, licencias de software y la falta de personal capacitado en estas tecnologías [10–12]. La mayoría de las operaciones medianas y pequeñas, en cambio, siguen trabajando con las metodologías de siempre, lo que dificulta cualquier intento de transición hacia esquemas más modernos como los que plantea la Minería 4.0 [13].

Frente a esto, los modelos de inteligencia artificial generativa, particularmente los modelos GPT (Generative Pre-trained Transformer), ofrecen una alternativa que merece atención. A diferencia de los modelos de machine learning, estos sistemas no necesitan pasar por entrenamientos complejos ni exigen infraestructura especializada para cada aplicación particular. Los GPT trabajan con arquitecturas de lenguaje que ya han

sido preentrenadas con cantidades masivas de información, lo que les permite interpretar instrucciones técnicas, analizar datos de distintas naturalezas y producir respuestas razonablemente coherentes en varios campos de la ingeniería [14–16]. Cuando se les incorporan capacidades de visión y lenguaje (Vision-Language Models, VLM), su potencial crece todavía más, ya que pueden procesar imágenes y texto al mismo tiempo, algo útil cuando se trabaja con fotografías de campo de macizos rocosos [17,18].

Algunos estudios recientes han encontrado que estos modelos preentrenados logran generar descripciones técnicas de imágenes con un nivel de detalle comparable al de especialistas humanos, sin requerir ajustes adicionales [19]. A pesar de esto, hay pocos trabajos que evalúen de manera sistemática qué tan bien funcionan los modelos GPT personalizados como apoyo en la clasificación geomecánica, sobre todo cuando se trata de imágenes capturadas directamente en campo.

Un punto que resulta fundamental al evaluar este tipo de sistemas es cómo se mide realmente su desempeño. La exactitud, por sí sola, no dice mucho; hace falta incorporar métricas que reflejen el grado de concordancia y confiabilidad frente a las evaluaciones realizadas por personas [20–22]. Métricas como el coeficiente Kappa, el porcentaje de acuerdo y el porcentaje de acierto se han usado ampliamente para este propósito, tanto en ingeniería como en otras disciplinas [23–25].

Con base en estas consideraciones, el objetivo de este estudio es evaluar de forma preliminar qué tan bien se desempeña un modelo GPT personalizado en la clasificación geomecánica de macizos rocosos. Se trabaja con imágenes y datos de entrada, y los resultados se comparan con clasificaciones hechas en campo usando el sistema Rock Mass Rating (RMR). El presente estudio no busca sustituir los métodos tradicionales, sino explorar si la inteligencia artificial generativa puede servir como herramienta de apoyo que ayude a reducir la subjetividad, acorte tiempos de análisis y contribuya a estandarizar las etapas iniciales de la caracterización geomecánica, teniendo en cuenta las limitaciones y necesidades propias del contexto minero peruano.

II. METODOLOGÍA

A. Enfoque y diseño del estudio

El estudio se desarrolló bajo un enfoque cuantitativo, orientado a la recolección y análisis de datos estructurados con el fin de evaluar, de manera numérica, el desempeño de un modelo GPT personalizado en la clasificación geomecánica de macizos rocosos. Este enfoque permitió realizar comparaciones sistemáticas entre la clasificación obtenida en campo y la generada por el modelo, así como aplicar

indicadores estadísticos de precisión, concordancia y confiabilidad [26].

El nivel de investigación fue aplicado, dado que se buscó resolver un problema específico de la ingeniería geomecánica mediante el uso de inteligencia artificial como herramienta de apoyo en la clasificación del macizo rocoso [27]. El alcance fue exploratorio–descriptivo: exploratorio, al tratarse de una evaluación preliminar de la viabilidad del uso de modelos GPT en este contexto; y descriptivo, porque se analizaron las categorías de clasificación RMR y sus coincidencias sin establecer relaciones causales [26].

El diseño fue no experimental y transversal, observándose el fenómeno en su contexto real, sin manipulación de variables y con recolección de datos en un único momento temporal [28].

B. Área de estudio y contexto geológico

La población de estudio estuvo conformada por macizos rocosos aflorantes en taludes de carreteras de la región Cajamarca, Perú. Los puntos de observación se localizaron en los tramos Quinamayó–desvío Michiquillay, Puente Muyuna–Tabacal, La Mónica–Llallán y Puente Yonán–Tembladera.

Los afloramientos evaluados corresponden a distintas unidades geológicas, incluyendo los grupos Pillucana (Ks-pu), Farrat (Ki-f) y Quilquiñan (Ks-qu), así como las formaciones Carhuaz (Ki-ca), Llama (P-ll) y Santa (Ki-s). Estas unidades presentan litologías variadas, tales como calizas, lutitas, areniscas, cuarcitas, conglomerados y tobas volcánicas, con diferentes grados de fracturamiento y alteración, lo que permitió abarcar un rango amplio de calidades geomecánicas para su clasificación mediante el sistema Rock Mass Rating (RMR) [1].

C. Muestra y selección de puntos de observación

La muestra estuvo conformada por 30 puntos de observación geomecánica, seleccionados mediante muestreo por conveniencia en afloramientos accesibles de los taludes. En cada punto se realizó un mapeo geomecánico completo y se obtuvo evidencia fotográfica del macizo rocoso.

El tamaño muestral se definió considerando criterios metodológicos para la aplicación del coeficiente Kappa. Estudios recientes señalan que, en investigaciones exploratorias con variables categóricas de múltiples niveles, un número de entre 25 y 30 observaciones proporciona mayor estabilidad estadística y robustez en la estimación de la concordancia [29,30].

D. Técnicas e instrumentos de recolección de datos

En campo se empleó la observación directa estructurada como técnica principal, aplicada durante el mapeo geomecánico para registrar los parámetros del sistema RMR básico [4]. Este procedimiento se complementó con el registro fotográfico sistemático de cada afloramiento, constituyendo la referencia humana frente a la cual se evaluó el modelo GPT.

El instrumento principal fue una ficha de registro geomecánico diseñada específicamente para el estudio, que permitió organizar de manera homogénea los datos obtenidos en cada punto de observación. Como herramientas auxiliares

se utilizaron un teléfono móvil con cámara de alta resolución, un flexómetro, un martillo geológico y un peine de Barton para la evaluación de la rugosidad de las discontinuidades.

E. Procedimiento de mapeo geomecánico en campo

El procedimiento se inició con una revisión bibliográfica especializada en mecánica de rocas y sistemas de clasificación geomecánica, con énfasis en el método RMR, a fin de asegurar la correcta aplicación de los criterios durante el trabajo de campo.

En cada punto de observación se realizó un mapeo geomecánico por celda de 3×3 m. La resistencia de la roca se estimó de manera indirecta mediante observación de su dureza, textura y grado de alteración. El RQD se calculó aplicando la formulación de Priest y Hudson, basada en el conteo de fracturas visibles por metro. El espaciamiento de discontinuidades se midió directamente y se utilizó el valor promedio de las mediciones representativas.

Las condiciones de las discontinuidades se evaluaron considerando la apertura, la rugosidad y el tipo de relleno, mientras que el estado de humedad se registró mediante observación directa. Todos los parámetros fueron valorados según las escalas propuestas por Bieniawski [1], obteniéndose el valor de RMR básico para cada estación. No se aplicaron correcciones por orientación de discontinuidades, con el fin de mantener la objetividad y facilitar la comparación directa con los resultados del modelo GPT.

Los mapeos y las clasificaciones finales fueron revisados por un ingeniero con experiencia en geotecnia, garantizando la validez y consistencia de los registros.

F. Configuración y aplicación del modelo GPT personalizado

Finalizada la fase de campo, se construyó una base de datos estructurada que vinculó cada imagen con sus parámetros registrados y su clasificación RMR correspondiente. Esta base de datos constituyó el insumo para el análisis automatizado.

El modelo GPT personalizado fue desarrollado en la plataforma OpenAI (ChatGPT Plus) y configurado mediante la incorporación de información teórica digitalizada del sistema RMR, incluyendo tablas de valoración y criterios técnicos. La interacción con el modelo se realizó a través de instrucciones de configuración y prompts estructurados, en los cuales se ingresaron las imágenes de los afloramientos junto con información contextual mínima.

Cada imagen fue procesada en tres pasadas independientes, permitiendo generar múltiples clasificaciones para los mismos datos de entrada. Si bien el modelo produjo valores individuales para cada parámetro del RMR, el análisis se centró exclusivamente en la categoría final de clasificación, por ser el insumo principal utilizado en la práctica profesional para la toma de decisiones.

G. Análisis estadístico

El desempeño del modelo se evaluó considerando tres dimensiones: concordancia, precisión y confiabilidad. La concordancia entre la clasificación de campo y la generada por el modelo GPT se analizó mediante matrices de confusión y se

cuantificó utilizando el coeficiente Kappa simple, el Kappa ponderado cuadrático y el porcentaje de acuerdo [23,24].

La precisión se definió como el porcentaje de aciertos globales del modelo frente a la referencia humana, mientras que la confiabilidad intra-modelo se estimó comparando las tres pasadas independientes del GPT mediante el porcentaje de acuerdo y los coeficientes Kappa obtenidos en comparaciones por pares.

Los cálculos se realizaron en Microsoft Excel mediante una plantilla de análisis con fórmulas verificadas. Para asegurar la exactitud de los coeficientes Kappa, los resultados fueron contrastados con una calculadora estadística independiente (MedCalc). La interpretación de los valores de Kappa se realizó siguiendo los criterios propuestos por Vanbelle et al. [25], mientras que para el porcentaje de acuerdo y la precisión se consideró como aceptable un umbral mínimo del 75 % [31]. Dado el carácter exploratorio del estudio, no se aplicaron pruebas de significancia estadística.

H. Consideraciones éticas

La investigación se desarrolló conforme a los lineamientos del Código de Ética para la Investigación Científica de la Universidad Privada del Norte. Se garantizaron los principios de autonomía, beneficencia, justicia, honestidad y transparencia, así como el respeto a la propiedad intelectual mediante la correcta citación de las fuentes utilizadas. Asimismo, se aseguró la originalidad del trabajo, manteniendo un nivel de similitud acorde con los criterios institucionales establecidos.

III. RESULTADOS

El análisis comparativo entre las clasificaciones realizadas en campo y las generadas por el modelo GPT personalizado permitió evaluar su precisión, concordancia y confiabilidad.

A. Clasificación RMR en campo

En los 30 puntos de observación evaluados, los valores de RMR variaron entre 38 y 89, abarcando desde macizos malos hasta muy buenos.

La Tabla 1 resume la distribución de categorías obtenidas en campo.

TABLA I
CLASIFICACIÓN GEOMECAÁNICA DE CAMPO SEGÚN RMR

Categoría RMR	Rango de RMR	Frecuencia	Porcentaje (%)
I – Muy buena	81 – 100	7	23%
II – Buena	61 – 80	12	40%
III – Regular	41 – 60	9	30%
IV – Mala	21 – 40	2	7%

B. Configuración del modelo GPT

Bloque	Elemento	Descripción
Identidad del modelo	Nombre y logo del modelo GPT	Elementos visuales que distinguen al modelo personalizado.
Configuración interna	Instrucciones de configuración	Guías iniciales que integran criterios técnicos del sistema RMR.
	Tablas de referencia RMR	Rangos de puntuación para resistencia, RQD, espaciamiento, discontinuidades y agua.
Datos de entrada (inputs)	Parámetro complementario	Resistencia de la roca intacta ingresada manualmente.
	Tipo de entrada (prompt)	Instrucciones que combinan imagen del afloramiento + resistencia + litología.
Resultados (outputs)	Tipo de salida	Categoría RMR y puntuación correspondiente generada por el modelo.
Acceso al modelo	Enlace de acceso	https://chatgpt.com/g-689d6567e60481918dadbca8f349a469-rockscan-rmr-gpt

C. Desempeño del modelo GPT personalizado (precisión y concordancia)

TABLA II
MATRIZ DE CONFUSIÓN: HUMANO VS GPT-1 (PASADA 1)

Humano / GPT1	I-Muy buena	II-Buena	III-Regular	IV - Mala	V-Muy mala	Total
I-Muy buena)	7	0	0	0	0	7
II-Buena	0	8	4	0	0	12
III-Regular	0	0	9	0	0	9
IV -Mala	0	0	0	2	0	2
V-Muy mala	0	0	0	0	0	0
Total	7	8	13	2	0	30

En la comparación entre la clasificación realizada en campo y la obtenida en la primera ejecución del modelo GPT, se observó que 26 de los 30 casos coincidieron de manera exacta (2 en la categoría IV–Mala, 9 en III–Regular, 8 en II–Buena y 7 en I–Muy buena), que representan un 86.7% de aciertos (precisión). Con un Kappa simple de 0.811 y un Kappa cuadrático de 0.917, lo que indica una concordancia sustancial a casi perfecta.

TABLA III
MATRIZ DE CONFUSIÓN: HUMANO VS MODELO GPT-2 (PASADA 2)

Humano \ GPT2	I-Muy buena	II-Buena	III-Regular	IV - Mala	V-Muy mala	Total
I-Muy buena)	4	3	0	0	0	7
II-Buena	2	9	1	0	0	12
III-Regular	0	0	9	0	0	9
IV -Mala	0	0	1	1	0	2
V-Muy mala	0	0	0	0	0	0
Total	6	12	11	1	0	30

En la comparación entre la clasificación realizada en campo y la obtenida en la segunda ejecución del modelo GPT, se observó que 23 coincidencias exactas (1 en IV–Mala, 9 en III–Regular, 9 en II–Buena y 4 en I–Muy buena), que representan un 76.7% de aciertos (precisión). Este nivel de coincidencias equivale a un 76.7% de acuerdo global (concordancia), con un Kappa simple de 0.657 y un Kappa cuadrático de 0.834, lo que indica una concordancia sustancial.

TABLA IV
MATRIZ DE CONFUSIÓN: HUMANO VS MODELO GPT-3 (PASADA 3)

Humano \ GPT3	I-Muy buena	II-Buena	III-Regular	IV - Mala	V-Muy mala	Total
I-Muy buena)	6	1	0	0	0	7
II-Buena	0	11	1	0	0	12
III-Regular	0	0	9	0	0	9
IV -Mala	0	0	2	0	0	2
V-Muy mala	0	0	0	0	0	0
Total	6	12	12	0	0	30

En la comparación entre la clasificación realizada en campo y la obtenida en la tercera ejecución del modelo GPT, se observó que 26 coincidencias exactas (9 en III–Regular, 11 en II–Buena y 6 en I–Muy buena), que representan un 86.7% de aciertos (precisión). Este nivel de coincidencias corresponde también a un 86.7% de acuerdo global (concordancia), con un Kappa simple de 0.802 y un Kappa cuadrático de 0.899, lo que indica una concordancia sustancial a casi perfecta.

TABLA V
INDICADORES DE DESEMPEÑO DEL MODELO GPT PERSONALIZADO

Ejecución	Precisión (%)	Kappa simple	Kappa ponderado cuadrático
GPT–1	86.7	0.81	0.92
GPT–2	76.7	0.66	0.83
GPT–3	86.7	0.80	0.89
Promedio	83.4	0.76	0.88

El modelo logro alcanzar una precisión promedio de 83.4 % y kappa simple promedio de 0.76 y cuadrático de 0.88 correspondientes a una concordancia sustancial a casi perfecta.

D. Confiabilidad entre pasadas

TABLA VI
MATRIZ DE CONFUSIÓN: PASADA-1 VS PASADA-2

GPT-1 \ GPT-2	I-Muy buena	II- Buena	III- Regular	IV - Mala	V-Muy mala	Total
I-Muy buena)	4	3	0	0	0	7
II-Buena	2	5	1	0	0	8
III-Regular	0	3	10	0	0	13
IV -Mala	0	0	1	1	0	2
V-Muy mala	0	0	0	0	0	0
Total	6	11	12	1	0	30

En la comparación entre las pasadas 1 y 2, se evidencia que el modelo volvió a clasificar de forma idéntica en 20 de los 30 casos, que es igual a 66,7 % de acuerdo. El Kappa simple fue de 0.51 y el cuadrático de 0.78 indican una concordancia moderada, lo que muestra que el modelo no siempre reproduce sus decisiones iniciales cuando repite el análisis.

TABLA VII
MATRIZ DE CONFUSIÓN: PASADA-1 VS PASADA-3

GPT-1 \ GPT-3	I-Muy buena	II- Buena	III- Regular	IV - Mala	V- Muy mala	Total
I-Muy buena)	6	1	0	0	0	7
II-Buena	0	8	0	0	0	8
III-Regular	0	3	10	0	0	13
IV -Mala	0	0	2	0	0	2
V-Muy mala	0	0	0	0	0	0
Total	6	12	12	0	0	30

En la comparación de las pasadas 1 y 3, el modelo mantuvo la misma clasificación en 24 de los 30 casos (80 %), reflejando un nivel mayor de estabilidad. El Kappa simple (0.70) y el cuadrático (0.86) evidencian una concordancia sustancial, lo que indica que el modelo tiende a reproducir sus decisiones al repetir la evaluación.

TABLA VIII
MATRIZ DE CONFUSIÓN: PASADA-2 VS PASADA-3

GPT-2 \ GPT-3	I-Muy buena	II- Buena	III- Regular	IV - Mala	V- Muy mala	Total
I-Muy buena)	4	2	0	0	0	6
II-Buena	2	9	0	0	0	11
III-Regular	0	1	11	0	0	12
IV -Mala	0	0	1	0	0	1
V-Muy mala	0	0	0	0	0	0
Total	6	12	12	0	0	30

Se evidencia que también se alcanzaron 24 coincidencias (80 %), lo que confirma que en la mayoría de los casos el modelo volvió a asignar la misma categoría en pasadas consecutivas. Los valores de Kappa (0.69 simple y 0.84 cuadrático) confirman una concordancia sustancial, aunque ligeramente inferior a GPT1 vs GPT3 por la mayor concentración de etiquetas en una sola categoría.

TABLA IX
RESUMEN DE LA CONFIABILIDAD ENTRE PASADAS

Comparación	% Acuerdo	Kappa simple	Kappa cuadrática
GPT-1 vs GPT-2	66.70%	0.51	0.78
GPT-1 vs GPT-3	80.00%	0.70	0.86
GPT-2 vs GPT-3	80.00%	0.69	0.84
Media	75.6 %	0.64	0.82

En promedio, se obtuvo un $75.6\% \pm 7.7$ de acuerdo, con un Kappa simple de 0.636 ± 0.11 y un Kappa cuadrático de 0.823 ± 0.04 . Estos valores reflejan un nivel de confiabilidad aceptable, aunque con menor consistencia entre las pasadas 1 y 2.

IV. DISCUSIÓN

El objetivo general de este estudio fue evaluar de manera preliminar el desempeño de un modelo GPT personalizado para la clasificación geomecánica de macizos rocosos a partir de imágenes y parámetros de entrada. Los resultados obtenidos evidencian un desempeño estadísticamente sólido, con una precisión promedio del 83.4 %, un coeficiente Kappa simple de 0.757 y un Kappa ponderado cuadrático de 0.883, lo que corresponde a niveles de concordancia que varían de sustancial a casi perfecta según los criterios establecidos para datos ordinales [25]. Estos valores indican que el modelo fue capaz de reproducir, con un alto grado de coherencia, los criterios de clasificación aplicados por el especialista humano.

Desde el punto de vista de la confiabilidad, las tres pasadas del modelo mostraron un acuerdo promedio del 75.6 %, lo que refleja una estabilidad aceptable en la repetición de las clasificaciones. Este comportamiento es consistente con estudios que evalúan la concordancia entre sistemas de inteligencia artificial y evaluaciones humanas, donde se reportan niveles sustanciales de acuerdo como indicadores adecuados de desempeño [24]. Asimismo, el análisis conjunto de precisión, concordancia y estabilidad responde a las recomendaciones metodológicas actuales para la evaluación integral de sistemas de inteligencia artificial [20].

No obstante, se identificó una tendencia del modelo a reclasificar observaciones entre categorías geomecánicas cercanas, lo que sugiere una sensibilidad limitada para discriminar límites finos entre clases adyacentes del sistema RMR. Este comportamiento es coherente con lo reportado en la literatura sobre modelos de inteligencia artificial que carecen de mecanismos avanzados de calibración o control

explícito de la aleatoriedad [21,22]. En este sentido, aunque los resultados fueron positivos, se evidencia la necesidad de ampliar el número de muestras y optimizar las instrucciones para mejorar la estabilidad del modelo.

El establecimiento de una línea base mediante el mapeo geomecánico en campo, aplicando el sistema RMR, permitió contar con una referencia confiable para la comparación entre la clasificación humana y la clasificación automatizada. Los valores RMR obtenidos abarcaron las categorías I a IV, con predominio de la categoría II (Buena). El uso del sistema RMR, ampliamente validado en la ingeniería geotécnica [1,5], aseguró la solidez metodológica del estudio. Sin embargo, la ausencia de la categoría V (Muy mala) limitó la cobertura total del rango de calidad del macizo rocoso, reduciendo parcialmente la representatividad de la muestra.

La configuración del modelo GPT se realizó mediante la incorporación de criterios técnicos, tablas de referencia y rangos de puntuación del sistema RMR, lo que permitió estructurar su razonamiento conforme a los principios propuestos por Bieniawski [1]. Este enfoque es coherente con los avances recientes en el uso de modelos de inteligencia artificial aplicados a la clasificación geomecánica, los cuales destacan la utilidad de herramientas basadas en inteligencia artificial para apoyar la toma de decisiones sin requerir entrenamientos supervisados extensivos [6,7]. Asimismo, el uso de modelos de visión-lenguaje representa una alternativa práctica frente a métodos tradicionales que demandan grandes volúmenes de datos [17-19].

En términos de concordancia, los valores de Kappa simple y ponderado confirman que el modelo GPT logró reproducir de manera consistente las decisiones del evaluador humano. La aplicación del coeficiente Kappa ponderado es particularmente adecuada en clasificaciones ordinales como el RMR, ya que permite considerar la magnitud de los desacuerdos entre categorías [23,25]. Además, la evaluación de la confiabilidad intra-modelo mediante pasadas repetidas concuerda con la importancia atribuida a la repetibilidad como indicador clave del desempeño de sistemas de inteligencia artificial [20].

Entre las principales limitaciones del estudio se encuentran el tamaño reducido de la muestra, la concentración de los puntos en una sola región geológica, la imposibilidad de aplicar otros sistemas de clasificación como Q o GSI [5], y la ausencia de antecedentes específicos sobre el uso de modelos GPT en geomecánica. Asimismo, el modelo utilizado depende de conectividad a internet y de una suscripción comercial, lo que puede restringir su aplicación operativa. A pesar de estas limitaciones, los resultados obtenidos constituyen una base sólida y demuestran que los modelos GPT personalizados pueden emplearse como herramientas de apoyo en la clasificación geomecánica.

Desde una perspectiva académica y práctica, este estudio aporta evidencia preliminar sobre el potencial de la inteligencia artificial en la caracterización de macizos rocosos, alineándose con las tendencias actuales de la minería 4.0 y la digitalización de procesos geotécnicos [11-13]. La ampliación

futura del número de muestras, la inclusión de otros sistemas de clasificación y la aplicación del modelo en diferentes contextos geológicos permitirán fortalecer la validez y aplicabilidad de este enfoque.

V. CONCLUSIONES

El modelo GPT personalizado evaluado en este estudio mostró un desempeño preliminar sólido en la clasificación geomecánica de macizos rocosos a partir de imágenes y datos de entrada. Se obtuvo una precisión del 83.4 %, con niveles de concordancia que varían de sustancial a casi perfecta (Kappa simple = 0.757; Kappa ponderado cuadrático = 0.883) y una confiabilidad global promedio del 75.6 %. Estos resultados evidencian la capacidad del modelo para reproducir de manera coherente la lógica de clasificación aplicada por especialistas humanos, aun cuando el análisis se basa en fotografías complementadas con información geomecánica básica.

El mapeo geomecánico realizado en campo permitió establecer una línea base de referencia a partir de 30 puntos evaluados mediante el sistema RMR, acompañados de registro fotográfico sistemático. Esta estrategia proporcionó una base objetiva y consistente para contrastar las clasificaciones generadas por el modelo GPT. No obstante, la ausencia de observaciones en la categoría “Muy mala” y el tamaño limitado de la muestra reducen la representatividad del conjunto de datos, lo que confirma la necesidad de ampliar futuras campañas de campo y abarcar una mayor diversidad de condiciones geológicas.

La configuración del modelo GPT mediante instrucciones y prompts específicos permitió su adaptación a la tarea de clasificación geomecánica sin requerir entrenamiento supervisado. Sin embargo, se evidenció que el rendimiento del modelo estuvo fuertemente condicionado por la claridad, precisión y estructura de las instrucciones proporcionadas. Las discrepancias observadas entre pasadas se explican, en parte, por la imposibilidad de ajustar parámetros avanzados como la temperatura o el nivel de aleatoriedad, lo que resalta la importancia de perfeccionar y estandarizar los prompts en futuras aplicaciones.

El modelo GPT personalizado demostró ser capaz de aplicar de manera consistente el sistema RMR para la clasificación geomecánica, manteniendo estabilidad metodológica en sus resultados. Este comportamiento pone en evidencia el potencial de los modelos generativos como herramientas de apoyo para la caracterización geomecánica, especialmente en contextos donde se busca optimizar tiempos de análisis. No obstante, la ligera tendencia a reclasificar entre categorías vecinas indica la necesidad de optimizar las instrucciones y ampliar las referencias visuales para mejorar la precisión en estudios posteriores.

Finalmente, el análisis conjunto de concordancia, precisión y confiabilidad entre las clasificaciones generadas por el modelo GPT y las obtenidas en campo mostró acuerdos que oscilaron entre el 66.7 % y el 86.7 %, con valores de Kappa simple y ponderado en rangos estadísticamente

aceptables. La comparación entre las distintas pasadas evidenció una estabilidad interna adecuada del modelo, aunque con menor consistencia en la segunda corrida. Este comportamiento refuerza la necesidad de incrementar el número de pruebas y diversificar los contextos geológicos evaluados, con el fin de fortalecer la validez y robustez de futuras aplicaciones.

REFERENCIAS

- [1] Bieniawski ZT. Engineering rock mass classifications: a complete manual for engineers and geologists in mining, civil, and petroleum engineering. 1989.
- [2] Erhardt GH. Rock mass structure characterization considering finite and folded discontinuities: a parametric study. *Rock Mech Rock Eng.* 2024;57(7):5229-5249.
- [3] Liu Y, Hua W, Chen Q, Liu X. Characterization of complex rock mass discontinuities from lidar point clouds. *Remote Sens.* 2024;16(17):3291.
- [4] Bravo MF, Muñoz E, Lapeña MP, Montenegro CJ, King RW. Analysis of the influence of geomechanical parameters and geometry on slope stability in granitic residual soils. *Appl Sci.* 2022;12(11):5574.
- [5] Rodríguez G. Clasificación geomecánica de macizos rocosos: RMR (Bieniawski), Q (Barton) y GSI (Hoek) - comparación, limitaciones y aplicaciones. Centro Geotécnico. 2025 Jan 30. Available from: <https://www.centrogeotecnico.com/blog-geotecnia-geomecanica/clasificacion-geomecanica-de-macizos-rocosos-rmr-bieniawski-q-barton-y-gsi-hoek-comparacion-limitaciones-y-aplicaciones.html>
- [6] Saadati G, Javankhoshdel S, Mohebbi Najm Abad J, Mett M, Kontrus H, Schneider-Muntau B. AI-powered geotechnics: enhancing rock mass classification for safer engineering practices. *Rock Mech Rock Eng.* 2024. doi:10.1007/s00603-024-04189-7.
- [7] Niu G, He X, Xu H, Dai S. Development of rock classification systems: a comprehensive review with emphasis on artificial intelligence techniques. *Eng.* 2024;5(1):217-245.
- [8] Garzón-Roca J, Torrijo FJ, Rodríguez-Peces MJ, Ramos A. Geological strength index quantitative estimation of flysch rock masses from on-field geomechanical data based on AI tools. *Bull Eng Geol Environ.* 2025;84(6).
- [9] Castro Huaman KV, Brousset Cornejo JG. Propuesta de red neuronal para caracterizar el macizo rocoso en una mina subterránea en el centro del Perú. 2023 Nov 10. Available from: <http://hdl.handle.net/10757/670755>
- [10] Fernández J, Espinoza A, Quispe C, Asencio E, Aponte R. Importancia del desarrollo de modelos geomecánicos 3D en entornos geológicos complejos, aplicado a la mina Catalina Huanca, Aya. *Rev Minería.* 2022. Available from: https://revistamineria.com.pe/tecnico-cientifico/importancia-del-desarrollo-de-modelos-geomecanicos-3d-en-entornos-geologicos-complejos%2C-aplicado-a-la-mina-catalina-huanca%2C-aya?utm_source
- [11] Carrasco MYP. Mining 4.0: a digital transformation approach to mining sector: a peruvian case study. *Portland Int Conf Manag Eng Technol (PICMET).* 2023;1-4.
- [12] Medina-Tasilla DY, Jacobo-Barreto WJ, Mendoza-Cuti AO, Pinto-Huisacayna KL, Torres-Guerra JA, Soto-Juscamayta LM, et al. Optimizing mineral extraction in Peru: integrating geometallurgical planning with mining 4.0 technologies. *Int J Design Nature Ecodynamics.* 2024;19(4):1457-1468.
- [13] Alvia E. Inteligencia artificial en minería. *Rev Execution.* 2025 Mar 17. Available from: https://www.ey.com/es_pe/insights/revista-execution/disrupcion/inteligencia-artificial-industria-minera
- [14] Morandín AF. What is artificial intelligence? *Int J Res Publ Rev.* 2022;3(12):1947-1951.
- [15] Belcic I, Stryker C. Gpt. IBM. Accessed 2025 Sep 20. Available from: <https://www.ibm.com/mx-es/think/topics/gpt>
- [16] Lopezosa C. La inteligencia artificial generativa en la comunicación científica: retos y oportunidades. *Rev Invest Innov Cienc Salud.* 2023;5(1):1-5.
- [17] Caballar R, Stryker C. What are vision language models (VLMs)? IBM. 2025 Feb 25. Available from: https://www.ibm.com/think/topics/vision-language-models?utm_source
- [18] Merve, Beeching E. Vision language models explained. 2024 Apr 11. Available from: https://huggingface.co/blog/vlms?utm_source
- [19] Bazrafshan P, Melag K, Ebrahimkhanlou A. Semantic and lexical analysis of pre-trained vision language artificial intelligence models for automated image descriptions in civil engineering. *AI Civ Eng.* 2025;4(1).
- [20] Corso A, Karamadian D, Valentin R, Cooper M, Kochenderfer MJ. A holistic assessment of the reliability of machine learning systems. *arXiv.* 2023 Jul 20. doi:10.48550/arXiv.2307.10586.
- [21] Wang C. Calibration in deep learning: a survey of the state-of-the-art. *arXiv.* 2023 Aug 2. Available from: <https://arxiv.org/abs/2308.01222>
- [22] Agdestein I. Validación de modelos de IA: garantizar la precisión y la fiabilidad. *Focalx.* 2025 Feb 27. Available from: https://focalx.ai/es/inteligencia-artificial-es/validacion-de-modelos-de-ia-garantizar-la-precision-y-la-fiabilidad/?utm_source=chatgpt.com#elementor-toc_heading-anchor-6
- [23] Li M, Gao Q, Yu T. Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. *BMC Cancer.* 2023;23(1).
- [24] Olsen GH, Goodman ED, Aklilu JG, Bartoletti S, Hung KS, Yang JH, et al. Using artificial intelligence to model expert panel diagnosis of cholecystitis severity. *Surg Endosc.* 2025.
- [25] Vanbelle S, Hernandez Engelhart C, Blix E. A comprehensive guide to study the agreement and reliability of multi-observer ordinal data. *BMC Med Res Methodol.* 2024;24(1).
- [26] Hernández Sampieri R, Fernández Collado C, Baptista Lucio P. Metodología de la investigación. 2014. Available from: <https://dialnet.unirioja.es/servlet/libro?codigo=775008>
- [27] Müggenburg Rodríguez MC, Pérez Cabrera I. Tipos de estudio en el enfoque de investigación cuantitativa. *Enferm Univ.* 4(1):35-38.
- [28] Tarrillo Saldaña O, Mejía Huamán J, Dávila Mego JS, Chilón Camacho WM, Pintado Castillo CA, Tapia Idrogo CE, et al. Metodología de la investigación una mirada global ejemplos prácticos. 2024. doi:10.37811/cli_w1078.
- [29] Bujang MA, Omar ED, Foo DHP, Hon YK. Sample size determination for conducting a pilot study to assess reliability of a questionnaire. *Restor Dent Endod.* 2024;49(1):e3.
- [30] Totton N, Lin J, Julious S, Chowdhury M, Brand A. A review of sample sizes for UK pilot and feasibility studies on the ISRCTN registry from 2013 to 2020. *Pilot Feasibility Stud.* 2023;9(1).
- [31] Cabitza F, Campagner A, Zotti FD, Sternini F. All you need is higher accuracy? On the quest for minimum acceptable accuracy for medical artificial intelligence. 2020 Apr 1. Available from: https://www.researchgate.net/publication/348199006_All_you_need_is_higher_accuracy_On_the_quest_for_minimum_acceptable_accuracy_for_Medical_Artificial_Intelligence