GPT-4V exhibits human-like performance in biomedical image classification

Wenpin Hou[1,†] **and Zhicheng Ji**[2,†]

[1]Department of Biostatistics, The Mailman School of Public Health, Columbia University, New York City, NY, USA
[2]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA.
[†]Corresponding author. E-mail: wh2526@cumc.columbia.edu; zhicheng.ji@duke.edu

ABSTRACT

We demonstrate that GPT-4V(ision), a large multimodal model, exhibits strong one-shot learning ability, generalizability, and natural language interpretability in various biomedical image classification tasks, including classifying cell types, tissues, cell states, and disease status. Such features resemble human-like performance and distinguish GPT-4V from conventional image classification methods, which typically require large cohorts of training data and lack interpretability.

Main

The discourse surrounding human cognition compared to machine capability has long been a focal point of both scientific and philosophical exploration[1]. A salient feature distinguishing human cognition is the capacity to recognize and generalize patterns from limited data. Machines, particularly those grounded in deep learning (DL), traditionally necessitate extensive training sets for comparable proficiency[2]. The question arises: what engenders this differential in data efficiency? One hypothesis centers on the human capability for multimodal sensory processing[3,4]. Humans integrate diverse sensory inputs, such as visual, auditory, and tactile inputs, with higher cognitive functions, notably language-based reasoning[5,6]. This ability facilitates comprehensive data abstraction and generalization. Conversely, many conventional machine learning (ML) and DL methodologies were initially designed for unimodal tasks, focusing on individual data types such as images or textual information[7–10]. Language emerges as a critical factor in this differentiation. As a mechanism to structure and abstract information, language enables intricate data abstraction. For instance, when analyzing a tissue sample, a descriptor like "aggregated" encapsulates a spectrum of human understanding. Traditionally, machines operating on deterministic algorithms lacked this depth of linguistic interpretation. In comparison, GPT-4V(ision), a large multimodal model evolving from the GPT-4 model, is able to capture such comprehensive interpretation in language form, for example “clustered or aggregated distribution" (Fig. 1a).

This comprehensive understanding becomes particularly critical when considering biomedical imaging classification. The dependency on vast datasets presents challenges in sample sizes, labor-intensive labeling, and risks of over-reliance, especially for rare diseases. It is in this setting that GPT-4V illuminates the scene. GPT-4V is distinctive due to its integration of natural language reasoning, mirroring aspects of human cognitive processing. This transition from standard algorithmic processing to natural language-centric reasoning positions GPT-4V in a novel domain, distinct from its predecessors. Methodologically, three attributes underscore this advancement. The first is interpretability. Beyond mere classification, GPT-4V offers rationale in human-accessible language. Although the intrinsic interpretability is not yet accessible, such *post hoc* interpretability has made a significant leap[11]. In critical fields such as medicine, the rationale is often as pivotal as the result itself. The second is data efficiency. While conventional DL models demand extensive training data, which are often challenging to obtain, GPT-4V demonstrates effectiveness with limited reference images even as few as one, a concept known as *one-shot learning* that has been demonstrated in humans[12]. This can substantially benefit applications in medical diagnostics and rare diseases where reference data are scarce. The third is generalizability. GPT-4V generalizes better to testing data less consistent with training data, such as data with different resolutions or from different studies, compared to existing deep learning methods, offering high flexibility in various settings. Additionally, while existing deep learning methods often require significant machine learning and coding expertise, GPT-4V’s user-friendly interface substantially lowers the barrier to performing image classification, democratizing artificial intelligence usage in biomedical image analysis.

In this study, we systematically evaluated GPT-4V’s performance on four biomedical image classification tasks (Fig. 2a). These tasks include classifying two tissues of artery and tibial nerve using hematoxylin and eosin stain (H&E) images, classifying two cell types of microglial (BV-2 cell line) and neuroblastoma (SH-SY5Y cell line) using light microscopy images, classifying two cell states of normal and senescent cells using DAPI fluorescent stain images, and classifying two disease

conditions of normal retina and retina with diabetic macular edema (DME) using optical coherence tomography (OCT) images. To test generalizability, we created two scenarios for each task, where training and testing data are either highly consistent or less consistent (Methods). We evaluated GPT-4V and ten competing image classification methods for one-shot learning, with one image provided for each image category as training data. We used a transfer learning approach for the ten competing methods, including AlexNet[7], DenseNet[13], EfficientNetV2[14], GoogLeNet[15], Inception V3[16], ResNet[17], ResNeXt[18], SwinTransformer[19], VGG[20], and VisionTransformer[21]. These competing models were all pretrained with ImageNet1K[22] and then fine-tuned with the training data provided in this study (Methods).

We developed a prompting strategy for GPT-4V image classification (Fig. 1b). We first created a reference image array with two rows and two columns. The two rows represent two image categories. The first column indicates “Row 1” and “Row 2”. The second column contains two training images, one for each category. The reference image array can be easily created with PowerPoint, Keynote, or other similar software. A dictionary of tissues, cell types, or disease conditions corresponding to the two rows is stored elsewhere and hidden from GPT-4V. When given a testing image, GPT-4V was asked to identify which row in the reference image array most resembles the testing image. One can then refer to the dictionary to determine the actual category. This strategy offers several advantages. First, additional rows and columns can be added to represent more image categories or additional training images within each category, thus avoiding the restriction of a maximum number of input image files allowed by GPT-4V (at most 10 image files as of Dec 25, 2023). Second, compared to providing independent image files for each category, a reference image array simplifies communication for both the user and GPT-4V. Third, sensitive information like patient data can be stored in the separate dictionary for increased confidentiality, which is important in clinical settings. This design also minimizes the chance that GPT-4V learns from its prior knowledge of image categories (e.g., cell type names), ensuring a fair comparison with other methods.

We compared the accuracy of GPT-4V with that of ten competing methods (Fig. 2b). Accuracy is defined as the percentage of testing images correctly categorized by each method (Methods). The accuracy and responses of GPT-4V are included in Supplementary Table 1. GPT-4V has the best performance, reaching nearly 100% accuracy in all scenarios. In comparison, the performance of competing methods drop substantially when the training data are less consistent with the testing data. For example, none of the competing methods achieved an accuracy greater than 70% in tissue and cell type classifications when training and testing data are less consistent. In addition, the competing methods also perform poorly in disease diagnostics where the two image categories are less distinct. SwinTransformer, which is the best among the ten competing methods, reaches 72% and 80% accuracy, substantially lower than the 99% and 96% accuracy of GPT-4V. These results suggest that GPT-4V has substantially stronger one-shot learning ability and generalizability compared to the conventional image classification methods.

Beyond accuracy, a distinct feature that makes GPT-4V stand out is its text output interpreting how conclusions are drawn. To demonstrate its rationality, we summarized the keywords in GPT-4V’s responses describing differences between two image categories, and their enrichment is shown as word clouds (Fig. 2c, Methods). The keywords align well with features distinguishing the two image categories. For example, "dense" and "homogeneous" describe the image features of an artery, while "round" and "circular" describe those of a tibial nerve in tissue classification. In disease diagnostics, "uniform" and "smooth" describe image features of a normal retina, and "peak" and "irregular" those of DME. These results highlight GPT-4V’s interpretability to explain the reasoning behind image classification, an ability not seen in other methods. In addition, keywords given by GPT-4V are highly consistent regardless of the consistency between training and testing data in each classification task, suggesting that GPT-4V is able to capture the most relevant information in image classification tasks, a potential reason for its better generalizability. In comparison, other image classification methods could be mislead by irrelevant information in the training data that cannot be transferred to testing data.

To further quantify GPT-4V’s interpretability, we designed a unique text-based image classification method that uses GPT-4V’s text-based output from a previous conversation as training information in a completely new conversation (Fig. 1c-d). First, GPT-4V was asked to summarize the main differences between the two image categories in the reference image array (Fig. 1c). Next, in a completely new conversation window, GPT-4V was queried to identify the image category of a testing image using this summarized text information (Fig. 1d). In the new conversation window, GPT-4V does not have access to the training images when performing image classification on the testing image. The natural language that GPT-4V itself used to differentiate image categories serves as the sole information carrier of the training images. The performance, labeled as “GPT-4V (text)” in Fig. 2b, is strikingly better than most competing methods, reaching a reliable accuracy of at least 89% in all tasks. These results demonstrate GPT-4V’s ability to convert imaging features into concepts in natural language that can be further applied to classifying testing images, even in a one-shot learning setting. Additionally, GPT-4V’s ability to utilize natural language in the classification process paves the way for widespread applications, potentially in clinical settings.

In summary, our study illustrates that GPT-4V leverages natural language as a foundational component for reasoning in imaging classification, mirroring human cognitive processes. This feature substantially enhances GPT-4V’s capabilities in one-shot learning, generalizability, interpretability, and text-based image classification, surpassing conventional image classification methods. These findings indicate that large multimodal models, which integrate natural language with visual data,

have potential advantages over unimodal methods that rely solely on images. This insight could inform future research directions in computer vision, highlighting the benefits of combining linguistic and visual data. Moreover, the text-based classification approach used by GPT-4V suggests the possibility of developing new image analysis methods that utilize language-based communication between artificial intelligence systems. This approach could offer a more comprehensive way of interpreting and interacting with complex visual data, potentially enhancing the utility and applicability of AI in various domains that can bridge the gap between human interpretative skills and machine efficiency.

Methods

Imaging data

For tissue classification, H&E images of artery and tibial nerve tissues were downloaded from the GTEx portal[23]. To create a scenario where training and testing images are highly consistent, we collected fifteen images of artery from different randomly chosen samples and fifteen images of tibial nerve from different randomly chosen samples. We randomly zoomed in on each of the original image provided by GTEx to display a part of the tissue, and all images have the same zoom level. We created five sets of training images, each set containing one artery image and one tibial nerve image. All the remaining images were used for testing. To create a scenario where training and testing images are less consistent, we further zoomed in on each image of the same training set to display a smaller region of the tissue. The testing set remained the same.

For cell type classification, light microscopy images of BV-2 and SH-SY5Y cell lines were downloaded from the LIVECell dataset[24]. To create a scenario where training and testing images are highly consistent, fifteen images from each of the two cell lines were randomly chosen. Five images from each cell line were used to create five sets of training images, and the remaining images were used for testing. To create a scenario where training and testing images are less consistent, we downloaded light microscopy images of BV-2 and SH-SY5Y from two additional studies[25,26]. For each image, we randomly cropped five different regions of the original image to obtain five new images for training. The testing set remained the same.

For cell state classification, DAPI fluorescence images were downloaded from a previous study[27]. To create a scenario where training and testing images are highly consistent, for each cell state of normal or senescence, fifteen images of single cells were randomly cropped from the original images in the study. Five images from each cell state were used to create five sets of training images, and the remaining images were used for testing. To create a scenario where training and testing images are less consistent, we downloaded DAPI fluorescence images from another study[28]. For each cell state, five images of single cells were randomly cropped from the original images and five new sets of training images were created. The testing set remained the same.

For disease diagnostics, OCT images were downloaded from a previous study[29]. To create a scenario where training and testing images are highly consistent, fifteen images were randomly selected for each disease status. Five images for each disease status were used for training and the remaining images were used for testing. To create a scenario where training and testing images are less consistent, we downloaded OCT images from another study[30]. For each disease status, five images were randomly selected for creating five sets of training images. The testing set remained the same.

Image classification with GPT-4V

GPT-4V was performed with the `gpt-4-1106-vision-preview` model via the `openai` python package (version 1.3.5) provided by OpenAI.

For GPT-4V image-based classification, GPT-4V was given a reference image array consisting of two training images as the first image, the testing image as the second image, and queried with the following prompt message for tissue classification:

"There are 2 rows in total for the first image. Each row represents example images from a tissue. The second image is another image of a tissue. Which row in the first image is more similar to the second image and why?"

For cell type classification, cell state classification, and disease diagnostics, the word "tissue" was replaced with "cell type", "cell state", and "status of a tissue" respectively in the previous prompt message.

For GPT-4V text-based classification, GPT-4V was first given a reference image array consisting of two training images and queried with the following prompt message:

"What are the most distinct features differentiating the two rows of images? Provide detailed explanations."

The response by GPT-4V was recorded, and attached to the bottom of the following sentence to form a new prompt message.

"Is the given image more consistent with row 1 or row 2 based on the descriptions below?"

In a completely new conversation window, GPT-4V was then given the testing image and queried with this new prompt message.

The responses by GPT-4V for both image-based and text-based classifications were recorded and manually evaluated for correctness.

Competing methods for image classification

The Python `torchvision` package[31] (version 0.14.1), part of the `PyTorch` project[32], was used to implement the ten competing methods for image classification with pre-trained weights. Specifically, the following models and their pre-trained weights were used: `alexnet` model with `AlexNet_Weights`, `densenet201` model with `DenseNet201_Weights`, `efficientnet_v2_m` model with `EfficientNet_V2_M_Weights`, `googlenet` model with `GoogLeNet_Weights`, `inception_v3` model with `Inception_V3_Weights`, `resnet50` model with `ResNet50_Weights`, `resnext101_32x8d` model with `ResNeXt101_32X8D_Weights`, `swin_b` model with `Swin_B_Weights`, `vgg16` model with `VGG16_Weights`, and `vit_b_32` model with `ViT_B_32_Weights`. The pretrained models were fine-tuned with training images provided in this study with a cross entropy loss, Adam optimizer[33], and a learning rate of 0.0001. The model training was stopped when the training loss is smaller than 0.00001 or when the training exceeds 50 epochs, whichever happens first.

Performance evaluation

For an image classification task, an image classification method was trained on each of the five reference image arrays and tested on each of the 20 testing images, resulting in 100 predicted results. The accuracy is defined as the percentage of the 100 predicted results that agree with the correct answers.

More formally, denote p_{ij} as the predicted binary label when a method was trained on ith ($i = 1,2,3,4,5$) reference image array and tested on jth ($j = 1,...,20$) testing image. Denote l_j as the true binary label for jth testing image. The accuracy is defined as $\sum_{i,j}(p_{ij} = l_j)/100$. The R package `ggplot2` (v.3.3.0)[34] was used for visualizing the performance.

GPT-4V word cloud

For each classification task, we calculated the number of times each word occur in the 100 responses by GPT-4V, after removing special characters. A word cloud is created using the `wordcloud` R package (version 2.6)[35] for words that occur in at least 10% of the 100 GPT-4V responses.

Acknowledgments

Z.J. was supported by the National Institutes of Health under Award Number 1U54AG075936-01. W.H. was supported by the General Fund at Columbia University Department of Biostatistics and by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R35GM150887. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

Both authors conceived the study, conducted the analysis, and wrote the manuscript.

Competing interests

All authors declare no competing interests.

Data availability

H&E images for tissue classification were downloaded from the GTEx Histology Viewer of the GTEx portal[23] (https://gtexportal.org/home/histologyPage). Light microscopy images for cell type classification were downloaded from previous studies[24–26]. DAPI fluorescence images for cell state classification were downloaded from previous studies[27,28]. OCT images for disease diagnostics were downloaded from previous studies[29,30]. Responses of GPT-4V in this study are available as supplementary materials. All training and testing images are available at Figshare (https://figshare.com/articles/figure/Images_for_assessing_GPT-4V_s_performance_in_image_classification/24921090).

Code availability

All codes to reproduce the presented analyses are publicly available in Github repository https://github.com/Winnie09/gptimage.

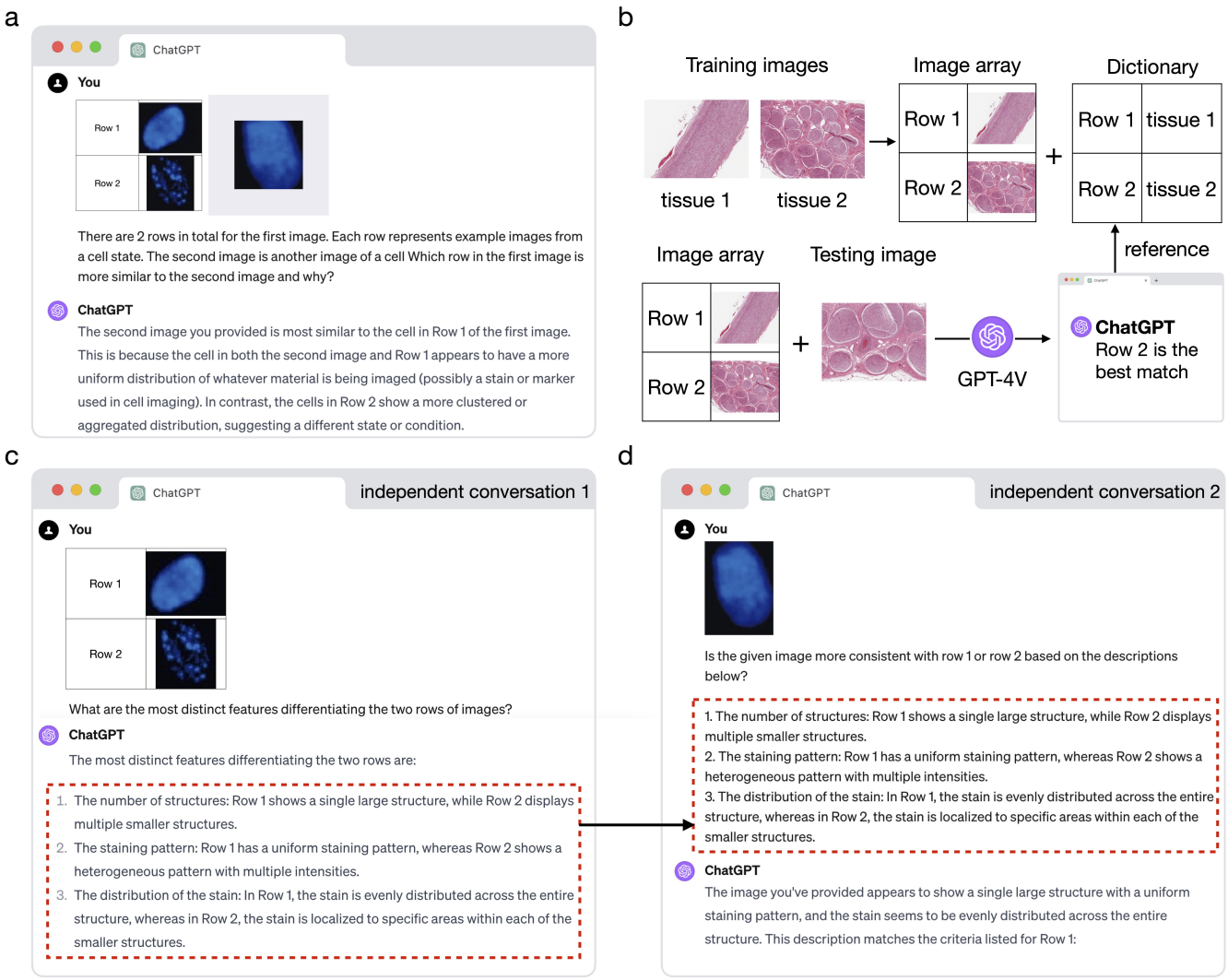

Figure 1. Demonstration of GPT-4V image classification. **a**, An example conversation with GPT-4V to perform image-based image classification. **b**, A schematic of prompting strategy for GPT-4V image-based image classification. **c, d**, An example process with GPT-4V to perform text-based image classification. The process consists of two independent conversations for obtaining the text description of training image's distinct features (**c**) and querying GPT-4V for classifying a testing image (**d**).

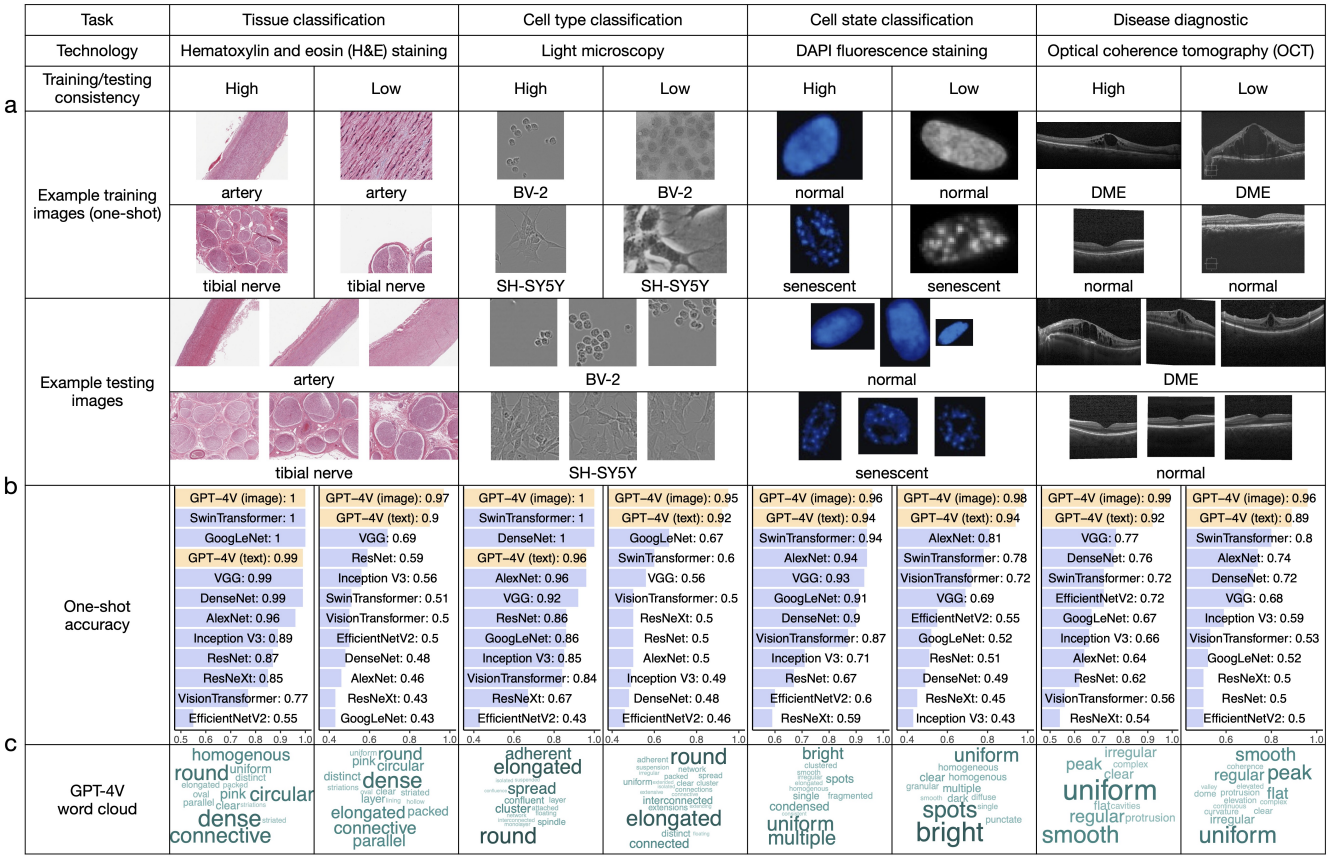

Figure 2. Performance evaluation of GPT-4V and ten competing methods. **a**, Example training and testing images for four image classification tasks. **b**, Accuracy of one-shot learning for GPT-4V and ten competing methods. "GPT-4V (image)" indicates image-based image classification with GPT-4V and "GPT-4V (text)" indicates text-based image classification with GPT-4V. **c**, Word cloud showing keywords in GPT-4V's responses.

References

1. Siemens, G. *et al.* Human and artificial cognition. *Comput. Educ. Artif. Intell.* **3**, 100107 (2022).
2. Wang, F., Kaushal, R. & Khullar, D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? (2020).
3. Meredith, M. A. & Stein, B. E. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *J. neurophysiology* **56**, 640–662 (1986).
4. Foxe, J. J. *et al.* Multisensory auditory–somatosensory interactions in early cortical processing revealed by high-density electrical mapping. *Cogn. Brain Res.* **10**, 77–83 (2000).
5. Anderson, J. R. *The architecture of cognition*, vol. 5 (Psychology Press, 1996).
6. Anderson, J. R. *Rules of the mind* (Psychology Press, 2014).
7. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. neural information processing systems* **25** (2012).
8. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–1780 (1997).
9. Vaswani, A. *et al.* Attention is all you need. *Adv. neural information processing systems* **30** (2017).
10. Graves, A., Mohamed, A.-r. & Hinton, G. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, 6645–6649 (Ieee, 2013).
11. Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W. & Wallach, H. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–52 (2021).
12. Fei-Fei, L., Fergus, R. & Perona, P. One-shot learning of object categories. *IEEE transactions on pattern analysis machine intelligence* **28**, 594–611 (2006).
13. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).
14. Tan, M. & Le, Q. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, 10096–10106 (PMLR, 2021).
15. Szegedy, C. *et al.* Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9 (2015).
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826 (2016).
17. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
18. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500 (2017).
19. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).
20. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
21. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
22. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).
23. GTEx Consortium. The Genotype-Tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
24. Edlund, C. *et al.* LIVECell-A large-scale dataset for label-free live cell segmentation. *Nat. Methods* **18**, 1038–1045 (2021).
25. Saleppico, S. *et al.* Iron regulates microglial cell-mediated secretory and effector functions. *Cell. Immunol.* **170**, 251–259 (1996).
26. Ross, R. A., Spengler, B. A. & Biedler, J. L. Coordinate morphological and biochemical interconversion of human neuroblastoma cells. *J. Natl. Cancer Inst.* **71**, 741–747 (1983).

27. Narita, M. *et al.* Rb-mediated heterochromatin formation and silencing of E2F target genes during cellular senescence. *Cell* **113**, 703–716 (2003).

28. Nelson, D. M. *et al.* Mapping H4K20me3 onto the chromatin landscape of senescent cells indicates a function in control of cell senescence and tumor suppression through preservation of genetic and epigenetic stability. *Genome Biol.* **17**, 158 (2016).

29. Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131.e9 (2018).

30. Gholami, P., Roy, P., Parthasarathy, M. K. & Lakshminarayanan, V. Octid: Optical coherence tomography image database. *Comput. Electr. Eng.* **81**, 106532 (2020).

31. Marcel, S. & Rodriguez, Y. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, 1485–1488 (2010).

32. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. neural information processing systems* **32** (2019).

33. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

34. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. Use R! (Springer International Publishing, Switzerland, 2016).

35. Oesper, L., Merico, D., Isserlin, R. & Bader, G. D. Wordcloud: a cytoscape plugin to create a visual semantic summary of networks. *Source code for biology medicine* **6**, 7 (2011).