nature portfolio

Corresponding author(s): Xuhai Xu, Marzyeh Ghassemi

Last updated by author(s): May 12, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For the data collection for ML model training, we merged 16 datasets (15 public access, 1 private access, see the list in Stable. 4). Using these datasets, we trained ML models as described in the Method section. As for the data collection for the human-AI collaboration study, we adopted Qualtrics as the platform for both human-subject studies.
Data analysis	We used python (specifically PyTorch 1.13) to write custom code to train ML models. For the human subject study results, we used python and Jupyter Notebook for data analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

For ML model training, our datasets are organized and available in this link. Its description is available in the README file in the Figshare subfolder of ml-model (https://figshare.com/s/eaf2110da509abbe9a6a). For the two human-subject studies, the images used in the human experiment and the study results are organized and available in the Figshare subfolder of human-study-results-analysis (https://figshare.com/s/eaf2110da509abbe9a6a).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Study 1 included 623 participants from the general public, consisting of 314 females, 305 males, and 4 who identified as non-binary or other. Study 2 involved 153 primary care physicians (PCPs), of whom 66 were female, 84 were male, and 3 identified as non-binary or other. An additional cohort of 320 medical students was recruited, which included 171 females, 141 males, and 8 who identified as non-binary or other. While these data were collected for demographic purposes, sex- and gender-based analyses were not the focus of this study, and therefore were not performed. The primary analyses centered on the effects of medical expertise, XAI methods, and human-AI decision paradigms on diagnostic performance.
Reporting on race, ethnicity, or other socially relevant groupings	Participant race and ethnicity were collected via self-report. The distribution of participants is as follows: - Study 1 (General Public, N=623): White (402), American Indian or Alaska Native (61), Black or African American (45), Asian (35), Hispanic or Latino or Spanish Origin (26), Native Hawaiian or Other Pacific Islander (14), Middle Eastern or North African (4), and other (10). 18 participants identified as multi-racial, and 8 preferred not to answer. - Study 2 (PCPs, N=153): Asian (50), White (46), Black or African American (19), Middle Eastern or North African (14), Hispanic or Latino or Spanish Origin (7), American Indian or Alaska Native (3), and other (5). 6 participants identified as multi-racial, and 3 preferred not to answer. - Study 2 (Medical Students, N=320): Asian (129), Black or African American (70), White (46), Middle Eastern or North African (25), Hispanic or Latino or Spanish Origin (9), American Indian or Alaska Native (1), and other (11). 17 participants identified as multi-racial, and 12 preferred not to answer. The primary socially relevant variable analyzed in this study was patient skin tone within the clinical images (not the human subjects), used to investigate diagnostic disparities and the fairness of AI assistance. Human subjects' info was collected solely to describe cohort diversity and to assess potential performance equity; they were not used as proxies for socioeconomic status or other constructs.
Population characteristics	See above.
Recruitment	Participants in Study 1 (general public) were recruited from the online platform Prolific. Participants in Study 2 (PCPs and medical students) were recruited from several clinical networks and platforms, including local networks at Boston Medical Center and Stanford University, as well as Centaur Labs, MedShr, Pathway, and XPC. This recruitment strategy may introduce a self-selection bias. Participants from online platforms like Prolific may be more digitally literate than the general population. Similarly, clinicians who opted into the study through professional networks may have a pre-existing interest in artificial intelligence or digital health technologies. This potential bias could mean that the observed effects of AI collaboration might be more pronounced than in a broader, less technologically inclined population.
Ethics oversight	The study protocol was reviewed and approved by the Massachusetts Institute of Technology's (MIT) Committee on the Use of Humans as Experimental Subjects as Exempt Category 3—Benign Behavioral Intervention. All participants were informed about the study's procedures and the nature of the visual content, and all provided informed consent before beginning the experiment. Participants were explicitly informed that they could withdraw from the study at any time.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This was a quantitative, experimental study. It consisted of two large-scale, digital experiments that used a between-subject factorial design (4 XAI methods x 2 decision paradigms) to assess human-AI collaborative diagnostic performance. Data collected included quantitative measures of diagnostic accuracy (binary and free-text), decision confidence ratings, and responses to validated psychometric questionnaires.
Research sample	The study involved two primary research samples chosen to represent different levels of medical expertise: Study 1: 623 participants from the general public with no medical background; Study 2: 153 PCPs and an additional cohort of 320 medical students. The demographic details of these cohorts can be found in the materials above. The samples were collected across diverse populations from online platforms and clinical networks. The rationale for choosing these distinct groups was to systematically investigate how diagnostic performance with AI assistance varies across different levels of expertise.
Sampling strategy	Recruitment was convenience sampling: Prolific's pre-screened panels and professional e-mail/list-serv networks for clinicians and students. For Study 1, an a priori power analysis was conducted to determine a sufficient sample size for detecting differences in diagnostic accuracy across the four XAI explanation types (Basic, GradCAM, CBIR, LLM). The analysis was based on a one-way, between-subjects ANOVA with four groups. The significance level was set at $\alpha = 0.05$, with a target power of 0.80. Based on prior literature, we estimated a small to medium effect size, Cohen's $f = 0.15$, to reflect these modest expected differences in a general population. The analysis indicated that a total sample size of at least 492 participants would be required to detect an effect of this magnitude under the specified parameters. Therefore, our final sample of 623 participants was sufficient to detect the hypothesized differences between XAI conditions. We conducted a similar power analysis for Study 2. We estimated Cohen's $f = 0.30$ for PCPs and 0.20 for medical students -- we hypothesized that the utility of different explanation types could lead to more pronounced differences in performance compared to the general public. This leads to a total sample size of at least 128 for PCPs and 280 for students would be required. Our final samples were considered sufficient.
Data collection	All sessions ran in a custom Qualtrics interface. Participants reviewed 12 clinical images of human skin, entered diagnoses and confidence, viewed AI outputs/explanations according to assignment, and repeated the decision once. Responses and timings were recorded automatically; no experimenter was present, and researchers were blind to condition during data collection. Attention-check items and per-image time windows (10 s – 5 min) enforced data quality.
Timing	Study 1 was conducted during March 20 – May 01, 2024; Study 2 was conducted during June 10 - Sept 3, 2024.
Data exclusions	Data were excluded from the final analysis based on pre-established quality control criteria. The study embedded attention check questions, and participants who failed them were filtered out. Additionally, responses with abnormal durations (averaging less than 10 seconds or more than 5 minutes per image) were excluded. In Study 1, 237 participants were filtered and excluded. In Study 2, 48 PCPs and 89 medical students were excluded. Moreover, 289 participants were excluded due to their medical jobs (not PCPs or students).
Non-participation	Participants who didn't complete the Qualtrics survey were not recorded.
Randomization	Participants were randomly allocated to experimental groups. Upon entering the study, each participant was randomly assigned to one of eight conditions, representing a factorial combination of four different XAI methods (Basic, GradCAM, CBIR, or LLM) and two human-AI decision paradigms (Human-First or AI-First). Within each participant, the 12 images were presented in random order, and AI correctness was preset but balanced across skin tones to avoid systematic covariates.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.