The Innovation, Volume *5*

Supplemental Information

A multimodal integration pipeline for accurate diagnosis, pathogen identification, and prognosis prediction of pulmonary infections

Jun Shao, Jiechao Ma, Yizhou Yu, Shu Zhang, Wenyang Wang, Weimin Li, and Chengdi Wang

Supplemental Information

A multimodal integration pipeline for accurate diagnosis, pathogen identification, and prognosis prediction of pulmonary infections

Jun Shao, Jiechao Ma, Yizhou Yu, Shu Zhang, Wenyang Wang, Weimin Li, Chengdi Wang

Table of Contents

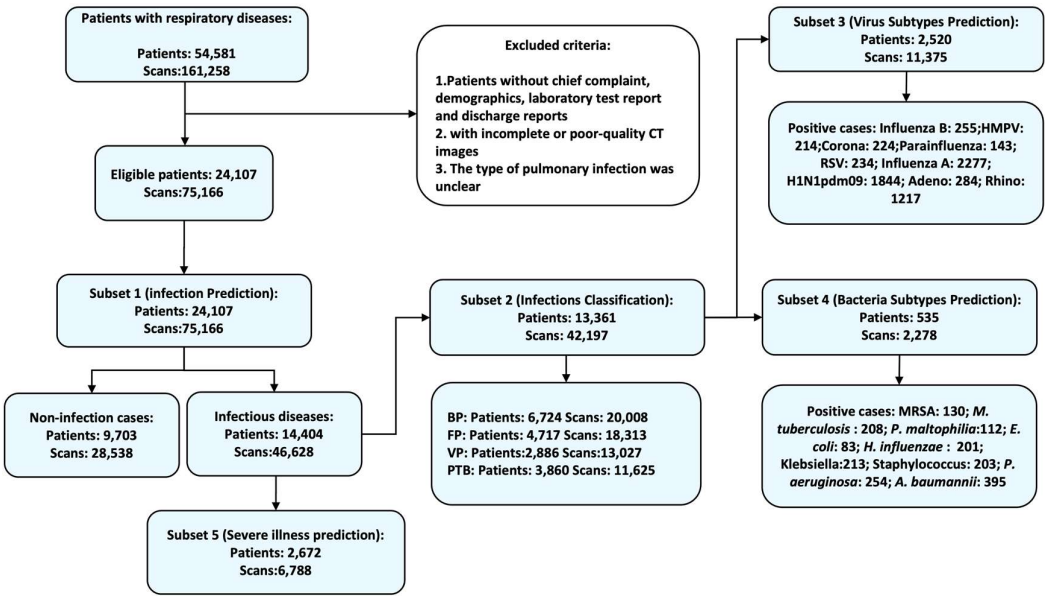

Figure S1. Overview of patient selection and data categorization. The study flow diagram presented the screening and categorization of patients with respiratory conditions at West China Hospital of Sichuan University and Chengdu ShangJin Nanfu Hospital. The process involved exclusion criteria application, resulting in 24,107 eligible patients. Data were subdivided for various analyses: primary prediction of respiratory diseases, classification of infections, virus and bacteria prediction, and severe pneumonia identification.

Abbreviations: *A.baumannii*, *Acinetobacter baumannii*; BP, bacterial pneumonia; *E.coli*, *Escherichia coli*; FP, fungal pneumonia; HMPV, human metapneumovirus; *H.influenzae*, *Haemophilus influenzae*; MRSA, Methicillin-resistant *Staphylococcus aureus*; *M. tuberculosis*, *Mycobacterium tuberculosis*; *P.maltophilia*, *Pseudomonas maltophilia*; *P.aeruginosa*, *Pseudomonasaeruginosa*; PTB, pulmonary tuberculosis; RSV, respiratory syncytial virus; VP, viral pneumonia.

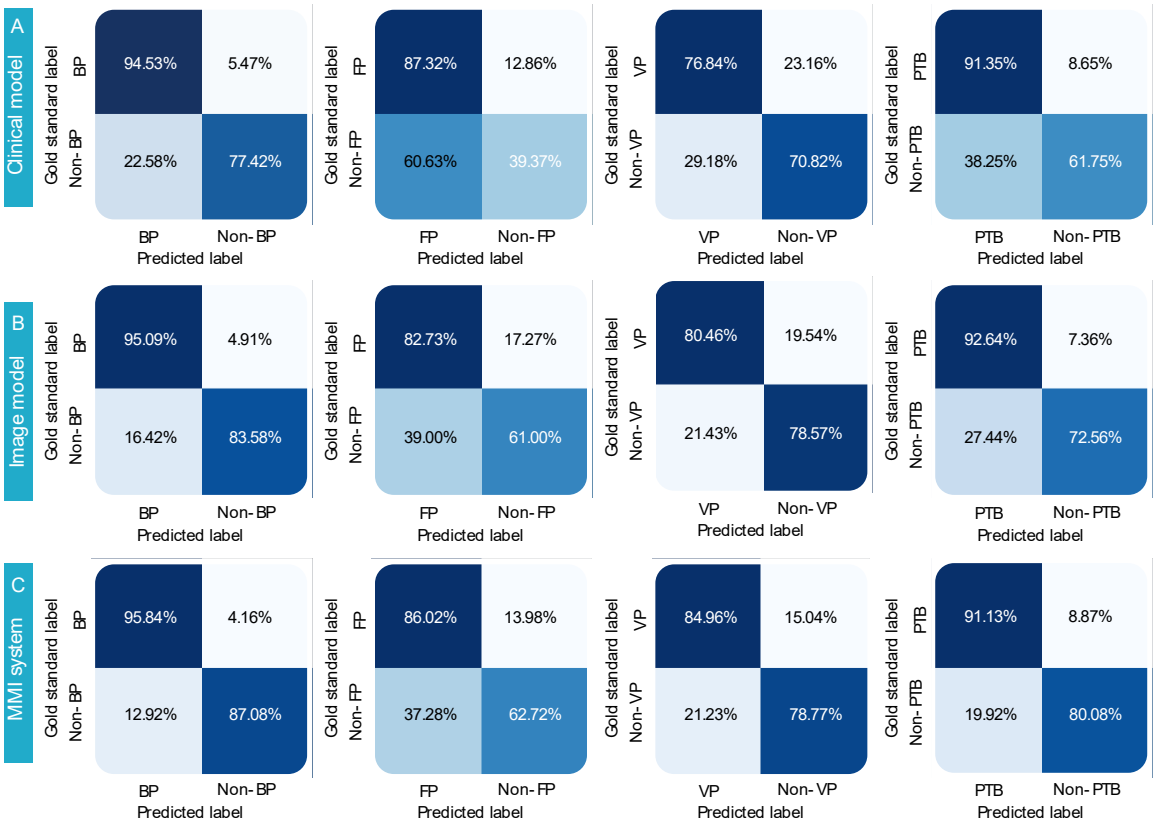

Figure S2 | Performance of the MMI system for identifying four categories pneumonia in the internal testing sets. A-C, The confusion matrix for identifying pneumonia based on clinical model (A), image model (B) and MMI system (C) in the internal testing set.

Abbreviations: BP, bacterial pneumonia; FP, fungal pneumonia; MMI, multimodal integration; PTB, pulmonary tuberculosis; VP, viral pneumonia.

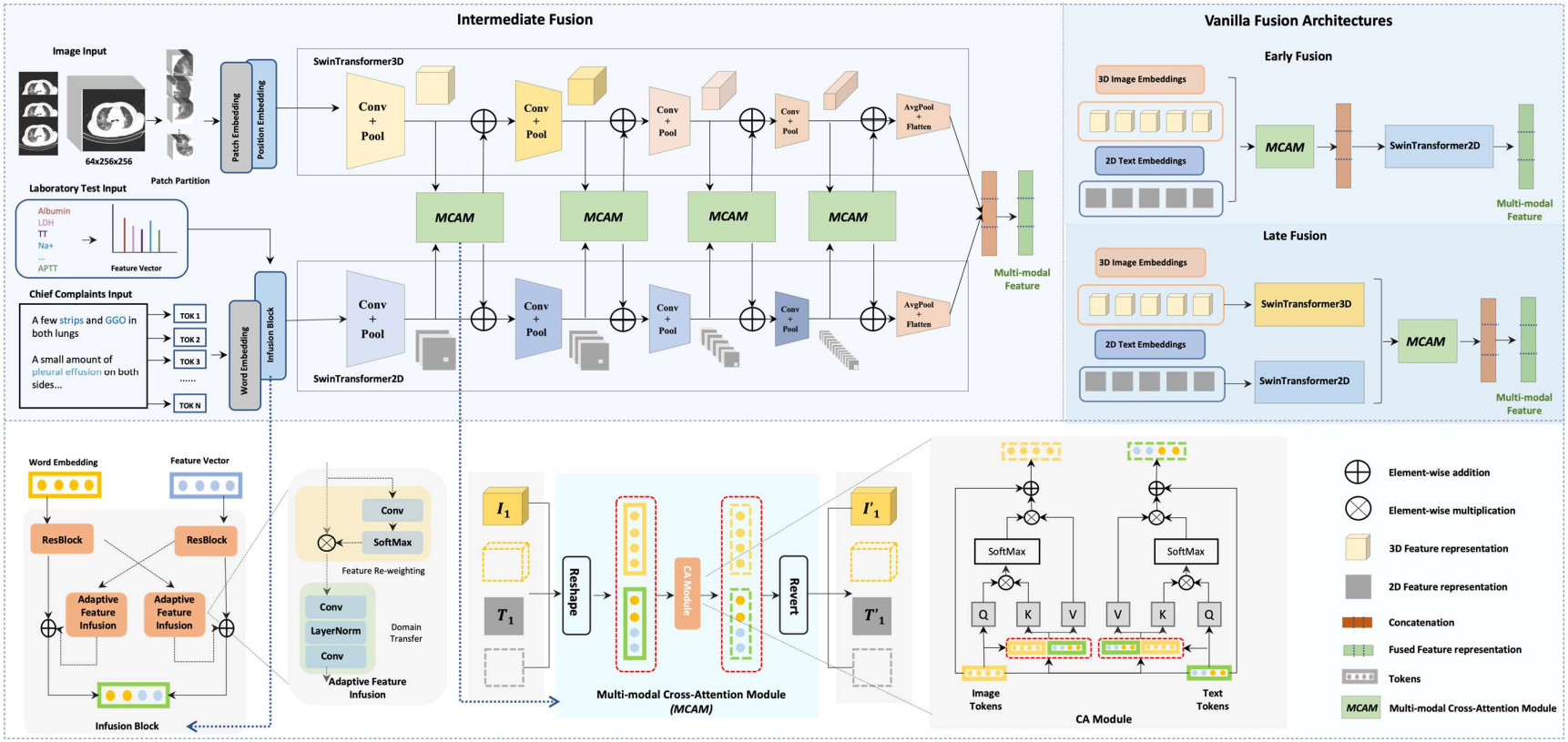

Figure S3 | Multimodal data fusion architecture. **Upper left panel:** Pre-processing and feature extraction stages for image, laboratory, and clinical text data inputs using Swin-Transformers and convolutional operations. **Upper right panel:** Depiction of the fusion process utilizing the Multimodal Cross-Attention Modules (MCAM), comparing early and late fusion methodologies. **Lower panel:** Detailed internal structure of the MCAM.

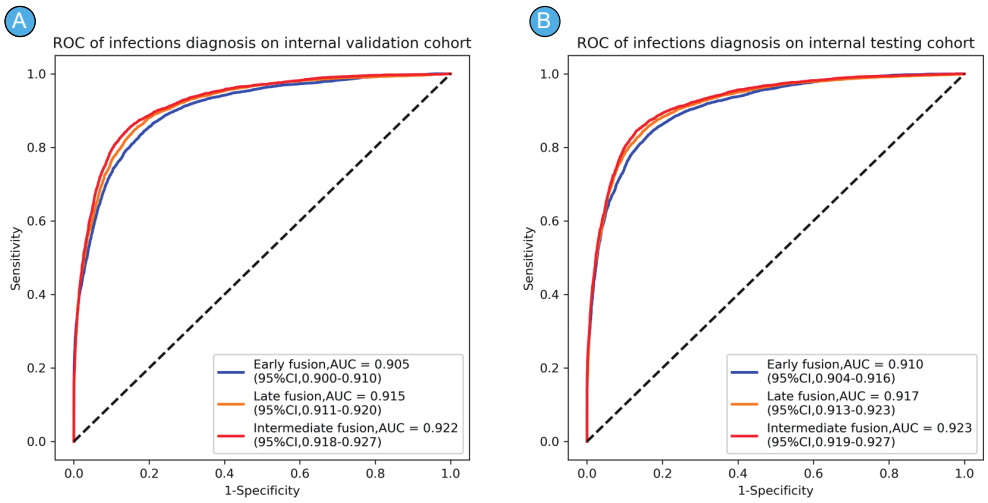

Figure S4 | Performance of different fusion methods in the validation and internal testing datasets. A-B, The ROC curves of different fusion methods for identifying pulmonary infections in internal validation cohort (**A**) and internal testing cohort (**B**).

Supplementary Table 1. Summary of clinical characteristics of enrolled patients for the training, validation, internal testing and external testing datasets.

	Subset 1 (N=24,107)			
Demographics	**Training (N=19,046)**	**Validation (N=2,432)**	**Internal Testing (N=2,433)**	**External Testing (N=196)**
Age (years)	55.53±18.99	56.21±19.00	56.09±18.72	57.71±19.43
Sex (male)	11,620(61.0%)	1,449(59.6%)	1,464(60.2%)	112(57.1%)
Scans	59,490	7,497	7,554	625
Infections				
Yes	11,439(60.1%)	1,389(57.1%)	1,422(58.4%)	154(78.6%)
No	7,607(39.9%)	1,043(42.9%)	1,011(41.6%)	42(21.4%)
	Subset 2 (N=13,361)			
Demographics	**Training (N=10,578)**	**Validation (N=1,325)**	**Internal Testing (N=1,325)**	**External Testing (N=133)**
Age (years)	52.33±20.07	52.67±20.07	52.34±20.03	53.13±19.12
Sex (male)	6,391(60.0%)	801(60.5%)	774(58.4%)	79(59.4%)
Scans	33,411	4,267	4,094	425
Infections types				
BP	5,278(49.5%)	686(51.8%)	688(51.9%)	72(54.1%)
FP	3,702(35.7%)	484(36.5%)	480(36.2%)	51(38.3%)
VP	2,321(21.8%)	266(20.1%)	270(20.4%)	29(21.8%)
PTB	3,103(29.1%)	370(27.9%)	355(26.8%)	32(24.1%)

Abbreviations: BP, bacterial pneumonia; FP, fungal pneumonia; VP, viral pneumonia; PTB, pulmonary tuberculosis.

Supplementary Table 2. Performance of MMI system in identifying pulmonary infections.

	Datasets	Sensitivity (95%CI)	Specificity (95%CI)	Accuracy (95%CI)	AUC (95%CI)
Clinical model	Validation	0.808(0.797–0.818)	0.765(0.754–0.779)	0.782(0.775–0.790)	0.868(0.861–0.875)
	Internal testing	0.787(0.776–0.798)	0.795(0.784–0.806)	0.792(0.784–0.800)	0.879(0.870–0.885)
	External testing	0.624(0.582–0.685)	0.775(0.732–0.822)	0.692(0.665–0.735)	0.770(0.737–0.815)
Image model	Validation	0.852(0.841–0.862)	0.831(0.819–0.842)	0.835(0.828–0.842)	0.918(0.913–0.923)
	Internal testing	0.845(0.836–0.855)	0.848(0.839–0.857)	0.836(0.830–0.842)	0.926(0.922–0.930)
	External testing	0.770(0.721–0.815)	0.777(0.732–0.821)	0.759(0.736–0.791)	0.830(0.792–0.867)
MMI system	Validation	0.864(0.855–0.872)	0.840(0.829–0.849)	0.846(0.839–0.852)	0.930(0.925–0.934)
	Internal testing	0.866(0.857–0.874)	0.838(0.829–0.848)	0.849(0.844–0.855)	0.935(0.932–0.939)
	External testing	0.852(0.813–0.889)	0.853(0.814–0.891)	0.919(0.898–0.937)	0.888(0.856–0.916)

Supplementary Table 3. Performance of MMI system in identifying single infection and mixed infections.

	Datasets	Sensitivity (95%CI)	Specificity (95%CI)	Accuracy (95%CI)	AUC (95%CI)
Single infection	Internal testing	0.864(0.841–0.890)	0.915(0.906–0.924)	0.904(0.895–0.912)	0.949(0.943–0.954)
Mixed infections	Internal testing	0.864(0.836–0.896)	0.765(0.728–0.805)	0.852(0.837–0.868)	0.876(0.861–0.890)

Supplementary Table 4. Performance of MMI system in identifying various pulmonary infections based on different fusion methods.

	Datasets	Sensitivity (95%CI)	Specificity (95%CI)	Accuracy (95%CI)	AUC (95%CI)
Early fusion	Validation	0.836(0.820–0.853)	0.850(0.840–0.861)	0.846(0.838–0.855)	0.905(0.900–0.910)
	Internal testing	0.846(0.830–0.860)	0.847(0.837–0.858)	0.848(0.838–0.856)	0.910(0.904–0.916)
Intermediate fusion	Validation	0.852(0.838–0.868)	0.879(0.869–0.889)	0.870(0.862–0.879)	0.922(0.918–0.927)
	Internal testing	0.849(0.833–0.863)	0.882(0.874–0.892)	0.870(0.862–0.879)	0.923(0.919–0.927)
Late fusion	Validation	0.867(0.852–0.883)	0.842(0.832–0.853)	0.851(0.843–0.860)	0.915(0.911–0.920)
	Internal testing	0.849(0.834–0.863)	0.865(0.856–0.875)	0.859(0.852–0.868)	0.917(0.913–0.923)

Supplementary Table 5. Weighted error results of the MMI system vs. physicians in diagnosing pulmonary infections.

Weighted errors	Junior physicians	Senior physicians	MMI system
Mean	24.10%	8.98%	13.52%
Physician 1	23.03%	11.51%	-
Physician 2	25.17%	6.45%	-

Supplementary Table 6. Performance of different architectures in identifying pulmonary infections.

	Sensitivity (95%CI)	Specificity (95%CI)	Accuracy (95%CI)	AUC (95%CI)
ResNet	0.793(0.782–0.806)	0.787(0.776–0.799)	0.786(0.778–0.794)	0.873(0.867–0.879)
DenseNet	0.909(0.902–0.917)	0.745(0.735–0.753)	0.803(0.797–0.809)	0.881(0.877–0.886)
Swin-Transformer	0.832(0.822–0.842)	0.849(0.840–0.859)	0.829(0.823–0.836)	0.927(0.923–0.931)
Swin-Transformer with cross-shaped	0.866(0.857–0.874)	0.838(0.829–0.848)	0.849(0.844–0.855)	0.935(0.932–0.939)

MATERIALS AND METHODS

Data acquisition

In this study, a comprehensive analysis was conducted utilizing data from hospitalized inpatients who were admitted to West China Hospital (WCH) of Sichuan University and Chengdu ShangJin Nanfu Hospital (CSJH). The inclusion criteria were as follows: (1) over the age of 18 years old; (2) with clear diagnosis regarding the presence or absence of pulmonary infection; (3) with complete medical information, inclusive of chest CT scans. The exclusion criteria were as follows: (1) patients without chief complaint, demographics, laboratory test reports and discharge reports; (2) with incomplete or poor-quality CT images, such as scans < 25 slices, motion artifacts or significant resolution reductions; (3) the type of pulmonary infection was unclear. The studies involving human participants were reviewed and approved by the Institutional Review Board and Ethics Committee of West China Hospital.

The dataset consisted of CT images acquired in the axial direction at a resolution of 512×512 pixels. The slice spacing varied ranged from 0.625 to 5 mm. These images were procured utilizing apparatuses furnished by illustrious entities such as Philips, GE Healthcare, United Imaging, and Siemens Healthineers. During the CT examinations, a tube voltage of 120 kilovolts peak (kVp) was consistently employed. To optimize image quality and minimize radiation exposure, an automatic tube current modulation technique was employed to modulate the tube current. The range of the tube currents used was 30 to 70 milliamperes (mAs). A stringent quality control procedure was implemented to ensure the integrity and reliability of the collected data.

Pre-processing

Furthermore, to ensure uniformity and enhance the quality of the CT scans, standardized image pre-processing protocols were instituted. These corrective interventions were enacted to attenuate any potential variations or biases resulting from the imaging process or equipment used. This study adopted a two-step process for analysing the CT scans obtained during the same patient admission, with a specific focus on the chest sequences. First, an evaluation of the convolution kernel utilized to fabricate each set of CT scans was conducted. This analysis aimed to elucidate and compensate for the variations or disparities stemming from the specific convolution kernel utilized. To ensure optimal resolution, all radiographs were initially screened, eliminating low-quality scans or discontinuities. Subsequently, all the continuous DICOM sequences were merged to generate a cohesive three-dimensional (3D) volume representation of the scans.[75] This merging process allowed the consolidation of

multiple sequences into a single comprehensive dataset. To meet the input requirements of the model, the dimensions of the resulting 3D volume were modified to 64 × 256 × 256. This resizing ensured compatibility and consistency across all the scans. By adhering to these protocols, the objective was to standardize the data and prepare it for further analysis, thereby guaranteeing that the input to the model remained uniform while focusing on the chest region.

In contrast, clinical text data of each patient were extensively collected. This comprehensive dataset encompassed a myriad of aspects concerning patient health records. Basic demographic information was assembled such as age, sex, and the highest body temperature recorded at the time of admission. Furthermore, the chief complaints reported by the patients upon admission were diligently documented providing insight into their specific symptoms or concerns. This information provided a rich contextual backdrop for analyzing their health conditions. In addition to the patient demographic details and chief complaints, their laboratory test results were also collected. These laboratory test results covered various markers pertaining to different aspects of health evaluation. For instance, liver biochemical markers, including albumin, serum lactic dehydrogenase (LDH), and indirect bilirubin were recorded. Moreover, coagulation markers such as thrombin time (TT), activated partial thromboplastin time (APTT), and platelet count were analyzed. These markers provided insights into the blood coagulation abilities and potential clotting disorders. To acquire a holistic view of the patients' health status, electrolyte and acid-base balance markers were also recorded such as Na^+, K^+, and HCO_3^-. These markers were instrumental in assessing the patients' overall electrolyte levels and acid-base equilibrium. To assess the inflammatory response, inflammatory markers were incorporated into the dataset. These included C-reactive protein (CRP) level, white blood cell count, lymphocyte count, and neutrophil count. Additionally, procalcitonin (PCT) and interleukin 6 (IL-6) levels as indicators of inflammation were measured, which provided insightful information regarding the patients' immune responses.

In the pre-processing of the structured data, a normalized approach was employed to capture and quantify over 50 factors that played a role in determining whether a patient had severe pneumonia. For the laboratory data, a median imputation technique was utilized to address missing values within the factors. When the missing rate for a specific marker was more than 50%, the factor was either excluded, its influence significantly diminished, or compensation for the absent markers was applied. By leveraging the median values of the available data for a particular marker, the missing values were effectively imputed, ensuring that the dataset remained as complete as possible for subsequent analysis. This approach helped mitigate the potential biases

introduced by missing data and preserved the integrity and comprehensiveness of the dataset. On the other hand, when it came to unstructured data such as the chief complaints recorded in free-text format, a robust natural language processing (NLP) algorithm was leveraged to extract the corresponding tokens. This NLP algorithm was able to process and parse the textual data, extracting relevant information and converting it into a structured format suitable for further analysis. By employing this NLP technique, the unstructured data was effectively harnessed, extracting valuable insights to augment the analysis. By combining a normalized approach for structured data and leveraging NLP algorithms for unstructured data,[76] the accuracy and completeness of the clinical record dataset information was ensured.

Microbiological analysis

To thoroughly investigate the various types of pulmonary infections in this study, the laboratory test results and benchmarked clinical diagnosis were comprehensively analyzed as the gold standard. To diagnose the specific viral subtypes, nucleic acid tests for respiratory pathogens were executed. These tests facilitated the discernment of an array of respiratory viruses, including influenza B virus, human metapneumovirus (HMPV), coronavirus, parainfluenza virus, respiratory syncytial virus (RSV), influenza A virus, H1N1pdm09, adenovirus, and rhinovirus. Similarly, for diagnosing bacterial pathogens, combined nucleic acid tests for respiratory pathogens were employed to analyze the distribution of different bacteria and isolate them for identification. This approach enabled the concurrent detection of clinically common lower respiratory tract bacterial pathogens, namely methicillin-resistant *Staphylococcus aureus* (MRSA), *Mycobacterium tuberculosis*, *Pseudomonas maltophilia*, *Escherichia coli*, *Haemophilus influenzae*, *Klebsiella pneumoniae*, *Staphylococcus*, *Pseudomonas aeruginosa*, and *Acinetobacter baumannii*.

Schema design

The schema employed in this study encompassed a series of modules that replicated the sequential diagnostic process undertaken by clinicians in real-world clinical settings. This schema was architected to extract relevant information from the symptoms, CT scans, and laboratory assay results to aid in diagnosing patients. The overall goal was to maximize data interoperability across diverse medical facilities for future research purposes. The diagnostic process began with the admission of a patient, whereupon the clinician initially assessed the patient's basic condition and laboratory examination results (subset 1). Based on this information, the clinicians determined whether the patient had an infectious disease. If the answer was affirmative, indicating the presence

of an infectious disease, the diagnostic process proceeded further. In the next step, the clinicians focused on determining the specific categorical infection type for the patient and prescribing the appropriate antibiotic treatment (subset 2). Subsequently, if the patient was diagnosed with viral or bacterial pneumonia, the schema applied additional modules to identify more refined subtypes of infections (subset 3 and subset 4). This step enabled a more granular classification of the pneumonia subtype, which could guide treatment decisions and further inform the clinical management of the patient. Moreover, for patients confirmed to be infected, the schema incorporated a prospective prediction module to estimate the likelihood of progression into severe pneumonia (subset 5). This predictive analysis would serve as a valuable tool for assessing the potential severity of the infections, enabling proactive intervention strategies to prevent or manage the development of severe illness.

Diagnosis system and network architectures

The deep-learning model employed for subtyping of infectious diseases was based on the Swin-Transformer architecture.[77] The model structure comprised multiple components, including a token embedding layer and four stage blocks. Each stage block was interfaced to a convolutional layer that performed subsampling of the feature maps. This design followed a similar pattern to a typical ResNet-50 architecture. The model harnessed a token-embedding layer to represent the input data in a suitable format for deep learning computations.[78] The token embeddings captured the essential information from the input, serving as the input for subsequent stages of the model. To enhance the model's performance further, convolutional layers were strategically situated subsequent to each stage block. These convolutional layers undertook subsampling of the feature maps, thereby reducing their spatial dimensions while increasing the number of feature channels. This down-sampling process fortified the model's receptive scope and enhanced its ability to capture and characterize relevant features. The model architecture exhibited a methodical escalation in the number of dimensions after each down-sampling operation. This increase in dimensions contributed to the expansion of expressive capacity of the model and allowed for better feature representation and discrimination.

Next, the pre-processed normalized 3D volume ($64 \times 256 \times 256$) was input into the convolutional token embedding (CTE) module. To optimize computational efficiency, a $2 \times 7 \times 7$ convolution kernel with a stride of four was opted. This convolution operation directly embedded the input volume, thereby alleviating the computational burden while preserving the essential information within the data. Within each stage block, two stacked pre-normalization were incorporated to enhance the learning

capability of the model. The first pre-normalization consisted of LayerNorm and Cross-shaped window self-attention operations, along with a shortcut connection. The second pre-normalization step comprised LayerNorm and a multi-layer perceptron (MLP). Compared with the traditional Swin-Transformer architecture, the cross-shaped window blocks utilized in the model were designed to be computationally efficient (Table S6). By incorporating these pre-normalization layers and carefully managing the connections between them, a model that required fewer computations was achieved while maintaining strong representation and learning capabilities.

The diagnostic results derived from the CT scans, along with the corresponding multimodal input data, were fed into the subtype diagnosis multilabel classification module to obtain predictions for a spectrum of pneumonia subtypes, including bacterial pneumonia (BP), fungal pneumonia (FP), viral pneumonia (VP), and PTB. To effectively capture discriminative features from both the image data and text data, the cross-attention mechanism was employed. This attention mechanism conferred the model to selectively concentrate on the relevant regions and textual information contributing to the subtype diagnosis. By attending to specific regions of the radiologic volume and relevant textual features, the model could learn the distinctive patterns and characteristics associated with different pneumonia subtypes.

NLP model development

Then a free-text information extraction model was developed to extract and reformat the chief complaint and history of present illness features from unstructured text data. This model employed NLP techniques (such as BERT) to analyze and extract relevant information from the textual input.[79,80] BERT is trained on an expansive corpus of text data, enabling the generation of high-quality contextualized word embeddings. These embeddings were utilized for pre-processing and initial feature learning in this study. To manipulate the structured data, such as laboratory test results, a normalization technique was employed to generate vector representations for specific factors (such as CRP). This normalization process contributed to standardizing the data and rendering them suitable for analysis. Furthermore, to enhance the analysis, a multi-layer fusion module was introduced. This module facilitates the bidirectional feature embedding of structured laboratory features and unstructured medical record features. By leveraging this mechanism, the interdependencies and relationships between different data elements were captured. Additionally, a structured data extraction model was implemented, specifically designed for extracting features from laboratory examinations and basic demographic information. This model processed the structured data to extract and normalized meaningful features that were relevant to the diagnosis

and classification of pneumonia. The combination of these information extraction models was able to transform unstructured free-text data and structured laboratory data into more structured and usable formats.

Explicitly, the model accepted either the free-text input of the chief complaint and history of the present illness or the structured-text input of laboratory data. It processed these inputs and generated multiple discrete vector features as outputs. Patient records could vary significantly in terms of length and the density of data points. To ensure consistent and efficient processing, the data was vectorized into a structured format with multiple lines. Each line had a specified length of 200, which allowed for better data organization and handling. This vectorization approach was able to handle variable-length input data in a consistent manner, ensuring compatibility and ease of processing. The NLP model, with its vectorization scheme, affords the efficient extraction of features from the chief complaint, history of the present illness, and laboratory data, delivering valuable and fixed-length inputs for downstream tasks in pulmonary infections diagnosis and classification.

Multimodal data fusion

To enhance diagnostic accuracy and robustness, multimodal data fusion techniques have been utilized to combine multiple modalities, such as CT scans, chief complaints, and laboratory testing, to enhance diagnostic accuracy and robustness. In the infection diagnosis pipeline, different approaches were adopted based on the fusion level. These approaches encompassed early fusion, in which the raw modalities were combined before feature extraction; intermediate fusion, where the features from each modality were concatenated before classification; and late fusion, where the classification results from each modality were combined (Figure S4). To integrate the two modalities, an attention-based structure known as cross attention was also employed. This approach facilitated the efficacious capitalization of the complementary information in multimodal data, culminating in efficiency and reduced computational complexity. However, although the aforementioned self-attention module effectively captured intramodality relationships, it did not explore the inter-modality relationships, such as the relationship between image regions and sentence words. Therefore, the Cross-Attention Module was utilized in this study, which modeled both the inter-modality and intra-modality relationships within a harmonized framework.[81]

Network training strategy

During the training process, the parameters of the Transformer model underwent initial pre-training using the unsupervised learning of visual features. This pre-training phase

involved contrasting cluster assignments, allowing the model to forge meaningful representations from the input data without explicit labels or annotations. The goal was to capture rich visual features that could be leveraged in ensuing supervised tasks. To train and test this model, the PyTorch deep-learning framework was run on a system equipped with 8 NVIDIA TITAN RTX GPUs. The AdamW optimizer was employed to train the model, incorporating a weight decay of 0.0001, to train the model. The learning rate was initialized to 0.001, which was then decayed by a factor of 10 after the 35th, 40th, and 50th epochs to fine-tune the training process. All the models were trained for 60 epochs. Constrained by GPU memory limits, the batch sizes for optimal performance were adjusted. Specifically, the batch size of each GPU was set to 16. These batch-size configurations allowed for efficient processing and training of the model while maximizing the utilization of available computational resources.

For the models based on subset 1 and subset 2, patient cases were randomly divided into two sets: a training set comprising 80% of the cases and a test set comprising the remaining 20%. These sets were utilized to train the models and evaluate their performance. Random splitting ensured the unbiased distribution of cases across the training and test sets. In the context of subset 3 and subset 4, where the focus was on less frequent diseases, additional measures were taken to account for the rarity of these conditions and to enhance the robustness of the AI system's identification capabilities. To achieve this, the representation of these rare diseases within the validation and testing sets were deliberately augmented. Particularly, in the testing subset, the rare diseases were represented 40%, exceeding their prevalence in the overall patient population. This strategy was devised to present a more challenging evaluation scenario and validate the ability of the model to accurately identify and classify these less frequent diseases. To broaden validation and generalize final results, a five-fold cross-validation approach was employed. The experiment was replicated five times for each disease model.

Comparison of AI and physicians

Then we compared the performance of an AI framework with that of physicians in analyzing CT scans, chief complaints, and laboratory tests from electronic health records (EHR) to diagnose infections. The gold standard for diagnosis was established on sputum culture, polymerase chain reaction (PCR) or molecular testing results. To ensure a fair comparison, four practicing physicians were recruited to partake in the study. The physicians were categorized into two groups based on their level of clinical tenure: a junior group, consisting of physicians with less than 10 years of experience, and a senior group, consisting of physicians with over 10 years of experience. The

performances of the AI framework and the human physicians were evaluated using a weighted errors metric based on penalty scores. This evaluation metric was contrived to reflect the clinical performance of the AI system and physician expertise. During the testing phase, the AI framework and the physicians were furnished with the identical dataset, which comprised CT scans, corresponding chief complaints, and laboratory testing results from the electronic health records. The performance of each entity was assessed against the gold standard, evaluating their competence to gold standard accurately.

Prognosis analysis for integrating multimodal features

To decipher the influence of each factor on severe pneumonia, a machine-learning approach was employed to extract quantized factors and non-quantized multimodal feature (M-score) from clinical texts, images. These features were subsequently used in a prognostic prediction model, employing the widely recognized gradient-boosting decision tree algorithm (GBDT) as the classifier [82]. To construct a comprehensive predictive score for the clinical outcome, the image features extracted by the AI system were combined with relevant clinical parameters, such as age, albumin levels, blood oxygen saturation, CRP, and other pertinent factors. This composite score was applied to predict the progression to critical illness, measured by the need for intensive care unit (ICU) transfer, mechanical ventilation, or death, and also considering the time elapsed since the initial hospital admission. Clinical and radiological features were selected predominantly based on their correlation with the severity status. The importance of these features was appraised by examining the magnitude of the log-rank test statistics using the Shapley Additive exPlanation (SHAP) method. This enabled physicians to visualize the impact of the relevant risk factors on the prognostic prediction of critical illnesses, providing valuable insights into the factors that influence disease progression. To ensure the robustness and reliability of the model, its performance was validated using a five-fold cross-validation approach, which allowed physicians to tune the optimal hyperparameters and assess the consistency and accuracy of the model across different data subsets.

The random survival forest method was suitable for integrating high-dimensional features. In this study, this method was employed to analyse the data and engender a multi-model score ranging from 0 to 1. This score epitomizes the average expected number of events across all the random survival forest model trees. By instituting a cut-off score of 0.5, patients were classified into two distinct groups: a high-risk group (with a score greater than 0.5) and a low-risk group (with a score less than 0.5). This stratification facilitated the differentiation of patients according to their predicted risks

of adverse outcomes. To delve deeper into stratified groups, the Kaplan-Meier estimator was utilized to calculate the survival times for high-risk and low-risk groups. Additionally, a log-rank test was conducted to evaluate the statistical significance of the differences observed between the two groups regarding survival outcomes.

Quantification and statistical analysis

The MMI system was architected to perform multilabel classification and prognostic prediction tasks. To evaluate classification performance, the mean macro area under the receiver operating characteristic curve (AUC) was employed as a performance metric. Confidence intervals (CIs) were computed using a bootstrapping approach with nonparametric, unstratified resampling (1000 times) to estimate the uncertainty in the AUC estimates. Diagnostic performance of the system was quantified through the metrics of its sensitivity, specificity, and accuracy at the selected operating points. The operating point was selected to strike a balance between a low false negative diagnostic rate (sensitivity) and a low positive rate (1-specificity), with the thresholds adjusted accordingly. For statistical correlation significance, Pearson's and Spearman's correlation tests were used, supplemented by Holm-Bonferroni method for multiple comparisons. Normally distributed data were described using the mean and standard deviation (SD), while non-normally distributed data were described using the median and interquartile range (IQR). Categorical variables were presented as numbers and percentages. The deep learning models were trained, validated, and tested using PyTorch (v1.11.0), a renowned deep learning framework. For the data analysis, the scikit-learn library was utilized in Python. Graphs and visualizations were crafted using Python libraries (Matplotlib and Seaborn). Kaplan-Meier survival curves were generated to approximate the diagnosis time based on follow-up visits. The log-rank test compared the survival curves between the subgroups, allowing physicians to assess any significant differences in the time to diagnosis. The codes that support the findings of this study were available as follows: https://github.com/chiehchiu/MMI

References

75. Zhang, S., Xu, J., Chen, Y.-C., et al. (2020). Revisiting 3D context modeling with supervised pre-training for universal lesion detection in CT slices. Medical Image Computing and Computer Assisted Intervention (MICCAI).
76. Xiang, L., Ma, J., and Li, H. (2019). Invasiveness prediction of pulmonary adenocarcinomas using deep feature fusion networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1909.09837.
77. Zhang, S., Li, Z., Zhou, H.-Y., et al. (2023). Advancing 3D medical image analysis with variable dimension transform based supervised 3D pre-training. Neurocomputing **529**, 11-22. DOI: 10.1016/j.neucom.2023.01.012.
78. Wang, H., Li R., Jiang H., et al. (2023). LightToken: a task and model-agnostic lightweight token embedding framework for pre-trained language models. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2302-2313.
79. Liu, P., Yuan, W., Fu J., et al. (2023). Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Computing Surveys **55** (9), 1-35.
80. Thoppilan, R., Freitas, D., Hall, J., et al. (2022). Lamda: language models for dialog applications. Preprint at arXiv. https://doi.org/10.48550/arXiv.2201.08239
81. Ma, J., Li, X., Li, H., et al. (2021). Cross-view relation networks for mammogram mass detection. International Conference on Pattern Recognition (ICPR).
82. Ke, G., Xu Z., Zhang J., et al. (2019). DeepGBM: a deep learning framework distilled by GBDT for online prediction tasks. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 384-394.