nature biomedical engineering

Article

https://doi.org/10.1038/s41551-023-01045-x

A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics

Received: 31 August 2022

Accepted: 26 April 2023

Published online: 12 June 2023

Check for updates

Hong-Yu Zhou[1,9], Yizhou Yu[1,9] ✉, Chengdi Wang[2,9] ✉, Shu Zhang[3], Yuanxu Gao[4], Jia Pan[1], Jun Shao[2], Guangming Lu[5], Kang Zhang[6,7,8] ✉ & Weimin Li[2] ✉

During the diagnostic process, clinicians leverage multimodal information, such as the chief complaint, medical images and laboratory test results. Deep-learning models for aiding diagnosis have yet to meet this requirement of leveraging multimodal information. Here we report a transformer-based representation-learning model as a clinical diagnostic aid that processes multimodal input in a unified manner. Rather than learning modality-specific features, the model leverages embedding layers to convert images and unstructured and structured text into visual tokens and text tokens, and uses bidirectional blocks with intramodal and intermodal attention to learn holistic representations of radiographs, the unstructured chief complaint and clinical history, and structured clinical information such as laboratory test results and patient demographic information. The unified model outperformed an image-only model and non-unified multimodal diagnosis models in the identification of pulmonary disease (by 12% and 9%, respectively) and in the prediction of adverse clinical outcomes in patients with COVID-19 (by 29% and 7%, respectively). Unified multimodal transformer-based models may help streamline the triaging of patients and facilitate the clinical decision-making process.

It has been common practice in modern medicine to use multimodal clinical information for medical diagnosis. For instance, apart from chest radiographs, thoracic physicians need to take into account each patient's demographics (such as age and gender), the chief complaint (such as history of present and past illness) and laboratory test reports to make accurate diagnostic decisions. In practice, abnormal radiographic patterns are first associated with symptoms mentioned in the chief complaint or abnormal results in the laboratory test report. Then, physicians rely on their rich domain knowledge and years of training to make optimal diagnoses by jointly interpreting such multimodal data[1,2].

[1]Department of Computer Science, The University of Hong Kong, Pokfulam, China. [2]Department of Pulmonary and Critical Care Medicine, Med-X Center for Manufacturing, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu, China. [3]AI Lab, Deepwise Healthcare, Beijing, China. [4]Guangzhou Laboratory, Guangzhou, China. [5]Department of Medical Imaging, Jinling Hospital, Nanjing University School of Medicine, Nanjing, China. [6]Zhuhai International Eye Center and Provincial Key Laboratory of Tumor Interventional Diagnosis and Treatment, Zhuhai People's Hospital and the First Affiliated Hospital of Faculty of Medicine, Macau University of Science and Technology and University Hospital, Guangdong, China. [7]Department of Big Data and Biomedical Artificial Intelligence, National Biomedical Imaging Center, College of Future Technology, Peking University, Beijing, China. [8]Clinical Translational Research Center, West China Hospital, Sichuan University, Chengdu, China. [9]These authors contributed equally: Hong-Yu Zhou, Yizhou Yu, Chengdi Wang. ✉e-mail: yizhouy@acm.org; chengdi_wang@scu.edu.cn; kang.zhang@gmail.com; weimi003@scu.edu.cn

The importance of exploiting multimodal clinical information has been extensively verified in the literature[3–10] in different specialties, including but not limited to radiology, dermatology and ophthalmology.

The above multimodal diagnostic workflow requires substantial expertise, which may not be available in geographic regions with limited medical resources. Meanwhile, simply increasing the workload of experienced physicians and radiologists would inevitably exhaust their energy and thus increase the risk of misdiagnosis. To meet the increasing demand for precision medicine, machine-learning techniques[11] have become the de facto choice for automatic yet intelligent medical diagnosis. Among these techniques, the development of deep learning[12,13] endows machine-learning models with the ability to detect diseases from medical images near or at the level of human experts[14–18].

Although artificial intelligence (AI)-based medical image diagnosis has achieved tremendous progress in recent years, how to jointly interpret medical images and their associated clinical context remains a challenge. As illustrated in Fig. 1a, current multimodal clinical decision support systems[19–23] mostly lean on a non-unified way to fuse information from multiple sources. Given a set of input data from different sources, these approaches first roughly divide them into three basic modalities, that is, images, narrative text (such as the chief complaint, which includes the history of present and past illness) and structured fields (for example, demographics and laboratory test results). Next, a text structuralization process is introduced to transform the narrative text into structured tokens. Then, data in different modalities are fed to different machine-learning models to produce modality-specific features or predictions. Finally, a fusion module is employed to unify these modality-specific features or predictions for making final diagnostic decisions. In practice, according to whether multiple input modalities are fused at the feature or prediction level, these non-unified methods can be further categorized into early[19–22] or late fusion[23] methods.

One glaring issue with early and late fusion methods is that they separate the multimodal diagnostic process into two relatively independent stages: modality-specific model training and diagnosis-oriented fusion. Such a design has one obvious limitation: the inability to encode the connections and associations among different modalities. Another non-negligible drawback of these non-unified approaches lies in the text structuralization process, which is cumbersome and still labour-intensive, even with the assistance of modern natural language processing (NLP) tools. On the other hand, transformer-based architectures[24] are poised to broadly reshape NLP[25] and computer vision[26]. Compared with convolutional neural networks[27] and word embedding algorithms[28,29], transformers[24] impose few assumptions about the input data form and thus have the potential to learn higher-quality feature representations from multimodal input data. More importantly, the basic architectural component in transformers (that is, the self-attention block) remains nearly unchanged across different modalities[25,26], providing an opportunity to build a unified yet flexible model to conduct representation learning on multimodal clinical information.

In this paper, we present IRENE, a unified AI-based medical diagnostic model designed to make decisions by jointly learning holistic representations of medical images, unstructured chief complaint and structured clinical information. To the best of our knowledge, IRENE is presumably the first medical diagnostic approach that uses a single, unified AI model to conduct holistic representation learning on multimodal clinical information simultaneously, as shown in Fig. 1a. At the core of IRENE are the unified multimodal diagnostic transformer (MDT) and bidirectional multimodal attention blocks. MDT is a new transformer stack that directly produces diagnostic results from multimodal input data. This new algorithm enables IRENE to take a different approach from previous non-unified methods by learning holistic representations from multimodal clinical information progressively while eliminating separate paths for learning modality-specific features. In addition, MDT endows IRENE with the ability to perform representation learning on top of unstructured raw text, which avoids tedious text structuralization steps in non-unified approaches. For better handling of the differences among modalities, IRENE introduces bidirectional multimodal attention to bridge the gap between token-level modality-specific features and high-level diagnosis-oriented holistic representations by explicitly encoding the interconnections among different modalities. This explicit encoding process can be regarded as a complement to the holistic multimodal representation learning process in MDT.

As shown in Fig. 2a, MDT is primarily composed of embedding layers, bidirectional multimodal blocks and self-attention blocks. Because of the MDT, IRENE has the ability to jointly interpret multimodal clinical information simultaneously. Specifically, a free-form embedding layer is employed to convert unstructured and structured texts into uniform text tokens (Fig. 2b). Meanwhile, a similar tokenization procedure is also applied to each input image (Fig. 2c). Next, two bidirectional multimodal blocks (Fig. 2d) are stacked to learn fused mid-level representations across multiple modalities. In addition to computing intramodal attention among tokens from the same modality, these blocks also explicitly compute intermodal attention among tokens across different modalities (Fig. 2e). These intra- and intermodal attentional operations are consistent with daily clinical practices, where physicians need to discover interconnected information within the same modality as well as across different modalities. In reality, these connections are often hidden among local patterns, such as words in the chief complaint and image regions in radiographs, and different local patterns may refer to the same lesion or the same disease. Therefore, such connections provide mutual confirmations of clinical evidence and are helpful to both clinical and AI-based diagnosis. In bidirectional multimodal attention, each token can be regarded as the representation of a local pattern, and token-level intra- and intermodal attention respectively capture the interconnections among local patterns from the same modality and across different modalities. In comparison, previous non-unified methods make diagnoses on top of separate global representations of input data in different modalities and thus cannot exploit the underlying local interconnections. Finally, we stack ten self-attention blocks (Fig. 2f) to learn multimodal representations.

IRENE shares some common traits with vision–language fusion models[29–33], both of which aim to learn a joint multimodal representation. However, one most noticeable difference exists in the roles of different modalities. IRENE is designed for a scenario where multiple modalities supply complementary semantic information, which can be fused and used to improve prediction performance. In contrast, recent vision–language fusion approaches[31–33] heavily rely on the distillation and exploitation of common semantic information among different modalities to provide supervision for model training.

We validated the effectiveness of IRENE on two tasks (Fig. 1b): (1) pulmonary disease identification and (2) adverse clinical outcome prediction in patients with COVID-19. In the first task, IRENE outperformed previous image-only and non-unified diagnostic counterparts by approximately 12% and 9% (Fig. 1c), respectively. In the second task, we employed IRENE to predict adverse clinical events in patients with COVID-19, that is, admission to the intensive care unit (ICU), mechanical ventilation (MV) therapy and death. Different from the first task, the second task relies more on textual clinical information. In this scenario, IRENE significantly outperformed non-unified approaches by over 7% (Fig. 1d). Particularly noteworthy is the nearly 10% improvement that IRENE achieved on death prediction, demonstrating the potential for assisting doctors in taking immediate steps to save patients with COVID-19. When compared to human experts (Fig. 1e) in pulmonary disease identification, IRENE clearly surpassed junior physicians (with <7 yr of experience) in the diagnosis of all eight diseases and delivered a performance comparable to or better than that of senior physicians (with >7 yr of experience) on six diseases.

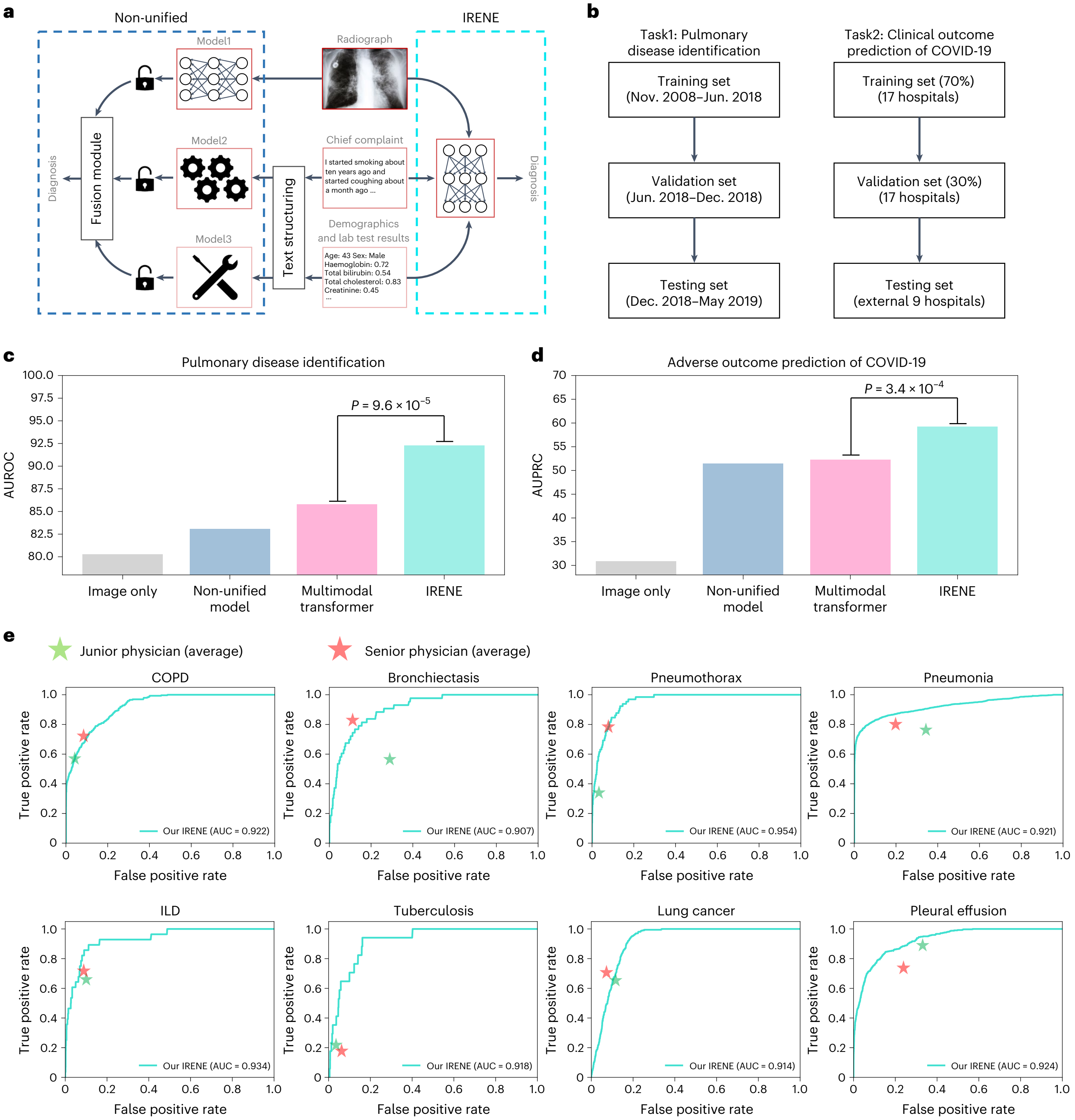


Fig. 1 | IRENE. a, Contrasting the previous non-unified multimodal diagnosis paradigm with IRENE. IRENE eliminates the tedious text structuralization process, separate paths for modality-specific feature extraction and the multimodal feature fusion module in traditional non-unified approaches. Instead, IRENE performs multimodal diagnosis with a single unified transformer. **b**, Scheme for splitting an original dataset into training, validation and testing sets for pulmonary disease identification (left) and adverse clinical outcome prediction of COVID-19 (right). **c**,**d**, Comparison of the experimental results from the image-only models, non-unified early fusion methods, multimodal transformer (that is, Perceiver) and IRENE in the two tasks in **b**. We compared the mean performance of IRENE and the multimodal transformer using independent two-sample *t*-test (two-sided). Specifically, we repeated each experiment ten times with different random seeds, after which *P* values were calculated. **e**, Comparison of IRENE with junior (<7 yr of experience, $n = 2$) and senior (>7 yr of experience, $n = 2$) physicians; average performance reported for each group. IRENE surpasses the diagnosis performance of junior physicians while performing competitively with senior experts. AUC, area under the curve.

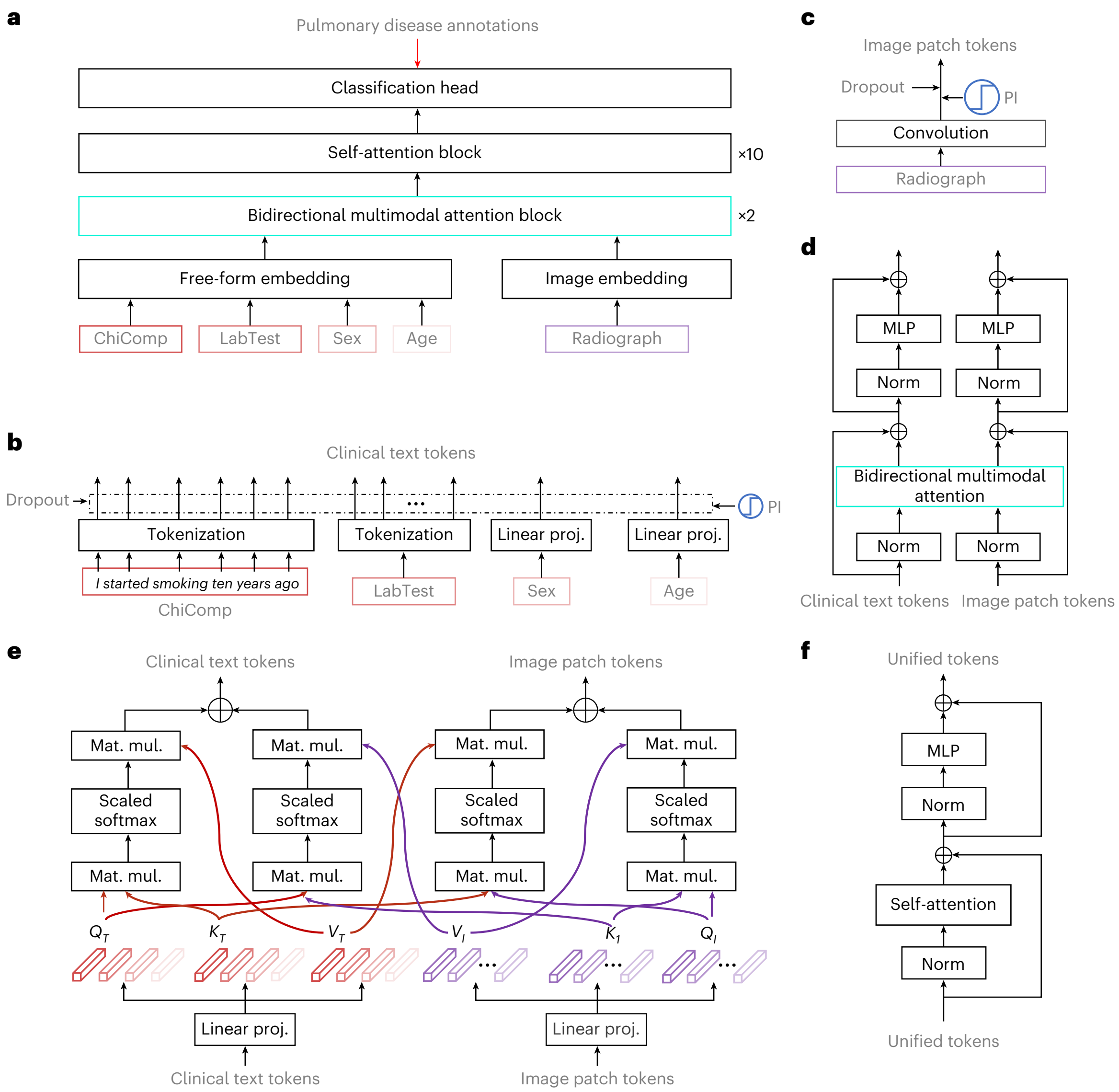


Fig. 2 | Network architecture of IRENE. a, Overall workflow of IRENE in the first task, that is, pulmonary disease identification. The input data consist of five parts: the chief complaint (ChiComp), laboratory test results (LabTest), demographics (sex and age) and radiograph. Our MDT includes two bidirectional multimodal attention blocks and ten self-attention blocks. The training process is guided by pulmonary disease annotations provided by human experts. **b**, Encoding different types of clinical text in the free-form embedding. Specifically, IRENE accepts unstructured chief complaints as part of the input. **c**, Encoding a radiograph as a sequence of image patch tokens. **d**, Detailed design of a bidirectional multimodal attention block, which consists of two-layer normalization layers (Norm), one bidirectional multimodal attention layer and one MLP. **e**, Detailed attention operations in the bidirectional multimodal attention layer, where representations across multiple modalities are learned and fused simultaneously. **f**, Detailed architecture of a self-attention block. PI, position injection.

Results

Dataset characteristics for multimodal diagnosis

The first dataset focused on pulmonary diseases. We retrospectively collected consecutive chest X-rays from 51,511 patients between 27 November 2008 and 31 May 2019 at West China Hospital, which is the largest tertiary medical centre in western China covering a 100 million population. Each patient is associated with at least one radiograph, a short piece of unstructured chief complaint, history of present and past illness, demographics and a complete laboratory test report. The dataset is built for eight pulmonary diseases: chronic obstructive pulmonary disease (COPD), bronchiectasis, pneumothorax, pneumonia, interstitial lung disease (ILD), tuberculosis, lung cancer and pleural effusion. Discharge diagnoses were extracted from discharge summary reports following a standard process described in a previous study[16] and taken as the ground-truth disease labels. The discharge summary reports were produced as follows. An initial report was written by a junior physician, which was then reviewed and confirmed by a senior physician. In case of any disagreement, the final decision was made by a departmental committee comprising at least three senior physicians.

The built dataset consisted of 72,283 data samples, among which 40,126 samples were normal. The distribution of diseases (that is, the number of relevant cases) is as follows: COPD (4,912), bronchiectasis (676), pneumothorax (2,538), pneumonia (21,409), ILD (3,283), tuberculosis (938), lung cancer (2,651) and pleural effusion (4,713). The performance metric is the area under the receiver operating characteristic curve (AUROC). We split this dataset into training, validation and testing sets according to each patient's admission date. Specifically, the training set included 44,628 patients admitted between 27 November

Table 1 | Comparison with baseline models in the task of pulmonary disease identification

Method	Mean	COPD	Bronchiectasis	Pneumothorax	Pneumonia	ILD	Tuberculosis	Lung cancer	Pleural effusion
Image-only	0.805 (0.802, 0.808)	0.847 (0.845, 0.851)	0.746 (0.743, 0.748)	0.789 (0.786, 0.791)	0.845 (0.843, 0.848)	0.799 (0.796, 0.801)	0.769 (0.765, 0.772)	0.825 (0.821, 0.830)	0.819 (0.817, 0.822)
Early fusion	0.835 (0.832, 0.839)	0.895 (0.893, 0.898)	0.772 (0.768, 0.775)	0.810 (0.807, 0.812)	0.873 (0.870, 0.875)	0.824 (0.822, 0.827)	0.793 (0.791, 0.796)	0.871 (0.868, 0.875)	0.842 (0.839, 0.845)
Late fusion	0.826 (0.823, 0.828)	0.888 (0.885, 0.890)	0.765 (0.763, 0.767)	0.822 (0.820, 0.825)	0.870 (0.868, 0.872)	0.804 (0.802, 0.805)	0.770 (0.767, 0.772)	0.839 (0.836, 0.841)	0.850 (0.847, 0.852)
GIT	0.848 (0.844, 0.850)	0.911 (0.907, 0.913)	0.798 (0.796, 0.800)	0.824 (0.821, 0.827)	0.895 (0.893, 0.898)	0.819 (0.816, 0.821)	0.807 (0.804, 0.810)	0.872 (0.871, 0.873)	0.858 (0.855, 0.860)
Perceiver	0.858 (0.855, 0.861)	0.910 (0.907, 0.912)	0.788 (0.784, 0.791)	0.846 (0.842, 0.850)	0.903 (0.901, 0.906)	0.830 (0.827, 0.833)	0.825 (0.823, 0.828)	0.890 (0.887, 0.892)	0.872 (0.869, 0.874)
IRENE	0.924 (0.921, 0.926)	0.922 (0.920, 0.925)	0.907 (0.903, 0.910)	0.954 (0.952, 0.957)	0.921 (0.918, 0.923)	0.934 (0.929, 0.937)	0.918 (0.917, 0.921)	0.914 (0.911, 0.917)	0.924 (0.921, 0.926)

The baseline models include the image-only model, the early fusion method, the late fusion approach and two recent transformer-based multimodal classification models (that is, GIT and Perceiver). The evaluation metric is AUROC, with 95% confidence intervals in brackets.

2008 and 1 June 2018. The validation set included 3,325 patients admitted between 2 June 2018 and 1 December 2018. Finally, the trained and validated IRENE system was tested on 3,558 patients admitted between 2 December 2018 and 31 May 2019. Although this was a retrospective study, our data splitting scheme followed the practice of a prospective study, thus creating a more challenging and realistic setting to verify the effectiveness of different multimodal medical diagnosis systems, in comparison to data splitting schemes based on random sampling.

The second dataset, MMC (that is, the multimodal COVID-19 dataset)[19], on which IRENE was trained and evaluated, consisted of chest computed tomography (CT) scan images and structured clinical information (for example, chief complaint that comprises comorbidities and symptoms, demographics, laboratory test results and so on) collected from patients with COVID-19. The CT images were associated with inpatients with laboratory-confirmed COVID-19 infection between 27 December 2019 and 31 March 2020. There were three types of adverse event that could happen to patients in MMC, namely admission to ICU, MV and death. The training and validation sets came from 17 hospitals and the training set had 1,164 labelled cases (70%), while the validation set had 498 labelled ones (30%). Next, we chose the trained model with the best performance on the validation set and tested it on the independent testing set, which comprised 700 cases collected from 9 external medical centres. The distribution of the three events in the testing set was as follows: ICU (155), MV (94), death (59). This was an imbalanced classification problem where the majority of patients did not have any adverse outcomes. Against this background, we used the area under the precision-recall curve (AUPRC) instead of AUROC as the performance metric, as the former focused more on identifying adverse events (that is, ICU, MV and death).

Pulmonary disease identification

Table 1 and Fig. 3 present the experimental results from IRENE and other methods on the dataset for pulmonary disease identification. As shown in Table 1, IRENE significantly outperformed the image-only model, the traditional non-unified early[19] and late fusion[23] methods and two recent state-of-the-art transformer-based multimodal methods (that is, Perceiver[30] and GIT[33]) in identifying pulmonary diseases. Overall, IRENE achieved the highest mean AUROC of 0.924 (95% CI: 0.921, 0.927), about 12% higher than the image-only model (0.805, 95% CI: 0.802, 0.808) that only takes radiographs as the input. In comparison with diagnostic decisions made by non-unified early fusion (0.835, 95% CI: 0.832, 0.839) and late fusion (0.826, 95% CI: 0.823, 0.828) methods, IRENE maintained an advantage of at least 9%. Comparing IRENE to GIT (0.848, 95% CI: 0.844, 0.850), we observed an advantage of over 7%. Even when compared to Perceiver, the transformer-based multimodal classification model developed by DeepMind, IRENE still delivered competitive results, surpassing Perceiver (0.858, 95% CI: 0.855, 0.861) by over 6%. When carefully checking each disease and comparing IRENE against the previous best result among all five baselines, we observed that among all eight pulmonary diseases, IRENE achieved the largest improvements on bronchiectasis (12%), pneumothorax (10%), ILD (10%) and tuberculosis (9%).

We also compared IRENE against human experts who were divided into two groups: one group of two junior physicians (with <7 yr of experience) and a second group of two senior physicians (with ≥7 yr of experience). For better comparison, we present the average performance within each group in Fig. 1e. Specifically, we extracted annotations by human experts from electronic discharge diagnosis records. Notably, all physicians from the reader study did not participate in data annotation. We observed that IRENE exhibited advantages over the junior group on all eight pulmonary diseases, especially in the diagnosis of bronchiectasis (junior, false positive rate (FPR): 0.29, true positive rate (TPR): 0.58), pneumonia (junior, FPR: 0.37, TPR: 0.76), ILD (junior, FPR: 0.09, TPR: 0.63) and pleural effusion (junior, FPR: 0.35, TPR: 0.86). Compared with the senior group, IRENE was advantageous in the diagnosis of pneumonia (senior, FPR: 0.21, TPR: 0.80), tuberculosis (senior, FPR: 0.07, TPR: 0.17) and pleural effusion (senior, FPR: 0.25, TPR: 0.77). In addition, IRENE performed comparably with senior physicians on COPD (senior, FPR: 0.07, TPR: 0.76), ILD (senior, FPR: 0.09, TPR: 0.71) and pneumothorax (senior, FPR: 0.08, TPR: 0.79) while showing slightly worse performance on bronchiectasis (senior, FPR: 0.12, TPR: 0.82) and lung cancer (senior, FPR: 0.08, TPR: 0.73).

Adverse clinical outcome prediction in patients with COVID-19

Triage of patients with COVID-19 heavily depends on joint interpretation of chest CT scans and other non-imaging clinical information. In this scenario, IRENE exhibited even more advantages than it did in the pulmonary disease identification task. As shown in Table 2, IRENE consistently achieved impressive performance improvements on the prediction of the three adverse clinical outcomes for patients with COVID-19; that is, admission to ICU, MV and death. In terms of mean AUPRC, IRENE (0.592, 95% CI: 0.500, 0.682) outperformed the image-only model (0.307, 95% CI: 0.237, 0.391), early fusion model[22] (0.521, 95% CI: 0.435, 0.614) and late fusion model[23] (0.503, 95% CI: 0.422, 0.598) by nearly 29%, 7% and 9%, respectively. As for specific clinical outcomes, IRENE (0.712, 95% CI: 0.587, 0.834) achieved about 5% AUPRC gain over the non-unified early fusion method (0.665, 95% CI: 0.548, 0.774) in the prediction of admission to ICU. Similarly, in the prediction of MV, IRENE achieved a >6% performance improvement when compared with the early fusion model. Last but not least, IRENE (0.441, 5% CI: 0.270, 0.617) was much more capable of predicting death than the image-only model (0.192, 95% CI: 0.073, 0.333), early fusion

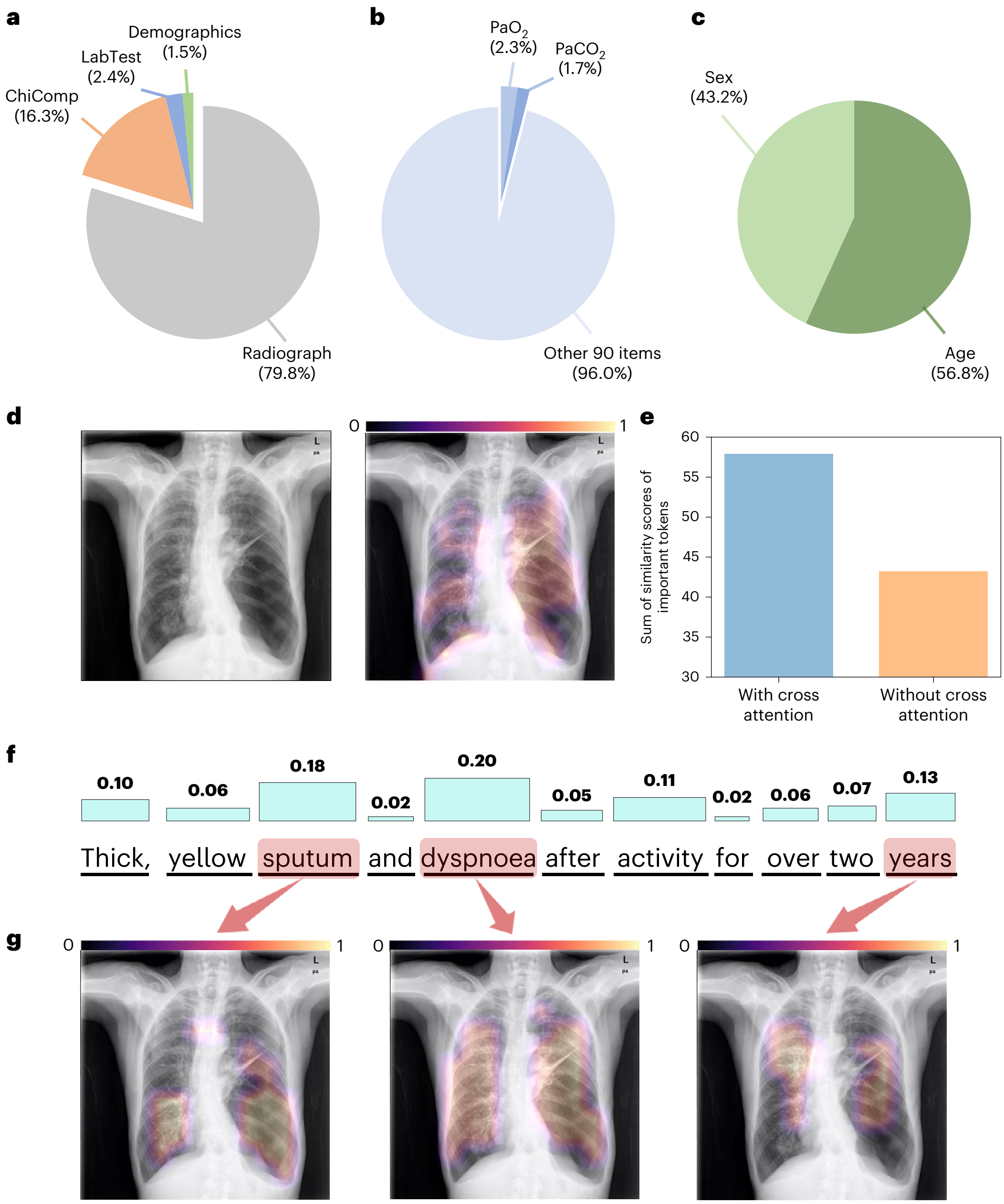


Fig. 3 | Attention analysis. a, Attention allocated to different types of input from a patient with COPD, that is, the radiograph, ChiComp, LabTest and demographics. **b**, Relative importance of laboratory test items. **c**, Comparison of the importance of sex and age in making a diagnostic decision. **d**, Visualization of the attention assigned to individual pixels in the radiograph. Left: input chest X-ray. Right: pixels with different attention values. **e**, The impact of cross attention on the relevance and importance of high-ranking words (from chief complaints) and image patches (from radiographs) in the pulmonary disease identification task. Specifically, we define high-ranking words and patches as those whose tokens have top 25% cosine similarity scores with the CLS token. **f**, Normalized importance of every word in the chief complaint. **g**, Visualization of the distribution of attention between every image patch and each of the top 3 ranked words. The colour bars in **d** and **g** illustrate the confidence of IRENE about a pixel being abnormal, where a bright colour stands for high confidence and a dark colour denotes low confidence.

model (0.346, 95%: 0.174, 0.544) and late fusion model (0.335, 95% CI: 0.168, 0.554). Compared with two transformer-based multimodal models (that is, GIT and Perceiver), we observed an advantage of over 6% on average.

Impact of different modules and modalities in IRENE

To investigate the impact of different modules and modalities, we conducted thorough ablative experiments and report their results in Table 3. First, we investigated the impact of bidirectional multimodal attention blocks (rows 0–2). We found that replacing all bidirectional multimodal attention blocks with self-attention blocks led to ~7% performance drop (from 0.924 to 0.858) in pulmonary disease identification. This phenomenon verified our intuition that directly learning progressively fused representations from raw data would deteriorate diagnosis performance. In contrast, simply increasing the number of bidirectional multimodal attention blocks from two to six did not bring obvious performance improvements (from 0.924 to 0.905), indicating that using two successive bidirectional multimodal attention blocks could be an optimal choice in IRENE. In row 3, we presented the result of using unidirectional attention (that is, text-to-image attention). Comparing row 0 with row 3,

Table 2 | Comparison with baseline models in the task of adverse clinical outcome prediction in patients with COVID-19

Method	Mean	Admission to ICU	Need for MV	Death
Image-only	0.307 (0.237, 0.391)	0.482 (0.355, 0.636)	0.247 (0.136, 0.398)	0.192 (0.073, 0.333)
Early fusion	0.521 (0.435, 0.614)	0.665 (0.548, 0.774)	0.551 (0.397, 0.699)	0.346 (0.174, 0.544)
Late fusion	0.503 (0.422, 0.598)	0.647 (0.535, 0.759)	0.533 (0.388, 0.685)	0.330 (0.164, 0.531)
GIT	0.514 (0.442, 0.605)	0.653 (0.546, 0.743)	0.554 (0.411, 0.702)	0.335 (0.168, 0.554)
Perceiver	0.526 (0.448, 0.611)	0.652 (0.529, 0.771)	0.566 (0.406, 0.715)	0.360 (0.201, 0.543)
IRENE	0.592 (0.500, 0.682)	0.712 (0.587, 0.834)	0.624 (0.473, 0.754)	0.441 (0.270, 0.617)

The evaluation metric is AUPRC, with 95% confidence intervals in brackets.

we observed that our bidirectional design brought a 4% performance gain (from 0.884 to 0.924). Next, we studied the impact of clinical texts (rows 4 and 5). The first observation was that using the complementary narrative chief complaint substantially boosted the diagnostic performance because removing chief complaint from the input data reduced model performance by 6% (from 0.924 to 0.860). Apart from the chief complaint, we also studied the impact of laboratory test results (row 5). We observed that including laboratory test results brought about a 4% performance gain (from 0.882 to 0.924). Then, we investigated the impact of tokenization procedures. We saw that modelling the chief complaint and laboratory test results of a patient as a sequence of tokens (row 0) did perform better than directly passing an averaged representation (row 6) to the model. This improvement brought by the tokenization of the chief complaint and laboratory test results verified the advantage of token-level intra- and intermodal bidirectional multimodal attention, which exploited local interconnections among the word tokens of the clinical text and the image patch tokens of the radiograph in the input data. Lastly, we investigated the impact of the input image in IRENE (row 7) and observed a substantial performance drop (from 0.924 to 0.543). This phenomenon indicated the vital role of the input radiograph in pulmonary disease identification. We then investigated the impact of chief complaints and laboratory test results on each respiratory disease (Extended Data Fig. 1). When we removed either chief complaints or the laboratory test results from the input, the performance decreased on each disease. Specifically, we found that introducing the chief complaint could be most helpful for the diagnosis of pneumothorax, lung cancer and pleural effusion, while the laboratory test results affected the diagnosis of bronchiectasis and tuberculosis the most. Clinical interpretations can be found in Supplementary Note 1.

Attention visualization results

Figure 3 provides attention visualization results for a case with COPD. In Fig. 3a, we see that the image modality (that is, the radiograph) played a significant role in the diagnostic process, and its weight was nearly 80% in the final decision. The chief complaint was the second most important factor, accounting for roughly 16% weight. As Fig. 3b shows, PaO_2 (oxygen pressure in arterial blood) and $PaCO_2$ (partial pressure of carbon dioxide in arterial blood) were the two most important laboratory test items, which are consistent with the observations reported in the literature[34]. Nonetheless, we see that the total weight of the remaining 90 test items was quite large, with distribution over these 90 laboratory test items being nearly uniform. The reason might be that these laboratory test items could help rule out other diseases. Figure 3c shows that from the perspective of IRENE, age was a more critical factor than sex. Figure 3d provides the attention map of the radiograph, implying that IRENE would refer to hilar enlargement, hyper-expansion and flattened diaphragm as the most important pieces of evidence for the diagnosis of COPD. In addition, IRENE could also identify large black areas due to bullae as relatively important evidence. Figure 3e summarizes the experimental results with and without cross attention, where we present the sum of similarity scores of important (top 25%) tokens (that is, words and image patches) with the CLS token which is the start token that aggregates the information of the rest tokens. We found that with cross attention, the sum of similarity scores became larger, indicating that cross attention has improved the identification of important tokens compared with the model without cross attention. In Fig. 3f, IRENE recognized ‘sputum’, ‘dyspnoea’ and ‘years’ as the three most important words in the chief complaint. Figure 3g provides the cross-attention maps between each of the top three important words and the image. The word ‘sputum’ is primarily associated with the trachea and the lower pulmonary lobes in the image. The high attention area of the trachea could be reasonable because trachea is often the location where sputum might occur. The high attention region in the left lower lobe had reduced vascular markings, while both the left and right lower lobes of the lungs were hyperinflated. Hyperinflated lungs and reduced vascular markings are common symptoms of COPD, which often has abnormal sputum production. Our model has also associated the word ‘dyspnoea’ with most areas of the lungs in the image because dyspnoea can be caused by a variety of pulmonary abnormalities that could occur anywhere in the lungs. Lastly, our model has identified the areas surrounding the bronchi as the image regions associated with the word ‘years’, which implies ‘years’ should be associated with chronic diseases, such as chronic bronchitis, which is often part of COPD.

Discussion

IRENE is more effective than the previous non-unified early and late fusion paradigm in multimodal medical diagnosis

This is the most prominent observation obtained from our experimental results, and it holds for the tasks of pulmonary disease identification and the triage of patients with COVID-19. Specifically, IRENE outperforms previous early fusion and late fusion methods by an average of 9% and 10%, respectively, for identifying pulmonary diseases. Moreover, IRENE achieves about 3% performance gains on all eight diseases and substantially improves the diagnostic performance on four diseases (that is, bronchiectasis, pneumothorax, ILD and tuberculosis) by boosting their AUROC by over 10%. We believe that these performance benefits are closely related to several capabilities of IRENE. First, IRENE is built on top of a unified transformer (that is, MDT). MDT directly produces diagnostic decisions from multimodal input data and learns holistic multimodal representations progressively and implicitly. In contrast, the traditional non-unified approach decomposes the diagnosis problem into several components which, in most cases, consist of data structuralization, modality-specific model training and diagnosis-oriented fusion. In practice, these components are hard to optimize and may prevent the model from learning holistic and diagnosis-oriented features. Second, inspired by the daily activities of physicians, IRENE applies intra-directional and bidirectional intermodal attention to tokenized multimodal data for exploiting the local interconnections among complementary modalities. In contrast, the previous non-unified paradigm directly makes use of the extracted global modality-specific representations or predictions for diagnosis. In practice, the token-level attentional operations in bidirectional multimodal attention helps capture and encode the interconnections among the local patterns of different modalities into the fused representations. Furthermore, IRENE is designed to conduct representation learning directly on unstructured raw texts. In contrast, the previous non-unified approach relies on non-clinically pre-trained NLP models to provide word embeddings, which inevitably distracts the diagnosis system from its intended functionality.

Table 3 | An ablation study of IRENE, removing or replacing individual components

Row	HA (2)	HA (0)	HA (6)	Unidirection	Image	ChiComp	LabTest	Tokenization	Mean
0	√				√	√	√	√	0.924 (0.921, 0.926)
1		√			√	√	√	√	0.858 (0.850, 0.867)
2			√		√	√	√	√	0.905 (0.899, 0.910)
3	√			√	√	√	√	√	0.884 (0.880, 0.888)
4	√				√		√	√	0.860 (0.855, 0.864)
5	√				√	√		√	0.882 (0.873, 0.891)
6	√				√	√	√		0.894 (0.886, 0. 900)
7	√					√	√	√	0.543 (0.525, 0.569)

HA (N) denotes the presence of N bidirectional multimodal attention block(s) in the MDT, while the remaining blocks are self-attention blocks (12 blocks in total). Image denotes the input radiograph. Unidirection means we only compute text-to-image attention in multimodal attention blocks. ChiComp stands for the chief complaint. LabTest denotes laboratory test results. Tokenization stands for the tokenization procedures for the chief complaint and laboratory test results. For each row in the table, check marks denote the associated modules are used in the model while blank spaces indicate the associated modules are removed. The evaluation metric is AUROC, with 95% confidence intervals in brackets.

The superiority of the aforementioned abilities has been partly verified in the second task: the prediction of adverse outcomes in patients with COVID-19. From Table 2, we see that IRENE holds a 7% average performance gain over the early fusion approach and an average of 9% advantage over the late fusion one. This performance gain is a little lower than that in the pulmonary disease identification task as there are no unstructured texts in the MMC dataset that IRENE can use. Nonetheless, IRENE can still leverage its unified and bidirectional multimodal attention mechanisms to better serve the goal of rapidly triaging patients with COVID-19. For example, IRENE boosts the performance of MV and death prediction by 7% and 10%, respectively. Such substantial performance improvements brought by IRENE are valuable in the real world for allocating appropriate medical resources to patients in a timely manner, as medical resources are usually limited during a pandemic.

IRENE provides a better transformer-based choice for jointly interpreting multimodal clinical information

We compared IRENE to GIT[33] and Perceiver[30], two representative transformer-based models that fuse multimodal information for classification. GIT performs multimodal pre-training on tens of millions of image-text pairs by using the common semantic information among different modalities as supervision signals. However, these characteristics have two obvious deficiencies in the medical diagnosis scenario. First, it is much harder to access multimodal medical data in the amount of the same order of magnitude. Second, multimodal data in the medical diagnosis scenario provide complementary instead of common semantic information. Thus, it is impractical to perform large-scale multimodal pre-training, as in GIT, using a limited amount of medical data. These deficiencies are also reflected in the experimental results. For instance, the average performance of GIT is about 7% and 8% lower than that of IRENE in the pulmonary disease identification task and adverse outcome prediction of COVID-19 task, respectively. These advantages show that token-level bidirectional multimodal attention in IRENE can effectively use a limited amount of multimodal medical data and exploit complementary semantic information.

Perceiver simply concatenates multimodal input data and takes the resulting one-dimensional (1D) sequence as the input instead of learning fused representations among modality-specific low-level embeddings as in IRENE. This poses a potential problem: the modality that makes up the majority of the input would have a larger impact on final diagnostic results. For example, since an image often has a much larger number of tokens than a text, Perceiver would inevitably assign more weight to the image instead of the text when making predictions. However, it is not always true that images play a more important role in daily clinical decisions. To some extent, this point is also reflected in our experimental observations. For example, Perceiver yields clear performance improvements (2% gain on average in Table 1) over the early fusion model in identifying pulmonary diseases where the input radiograph serves as the main information source. However, in the task of rapidly triaging patients with COVID-19, the performance of Perceiver is only comparable to that of the early fusion method. The underlying reason is that CT images are not as helpful in this task as radiographs in pulmonary disease identification. In contrast, IRENE demonstrates satisfactory performance in both tasks by learning holistic multimodal representations through bidirectional multimodal attention. Our method encourages features from different modalities to evenly blend into each other, which prevents the learned representations from being dominated by high-dimensional inputs.

IRENE helps reduce reliance on text structuralization in the traditional workflow

In traditional non-unified multimodal medical diagnosis methods, the usual way to deal with unstructured texts is text structuralization. Recent text structuralization pipelines in non-unified approaches[19–23] severely rely on artificial rules and the assistance of modern NLP tools. For example, text structuralization requires human annotators to manually define a list of alternate spellings, synonyms and abbreviations for structured labels. On top of these preparations, specialized NLP tools are developed and applied to extract structured fields from unstructured texts. As a result, text structuralization steps are not only cumbersome but also costly in terms of labour and time. In comparison, IRENE abandons such tedious structuralization steps by directly accepting unstructured clinical texts as part of the input.

Outlook

NLP technologies, particularly transformers, have contributed significantly to the latest AI diagnostic tools using either text-based electronic health records[35] or images[36]. We have described an AI framework consisting of a unified MDT and bidirectional multimodal attention blocks. IRENE is distinct from previous non-unified methods in that it progressively learns holistic representations of multimodal clinical data while avoiding separate paths for learning modality-specific features in non-unified techniques. This approach may be enhanced by the latest development of large language models[37,38].

In real-world scenarios, IRENE may help streamline patient care, such as triaging patients and differentiating between those patients who are likely to have a common cold from those who need urgent intervention for a more severe condition. Furthermore, as the algorithms become increasingly refined, these frameworks could become a diagnostic aid for physicians and assist in cases of diagnostic uncertainty or complexity, thus not only mimicking physician reasoning but also further enhancing it. The impact of our work may be most obvious in areas where there are few and uneven distribution of healthcare providers relative to the population.

There are several limitations that would need to be considered during the deployment of IRENE in clinical workflows. First, the currently used datasets are limited in both size and diversity. To resolve this issue, more data would need to be collected from additional medical institutions, medical devices, countries and ethnic groups, with which IRENE can be trained to enhance its generalization ability under a broader range of clinical settings. Second, the clinical benefits of IRENE need to be further verified. Thus, multi-institutional multinational studies would be needed to further validate the clinical utility of IRENE in real-world scenarios. Third, it is important to make IRENE adaptable to a changing environment, such as dealing with rapidly mutating SARS-CoV-2 viruses. To tackle this challenge, the model could be trained on multiple cohorts jointly or one could resort to other machine-learning technologies, such as online learning. Moreover, IRENE fails to consider the problem of modal deficiency, where one or more modalities may be unavailable. To deal with this problem, one can refer to masked modelling[25]. For instance, during the training stage, some modalities could be randomly masked to imitate the absence of these modalities in clinical workflows.

Methods

Image and textual clinical data

In the pulmonary disease identification task, chest X-ray (CXR) images were collected from West China Hospital. All CXRs were collected as part of the patients' routine clinical care. For the analysis of CXR images, all radiographs were first de-identified to remove any patient-related information. The CXR images consisted of both anterior and posterior views. There were three types of textual clinical data: the unstructured chief complaint (that is, history of present and past illness), demographics (age and gender) and laboratory test results. Specifically, the chief complaint is unstructured, while demographics and laboratory test results are structured. We set the maximum length of the chief complaint to 40. If a patient's chief complaint had more than 40 words, we only took the first 40; otherwise, zero padding was used to satisfy the length requirement. There were 92 results in each patient's laboratory test report (see Supplementary Note 2), most of which came from a blood test. We normalized every test result by minimum-maximum (min-max) scaling so that every normalized value was between 0 and 1, where the minimum and maximum values in min-max scaling were determined using the training set. In particular, −1 denoted missing values.

In the second task, that is, adverse clinical outcome prediction for patients with COVID-19, the available clinical data were divided into four categories: demographics (age and gender), the structured chief complaint consisting of comorbidities (7) and symptoms (9) and laboratory test results (19) (see Supplementary Note 3 for more details). We also applied median imputation to fill in missing values.

Institutional Review Board/Ethics Committees approvals were obtained from West China Hospital and all participating hospitals. All patients signed a consent form. The research was conducted in a manner compliant with the United States Health Insurance Portability and Accountability Act. It adhered to the tenets of the Declaration of Helsinki and complied with the Chinese Center for Disease Control and Prevention policy on reportable infectious diseases and the Chinese Health and Quarantine Law.

Baseline models

We include five baseline models in our experimental performance comparisons, including the diagnosis model purely based on medical images (denoted as Image-only), the traditional non-unified early and late fusion methods with multimodal input data and two recent state-of-the-art transformer-based multimodal classification methods (that is, GIT and Perceiver). Implementation details are discussed below.

Image-only. In the pulmonary disease identification task, we built the pure medical image-based diagnosis model on top of ViT[26], one of the most well-known and widely adopted transformer-based deep neural networks for image understanding. Our ViT-like network architecture had 12 blocks and each block consisted of one self-attention layer[24], one multilayer perceptron (MLP) and two-layer normalization layers[39]. There were two fully connected (FC) layers in each MLP, where the number of hidden nodes was 3,072. The input size of the first FC layer was 768. Between the two FC layers, we inserted a GeLU activation function[40]. After each FC layer, we added a dropout layer[41], where we set the dropout rate to 0.3. The output size of the second FC layer was also 768. Each input image was divided into a number of 16 × 16 patches. The output CLS token was used for performing the final classification. We used the binary cross-entropy loss as the cost function during the training stage. Note that before the training stage, we performed supervised ViT pre-training on MIMIC-CXR[42] to obtain visual representations with more generalization power. In the task of rapidly triaging patients with COVID-19, as in ref. 22, we first segmented pneumonia lesions from CT scans, then trained multiple machine-learning models (that is, logistic regression, random forest, support vector machine, MLP and LightGBM) using image features extracted from the segmented lesion areas and finally chose the optimal model according to their performance on the validation set.

Non-unified early and late fusion. There are a number of existing methods using the archetypical non-unified approach to fuse multimodal input data for diagnosis. For better adaptation to different scenarios, we adopted different non-unified models for different tasks. Specifically, we modified the previously reported early fusion method[19] for our first task (that is, pulmonary disease identification). In practice, a ViT model extracts image features from radiographs and the feature vector at its CLS token is taken as the representation of the input image. Similar to the image-only baseline, supervised pre-training on MIMIC-CXR[42] was applied to the ViT to obtain more powerful visual features before we carried out the formal task. To process the three types of clinical data (that is, the chief complaint, demographics and laboratory test results), we employed three independent MLPs to convert different types of textual clinical data to features, which were then concatenated with the image representation. The rationale is that both images and textual data should be represented in the same feature space for the purpose of cross referencing. Since the chief complaint includes unstructured texts, we first needed to transform them into structured items. To achieve this goal, we trained an entity recognition model to highlight relevant clinical symptoms in the chief complaint. Next, we used BERT[25] to extract features for all such symptoms, to which average pooling was applied to produce a holistic representation for each patient's chief complaint. Then, we used a three-layer MLP to further transform this holistic feature into a latent space similar to that of the image representation. The input size of this three-layer MLP was 768 and the output size was 512. The number of hidden nodes was 1,024. After each FC layer, we added a ReLU activation and a dropout layer, with the dropout rate set to 0.3. Likewise, for laboratory test results, we also applied an MLP with the same architecture but independent weight parameters to transform those test results into a 1D feature vector. The input size of this laboratory test MLP was 92 and the output size was 512. The MLP model for demographics had two FC layers, where the input size was 2 and the output size was 512. The hidden layer had 512 nodes. The feature fusion module included the concatenation operation and a three-layer MLP, with the number of hidden nodes set to 1,024. The output from the MLP in the feature fusion module was passed to the final classification layer for making diagnostic decisions. During the training stage, we jointly trained the ViT-like model and all MLPs using the binary cross-entropy loss. As for the late fusion baseline, we combined the predictions of the image- and text-based classifiers following ref. 23. Specifically, we trained a ViT model with radiographs and their associated labels. To construct the input to the text-based classifier, we concatenated laboratory test results, demographics and the holistic representation (obtained via averaging extracted features of symptoms, similar to the early fusion method) of the chief complaint. Then, we forwarded the constructed input through a three-layer MLP, whose input and output dimensions were 862 and 8, respectively. Then, we trained the MLP with the same labels used for training the ViT model. Finally, we averaged the predicted probabilities of the image- and text-based classifiers to obtain the final prediction.

In the second task, we followed a proposed early fusion method[22], where image features, structured chief complaint (comorbidities and symptoms) and laboratory test results had been concatenated as the input. Then, we trained multiple machine-learning models and chose the optimal model using previously introduced artificial rules[22]. For the late fusion baseline, we trained 5 machine-learning models (logistic regression, random forest, support vector machine, MLP and LightGBM) each for image features, structured chief complaints and laboratory test results following the protocol used in ref. 22. Then, we took the average of the predicted probabilities of these 15 machine-learning models as the adverse outcome prediction.

GIT. GIT[33] is a generative image-to-text transformer that unifies vision–language tasks. We took GIT-Base as a baseline in our comparisons. Its image encoder is a ViT-like transformer and its text decoder consists of six standard transformer blocks[24]. In practice, we fine-tuned the officially released pre-trained model on our own datasets. For fairness, we adopted the same set of fine-tuning hyperparameters used for IRENE. In the pulmonary disease identification task, we first forwarded each radiograph through the image encoder to extract an image feature. Next, we concatenated this image feature with the averaged word embedding (using BERT) of the chief complaint as well as the feature vectors of the demographics and laboratory test results. The concatenated features were then passed to the text decoder to make diagnostic predictions. In the task of adverse clinical outcome prediction for patients with COVID-19, we first averaged the image features of CT slices. Then, the averaged image feature was concatenated with the feature vectors of the clinical comorbidities and symptoms, laboratory test results and demographics. Next, we forwarded the concatenated multimodal features through the text decoder to predict adverse outcomes for patients with COVID-19.

Perceiver. This is a very recent state-of-the-art transformer-based model[30] from DeepMind, proposed for tackling the classification problem with multimodal input data. A variant of Perceiver[30], that is, Perceiver IO[43], introduces the output query on top of Perceiver to handle additional types of task. As making diagnostic decisions can be considered as a type of classification, we adopted Perceiver instead of Perceiver IO as one of our baseline models. Our Perceiver architecture followed the setting for ImageNet classification[30,44] and had six cross-attention modules. Each cross-attention module was followed by a latent transformer with six self-attention blocks. The input of Perceiver consists of two arrays: the latent array and byte array. Following ref. 30, we initialized the latent array using a truncated zero-mean normal distribution, with standard deviation set to 0.02 and truncation bounds set to (−2, 2). The byte array consisted of multimodal data. In the pulmonary disease identification task, we first flattened the input image into a 1D vector. Then, we concatenated it with the averaged word embedding (using BERT) of the chief complaint as well as 1D feature vectors of the input demographics and laboratory test results. This resulted in a long 1D vector, which was taken as the byte array. In the task of adverse clinical outcome prediction of COVID-19, we also flattened the input image into a 1D vector, which was then concatenated with the feature vectors of the clinical comorbidities and symptoms, laboratory test results and demographics. The learning process of Perceiver can be summarized as follows: the latent array evolves by iteratively extracting higher-quality features from the input byte array by alternating cross-attention and latent self-attention computations. Finally, the transformed latent array serves as the representation used for diagnosis. Note that similar to the image-only and non-unified baselines, we pre-trained Perceiver on MIMIC-CXR[42]. During pre-training, we used zero padding in the byte array for the non-existent clinical text in every multimodal input.

IRENE

In practice, we forwarded multimodal input data (that is, medical images and textual clinical information) to the MDT for acquiring prediction logits. During the training stage, we computed the binary cross-entropy loss between the logits and ground-truth labels. Specifically, we used pulmonary disease annotations (8 diseases) and real adverse clinical outcomes (3 clinical events) as the ground-truth labels in the first and second tasks, respectively.

MDT is a unified transformer, which primarily consists of two starting layers for embedding the tokens from the input image and text, respectively, two stacked bidirectional multimodal attention blocks for learning fused mid-level representations by capturing interconnections among tokens from the same modality and across different modalities, ten stacked self-attention blocks for learning holistic multimodal representations and enhancing their discriminative power, and one classification head for producing prediction logits.

The multimodal input data in the pulmonary disease identification task (that is, the first task) consisted of five parts: a radiograph, the unstructured chief complaint that includes history of present and past illness, laboratory test results, each patient's gender and age, which were denoted as $x^{\mathrm{I}}, x^{\mathrm{cc}}, x^{\mathrm{lab}}, x^{\mathrm{sex}}$ and x^{age}, respectively. We passed x^{I} to a convolutional layer, which produced a sequence of visual tokens. Next, we added standard learnable 1D positional embedding[21,23] and dropout to every visual token to obtain a sequence of image patch tokens $X^{\mathrm{I}}_{1:N}$. Meanwhile, we applied word tokenization to X^{cc} to encode each word from the unstructured chief complaint. Specifically, we used a pre-trained BERT[23] to generate an embedded feature vector for each word in x^{cc}, after which we obtained a sequence of word tokens $X^{\mathrm{cc}}_{1:N^{\mathrm{cc}}}$. We also applied a similar tokenization procedure to x^{lab}, where min-max scaling was first employed to normalize every component of x^{lab}. We then passed each normalized component to a shared linear projection layer to obtain a sequence of latent embeddings $X^{\mathrm{lab}}_{1:N^{\mathrm{lab}}}$. We also performed linear projections on x^{sex} and x^{age} to obtain encoded feature vectors X^{sex} and X^{age}. Subsequently, we concatenated $\{X^{\mathrm{cc}}_{1:N^{\mathrm{cc}}}, X^{\mathrm{lab}}_{1:N^{\mathrm{lab}}}, X^{\mathrm{sex}}, X^{\mathrm{age}}\}$ together to produce a sequence of clinical text tokens $X^{\mathrm{T}}_{1:\hat{N}}$, where $\hat{N} = N^{\mathrm{cc}} + N^{\mathrm{lab}} + 2$. In practice, we set N^{cc} and N^{lab} to 40 and 92, respectively.

As for the task of adverse clinical outcome prediction for patients with COVID-19, its multimodal input data also consisted of five parts: a set of CT slices, structured chief complaint (comorbidities and symptoms), laboratory test results, each patient's gender and age, which are denoted as $x^{\mathrm{I}}, x^{\mathrm{cc}}, x^{\mathrm{lab}}, x^{\mathrm{sex}}$ and x^{age}, respectively. Each CT slice was converted to a sequence of image patch tokens $X^{\mathrm{I}}_{1:N}$ as in the first task. Different from the first task, the chief complaint was structured. To convert x^{cc} to tokens, we conducted a shared linear projection to each component, which generated a sequence of embeddings $X^{\mathrm{cc}}_{1:N^{\mathrm{cc}}}$. A linear projection layer was applied to x^{lab} to acquire $X^{\mathrm{lab}}_{1:N^{\mathrm{lab}}}$. As for x^{sex} and x^{age}, we performed linear projections to obtain encoded X^{sex} and X^{age} as in the first task. Finally, we directly concatenated $\{X^{\mathrm{cc}}_{1:N^{\mathrm{cc}}}, X^{\mathrm{lab}}_{1:N^{\mathrm{lab}}}, X^{\mathrm{sex}}, X^{\mathrm{age}}\}$ to produce $\hat{N}$ clinical text tokens $X^{\mathrm{T}}_{1:\hat{N}}$, where $\hat{N} = N^{\mathrm{cc}} + N^{\mathrm{lab}} + 2$. We set N^{cc} and N^{lab} to 16 and 19, respectively.

The first two layers of MDT were two stacked bidirectional multimodal attention blocks. Suppose the input of the first bidirectional multimodal attention block consists of $\mathrm{X}^{l}_{\mathrm{I}}$ and $\mathrm{X}^{l}_{\mathrm{T}}$, where $\mathrm{l}\,(= 0)$ stands for the layer index, $X^{0}_{\mathrm{I}} = X^{\mathrm{I}}_{1:N}$ denotes the assembly of image patch tokens and $X^{0}_{\mathrm{T}} = X^{\mathrm{T}}_{1:\hat{N}}$ represents the bag of clinical text tokens. The process of generating the query, key and value matrices for each modality in the bidirectional multimodal attention block was as follows:

$$\mathrm{Q}_{\mathrm{I}}^{l},\mathrm{K}_{\mathrm{I}}^{l},\mathrm{V}_{\mathrm{I}}^{l}=\mathrm{LP}\left(\mathrm{Norm}\left(\mathrm{X}_{\mathrm{I}}^{l}\right)\right),$$

$$\mathrm{Q}_{\mathrm{T}}^{l},\mathrm{K}_{\mathrm{T}}^{l},\mathrm{V}_{\mathrm{T}}^{l}=\mathrm{LP}\left(\mathrm{Norm}\left(\mathrm{X}_{\mathrm{T}}^{l}\right)\right),$$

where $\mathrm{LP}(\cdot)$ and $\mathrm{Norm}(\cdot)$ represent linear projection and layer normalization, respectively. The forward pass inside a bidirectional multimodal attention block could be summarized as:

$$\mathfrak{x}_{\mathrm{I}}^{l}=\mathrm{Attention}\left(Q_{\mathrm{I}}^{l},K_{\mathrm{I}}^{l},V_{\mathrm{I}}^{l}\right)+\lambda\,\mathrm{Attention}\left(Q_{\mathrm{I}}^{l},K_{\mathrm{T}}^{l},V_{\mathrm{T}}^{l}\right),$$

$$\mathfrak{x}_{\mathrm{T}}^{l}=\mathrm{Attention}\left(Q_{\mathrm{T}}^{l},K_{\mathrm{T}}^{l},V_{\mathrm{T}}^{l}\right)+\lambda\,\mathrm{Attention}\left(Q_{\mathrm{T}}^{l},K_{\mathrm{I}}^{l},V_{\mathrm{I}}^{l}\right),$$

where $\mathrm{Attention}\left(Q_{\mathrm{I}}^{l},K_{\mathrm{I}}^{l},V_{\mathrm{I}}^{l}\right)$ and $\mathrm{Attention}\left(Q_{\mathrm{T}}^{l},K_{\mathrm{T}}^{l},V_{\mathrm{T}}^{l}\right)$ capture the intramodal connections in the image and text modalities, respectively. $\mathrm{Attention}\left(Q_{\mathrm{I}}^{l},K_{\mathrm{T}}^{l},V_{\mathrm{T}}^{l}\right)$ and $\mathrm{Attention}\left(Q_{\mathrm{T}}^{l},K_{\mathrm{I}}^{l},V_{\mathrm{I}}^{l}\right)$ dig out the intermodal connections between the image and text. Next, both intra- and intermodal connections were encoded into latent representations $\mathfrak{x}_{I}^{l}$ and $\mathfrak{x}_{T}^{l}$. We set λ to 1.0 as it gave rise to the best performance in our preliminary experiments. Attention (Q, K, V) included two matrix multiplications (mat. mul.) and one scaled softmax operation:

$$\mathrm{Attention}\,(Q,K,V)=\mathrm{softmax}(\frac{QK^{\top}}{\sqrt{d_k}}V),$$

where $\top$ stands for the matrix transpose operator, d_k is a scaling hyper-parameter, which was set to 64. Next, we introduced residual learning[45] and forwarded the resulting $\mathfrak{x}_{\mathrm{I}}^{l}$, $\mathfrak{x}_{\mathrm{T}}^{l}$ to the following normalization layer and MLP:

$$X_{\mathrm{I}}^{l+1}=\mathrm{MLP}\left(\mathrm{Norm}\left(\mathfrak{x}_{\mathrm{I}}^{l}\right)\right)++X_{\mathrm{I}}^{l},$$

$$X_{\mathrm{T}}^{l+1}=\mathrm{MLP}\left(\mathrm{Norm}\left(\mathfrak{x}_{\mathrm{T}}^{l}\right)\right)++X_{\mathrm{T}}^{l},$$

where X_{I}^{l+1} and X_{T}^{l+1} were passed to the next bidirectional multimodal attention block as the input, resulting in X_{I}^{l+2} and X_{T}^{l+2}. Then, we combined tokens in X_{I}^{l+2} and X_{T}^{l+2} to produce a bag of unified tokens, which were passed to the subsequent self-attention blocks[24]. We also allocated multiple heads[24] in both bidirectional multimodal attention and self-attention blocks, where the number of heads was set to 12. This multihead mechanism allowed the model to perform attention operations in multiple representation subspaces simultaneously and aggregate the results afterwards.

Lastly, we applied average pooling to the unified tokens generated from the last self-attention block to obtain a holistic multimodal representation for medical diagnosis. This representation was passed to a two-layer MLP to produce final prediction logits. During the training stage, we calculated the binary cross-entropy loss between these logits and their corresponding pulmonary disease annotations (the first task) or real adverse clinical outcomes (the second task). A loss function value was computed for every patient case. Specifically, in the first task, each patient case contained one radiograph and related textual clinical information. In the second task, each patient case involved multiple CT slices, and these CT slices shared the same textual clinical information. We forwarded each CT slice and its accompanying textual clinical information to MDT to obtain one holistic representation. Since we had multiple CT slices, we obtained a number of holistic representations (equal to the number of CT slices) for the same patient. Then, we performed average pooling over these holistic representations to compute an averaged representation, which was finally passed to a two-layer MLP and the binary cross-entropy loss.

Implementation details

For the pulmonary disease identification task, we first resized each radiograph to 256 × 256 pixels during the training stage, then cropped a random portion of each image, where the area ratio between the cropped patch and the original radiograph was randomly determined to be between 0.09 and 1.0. The cropped patch was resized to 224 × 224, after which a random horizontal flip was applied to increase the diversity of training data. In the validation and testing stages, each radiograph was first resized to 256 × 256 pixels, and then a square patch at the image centre was cropped. The size of the square crop was 224 × 224. The processed radiographs were finally passed to the image-only model, non-unified-chest, Perceiver and IRENE as input images. In the task of adverse clinical outcome prediction for patients with COVID-19, the input images were CT scans. We first used the lesion detection and segmentation methodologies proposed in ref. 46. This is a deep learning algorithm based on a multiview feature pyramid convolutional neural network[47,48], which performs lesion detection, segmentation and localization. This neural network was trained and validated on 14,435 participants with chest CT images and definite pathogen diagnosis. On a per-patient basis, the algorithm showed superior sensitivity of 1.00 (95% CI: 0.95, 1.00) and an F1-score of 0.97 in detecting lesions from CT images of patients with COVID-19 pneumonia. Adverse clinical outcomes of COVID-19 were presumed to be closely related to the characteristics of pneumonia lesion areas. For each patient's case, we cropped a 3D CT subvolume by computing the minimum 3D bounding box enclosing all pneumonia lesions. Next, we resized all 3D subvolumes from different patients to a uniform size, which was 224 × 224 × 64. Lastly, we sampled 16 evenly spaced slices from every 3D subvolume along its third dimension.

Before we performed the formal training procedure, we pre-trained our MDT on MIMIC-CXR[42], as what was done for the baseline models. Similar to Perceiver, during pre-training, we used zero padding for non-existent textual clinical information in every multimodal input. In the formal training stage, we used AdamW[49] as the default optimizer as we found empirically that it gave better performance on baseline models and IRENE. The initial learning rate was set to 3×10^{-5} and the weight decay was 1×10^{-2}. We trained each model for 30 epochs and decreased the initial learning rate by a factor of 10 at the 20th epoch. The batch size was set to 256 in the training stage of both tasks. It is worth noting that in the task of adverse clinical outcome prediction of COVID-19, we first extracted holistic feature representations from 16 CT slices (cropped and sampled from the same CT volume). Next, we applied average pooling to these 16 holistic features to obtain an averaged representation, which represented all pneumonia lesion areas in the entire CT volume. The binary cross-entropy loss was then computed on top of this averaged representation. During the training stage, we evaluated model performance on the validation set and calculated the validation loss after each epoch. The model checkpoint that produced the lowest validation loss was saved and then tested on the testing set. We employed learnable positional embeddings in all ViT models. IRENE was implemented using PyTorch[50] and the training stage was accelerated using NVIDIA Apex with the mixed-precision strategy[51]. In practice, we can finish the training stage of either task within 1 d using four NVIDIA GPUs.

We adopted the standard attention analysis strategy for vision transformers. For each layer in the transformer, we averaged the attention weights across multiple heads (as we used multihead self-attention in IRENE) to obtain an attention matrix. To account for residual connections, we added an identity matrix to each attention matrix and normalized the resulting weight matrices. Next, we recursively multiplied the weight matrices from different layers of the transformer. Finally, we obtained an attention map that included the similarity between every input token and the CLS token. Since the CLS token was used for diagnostic predictions, these similarities indicated the relevance between the input tokens and prediction results, which could then be used for visualization. For cross-attention results, we performed visualization with Grad-CAM[52].

Non-parametric bootstrap sampling was used to calculate 95% confidence intervals. Specifically, we repeatedly drew 1,000 bootstrap samples from the unseen test set. Each bootstrap sample was obtained through random sampling with replacement, and its size was the same as the size of the test set. We then computed AUROC (the first task) or AUPRC (the second task) on each bootstrap sample, after which we had 1,000 AUROC or AUPRC values. Finally, we sorted these performance results and report the values at 2.5 and 97.5 percentiles, respectively.

To demonstrate the statistical significance of our experimental results, we first repeated the experiments for IRENE and the best performing baseline (that is, Perceiver) five times with different random seeds. Then, we used independent two-sample *t*-test (two-sided) to compare the mean performance of IRENE and the best baseline results, and calculate *P* values.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Restrictions apply to the availability of the developmental and validation datasets, which were used with permission of the participants for the current study. De-identified data may be available for research purposes from the corresponding authors on reasonable request.

Code availability

The custom code is available at https://github.com/RL4M/IRENE.

References

1. He, J. et al. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **25**, 30–36 (2019).
2. Liang, H. et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat. Med.* **25**, 433–438 (2019).
3. Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J. & Shah, S. P. Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* **22**, 114–126 (2022).
4. Li, J., Shao, J., Wang, C. & Li, W. The epidemiology and therapeutic options for the COVID-19. *Precis. Clin. Med.* **3**, 71–84 (2020).
5. Comfere, N. I. et al. Provider-to-provider communication in dermatology and implications of missing clinical information in skin biopsy requisition forms: a systematic review. *Int. J. Dermatol.* **53**, 549–557 (2014).
6. Shao, J. et al. Radiogenomic system for non-invasive identification of multiple actionable mutations and PD-L1 expression in non-small cell lung cancer based on CT images. *Cancers* **14**, 4823 (2022).
7. Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digit. Med.* **3**, 136 (2020).
8. Wang, C. et al. Non-invasive measurement using deep learning algorithm based on multi-source features fusion to predict PD-L1 expression and survival in NSCLC. *Front. Immunol.* **13**, 828560 (2022).
9. Zhang, K. et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell* **181**, 1423–1433.e11 (2020).
10. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131.e29 (2018).
11. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
12. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
13. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
14. Wang, G. et al. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. *Nat. Biomed. Eng.* **5**, 509–521 (2021).
15. Zhou, H. Y. et al. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nat. Mach. Intell.* **4**, 32–40 (2022).
16. Tang, Y. X. et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digit. Med.* **3**, 70 (2020).
17. Wang, C. et al. Development and validation of an abnormality-derived deep-learning diagnostic system for major respiratory diseases. *npj Digit. Med.* **5**, 124 (2022).
18. Rajpurkar, P. et al. ChexNet: radiologist-level pneumonia detection on chest x-rays with deep learning. Preprint at https://arxiv.org/abs/1711.05225v3 (2017).
19. Mei, X. et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nat. Med.* **26**, 1224–1228 (2020).
20. Yala, A., Lehman, C., Schuster, T., Portnoi, T. & Barzilay, R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **292**, 60–66 (2019).
21. Zhang, K. et al. Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images. *Nat. Biomed. Eng.* **5**, 533–545 (2021).
22. Xu, Q. et al. AI-based analysis of CT images for rapid triage of COVID-19 patients. *npj Digit. Med.* **4**, 75 (2021).
23. Akselrod-Ballin, A. et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology* **292**, 331–342 (2019).
24. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
25. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at https://arxiv.org/abs/1810.04805v2 (2018).
26. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. Preprint at https://arxiv.org/abs/2010.11929v2 (2020).
27. LeCun, Y. et al. Handwritten digit recognition with a back-propagation network. *Adv. Neural Inf. Process. Syst.* **2**, 396–404 (1989).
28. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. Preprint at https://arxiv.org/abs/1301.3781v3 (2013).
29. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **26**, 3111–3119 (2013).
30. Jaegle, A. et al. Perceiver: general perception with iterative attention. In *Proc. 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) 4651–4663 (PMLR, 2021).
31. Li, J. et al. Align before fuse: vision and language representation learning with momentum distillation. *Adv. Neural Inf. Process. Syst.* **34**, 9694–9705 (2021).
32. Su, W. et al. VL-bert: pre-training of generic visual-linguistic representations. Preprint at https://arxiv.org/abs/1908.08530v4 (2020).

33. Wang, J. et al. GIT: A generative image-to-text transformer for vision and language. Preprint at https://arxiv.org/abs/2205.14100v5 (2022).
34. Pauwels, R. A., Buist, A. S., Calverley, P. M., Jenkins, C. R. & Hurd, S. S. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop summary. *Am. J. Respir. Crit. Care Med.* **163**, 1256–1276 (2001).
35. Li, Y. et al. BEHRT: transformer for electronic health records. *Sci. Rep.* **10**, 7155 (2020).
36. Xia, K. & Wang, J. Recent advances of transformers in medical image analysis: a comprehensive review. *MedComm Futur. Med.* **2**, e38 (2023).
37. Wang, D., Feng, L., Ye, J., Zou, J. & Zheng, Y. Accelerating the integration of ChatGPT and other large-scale AI models into biomedical research and healthcare. *MedComm-Future Med.* **2**, e43 (2023).
38. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
39. Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer normalization. Preprint at https://arxiv.org/abs/1607.06450v1 (2016).
40. Hendrycks, D. & Gimpel, K. Gaussian error linear units (GELUs). Preprint at https://arxiv.org/abs/1606.08415 (2016).
41. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
42. Johnson, A. E. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
43. Jaegle, A. et al. Perceiver IO: a general architecture for structured inputs & outputs. Preprint at https://arxiv.org/abs/2107.14795v1 (2021).
44. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
45. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
46. Ni, Q. et al. A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images. *Eur. Radiol.* **30**, 6517–6527 (2020).
47. Li, Z. et al. in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019* (eds Shen, D. et al.) 13–21 (Springer, 2019).
48. Zhao, G. et al. Diagnose like a radiologist: hybrid neuro-probabilistic reasoning for attribute-based medical image diagnosis. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7400–7416 (2022).
49. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Preprint at https://arxiv.org/abs/1711.05101 (2017).
50. Paszke, A. et al. Pytorch: an imperative style, high-performance deep learning library. *Adv. Neural. Inf. Process. Syst.* **32**, 8026–8037 (2019).
51. Micikevicius, P. et al. Mixed precision training. Preprint at https://arxiv.org/abs/1710.03740 (2017).
52. Selvaraju, R. R. et al. Grad-cam: visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision* 618–626 (IEEE, 2017).

Acknowledgements

This research was supported by the National Natural Science Foundation of China (82100119, 92159302, 91859203), Hong Kong Research Grants Council through General Research Fund (Grant 17207722), Macau Science and Technology Development Fund, Macau (0007/2020/AFJ, 0070/2020/A2, 0003/2021/AKP).

Author contributions

H.-Y.Z., Y.Y., K.Z. and W.L. conceived the idea and designed the experiments. H.-Y.Z., C.W. and S.Z. implemented and performed the experiments. H.-Y.Z., Y.Y., C.W., S.Z., J.P., J.S., Y.G., G.L., K.Z. and W.L. analysed the data and experimental results. H.-Y.Z., Y.Y., C.W., Y.G., K.Z. and W.L. wrote the paper. All authors commented on the paper.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41551-023-01045-x.

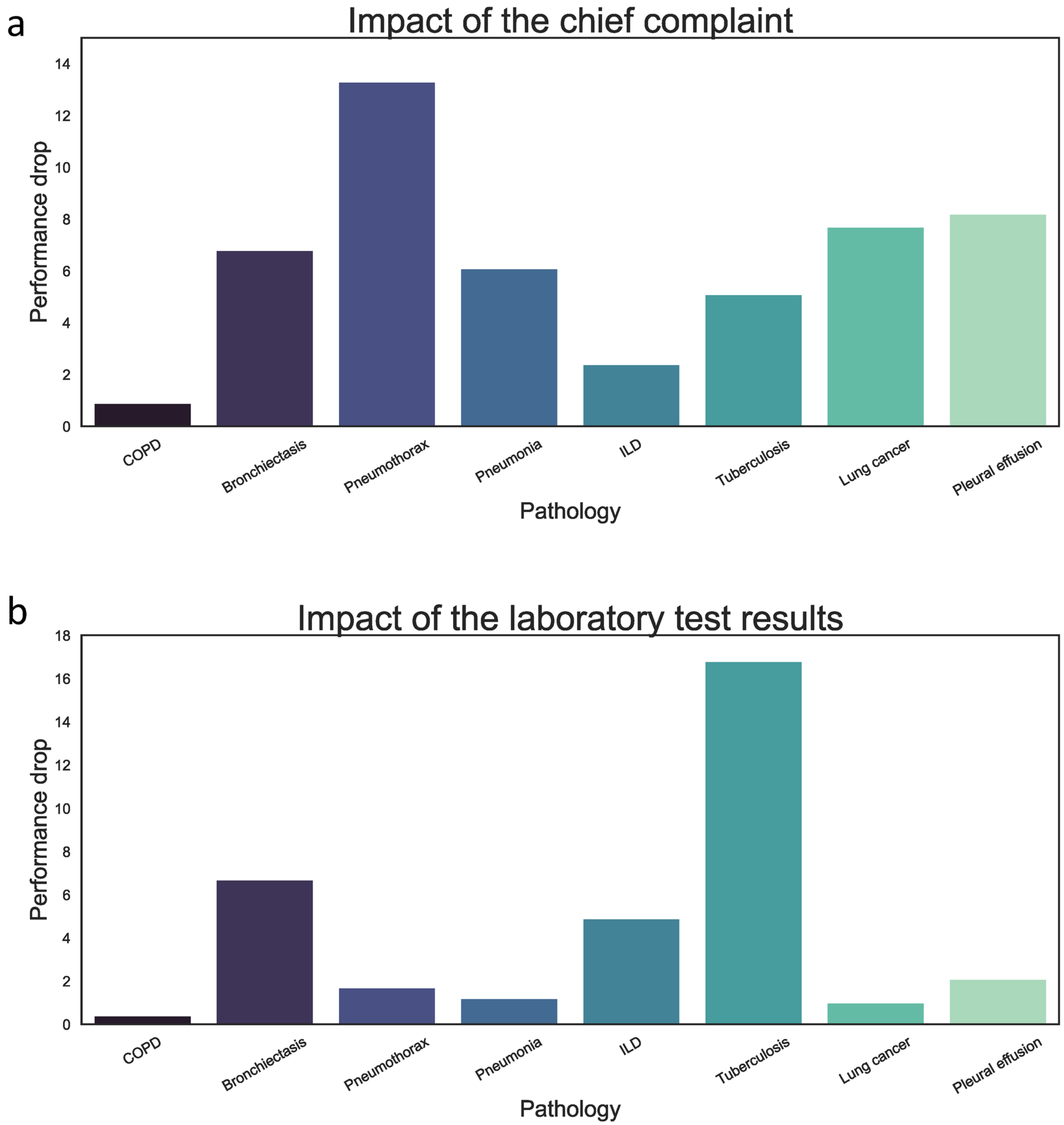
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41551-023-01045-x.

Correspondence and requests for materials should be addressed to Yizhou Yu, Chengdi Wang, Kang Zhang or Weimin Li.

Peer review information *Nature Biomedical Engineering* thanks Jong Chul Ye, Pranav Rajpurkar and Dawei Yang for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Impact of the chief complaint (a) or laboratory test results (b) on each respiratory disease. Specifically, we remove either the chief complaint or the laboratory test results from the input and report the performance drop on each disease. The evaluation metric is AUROC.

nature portfolio

Corresponding author(s):	Kang Zhang
Last updated by author(s):	Apr 18, 2023

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	No special software or code was used to collect the data.
Data analysis	Python 3.7.6, PyTorch 1.8.1, and CUDA 11.4. More information is provided at https://github.com/RL4M/IRENE.

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

Restrictions apply to the availability of the developmental and validation datasets, which were used with permission of the participants for the current study. Deidentified data may be available for research purposes from the corresponding authors on reasonable request.

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	Information on sex was collected and involved in the input to the neural network. Reported findings apply to both sexes. The sex of each patient was based on self-reporting. Sex- and gender-based analyses were not performed because they are not relevant to this study.
Reporting on race, ethnicity, or other socially relevant groupings	Information on race, ethnicity, or other socially relevant groupings was not collected.
Population characteristics	We retrospectively collected consecutive chest X-rays from 51,511 patients between November 27, 2008, and May 31, 2019, at West China Hospital, which is the largest tertiary medical center in western China covering a 100 million population. Each patient is associated with at least one radiograph, a short piece of unstructured chief complaint, history of present and past illness, demographics, and a complete laboratory test report. The dataset is built for eight pulmonary diseases, including chronic obstructive pulmonary disease (COPD), bronchiectasis, pneumothorax, pneumonia, interstitial lung disease (ILD), tuberculosis, lung cancer, and pleural effusion. The built dataset consists of 72,283 data samples, among which 40,126 samples are normal. The distribution of diseases (i.e., the number of relevant cases) is as follows: COPD (4,912), bronchiectasis (676), pneumothorax (2,538), pneumonia (21,409), ILD (3,283), tuberculosis (938), lung cancer (2,651) and pleural effusion (4,713). The performance metric is the area under the receiver operating characteristic curve (AUROC). We split this dataset into training, validation, and testing sets according to each patient's admission date. Specifically, the training set includes 44,628 patients admitted between November 27, 2008, and June 1, 2018. And the validation set includes 3,325 patients admitted between June 2, 2018 and December 01, 2018. Finally, the trained and validated IRENE system is tested on 3,558 patients admitted between December 02, 2018 and May 31, 2019. The second dataset MMC (i.e., multi-modal COVID-19 dataset), on which IRENE is trained and evaluated, consists of chest CT images and structured clinical information (e.g., chief complaint that comprises comorbidities and symptoms, demographics, laboratory test results, etc) collected from COVID-19 patients. The CT images are associated with inpatients with laboratory-confirmed COVID-19 infection between December 27, 2019 and March 31, 2020. There are three types of adverse events that could happen to patients in MMC, which are admission to ICU, mechanical ventilation (MV), and death. The training and validation sets came from 17 hospitals, and the training set has 1,164 labeled cases (70%) while the validation set has 498 labeled ones (30%). Next, we chose the trained model with the best performance on the validation set and test it on the independent testing set, which is comprised of 700 cases collected from 9 external medical centers. The distribution of the three events in the testing set is as follows: ICU (155), MV (94), Death (59).
Recruitment	No participants were recruited.
Ethics oversight	Institutional Review Board (IRB)/Ethics Committees approvals were obtained from West China Hospital and all participating hospitals. All patients signed a consent form. The research was conducted in a manner compliant with the United States Health Insurance Portability and Accountability Act (HIPAA). It was adherent to the tenets of the Declaration of Helsinki and in compliance with the Chinese CDC policy on reportable infectious diseases and the Chinese Health and Quarantine Law.

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

Life sciences study design

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

Sample size	The sample sizes for the pulmonary disease identification and adverse outcome prediction of COVID-19 are 72,283 and 2,362, respectively.
Data exclusions	No data were excluded
Replication	Replication was not relevant. We used independent validation cohorts.
Randomization	Samples were randomly allocated to the training, tuning and testing sets.
Blinding	During image processing, all images were first de-identified to remove any patient related information.

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