Journal of Medical Systems (2025) 49:120
https://doi.org/10.1007/s10916-025-02263-3

RESEARCH

A Pretraining Approach for Small-sample Training Employing Radiographs (PASTER): a Multimodal Transformer Trained by Chest Radiography and Free-text Reports

Kai-Chieh Chen[1] · Matthew Kuo[2] · Chun-Ho Lee[3] · Hao-Chun Liao[4] · Dung-Jang Tsai[5,6] · Shing-An Lin[6] · Chih-Wei Hsiang[7] · Cheng-Kuang Chang[7] · Kai-Hsiung Ko[7] · Yi-Chih Hsu[7] · Wei-Chou Chang[7] · Guo-Shu Huang[7] · Wen-Hui Fang[8] · Chin-Sheng Lin[5,6,9] · Shih-Hua Lin[10] · Yuan-Hao Chen[11] · Yi-Jen Hung[12] · Chien-Sung Tsai[13] · Chin Lin[1,5,6]

Received: 24 February 2025 / Accepted: 4 September 2025

Abstract

While deep convolutional neural networks (DCNNs) have achieved remarkable performance in chest X-ray interpretation, their success typically depends on access to large-scale, expertly annotated datasets. However, collecting such data in real-world clinical settings can be difficult because of limited labeling resources, privacy concerns, and patient variability. In this study, we applied a multimodal Transformer pretrained on free-text reports and their paired CXRs to evaluate the effectiveness of this method in settings with limited labeled data. Our dataset consisted of more than 1 million CXRs, each accompanied by reports from board-certified radiologists and 31 structured labels. The results indicated that a linear model trained on embeddings from the pretrained model achieved AUCs of 0.907 and 0.903 on internal and external test sets, respectively, using only 128 cases and 384 controls; the results were comparable those of DenseNet trained on the entire dataset, whose AUCs were 0.908 and 0.903, respectively. Additionally, we demonstrated similar results by extending the application of this approach to a subset annotated with structured echocardiographic reports. Furthermore, this multimodal model exhibited excellent small sample learning capabilities when tested on external validation sets such as CheXpert and ChestX-ray14. This research significantly reduces the sample size necessary for future artificial intelligence advancements in CXR interpretation.

Keywords Multimodal learning · Chest radiograph · Few-shot prediction · Small sample training · Foundation model · Transformer · Deep learning

Introduction

Chest X-rays (CXRs) are a commonly used imaging technique, with at least 2 billion global annual instances [1]. There is significant interest in the development of CXR analysis technology [2]. Deep learning techniques, particularly convolutional neural networks (CNNs) [3], have been found to achieve expert-level performance in interpreting medical images [4–6]. However, the success of these models heavily relies on substantial volumes of data and high-quality annotations, which often require collaboration among multiple experts [7, 8]. The current lack of large, well-annotated datasets significantly constrains supervised deep learning for medical image tasks [9, 10]. This limitation is particularly pronounced in real-world clinical settings, where data access, privacy concerns, and resource constraints hinder large-scale annotation efforts. The development of approaches to reduce this reliance on large-scale datasets with structured annotations is critical.

Transfer learning with pretrained models is a primary method for enhancing performance with limited samples and has gained widespread acceptance for medical image analysis processes [11]. In image classification tasks, several unsupervised pretrained algorithms exist [12–14]. Recently,

Extended author information available on the last page of the article

Published online: 30 September 2025

a multimodal model called contrastive language-image pre-training (CLIP), which is trained with text-and-image pairs, has demonstrated superior accuracy in downstream tasks [15]. In the medical field, especially for CXRs, electronic health records often include corresponding reports written by radiologists, making it feasible to adopt CLIP-like approaches for CXR analysis. Models such as ConVIRT, BioMedCLIP, and MedCLIP have used to apply similar techniques in this context, although most have been evaluated in large-scale or zero-shot scenarios [16–18]. While zero-shot prediction is quite convenient in applications, the CLIP model with linear probing, which uses more than 4 labeled training examples per class, surpasses the accuracy of zero-shot prediction [15]. Despite this progress, a gap remains in understanding how such models perform in genuinely small-sample regimes where data availability falls far short of the traditional supervised learning thresholds. However, for clinical validation and FDA approval, AI models require a minimum of several hundred samples for each specific indication [19]. Thus, it is essential to explore small sample learning using the CLIP model for CXR analysis.

Large-scale studies show promise in using multimodal models such as CLIP for CXR analysis, but small sample learning remains underexplored. Current research focuses on large datasets and lacks the benefits of smaller, manageable datasets. No comprehensive studies have evaluated pretrained multimodal models with small sample sizes for CXRs. While existing works have demonstrated strong zero-shot performances, little attention has been given to evaluating such models under constrained supervision. In this study, we address this gap by proposing and evaluating the pretraining approach for small-sample training employing radiographs (PASTER), which combines contrastive vision-language pretraining with linear probing to enable accurate CXR interpretation from limited data. We compare PASTER with traditional CNNs across various sample sizes, highlighting its potential for high accuracy with minimal data. The PASTER workflow is shown in Fig. 1: starting with contrastive learning using free-text reports and chest radiographs (Fig. 1A), the vision Transformer then extracts features from a limited sample pool, followed by a logistic regression analysis (Fig. 1B). Using only 128 cases with 384 controls achieved a highly satisfactory performance.

The main methodological contributions of this study are as follows:

1. We propose PASTER, a multimodal contrastive pretraining framework tailored for small-sample chest X-ray interpretation.
2. We evaluate its performance across varying dataset sizes and show that it achieves an accuracy comparable to that of fully supervised CNNs using only a fraction of the data.
3. We adopt linear probing as a lightweight adaptation method and demonstrate its effectiveness in leveraging pretrained representations under constrained supervision.

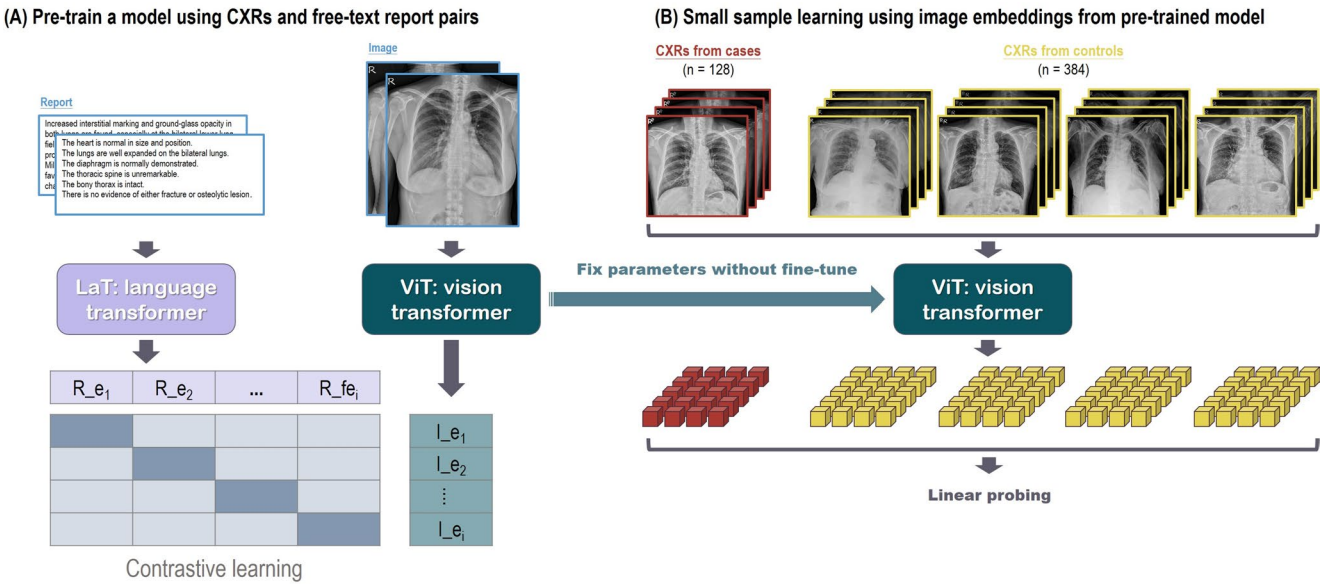

Fig. 1 PASTER training and application workflow. (**A**) Contrastive learning using free-text reports and chest radiographs to identify paired associations. (**B**) Vision Transformer extracts features from a limited sample pool, followed by logistic regression with regularization. Using 128 cases and 384 controls achieved satisfactory performance

Materials and methods

Datasets

This algorithm development study was based on data from the Tri-Service General Hospital in Taiwan to develop a deep learning model (DLM). The study was approved by the Institutional Review Board of the Tri-Service General Hospital, National Defense Medical Center (IRB NO. C20230519), in accordance with the Declaration of Helsinki. All the data were obtained from the hospital's quality control center, were fully anonymized prior to analysis, and were exempt from informed consent as approved by the IRB. Tri-Service General Hospital provided a private database of 1,105,436 AP or PA viewed CXRs from January 1, 2011, to February 28, 2022. The training set included 1,004,314 CXRs paired with radiological reports and annotated with 31 radiological labels. To prevent cross-contamination, a strict dataset split strategy was applied (Fig. 2). We subsampled the training set into five sizes: 2,000 (0.2%), 10,000 (1.0%), 50,000 (5.0%), 200,000 (20.0%), and 1,004,314 (100.0%). The internal and external test sets from hospitals A and B contained 101,122 and 81,614 CXRs, respectively. To avoid sample interdependence, only one CXR was used for each patient. The training dataset included more than one million CXRs from two hospitals. The average ages in the training, internal test, and external test sets were 58.0±21.2, 48.1±21.0, and 51.6±21.3 years, respectively, with male percentages of 55.1%, 51.6%, and 50.5%, respectively. The detailed patient characteristics and dataset distributions are shown in Table 1 and Extended Data Table 1. For a subset with echocardiographic labels, details are provided in Table 2 and Extended Data Table 2.

The CheXpert dataset, consisting of 224,316 CXRs from 65,240 patients at Stanford Hospital, was used to evaluate small-sample learning [20]. Only frontal CXRs were included, resulting in a training set of 191,027 CXRs. The dataset is labeled for 14 conditions, with a focus on five conditions: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. The test set included 668 CXRs from 500 patients. The distribution details are provided in Table 3.

The ChestX-ray14 dataset from the NIH Clinical Center includes 112,120 frontal CXRs from 30,805 patients [21]. The training set included 86,524 CXRs, and the test set included 25,596 CXRs. We selected eight labels for cross-database performance assessment: atelectasis, cardiomegaly, consolidation, edema, pleural effusion, emphysema, pneumonia, and pneumothorax. Detailed distribution information is available in Table 4.

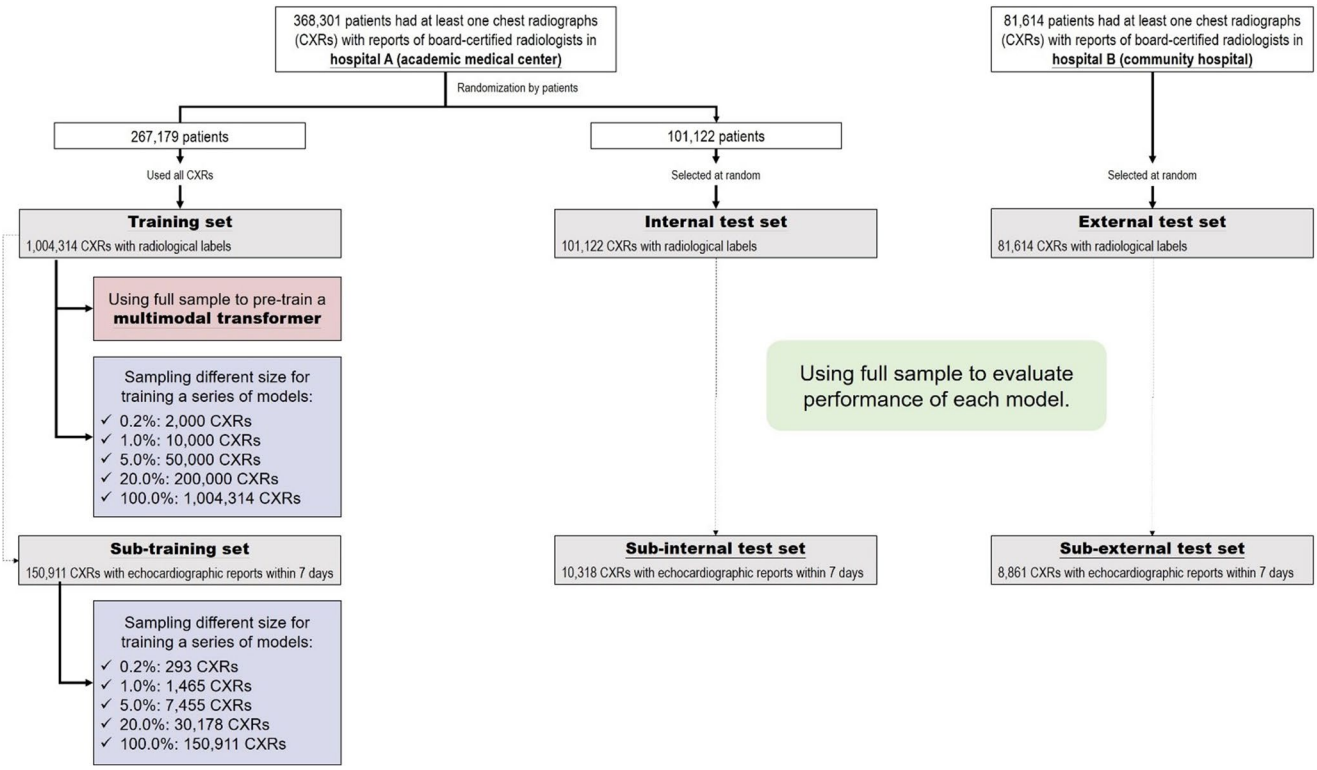

Fig. 2 Dataset creation and analysis strategy. Schematic ensuring robust training and testing by using patient data in only one set to avoid cross-contamination

Table 1 Distribution of structured radiological labels in the private dataset

	Training set	Internal test set	External test set
Radiological labels†	*n* = 1,004,314	*n* = 101,122	*n* = 81,614
Consolidation change	77,119(7.7%)	2332(2.3%)	1836(2.2%)
Pneumonia	22,606(2.3%)	798(0.8%)	550(0.7%)
Emphysematous change	21,163(2.1%)	1011(1.0%)	1114(1.4%)
Pneumothorax	13,346(1.3%)	399(0.4%)	282(0.3%)
Atelectasis	34,101(3.4%)	1060(1.0%)	839(1.0%)
Scalloping of the diaphragm	22,459(2.2%)	1812(1.8%)	1877(2.3%)
Costophrenic angle blunting	437,443(43.6%)	22,638(22.4%)	19,899(24.4%)
Pleural effusion	237,838(23.7%)	8727(8.6%)	7856(9.6%)
Atherosclerosis	505,309(50.3%)	32,690(32.3%)	31,520(38.6%)
Cardiomegaly	292,483(29.1%)	14,379(14.2%)	12,688(15.5%)
Prominence of hilar shadow	164,002(16.3%)	7010(6.9%)	5868(7.2%)
Pulmonary edema	53,235(5.3%)	1281(1.3%)	1054(1.3%)
Aneurysm	1787(0.2%)	64(0.1%)	53(0.1%)
Degenerative joint disease	452,835(45.1%)	29,814(29.5%)	28,911(35.4%)
Fracture	104,228(10.4%)	5236(5.2%)	4617(5.7%)
Spondylosis	299,812(29.9%)	16,830(16.6%)	16,652(20.4%)
Osteophyte formation	393,814(39.2%)	24,844(24.6%)	23,886(29.3%)
Osteoporosis	98,472(9.8%)	4735(4.7%)	5152(6.3%)
Osteoarthritis	228,868(22.8%)	12,197(12.1%)	11,597(14.2%)
Widening of the mediastinum	185,558(18.5%)	9540(9.4%)	9083(11.1%)
Malignancy	20,088(2.0%)	721(0.7%)	543(0.7%)
Inflammatory	326,035(32.5%)	15,823(15.6%)	14,017(17.2%)
Pigtail or drainage	41,408(4.1%)	942(0.9%)	591(0.7%)
Sternotomy	55,352(5.5%)	1174(1.2%)	1252(1.5%)
Port A implantation	72,678(7.2%)	2101(2.1%)	1714(2.1%)
Perm catheter insertion	33,975(3.4%)	734(0.7%)	584(0.7%)
Pacemaker	15,128(1.5%)	480(0.5%)	523(0.6%)
Tracheostomy	53,939(5.4%)	633(0.6%)	723(0.9%)
Vertebroplasty	12,062(1.2%)	436(0.4%)	423(0.5%)
Endotracheal tube	94,500(9.4%)	2153(2.1%)	1110(1.4%)
Nasogastric tube	192,923(19.2%)	4382(4.3%)	3044(3.7%)

†These structured radiological labels were preselected by each board-certified radiologist when free-text reports were composed. In theory, the content of the free-text reports is expected to encompass these labels, but they may also include free-text descriptions beyond the scope of these labels

Table 2 Distribution of structured echocardiographic labels in the private dataset

	Training set	Internal test set	External test set
Echocardiographic labels‡	*n* = 150,911	*n* = 10,318	*n* = 8861
Left ventricular dysfunction	11,637(7.7%)	377(3.7%)	268(3.0%)
Aortic stenosis	2255(1.5%)	93(0.9%)	95(1.1%)
Pulmonary arterial hypertension	18,729(12.4%)	681(6.6%)	621(7.0%)
Left atrial enlargement	29,644(19.6%)	1383(13.4%)	1319(14.9%)
Pericardial effusion	2312(1.5%)	81(0.8%)	58(0.7%)

‡These abnormal echocardiographic findings were extracted from structured cardiac ultrasound reports obtained within ±7 days of examination. Notably, the board-certified radiologists composing the free-text reports did not have access to these specific results during their reporting process. The definitions for these findings are as follows: left ventricular dysfunction, defined as a left ventricular ejection fraction≤35%; aortic stenosis, defined as moderate [jet velocity 3.0–4.0 m/s, mean gradient of 20–49 mmHg, or an aortic valve area of 1.1–1.5 cm^2] to severe [jet velocity≥4.0 m/s, mean gradient≥40 mmHg, a dimensionless velocity index (DVI) of ≤0.25, or an AVA of ≤1.0 cm^2]; (3) pulmonary arterial hypertension, defined as a systolic pulmonary artery pressure>50 mmHg (peak tricuspid regurgitation velocity>3.4 m/s); (4) left atrial enlargement, defined as a left atrial diameter>45 mm; and (5) pericardial effusion, defined as an effusion>1 cm. Left ventricular dysfunction [41] is considered challenging for radiologists to identify directly, despite prior deep learning research confirming the potential existence of subtle signs in CXR. Aortic stenosis [42], pulmonary arterial hypertension [43], left atrial enlargement [44], and pericardial effusion [45] are also difficult to interpret directly but can be indirectly inferred by radiologists through other features

Table 3 Distribution of structured labels in the chexpert dataset

	Training set	Validation set	Test set
CheXpert labels	*n* = 191,027	*n* = 202	*n* = 500
Atelectasis	112,217(58.7%)	75(37.1%)	153(30.6%)
Cardiomegaly	111,420(58.3%)	66(32.7%)	151(30.2%)
Consolidation	98,094(51.4%)	32(15.8%)	29(5.8%)
Edema	88,580(46.4%)	42(20.8%)	78(15.6%)
Pleural Effusion	93,189(48.8%)	64(31.7%)	104(20.8%)

Table 4 Distribution of structured labels in the ChestX-ray14 dataset

	Training set	Test set
ChestX-ray14 labels	*n* = 86,524	*n* = 25,596
Atelectasis	8280(9.6%)	3255(12.7%)
Cardiomegaly	1707(2.0%)	1065(4.2%)
Consolidation	2852(3.3%)	1815(7.1%)
Edema	1378(1.6%)	925(3.6%)
Pleural Effusion	8659(10.0%)	4648(18.2%)
Pneumothorax	2637(3.0%)	2661(10.4%)
Pneumonia	876(1.0%)	477(1.9%)
Emphysema	1423(1.6%)	1093(4.3%)

Foundation Model Pre-training

PASTER was developed using OpenAI's CLIP model, which incorporates an image encoder and a text encoder [15]. The image encoder uses the ViT-B/32 architecture, which is designed to process 256 × 256-sized images. Input images were initially divided into 32 × 32-sized image patches, which were then processed through 12 Transformer layers with a hidden size of 768, producing a compressed output as a 512-dimensional vector for each patch. The text encoder was a standard language Transformer with 512 token embeddings, a maximum token length of 256, 12 layers, a hidden size of 512, and 8 attention heads. Before the images and text were input into their respective encoders, an additional [CLS] token was added, and the output from the [CLS] token was used as an embedding for both the CXR and the report. During backpropagation, the inner product between these two embeddings was calculated, and the error was passed through softmax and cross-entropy loss for optimization. A schematic pseudocode illustrating the contrastive pretraining procedure is provided in Extended Fig. 1 [15].

All the technical details closely adhered to OpenAI's CLIP model, and the weights of this model were used as the initialization parameters [15]. During pretraining, we randomly split 1,004,314 CXRs from the Tri-Service General Hospital database into 90% for model fitting and 10% for validation. All network parameters were fine-tuned using the standard parameters of an SGD optimizer, with a batch size of 64, a learning rate of 0.0001, and a momentum of 0.9. Throughout the training, the images were randomly cropped to a size of 256 × 256 pixels. The model was trained for 50 epochs, with the validation loss computed at the end of each epoch to select the best-performing model. Model training was conducted in a Python environment, version 3.10.10, utilizing the "torch" package version 2.0.1.

Linear Probing

We conducted linear probing of CXR embeddings using the "glmnet" package version 2.0–16 in R software 3.4.4. This package uses the elastic net algorithm, which includes both L1 and L2 regularization terms within logistic regression. For each task, we oversampled the samples to ensure an equal number of cases and controls. Afterward, we divided the samples into four subsets for cross-validation, utilizing the "cv.glmnet" function. During cross-validation, we searched for the optimal values of lambda (the regularization hyperparameter) in the range from e^{-9} to e^{-1} (a logarithmically spaced sequence of length 45) and alpha (the ratio of L1 and L2 penalties) in the range from 0 to 1 (with increments of 0.1), aiming to maximize the cross-validation AUC. The selected model fit by the hyperparameter with the highest cross-validation AUC was subsequently applied directly to the test set. More details can be found in the code availability section. Note that owing to the high sensitivity of small sample training to random sample selection, all model fittings using linear probing were repeated 21 times. The results were then determined on the basis of the median accuracy across these 21 models.

Zero-shot Prediction

We conducted zero-shot experiments on each dataset using the labels in its test set to generate "< label>" and "no < label>" prompts for the softmax evaluation process.

Convolutional Neural Network

We used a 121-layer DenseNet architecture [22] to adapt to the training methodology for CXR from our previous study [23]. Our approach began with pretraining DenseNet on the ImageNet dataset. The parameters for each CXR network were initialized using parameters from the pretrained network of the ImageNet dataset. We updated the parameters of the DLMs to minimize cross-entropy losses, incorporating an oversampling technique based on class weights computed from the prevalence of each class in the training set. Because we initially attempted to train a single network for all labels but found that the accuracy was not satisfactory, we ultimately opted to train a separate network for each label.

During training, 90% of each subset was used for model fitting, and 10% was used for validation. This procedure was applied to all the datasets. For Tri-Service General Hospital, we used five training proportions, namely, 0.2%, 1.0%, 5.0%, 20.0% and 100.0%. For CheXpert and ChestX-ray14, we used three proportions, which were 1.0%, 10.0% and 100.0%. All network parameters were fine-tuned using the Adam optimizer with standard parameters, utilizing a batch size of 32 and an initial learning rate of 0.001. We reduced the learning rate by a factor of 10 whenever the validation loss plateaued after an epoch. To prevent overfitting, we implemented early stopping by saving the network after each epoch and selecting the saved network with the highest validation AUC. During training, we used random cropping of a 224 × 224-pixel region as input, with a 50% chance of applying a random horizontal flip. In the inference stage, we employed a 10-crop evaluation method to generate 10 probabilities for each CXR, and the final prediction was based on the average of these 10 probabilities.

The networks were trained for more than 50 epochs: including more than 30 epochs at a learning rate of 0.001 (Stage 1), more than 10 epochs at a learning rate of 0.0001

(Stage 2), and more than 10 epochs at a learning rate of 0.00001 (Stage 3). In most cases, the model with the highest AUC on the tuning subset was often found in Stages 2 and 3. The sole regularization technique used to prevent overfitting was a weight decay of 10^{-4} in this study. Model training was conducted using the R version 3.4.4 software environment with the "MXNet" package version 1.3.0.

Vision Transformer

Owing to prior research highlighting the advantages of vision Transformers (ViT) over convolutional neural networks in medical image analysis [24], we also trained a series of models based on the ViT-B/32 [25] architecture for comparison. Like in our CNNs experiments, we trained separate ViTs for each label and initialized the model weights using checkpoints pretrained on the ImageNet dataset. During training, 90% of each database subset was used for model fitting, and 10% was used for validation. We selected the model with the highest validation AUC for evaluation on the test set. Furthermore, we followed training details similar to those of the original ViT [25], including the use of the SGD optimizer with hyperparameters set to a cosine warmup learning rate starting from 3×10^{-3}, a momentum of 0.9, a batch size of 64, and a total number of epochs of 50. During training, we randomly cropped a 256×256-pixel region as input without applying horizontal flipping. In the inference stage, we took the central 256×256-pixel region as input. Model training was conducted in the Python version 3.10.10 software environment, utilizing the "torch" package version 2.0.1.

Models in Ablation Experiments

We sought to emphasize the importance of the hyperparameter searching approach in linear probing. To do this, we used the same PASTER model for extracting CXR embeddings and kept lambda = 0.001 and alpha = 0 as fixed hyperparameters in the elastic net algorithm for our initial experiment. The training process of this model resembled that of ViT, with the only difference being the weight initialization. The criterion for selection was based on the average AUC calculated for 31 radiological labels at the end of each epoch. We chose the epoch with the highest validation AUC for further CXR embedding extraction.

Since the architecture of PASTER is based on Transformers, and prior research has indicated that Transformers outperform convolutional neural networks in medical image analysis [24], we changed the image encoder of PASTER. Instead of ViT/B-32, we replaced it with a 121-layer DenseNet. Additionally, since an earlier study suggested there are benefits in using structured labels for supervised contrastive learning in small sample learning [26], we attempted to modify the text encoder of PASTER. We replaced it with a multilayer perceptron (MLP), which takes 31 labels as input. The first hidden layer contains 512 neurons, and subsequently, each hidden layer maintains 512 neurons while incorporating the concept of residual learning to establish shortcuts between layers. After five hidden layers, the model directly outputs the results. Finally, we compared our approach to a ViT trained for multilabel training using cross-entropy loss with direct utilization of 31 structured labels. This ViT is a separate model, distinct from the ViT trained individually for each label. We utilized its final layer's high-level features for linear probing comparisons. All these additional models were trained with identical hyperparameter settings, and we selected the one with the lowest validation loss after each epoch for further comparison.

We also compared the results obtained by initializing the weights using PASTER and conducting full-parameter fine-tuning. Since full-parameter fine-tuning is not suitable for small sample learning, this experiment was conducted only with sample sizes ranging from 0.2 to 100%.

Model Comparison

The proposed PASTER shares a similar training technique with previous models such as ConVIRT, which was trained on the publicly available MIMIC-CXR dataset. MedCLIP also adopts a similar vision-language contrastive pretraining framework; however, since its training involved the CheXpert dataset, we excluded it from CheXpert-related evaluations to avoid potential data contamination [18]. In addition to ConVIRT and MedCLIP, we also included other pretrained models that use different techniques, such as BioViL-T, which is based on series-alignment techniques such as those in the BioViL family; MoCo-CXR [27, 28], which modifies the contrastive learning framework Momentum Contrast (MoCo) for CXR interpretation; and SupCon [26], a contrastive learning method based on structured labels in CheXpert. Since ChestX-ray14 lacks results from small sample training, we utilized fully supervised learning results as a benchmark [21]. Given that these models do not have publicly available checkpoints for direct comparison, we extracted performance figures from published papers.

Cross-modal Retrieval

To evaluate the cross-modal retrieval capability of the proposed model, we conducted retrieval experiments using 1,000 randomly selected radiology reports as queries. All reports were sourced from private datasets (an internal test set and external test set), each paired with a corresponding

chest X-ray image. For each text query, the cosine similarity between the encoded text representation and all the image embeddings in the candidate pool was computed. The retrieved images were ranked on the basis of similarity scores. We reported Recall@1 and Recall@5, which measure the proportion of queries where the ground-truth image appears in the top 1 or top 5 retrieved results, respectively. All the retrieval experiments were conducted using fixed pretrained encoders without further fine-tuning.

Statistical Analysis

We chose the AUC as the primary metric for evaluating model accuracy. AUC is calculated on the basis of the receiver operating characteristic (ROC) curve, which compares sensitivity and specificity at different thresholds. Since all the datasets in this study are multilabel, and direct averaging does not align with a clinical perspective, our primary model performance comparison involved calculating AUC differences for each label between the two models. We present these differences as medians (interquartile ranges, IQRs) and illustrate the percentage of cases where the difference is >0 (indicating that the first model outperforms the second). Additionally, in the supplementary appendix, we provided accuracy comparisons for each label, including mean AUC comparisons. To assess model classification performance, we visualized confusion matrices for key radiological and echocardiographic labels, using rowwise normalized percentages to highlight sensitivity and specificity across internal and external test sets.

We employed a nonparametric bootstrap approach to generate these confidence intervals (CIs). Specifically, we repeatedly sampled random subsets of size n (matching the number of samples in the internal test set) from the original dataset, with replacement, for a total of 1,000 iterations. AUC values or differences were estimated for each of these bootstrap samples. The CIs were derived from the relative frequency distribution of these estimates across the resamples. We calculated the interval between the $100\times(\alpha/2)$ and $100\times(1-\alpha/2)$ percentiles, with α set to 0.05. For AUC differences, we also examined how many values in the same bootstrap distribution were more extreme than 0 to calculate a one-tailed p value. Finally, we presented the results as two-tailed p values for all the statistical tests.

We also assessed the correlation between the AUC obtained with different sample sizes and the AUC from the full sample using Pearson correlation coefficients. The specific approach involved using the AUC obtained during small sample learning as the input and the AUC from the full sample as the output, with each data point representing the results for each label. Additionally, we performed a linear regression, which can be used by future researchers to predict the ultimate accuracy of a task with a limited sample size.

To visualize the differences between different datasets, we extracted CXR embeddings using PASTER, and then reduced the original 512-dimensional vectors to 2 dimensions using the uniform manifold approximation and projection (UMAP) technique. It is essential to emphasize that this visualization method relies on highly nonlinear axes, rendering the assignment of interpretable units to either axis impossible. This analysis was performed using the "umap-learn" package version 0.5.5 in Python version 3.10.10 with default parameters.

Results

Supplementary Data 1 compares PASTER, DenseNet (a CNN), and the vision Transformer in terms of the AUC for all 31 radiological labels and the mean AUC. We initially explored the most suitable ratio of cases (with <label>) to controls (without <label>) for small sample learning. It is widely known that the optimal ratio is 1:1, but owing to the often limited number of cases, we attempted to increase the number of controls to enhance model performance. Detailed results for different numbers of cases are presented in Extended Fig. 2. We observed that the ratio of controls to cases reached the highest mean AUC (0.865/0.862 in internal/external test sets) when it was 3, with further increases in the number of control samples showing diminishing returns. We also provided results for cases with a count of 128, and similarly, we found that the efficiency decreased beyond a control-to-case ratio of 3. Therefore, for subsequent small sample learning experiments, we fixed the ratio of cases to controls at 1:3.

The average results are summarized in Fig. 3. PASTER achieved average AUCs of 0.907 and 0.903 on the internal and external test sets, respectively, when only 128 cases and 384 controls were used, closely matching DenseNet's performance with the full dataset ($n=1{,}004{,}314$). PASTER also outperformed ViT, which achieved average AUCs of 0.890 and 0.885 on the internal and external test sets, respectively. PASTER consistently outperformed DenseNet, even with larger training samples. MedCLIP achieved AUCs of 0.901 on the internal test set and 0.903 on the external test set, whereas BioViL-T showed lower performance, with AUCs of 0.859 on the internal test set and 0.861 on the external test set. Compared with both MedCLIP and BioViL-T, PASTER demonstrated superior performance across all the evaluations. Extended Fig. 3A shows the AUC differences between PASTER and DenseNet under the same sample sizes. Compared with DenseNet, PASTER achieves median AUC differences of 15.6%/14.2%,

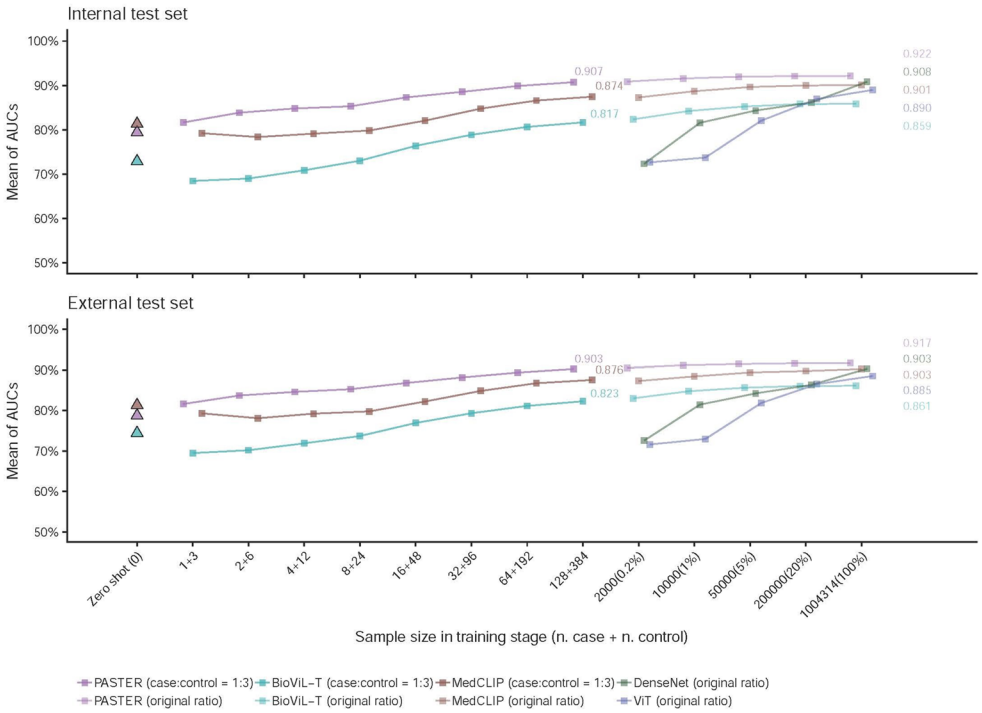

Fig. 3 Average AUC across 31 radiological labels using linear probing under varying sample sizes for PASTER, BioViL-T, MedCLIP, DenseNet, and ViT. Models were trained with increasing numbers of samples, and average AUC values were computed across 31 radiological labels

8.4%/7.0%, 7.1%/6.9%, 4.3%/4.5%, and 1.0%/0.9% in the internal/external test sets for training samples of 0.2%, 1%, 5%, 20%, and 100%, respectively. The results show that PASTER outperforms DenseNet on all radiological labels when the number of training samples is less than 200,000. Importantly, even when 100% of the samples were used, compared with DenseNet, PASTER maintained an advantage in 87.1%/83.9% of the radiological labels in the internal/external test sets. Extended Fig. 3B illustrates the comparison between PASTER using only 128 cases and 384 controls and DenseNet (20%), DenseNet (100%), and PASTER (100%). We observed that the performance of PASTER in small-sample learning closely matches that of DenseNet (100%), with only a slight difference (0.2%) in the median AUC difference. Furthermore, PASTER outperforms DenseNet (20%) significantly. Finally, when full samples were used for linear probing, PASTER performs better than small sample learning across all radiological labels did, albeit with minimal differences (median [internal]: −1.4% [IQR: −2.2%, −0.8%] and median [external]: −1.4% [IQR: −1.9%, −0.7%]). More detailed comparisons are presented in Supplementary Data 2. PASTER performs worse in zero-shot prediction than in small-sample learning for most labels. Confusion matrices for each radiological label are presented in Extended Fig. 2 to visualize classification patterns across internal and external test sets.

We further explored the performance of PASTER on five echocardiographic labels. Supplementary Data 3 provides a detailed comparison of PASTER, MedCLIP, BioViL-T, DenseNet, and ViT for each label's AUC and mean AUC, while Fig. 4 presents the average results. PASTER achieved average AUCs of 0.840 and 0.842 on the internal and external test sets, respectively, when only 128 cases and 384 controls were used, which are close to DenseNet's performance with 100% of the samples (n = 150,911), which achieved AUCs of 0.838 and 0.852, respectively. MedCLIP achieved higher AUCs of 0.885 on the internal test set and 0.883 on the external test set when it was trained with 100% of the samples. However, PASTER consistently outperforms MedCLIP under limited sample settings, demonstrating its advantage in small-sample learning scenarios. The performance of PASTER was evaluated using different sample sizes to understand the impact of training sample size on model performance. Extended Fig. 5 illustrates the differences in the AUCs between PASTER and DenseNet for the same sample sizes. We observe that the median AUC differences on the 5 echocardiographic labels increase by 5.7%/4.0%, 3.6%/2.9%, and 2.0%/2.0% in the internal/external test sets when 5%, 20%, and 100% of the training samples are used, respectively. When comparing PASTER using only 128 cases and 384 controls with DenseNet (20%), DenseNet (100%), and PASTER (100%), we note that PASTER achieves AUC differences with median values slightly lower by 0.2% (internal) and 0.3% (external) compared to DenseNet (100%). Similarly, PASTER outperforms DenseNet (20%), and the use of full samples still leads to improved accuracy for PASTER across all echocardiographic labels. More detailed comparisons are presented in Supplementary Data 4. We further assessed the correlation between the AUC obtained with different sample sizes

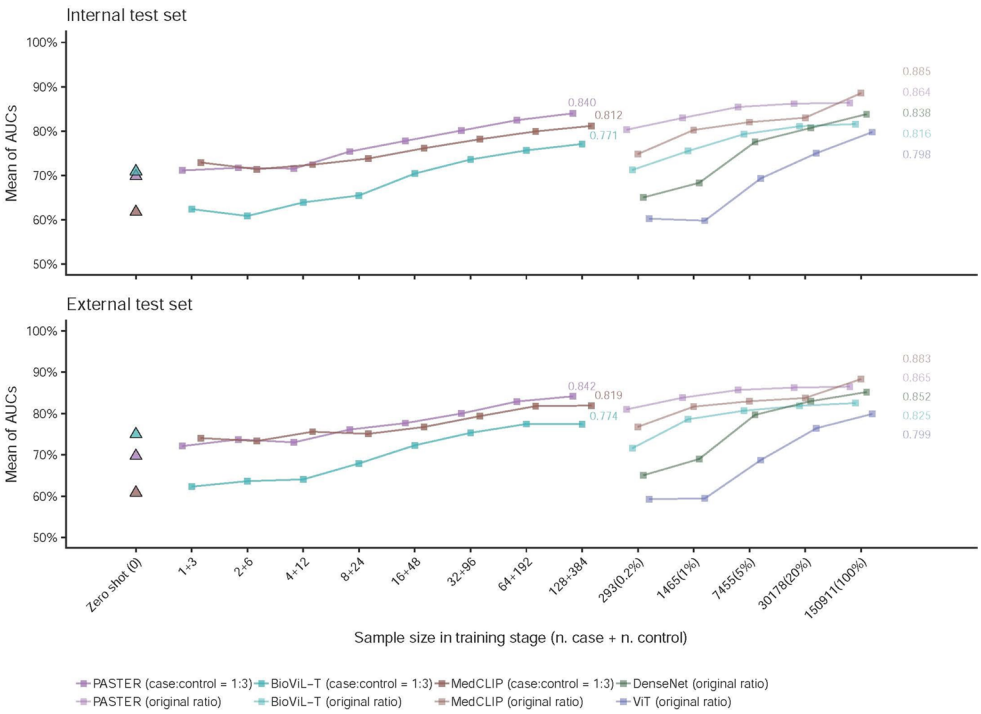

Fig. 4 Average AUC for five echocardiographic labels under varying sample sizes using linear probing for PASTER, BioViL-T, MedCLIP, DenseNet, and ViT. Models were trained with increasing numbers of samples, and average AUC values were computed across five echocardiographic labels

and the AUC from the full sample (Extended Fig. 6). The correlation on echocardiographic labels during zero-shot prediction was notably lower than that on radiological labels, with all correlation coefficients below 0.4. However, when

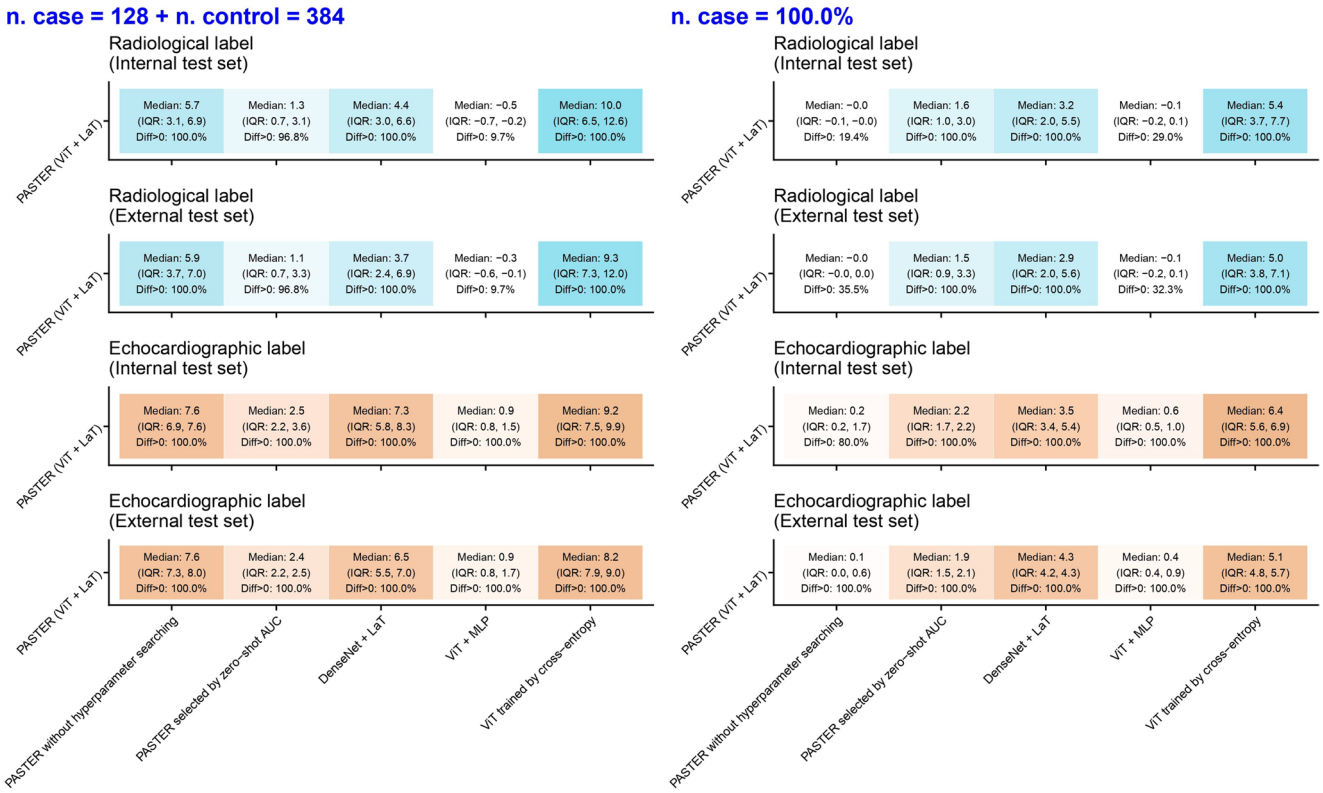

Fig. 5 Ablation study comparing architectural components and training strategies based on model variants of PASTER. The variants include combinations of different image and text encoders, loss functions, and training procedures. The results are shown as the median and interquartile range (IQR) of AUC differences at two settings: $n = 128$ cases (384 controls) and full data. Darker background colors indicate larger AUC differences

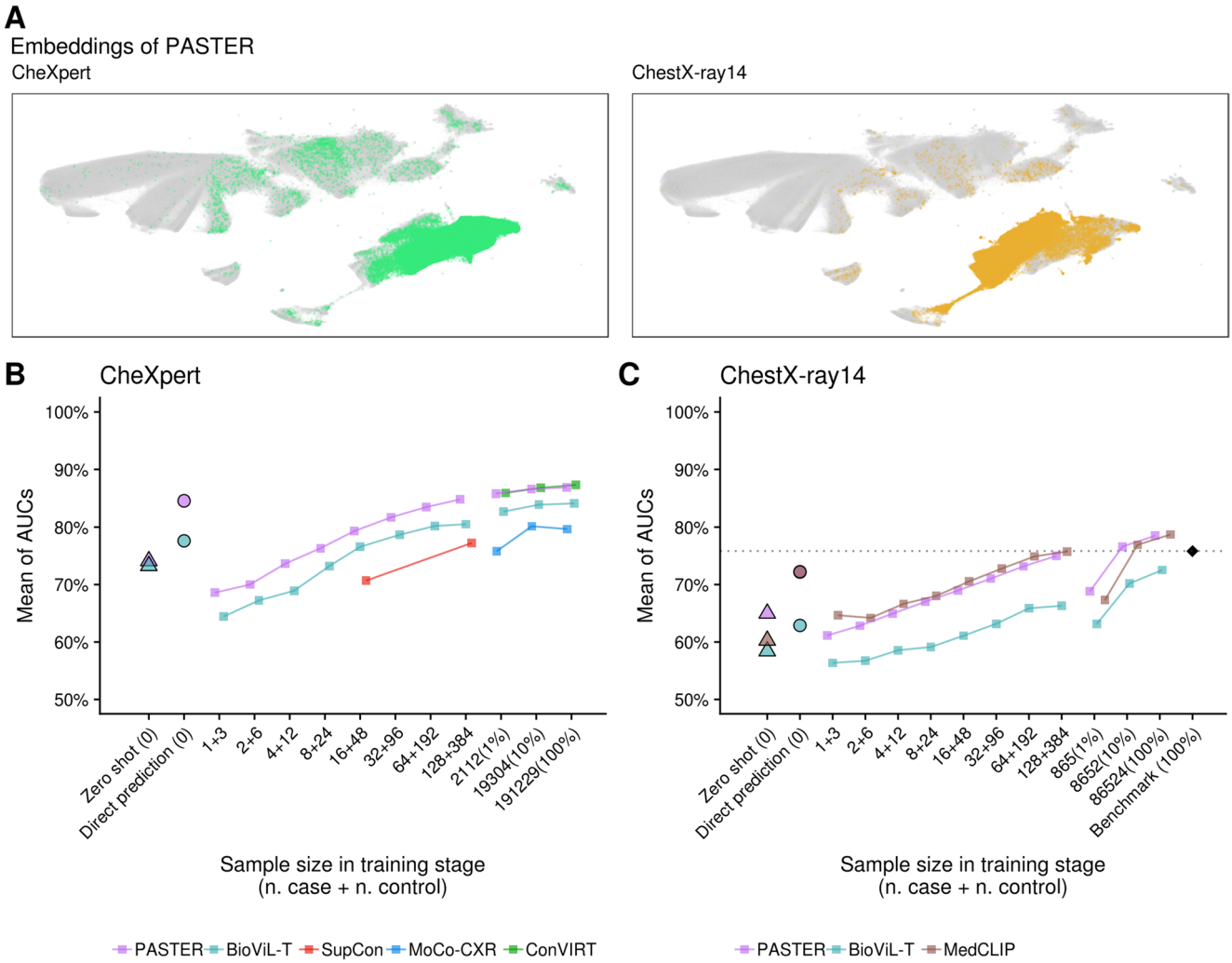

Fig. 6 Comparison of PASTER on public datasets (CheXpert [20] and ChestX-ray14 [21]) using UMAP visualization and AUC comparisons. (**A**) UMAP visualizations of image embeddings from PASTER. Each subplot shows samples from the specified dataset (green or brown), with other datasets shown in gray, including the private dataset. (**B**) Average AUC for five labels in the CheXpert dataset under varying training sample sizes. (**C**) Average AUC for eight labels in the ChestX-ray14 dataset under varying training sample sizes. The black diamonds and dashed lines indicate the benchmark results. PASTER, BioViL-T, and MedCLIP are shown for comparison [18, 27]

PASTER was trained with only 16 cases and 48 controls for linear probing, the correlations significantly increased to above 0.96. Extended Fig. 6B explains this phenomenon, taking left ventricular dysfunction (LVD) as an example. LVD is considered challenging for radiologists to identify directly from CXR [29]. Therefore, the zero-shot prediction of PASTER was inaccurate. However, using CXR to predict LVD actually yielded the highest ultimate AUC among all the echocardiographic labels (0.907/0.901 in the internal/external test set). Small-sample learning using only 16 cases and 48 controls can effectively reveal potential correlations in such cases. Confusion matrices for the five echocardiographic labels are presented in Extended Fig. 7, highlighting the classification performance of PASTER across the internal and external test sets.

We conducted ablation studies to understand the effects of different components and settings on the performance of PASTER. Supplementary Data 5 presents the results of using the PASTER technique with linear probing and full-parameter fine-tuning. The results are shown in Fig. 5. In small-sample learning, hyperparameter tuning was crucial, with default hyperparameters resulting in a median AUC decrease of 5.7–7.6% across both test sets and label types. When full samples were used, hyperparameter tuning was less critical. Furthermore, we chose models based on the zero-shot AUC from the validation set rather than validation loss (details in Extended Fig. 8). Selecting models based on the zero-shot AUC from the validation set instead of validation loss led to a median AUC decrease of 1.1–2.5%. Replacing the image encoder from ViT to DenseNet

Table 5 Quantitative and qualitative results of cross-modal retrieval

Model	Internal test set		External test set	
	Recall@1	Recall@5	Recall@1	Recall@5
PASTER	0.096	0.242	0.113	0.262
MedCLIP	0.001	0.005	0.001	0.005
BioViL-T	0.002	0.024	0.005	0.019

significantly decreased the AUC, whereas replacing the text encoder with a multilayer perceptron (MLP) slightly improved the radiological label accuracy but decreased echocardiographic label accuracy. Using structured labels for pretraining with the ViT trained by cross-entropy also reduced accuracy by more than 5%. Detailed comparisons are presented in Supplementary Data 6.

We applied PASTER to two publicly available datasets, CheXpert and ChestX-ray14, to evaluate their performance. The differences in the embeddings between the CXRs from CheXpert and ChestX-ray14 are shown in Fig. 6A. Additionally, we observed distinctions between our private dataset and these two public datasets (further details in Extended Fig. 9). Supplementary Data 7 provides a detailed comparison of the AUC for each label and the mean AUC for PASTER, BioViL-T, SupCon, and MoCo-CXR on CheXpert. As shown in Fig. 6B, PASTER maintained superior performance across different sample sizes, performing similarly to ConVIRT [16] and significantly outperformed the series-analysis based BioViL-T, the supervised contrastive learning-based SupCon [26], and the image contrastive learning-based MoCo-CXR [28]. Supplementary Data 8 provides a detailed comparison of the AUC for each label and mean AUC for PASTER, BioViL-T, MedCLIP, and the benchmark on ChestX-ray14. As shown in Fig. 6C, PASTER outperforms zero-shot predictions even when it is trained with only 128 cases and 384 controls and achieves further improvements when it is trained with 100% of the data [21]. Moreover, compared with MedCLIP, PASTER performed similarly and significantly outperformed BioViL-T. Extended Fig. 10A presents a comparison between PASTER on CheXpert with the same sample size. We compared the performance of PASTER for different sample sizes, and as previously observed, we found that 128 cases and 384 controls yield higher accuracy than the zero-shot predictions, although the accuracy slightly decrease when 100% of the samples are used. Notably, the model trained on the private dataset, referred to as "PASTER (direct predict)", achieved a level of accuracy similar to that of PASTER (128+384). Detailed comparisons of each label are presented in Supplementary Data 9. Extended Fig. 10B compared the results of PASTER on ChestX-ray14. We observed that compared with zero-shot predictions, PASTER achieved higher accuracy when 128 cases and 384 controls were used, and further improvements were seen when 100% of the samples were used. Notably, compared with PASTER (direct prediction), PASTER (128+384) performed better on ChestX-ray14. Detailed comparisons of each label are presented in Supplementary Data 10.

Table 5 shows the quantitative results of cross-modal retrieval. PASTER substantially outperformed both MedCLIP and BioViL-T in all the retrieval settings. On the internal test set, PASTER achieved a Recall@1 of 0.096 and a Recall@5 of 0.242, compared with 0.001/0.005 for MedCLIP and 0.002/0.024 for BioViL-T. Similar trends were observed on the external test set, where PASTER reached 0.113/0.262 for Recall@1 and Recall@5, while MedCLIP and BioViL-T remained below 0.02 in all the metrics. Extended Fig. 11 illustrates representative qualitative examples of the top-3 chest radiographs retrieved by PASTER for a given free-text report.

Discussion

PASTER achieved CNN-level accuracy with only 128 cases and 384 controls for linear probing. This highlights the strength of the model in low-data regimes and its potential value for real-world clinical settings where data collection is limited. Notably, even though the free-text reports of the pretraining dataset likely lacked descriptions of echocardiographic findings, the embeddings of CXRs extracted by PASTER still correlated strongly with these labels, highlighting its potential to transcend existing knowledge. We observed that PASTER's embeddings, although trained on radiological reports, also performed well on echocardiographic prediction tasks, suggesting that the learned representations capture clinically relevant anatomical or pathological patterns beyond the original supervision scope. Ablation experiments indicated that the success of PASTER is due primarily to effective contrastive learning. Replacing the language encoder with an MLP using 31 labels yielded similar performance. However, substituting the vision encoder (ViT) with a CNN led to a significant decrease in the AUC, highlighting the importance of Transformer-based representations. All the pretraining models based on CLIP technology (PASTER, ConVIRT, and MedCLIP) outperformed the other models in terms of small-sample learning, as validated across different datasets. These results collectively suggest that both architecture and pretraining strategy play crucial roles in enabling generalizable performance with minimal supervision.

In this study, pretrained models were applied to extremely small CXR datasets, and the results demonstrated that small-sample learning on local data could be beneficial, particularly as most radiology AI systems experience diminished performance during external validation [30]. We faced

similar challenges when we applied the model trained on our private dataset to CheXpert and ChestX-ray14. Previous research emphasized simpler "homegrown" models [31], making it more feasible to collect a few hundred samples for retraining with PASTER compared to the 20,000 samples required for CNN retraining in earlier studies [32]. These findings imply that the lightweight adaptation of pretrained multimodal models may offer a more scalable and accessible alternative for many institutions.

Zero-shot prediction contrasts with small sample-learning, as it requires no additional samples. However, our research revealed that PASTER, using a few dozen training samples for linear probing can be more accurate than zero-shot prediction while requiring substantially fewer samples than the hundreds typically reported in regulatory guidance for clinical validation [19]. Similarly, the results from MedCLIP and BioVi-T also revealed the importance of a few dozen training samples. Previous studies with general images also supported small-sample learning over zero-shot prediction [15]. Moreover, zero-shot prediction in medicine has several limitations: It struggles to predict pathologies not described in reports and still needs annotated samples to determine condition-specific probability thresholds [33]. Applying PASTER to zero-shot prediction for echocardiogram-related diseases resulted in significantly worse accuracy than ultimate the results. Therefore, in the context of medical imaging, particularly CXRs, our findings suggest prioritizing small-sample learning over zero-shot inference for practical and regulatory considerations.

PASTER demonstrated clear advantages over CNNs in small-sample learning and maintained better accuracy at the million-level training size, likely because of its well-trained Transformer using contrastive learning [24]. Training ViT for each label with the entire dataset yielded results inferior to those of CNNs, likely reflecting the limited availability of medical imaging data. Previous studies have shown that ViT requires more than 100 million samples to surpass CNNs [25, 34]. Notably, replacing PASTER's ViT with a CNN reduced accuracy, suggesting the effectiveness of ViT-based representations. Compared with direct linear probing, full fine-tuning did not yield better results, indicating the robustness of PASTER with smaller training sizes [35].

Pretraining with CLIP technology outperforms other models, leveraging free-text reports for higher accuracy [16]. SupCon [26], which uses limited structured labels, performed slightly worse. Both MedCLIP and ConVIRT, which adopt similar vision-language contrastive pretraining strategies, achieved comparable performance and consistently outperformed SupCon and MoCo-CXR, which rely on structured labels or image-only contrastive learning approaches [16, 18, 28, 36]. However, in our private dataset, the cross-modal retrieval results showed that MedCLIP performed worse than expected, possibly because the style of the CXR reports differed from those used in its training. This cross-dataset comparison suggests that PASTER may retain its performance across datasets, indicating potential robustness for clinical deployment in diverse settings. Considering our findings and the limited effectiveness of pretraining on general images for medical image analysis [16], as well as the privacy concerns associated with medical data [37], we suggest the need for a multinational, multicenter federated learning pre-training initiative using CXR and its reports to enhance CXR analysis.

AI models like PASTER have the potential to enhance scientific research by identifying previously underexplored correlations in CXR data [23]. Despite the systematic approaches used by radiologists, significant knowledge gaps remain, and several studies have reported that AI models can outperform radiologists in specific diagnostic tasks [38, 39]. High-quality annotations of CXR images can reveal unexpected applications, such as supporting the detection of heart failure, which can be challenging for radiologists [29, 40–42]. PASTER achieved high accuracy for diseases not described in CXR reports, suggesting its potential for further exploration of related CXR conditions. This could lead to "opportunistic screening" [43], predicting extensive nonadaptive disease risks from a single CXR, similar to the incidental findings in radiology [44]. PASTER could expand clinical indications beyond the current scope of CXR by addressing certain diagnostic gaps.

This study has several limitations. The pretraining dataset included only Taiwanese individuals, but further analysis on public datasets provided supportive evidence of performance on other populations. Owing to limited computational resources, we could not conduct extensive hyperparameter searches for CNN and ViT. PASTER was trained only on frontal CXRs, resulting in somewhat lower performance on datasets including lateral CXRs, such as CheXpert. The absence of multiple radiologists' collective annotations may introduce bias [8], although echocardiographic labels were treated as reference standards [7]. In addition, this study did not include retrieval-based or similarity-based downstream tasks. The probing results provide indirect evidence of semantic alignment but do not directly measure cross-modal retrieval ability. Future work could address these limitations by incorporating multi-institutional datasets, exploring federated learning strategies, validating multiview radiographs with broader label consensus, and expanding evaluation to include retrieval-based and interpretability analyses.

Conclusion

In this study, we explored the use of a pretrained multimodal model for analyzing chest X-rays in real-world settings where only limited labeled data are available. By leveraging large-scale image-text pretraining, we showed that strong performance can be achieved even with just a few hundred labeled samples. Given that chest X-rays and radiological reports are routinely available in clinical systems, such models may provide an efficient and scalable foundation for a wide range of downstream applications without the need for burdensome annotation efforts.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10916-025-02263-3.

Author Contributions C.L. contributed to the study conception and design. K.C.C. and C.L. drafted the initial manuscript. K.C.C., M.K., and C.H.L. trained the foundation model. K.C.C. and C.L. trained the supervised learning model. K.C.C. and S.A.L. provided the API for future use. K.C.C, D.J.T., and C.L. analyzed the model performance. H.C.L., W.H.F., and C.S.L. assisted with data interpretation. C.W.H., C.K.C., K.H.K., Y.C.H., W.C.C, G.S.H., Y.H.C., Y.J.H, and C.S.T. contributed to data acquisition. H.C.L., W.H.F., C.S.L., and S.H.L. revised the manuscript for important intellectual content. C.L. took final responsibility for this article and provided final approval of the version to be published.

Funding This study was supported by funding from the National Science and Technology Council, Taiwan (NSTC 114-2321-B-016-005 to Shih-Hua Lin) and the Medical Affairs Bureau (MND-MAB-C07-113021 to Chin Lin).

Data Availability The data from Tri-Service General Hospital utilized in this study are not publicly available because of patient privacy concerns. The proposed multimodal model with well-trained parameters may be released to other researchers on the IRB-approved agreement of the Tri-Service General Hospital. CheXpert data are available at https://aimi.stanford.edu/chexpert-chest-x-rays. The ChestX-ray14 data are available at https://nihcc.app.box.com/v/ChestXray-NIHCC?sortColumn=date&sortDirection=ASC.

Code availability The code used to train and evaluate the proposed multimodal model using text-and-image pairs is available on GitHub at https://github.com/ji9su/PASTER. Furthermore, the repository includes detailed instructions for future researchers on how to use the call service to obtain embeddings generated by PASTER by uploading chest radiographs. Additionally, we provide linear probing weights for predicting the 31 radiological labels and 5 echocardiographic labels, enabling future researchers to apply them directly.

Declarations

Competing Interests The authors declare no competing interests.

Institutional Review Board This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Tri-Service General Hospital (IRB NO. C20230519). The IRB approved the study protocol and waived the requirement for individual informed consent due to the use of fully anonymized and retrospective data.

Informed Consent All the data were obtained from the hospital's quality control center, fully anonymized prior to analysis, and exempt from informed consent as approved by the Institutional Review Board.

Conflict of interest The authors have no conflicts of interest to declare.

References

1. Raoof, S., et al., Interpretation of plain chest roentgenogram. *Chest*, 2012. 141(2): p. 545–558.
2. Çallı, E., et al., Deep learning for chest X-ray analysis: A survey. 2021. 72: p. 102125.
3. Litjens, G., et al., A survey on deep learning in medical image analysis. 2017. 42: p. 60–88.
4. Liu, X., et al., A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health*, 2019. 1(6): p. e271-e297.
5. Aggarwal, R., et al., Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit Med*, 2021. 4(1): p. 65.
6. Rajpurkar, P., et al., AI in health and medicine. *Nat Med*, 2022. 28(1): p. 31–38.
7. Willemink, M.J., et al., Preparing medical imaging data for machine learning. 2020. 295(1): p. 4–15.
8. Joskowicz, L., et al., Inter-observer variability of manual contour delineation of structures in CT. 2019. 29: p. 1391–1399.
9. Lutnick, B., et al., An integrated iterative annotation technique for easing neural network training in medical image analysis. 2019. 1(2): p. 112–119.
10. Esteva, A., et al., A guide to deep learning in healthcare. 2019. 25(1): p. 24–29.
11. Kim, H.E., et al., Transfer learning for medical image classification: a literature review. 2022. 22(1): p. 69.
12. Chen, T., et al. *A simple framework for contrastive learning of visual representations*. in *International conference on machine learning*. 2020. PMLR.
13. He, K., et al. *Momentum contrast for unsupervised visual representation learning*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
14. He, K., et al. *Masked autoencoders are scalable vision learners*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
15. Radford, A., et al. *Learning transferable visual models from natural language supervision*. in *International conference on machine learning*. 2021. PMLR.
16. Zhang, Y., et al. *Contrastive learning of medical visual representations from paired images and text*. in *Machine Learning for Healthcare Conference*. 2022. PMLR.
17. Zhang, S., et al., A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. *Nejm Ai*, 2025. 2(1).
18. Wang, Z., et al., MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. *Proc Conf Empir Methods Nat Lang Process*, 2022. 2022: p. 3876–3887.
19. Benjamens, S., P. Dhunnoo, and B.J.N.d.m. Meskó, The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. 2020. 3(1): p. 118.
20. Irvin, J., et al. *Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison*. in *Proceedings of the AAAI conference on artificial intelligence*. 2019.
21. Wang, X., et al. *Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and*

localization of common thorax diseases. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
22. Huang, G., et al. *Densely connected convolutional networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
23. Wang, H., et al., Scientific discovery in the age of artificial intelligence. 2023. 620(7972): p. 47–60.
24. Manzari, O.N., et al., MedViT: a robust vision transformer for generalized medical image classification. 2023. 157: p. 106791.
25. Dosovitskiy, A., et al., An image is worth 16x16 words: Transformers for image recognition at scale. 2020.
26. Sellergren, A.B., et al., Simplified transfer learning for chest radiography models using less data. 2022. 305(2): p. 454–465.
27. Bannur, S., et al. *Learning to exploit temporal structure for biomedical vision-language processing*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
28. Sowrirajan, H., et al. *Moco pretraining improves representation and transferability of chest x-ray models*. in *Medical Imaging with Deep Learning*. 2021. PMLR.
29. Lauzier, P.T. and B.J.W. Chow, Artificial Intelligence Detection of Left Ventricular Systolic Dysfunction Using Chest X-Rays: Prospective Validation, Please. *The Canadian journal of cardiology*, 2022. 38(6): p. 720–722.
30. Yu, A.C., B. Mohajer, and J.J.R.A.I. Eng, External validation of deep learning algorithms for radiologic diagnosis: a systematic review. 2022. 4(3): p. e210064.
31. van Ginneken, B.J.R., *Deep learning for triage of chest radiographs: should every institution train its own system?*Radiology, 2019. Radiological Society of North America: Radiology. p. 545–546.
32. Dunnmon, J.A., et al., Assessment of convolutional neural networks for automated classification of chest radiographs. 2019. 290(2): p. 537–544.
33. Tiu, E., et al., Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. 2022. 6(12): p. 1399–1406.
34. Vaid, A., et al., A foundational vision transformer improves diagnostic performance for electrocardiograms. 2023. 6(1): p. 108.
35. Kumar, A., et al. *Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution*. in *International Conference on Learning Representations*. 2022.
36. Sellergren, A.B., et al., Simplified transfer learning for chest radiography models using less data. *Radiology*, 2022. 305(2): p. 454–465.
37. Rieke, N., et al., The future of digital health with federated learning. 2020. 3(1): p. 1–7.
38. Liu, W.-T., et al., A deep-learning algorithm-enhanced system integrating electrocardiograms and chest X-rays for diagnosing aortic dissection. 2022. 38(2): p. 160–168.
39. Rajpurkar, P., et al., Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. 2018. 15(11): p. e1002686.
40. Bluemke, D.A., et al., *Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the radiology editorial board*.Radiology, 2020. Radiological Society of North America: Radiology. p. 487–489.
41. Hsiang, C., et al., Detection of Left Ventricular Systolic Dysfunction Using an Artificial Intelligence-Enabled Chest X-Ray. *The Canadian journal of cardiology*, 2022. 38(6): p. 763–773.
42. Seah, J.C., et al., Chest radiographs in congestive heart failure: visualizing neural network learning. 2019. 290(2): p. 514–522.
43. Pyrros, A., et al., Opportunistic detection of type 2 diabetes using deep learning from frontal chest radiographs. *Nat Commun*, 2023. 14(1): p. 4039.
44. Berland, L.L., et al., Managing incidental findings on abdominal CT: white paper of the ACR incidental findings committee. 2010. 7(10): p. 754–773.

Authors and Affiliations

Kai-Chieh Chen[1] · Matthew Kuo[2] · Chun-Ho Lee[3] · Hao-Chun Liao[4] · Dung-Jang Tsai[5,6] · Shing-An Lin[6] · Chih-Wei Hsiang[7] · Cheng-Kuang Chang[7] · Kai-Hsiung Ko[7] · Yi-Chih Hsu[7] · Wei-Chou Chang[7] · Guo-Shu Huang[7] · Wen-Hui Fang[8] · Chin-Sheng Lin[5,6,9] · Shih-Hua Lin[10] · Yuan-Hao Chen[11] · Yi-Jen Hung[12] · Chien-Sung Tsai[13] · Chin Lin[1,5,6]

✉ Chin Lin
xup6fup@mail.ndmctsgh.edu.tw

1 Graduate Institute of Life Sciences, College of Biomedical Sciences, National Defense Medical University, Taipei, Taiwan, R.O.C.

2 School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA

3 School of Public Health, College of Public Health, National Defense Medical University, Taipei, Taiwan, R.O.C.

4 Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, R.O.C.

5 Medical Technology Education Center, School of Medicine, College of Medicine, National Defense Medical University, Taipei, Taiwan, R.O.C.

6 Tri-Service General Hospital, Military Digital Medical Center, National Defense Medical University, Taipei, Taiwan, R.O.C.

7 Department of Radiology, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, R.O.C.

8 Department of Family and Community Medicine, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, R.O.C.

9 Division of Cardiology, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, R.O.C.

10 Division of Nephrology, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, R.O.C.

11 Department of Neurological Surgery, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, R.O.C.

12 Division of Endocrinology and Metabolism, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, R.O.C.

13 Division of Cardiovascular Surgery, Department of Surgery, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, R.O.C.