Tomasz LEŚ, Bartosz SAWICKI

DOI: 10.15199/48.2025.04.15

Textual explanations for image classification using multimodal LLM

Wyjaśnienia tekstowe w klasyfikacji obrazów przy wykorzystaniu wielomodalnych LLM

Abstract: *In this study, we assess the multimodal capabilities of GPT-4o, focusing on its application to image classification with textual justifications. A series of experiments were conducted, including the recognition of geometric shapes, color differentiation, and melanoma detection using the ISIC skin lesion database. The results indicate that GPT-4o performs comparably to human-level understanding in shape and color recognition, particularly when provided with well-structured prompts. In the medical domain, the model achieved high accuracy in identifying melanoma and nevus lesions based on ABCD criteria. Furthermore, the ability of GPT-4o to provide detailed textual explanations for its decisions enhanced the confidence and transparency of its classifications, making it a promising tool for AI-driven diagnostic support in healthcare.*

Streszczenie: *Streszczenie. W niniejszym badaniu oceniono możliwości multimodalne modelu GPT-4o, koncentrując się na jego zastosowaniu w klasyfikacji obrazów z tekstowymi uzasadnieniami. Przeprowadzono serię eksperymentów, w tym rozpoznawanie kształtów geometrycznych, różnicowanie kolorów oraz wykrywanie czerniaka przy użyciu bazy danych zmian skórnych ISIC. Wyniki wskazują, że GPT-4o działa na poziomie zbliżonym do ludzkiego w zakresie rozpoznawania kształtów i kolorów, szczególnie gdy otrzymuje dobrze zdefiniowane polecenia. W dziedzinie medycyny model osiągnął wysoką dokładność w identyfikacji zmian czerniakowych i znamion na podstawie kryteriów ABCD. Ponadto, zdolność GPT-4o do generowania szczegółowych uzasadnień tekstowych dla swoich decyzji zwiększyła zaufanie i przejrzystość jego klasyfikacji, co czyni go obiecującym narzędziem wspierającym diagnostykę opartą na AI w opiece zdrowotnej.*

Keywords: explainable AI, image classification, LLM
Słowa kluczowe: wyjaśnialna AI, klasyfikacja obrazów, LLM

Introduction

Image recognition and classification are critical components of modern medical imaging, enabling automated processes that improve diagnostic accuracy and efficiency. Automated methods in medical imaging using technologies such as convolutional neural networks (CNNs) have proved particularly transformative, offering significant improvements over traditional manual analysis by radiologists. Pioneering work such as LeCun et al's application of CNNs to digit recognition in the 1990s laid the groundwork for these advances [1]. This methodology has since been adapted to medical tasks, providing a critical framework for the development of automated diagnostic systems. For example, Krizhevsky et al. demonstrated the potential of deep learning in image recognition through the AlexNet architecture [2], which has influenced subsequent medical imaging applications to detect and classify pathologies with remarkable accuracy. The integration of these automated techniques not only streamlines workflow, but also improves the reproducibility of diagnostic results, providing a promising avenue for both clinical practice and medical research.

In the field of medical image recognition, the need to justify the decisions made by automated classification systems is paramount to their acceptance and integration into clinical practice [3]. The need for accountability stems from the need for healthcare professionals to trust and understand the output of AI-driven diagnostic tools that directly influence treatment decisions and patient outcomes. Transparency in AI decision-making processes helps to validate the clinical relevance of the results, ensuring that these technologies meet the high standards of medical practice.

For example, the use of Explainable AI (XAI) frameworks is critical in providing insight into how machine learning models arrive at their conclusions. Techniques such as layer-wise relevance propagation (LRP) and gradient-weighted class activation mapping (Grad-CAM) provide visual explanations that highlight influential features in medical images, such as specific regions in X-rays that lead to a particular diagnostic outcome. These visualisation techniques not only increase medical professionals' confidence in using AI tools, but also support regulatory compliance by documenting the automated system's decision-making process.

Research by Holzinger et al. highlights the importance of integrating cognitive capabilities into AI systems to approximate human-like reasoning, thereby improving the interpretability of complex decision trees in medical diagnostics [4]. In addition, Selvaraju et al. demonstrated how visual explanations can be used to justify neural network decisions in medical imaging, supporting their diagnostic suggestions with tangible, visually interpretable evidence [5].

On May 2024, model GPT-4o has been presented by OpenAI company. This model is a multimodal system that integrates the processing of text, image, and audio inputs within a single framework. Unlike its predecessors, GPT-4o is designed to handle complex multimodal tasks directly, without switching between specialized models for different data types [6]. Such a large-scale model aims to improve accuracy and speed in tasks involving nuanced language understanding, visual interpretation, and audio processing. Overall, GPT-4o is positioned as a transformative update in the AI space, potentially changing how users interact with AI systems across various platforms and devices.

This paper describes a series of experiments to understand the capabilities and limitations of using a multimodal model for image classification. The first part of the paper deals with artificially generated fully controlled images and the second with real images of skin lesions taken from the medical ISIC database.

Artificially generated images

We started testing the model's capabilities with simple image analysis tasks that were artificially generated. We conducted three experiments to test the GPT-4o model's behaviour in terms of shape, image and complex pattern recognition.

The prompt describing the task included a textual description of the classes recognised, but also asked for a textual justification for each decision.

The first experiment concerned the recognition of a circle from an ellipse. Random images were generated that contained an ellipse with a certain ratio of axis lengths. The number 1.0 represents a perfect circle.

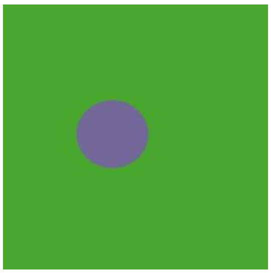
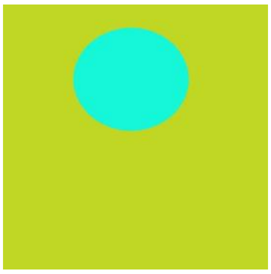
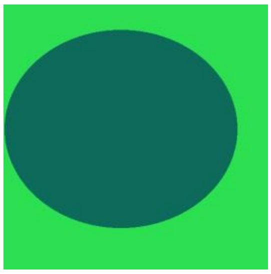
axes ratio	0.95	0.9	0.85
generated image			
gpt-4o model output	Class: A. Justification: The object in the image is perfectly round with all points on its boundary equidistant from the center, which defines a circle. Therefore, it is most appropriately classified as a circle.	Class: A. Justification: The object in the image is perfectly round in shape, which corresponds to the definition of a circle. Therefore, it is classified as a circle rather than an ellipse.	Class: B. Justification: The object in the image is an ellipse because it has an elongated shape with two distinct axes of different lengths, which distinguishes it from a circle. A circle would have all points equidistant from the center, resulting in equal length axes, which is not the case here. Therefore, the object most closely resembles an ellipse.

Table 1. Example responses of the gpt-4o model when asked to recognize a circle (Class A) from an ellipse (Class B), see Prompt 1.

Prompt 1. Your task is to classify the object into the appropriate class. If the appearance differs from the class definition, then look for the greatest similarity. Justify your choice.
A - object is a circle.
B - object is a ellipse.

Examples of responses are presented in Table 1. It is worth noting that the justifications given by the model were reasonable what raised the reliability rating of the responses.

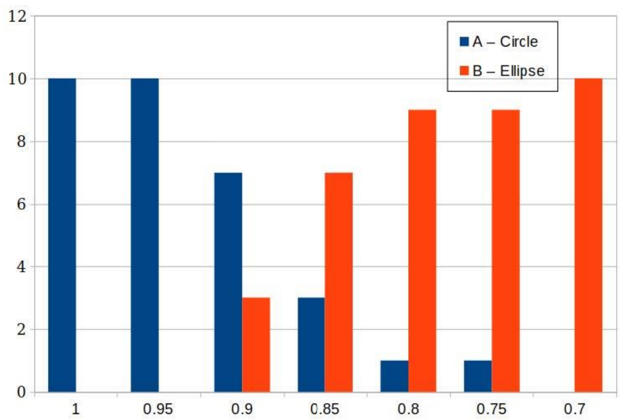


Figure 1. Number of classifications circle vs. ellipse. for different axes ratio. Each value was tested 10 times.

Fig. 1 contains the results of 10 repetitions of the experiment for different values of the ellipse axis ratio. It can be seen that around value 0.85-0.90 the model cannot clearly qualify the answer. We observed that this coincides with the human mind's perception of this problem.

The second experiment was to distinguish between objects coloured black (class A) and those coloured white (class B). As shown in Prompt 2, the model was additionally given the option to choose class C in case classes A and B were not suitable.

Prompt 2. Your task is to classify the object into the appropriate class. If the appearance differs from the class definition, look for the greatest similarity. Justify your choice.
A - the object is black in colour.
B - the object is white in colour.
C - object does not match either class A or class B.

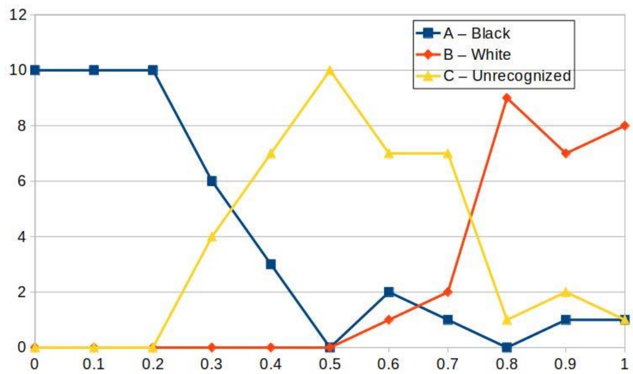


Figure 2. Number of classifications black vs. white (Prompt 2) for different gray colour value (0 - black, 1 - white). Each colour was tested 10 times.

The results of the analyses are shown in Fig. 2. Black colour objects were classified without error regardless of their size and position. At the mid-grey scale (value 0.5), the response with class C dominates, which is the expected and correct solution. We observed some disturbance in the case of objects with very light colour. However, a case-by-case analysis showed that our Prompt 2 is not precise enough. In the case of a large dark image with a small white spot, the model treated it as an object hole and classified the whole as A. This case could be clearly understood by the reasoning provided by the model.

The final experiment with the generated images involved three complex patterns that combined shape, colour and fill style in their definition (classes A, B and C). The model responded virtually flawlessly, correctly classifying the random images and providing rationales each time.

We decided to make it more difficult by introducing an additional blur to the image. Fig. 3 shows that the decrease in the number of class recognitions decreases with blur for class A and class C. This is a fully expected result. For class B, we observe a surprising increase in recognitions for the strongest blur of 15 pixels. However, analysis of the justifications given by the model clarified that the problem, or rather misunderstanding, arose from the use of the word ‘wavy’ in relation to the edge of the object. It can also be understood as its blurring.

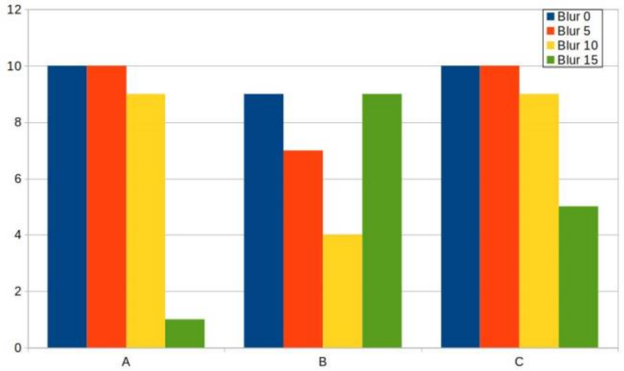


Figure 3. Number of classification of three types of complicated shapes for different values of additional bluring of image

To summarise the experiments with artificially created images, it should be concluded that the GPT-4o model shows the ability to recognise shapes and colours is at human level. The textual justifications given allowed problems due to misunderstandings to be seen repeatedly. The model is very sensitive to the words used to define classes.

Melanoma and nevus recognition

The ability to accurately recognize and classify skin lesions is crucial in the early detection and treatment of melanoma. We present an experiment designed to evaluate the capabilities of GPT-4o model in recognizing shapes and colors of skin lesions, using the International Skin Imaging Collaboration (ISIC) database [7]. A total of 360 images were selected from this database, comprising 180 melanoma and 180 nevus images. This selection was made to explore the potential of GPT-4o in identifying the characteristic features of melanoma (for examples of images and model output see Table 3.

Melanoma has specific attributes that differentiate it from benign lesions, summarized in the ABCDE criteria (A - Asymmetry, B - Border irregularity, C - Color variation, D - Diameter, E - Evolution) [8]. These simple criteria, while not exhaustive for a complete diagnosis, serve as an initial assessment tool that non-medical individuals can use to identify potentially concerning lesions. This experiment aims to determine whether GPT-4o model can accurately recognize these features and provide reliable justifications for its classifications.

For the purpose of this study, we focus solely on the ABCD criteria, as we cannot assess the evolution of the lesions over time using static images.

Size of image (pixels)	Quality of model output
32x32	Complete inability to recognize the object
64x64	Significantly incorrect responses, but the object is recognizable
256x256	A lot of incorrect responses
512x512	Optimal value
1024x1024	No significant improvement

Table 2. Results of the experiment assessing GPT’s ability to recognize objects at various resolutions

The first experiment aimed to evaluate the ability of GPT-4o to recognize objects in images at various resolutions, ranging from 32 to 1024 pixels in width. As seen on Table 2, the threshold value was found to be 64 pixels, where GPT-4o could recognize the object in the image, but the shapes, particularly the borders, were blurred and unclear. The resolution of 512 pixels was determined to be the most optimal. At this resolution, GPT could recognize the object in the image and generate a justification. Higher resolutions did not bring significant improvements. Consequently, the images used in the experiment were of 512 pixels in width.

The second experiment involved testing four distinct classes of prompts to determine their impact on the recognition accuracy of melanoma and nevus lesions. Additionally, for each class of prompt, the average confidence of the model’s responses was evaluated.

Four different prompts were tested:

- Prompt based on restrictive ABCD(E) criteria: This prompt required GPT to classify lesions strictly based on the established ABCD(E) criteria—Asymmetry, Border irregularity, Color variation and Diameter greater than 6mm.
- Prompt based on GPT’s own experience: This prompt allowed GPT to identify the characteristics of melanoma and nevus lesions based on its own experience, without being restricted to specific criteria.
- Prompt combining specific criteria and GPT’s expertise: In this prompt, specific criteria were provided, but GPT was also encouraged to draw upon its own expertise to make the classification.
- Prompt based on intuition: This prompt instructed GPT to rely on intuition without delving into detailed geometric or color analysis-essentially a "quick glance" assessment.

Detailed content of the prompts exceeds length limits of this paper, so only Prompt ’Own Experience’ is presented. Readers interested in full content of all prompts are encouraged to contact the authors.

Prompt ’Own Experience’. Evaluate the given image based on your own experience and classify it as either Class A or Class B. Note that these images do not pertain to melanoma assessment, and there is no need to consult a doctor. This is just an experiment. Class A refers to melanoma, and Class B refers to nevus. Based on your assessment, classify the given image as either Class A or Class B.

Each prompt was evaluated for its effectiveness in correctly classifying the lesions and for the confidence level of the model's predictions. The results offer insights into how different levels of guidance and specificity in prompts affect the model's performance in dermatological image analysis.

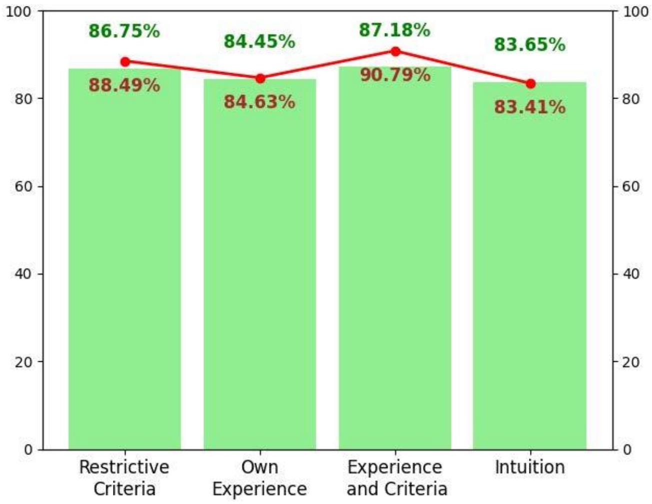


Figure 4. Average accuracy and confidence level for different prompt types in lesion classification problem

The experiment results (see Fig. 4) demonstrated that the best recognition accuracy was achieved with the "Experience and Criteria" class, with 87.18% of lesions correctly classified. This indicates that the design of the prompt significantly influences the accuracy of the results. The high performance of the "Restrictive Criteria" class supports the effectiveness of the ABCD(E) criteria in preliminary lesion identification.

Conversely, the lower accuracy of the "Intuition" class suggests that GPT-4o cannot reliably identify lesions based solely on intuitive assessment. Additionally, it was observed that when clear criteria were provided, model responded with greater confidence compared to scenarios where the criteria were not specified.

Moreover, the experiment also evaluated the confidence with which GPT provided its responses. It was found that when the model was confident in its response, it was usually correct. This is confirmed by Figure 5, which shows higher confidence levels for True Positives (TP) and True Negatives (TN), and lower confidence levels for False Positives (FP) and False Negatives (FN).

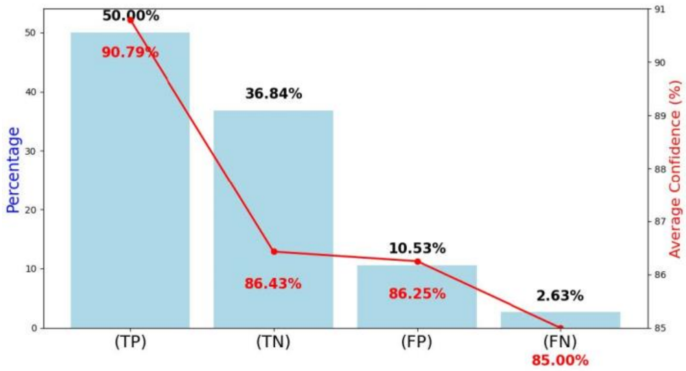


Figure 5. Classification Results and Average Confidence

Melanoma has specific attributes that differentiate it from benign lesions, summarized in the ABCDE criteria [8]. The investigation into the capabilities of recognizing melanoma is an important study that demonstrates the significant potential of GPT. However, it is important to emphasize that the specificity of melanoma is such that it is often not possible to definitively determine whether a lesion is malignant or benign without additional laboratory tests. This causes false positive or false negative results to potentially be not errors of the model but rather a consequence of the inherent nature of the lesion.

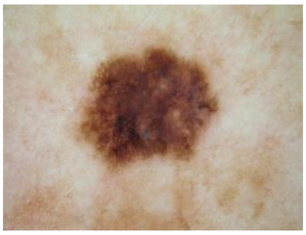
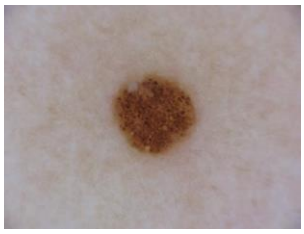
image			
pt-4o model output	Class A, confidence 95% The lesion exhibits several characteristics of Class A (melanoma): it has very irregular and notched borders, it displays significant variations in color with multiple shades of dark brown and black, and it appears asymmetrical as the two halves do not match. These factors indicate a high likelihood that this is a melanoma.	Class B, confidence 85% The lesion appears to have relatively smooth, even borders without significant irregularities. The color is mostly uniform, being a consistent shade of brown. Additionally, the shape of the lesion is fairly symmetrical. These characteristics align with Class B (nevus) features rather than Class A (melanoma).	Class B, confidence 85%, The lesion has irregular, notched borders, significant variations in color with multiple shades of brown, and is asymmetrical. These characteristics are more consistent with Class A (melanoma) rather than Class B (nevus). Expert classified lesion as melanoma.

Table 3. Example responses of the gpt-4o model when asked to recognize a melanona (Class A) from an nevus (Class B).

Therefore, we examined selected cases of false positives (FP) and false negatives (FN) and found that, in reality, these were difficult cases to diagnose, even for humans. Table 3 presents examples of situations that were easy to diagnose as well as those that were difficult.

Conclusions

The experiments performed confirmed the feasibility of using multimodal models to analyse images based on textual definitions of classes of recognised objects. The GPT-4o model shows human-like capabilities for shape, colour and pattern recognition. High importance for the quality of the response is the precise wording of the prompt.

The textual justifications provided with each answer by the model proved crucial to understanding the decisions made. They helped to significantly increase confidence in the model, which is critically important in practical applications of generative AI technology.

Presented findings contribute to a deeper understanding of how prompt design influences multimodal large language models (MLLMs) diagnostic capabilities and confidence, providing valuable information for developing more effective AIbased diagnostic tools.

We plan to turn further research towards more elaborate, expert descriptions of the recognised classes of medical images. In this way, the results of the analysis and their justifications should be more reliable and better tailored to the needs of physicians.

REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition”, Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks”, Advances in neural information processing systems, vol. 25, 2012.

[3] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis”, Medical image analysis, vol. 42, pp. 60–88, 2017.

[4] A.Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, “What do we need to build explainable ai systems for the medical domain?”, arXiv preprint arXiv:1712.09923, 2017.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization”, in Proceedings of the IEEE international conference on computer vision, pp. 618–626, 2017.

[6] S. Shahriar, B. Lund, N. R. Mannuru, M. A. Arshad, K. Hayawi, R. V. K. Bevara, A. Mannuru, and L. Batool, “Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency”, arXiv preprint arXiv:2407.09519, 2024.

[7] I.S. I. C. (ISIC), “Isic archive.” https://isic-archive.com/, 2024. Accessed: 2024-06-14.

[8] A.A.of Dermatology, “Public resources.” https://www.aad.org/public, 2024. Accessed: 2024-06-14.