Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CUPID-Med: Conformal Uncertainty-aware Prototype Discovery for Clinically Interpretable Deep Medical Clustering

Mojtaba Jahanian

mojtaba160672000@aut.ac.ir

Islamic Azad University of Hamedan

Abbas Karimi

Islamic Azad University, Karaj

Nafiseh Osati Eraghi

Islamic Azad University, Arak

Faraneh Zarafshan

Islamic Azad University, Karaj

Article

Keywords: deep medical clustering, medical foundation models, conformal prediction, uncertainty quantification, prototype learning, interpretable medical AI, self-supervised learning, CheXpert.

Posted Date: July 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10047883/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on October 19th, 2026. See the published version at https://doi.org/10.1038/s41598-026-75016-y.

CUPID-Med: Conformal Uncertainty-aware Prototype Discovery for Clinically Interpretable Deep Medical Clustering

Mojtaba Jahanian[1,2,*], Abbas Karimi[4,5,*], Nafiseh Osati Eraghi[3], and Faraneh Zarafshan[4,5]

[1]Department of Computer Engineering, National University of Skills (NUS), Tehran, Iran
[2]Department of Computer Engineering, Ha.C., Islamic Azad University, Hamedan, Iran
[3]Department of Computer, Arak Branch, Islamic Azad University, Arak, Iran
[4]Department of Computer Engineering, Karaj Branch, Islamic Azad University, Karaj, Iran
[5]Institute of Artificial Intelligence and Social and Advanced Technologies, Karaj Branch, Islamic Azad University, Karaj, Iran
[*]Corresponding authors: Mojtaba Jahanian (Mojtaba160672000@aut.ac.ir) and Abbas Karimi (Akarimi@iau.ac.ir)

Abstract

Deep medical clustering aims to discover latent disease subgroups from unlabeled medical data. However, most existing approaches produce deterministic cluster assignments and provide limited mechanisms for calibrated uncertainty communication or clinical interpretation. This paper proposes CUPID-Med, a conformal uncertainty-aware prototype discovery framework for clinically interpretable deep medical clustering. CUPID-Med integrates medical foundation encoders, distributional latent representation learning, prototype-based deep clustering, conformal uncertainty calibration, and clinically interpretable outputs. Rather than assigning every sample to a single hard cluster, CUPID-Med generates calibrated set-valued cluster predictions and flags ambiguous cases for expert review. The framework was evaluated on a 5,000-image frontal CheXpert subset using a no-label training protocol, where disease labels were excluded from training and used only for post-hoc clinical interpretation. Preliminary clustering baselines applied to DenseNet121 embeddings produced weak separation, with the best baseline Silhouette score of 0.0510. In contrast, the primary CUPID-Med run achieved a Silhouette score of 0.5175, a Davies–Bouldin index of 2.7809, and a Calinski–Harabasz score of 29,733.16, while also producing prototype assignments, confidence estimates, and latent uncertainty scores. These results suggest that distributional prototype learning can produce more structured and interpretable medical image subgroups than deterministic clustering of fixed deep embeddings.

Keywords: deep medical clustering; medical foundation models; conformal prediction; uncertainty quantification; prototype learning; interpretable medical AI; self-supervised learning; CheXpert.

“‘latex

1 Introduction

The rapid expansion of medical imaging, electronic health records (EHRs), and multimodal clinical repositories has created major opportunities for data-driven discovery of disease phenotypes, patient subgroups, and latent clinical trajectories. Modern healthcare systems routinely generate large volumes of radiological images, laboratory measurements, demographic variables, physiological time series, medication records, and diagnostic reports. However, only a limited portion

of these data is annotated at the level required for supervised deep learning. Expert annotation remains expensive, time-consuming, institution-dependent, and susceptible to inter-observer variability. This limitation is particularly important in medical imaging and clinical decision support, where disease manifestations are often heterogeneous, overlapping, and continuous rather than clearly separable into predefined diagnostic classes. Consequently, unsupervised and self-supervised learning have become increasingly important for extracting clinically meaningful structure from large-scale unlabeled medical data [5, 6, 7].

Deep clustering aims to jointly learn discriminative representations and cluster assignments in a unified latent space. Unlike conventional clustering methods such as k-means, Gaussian mixture models, or spectral clustering, which operate on fixed handcrafted or pretrained features, deep clustering methods optimize representation learning and grouping objectives simultaneously. Early methods such as Deep Embedded Clustering (DEC) demonstrated that representation learning and clustering can be coupled through a self-training objective based on target-distribution refinement [1]. Subsequent approaches have incorporated contrastive learning, pseudo-labeling, graph structure, and prototype-based learning to improve clustering quality and robustness [5, 8, 9]. Recent surveys emphasize that the central challenge in deep image clustering is no longer merely the selection of a clustering algorithm, but rather the interaction between representation learning, semantic grouping, assignment refinement, and stability under perturbations [6, 7].

In the medical domain, deep clustering is particularly relevant because clinically meaningful labels are often incomplete, noisy, delayed, or unavailable. Medical subgroups may correspond to latent disease phenotypes, imaging patterns, risk profiles, treatment response groups, or progression trajectories that are not explicitly encoded in standard diagnostic labels. Clustering-based analyses have been used to stratify patient cohorts from EHR data and to reveal clinically interpretable subtypes with distinct risk characteristics [10]. In medical imaging, clustering can support cohort discovery, dataset organization, weak-label refinement, anomaly detection, and exploratory phenotype analysis. Nevertheless, medical deep clustering remains challenging due to high intra-class variability, subtle inter-class differences, acquisition heterogeneity, class imbalance, missing modalities, and the frequent absence of reliable ground-truth cluster labels.

The emergence of medical foundation models provides a promising direction for addressing some of these limitations. Foundation models pretrained on large-scale medical image collections or multimodal radiology data can learn general-purpose representations that are transferable across downstream tasks with limited supervision [12, 13]. In radiology, multimodal foundation models have been developed to integrate imaging data with clinical text and other patient-level information, enabling flexible representation learning across imaging modalities and clinical contexts [14, 15]. However, despite their representational power, foundation models do not by themselves solve clinically reliable unsupervised subgroup discovery. Their embeddings may encode acquisition artifacts, demographic confounders, institution-specific biases, or non-clinical variation. Therefore, a medical clustering framework must not only exploit strong pretrained representations, but also impose clustering-specific structure, uncertainty awareness, and clinical interpretability.

Uncertainty quantification is essential for trustworthy medical clustering. In clinical practice, uncertainty is not a secondary property but a fundamental component of decision-making. Ambiguous imaging findings, borderline laboratory values, overlapping disease phenotypes, and incomplete clinical histories often require expert review rather than automatic hard assignment. Standard deep clustering methods typically produce a single cluster label or a soft assignment vector, but rarely provide calibrated uncertainty estimates or formal mechanisms for identifying uncertain samples. Conformal prediction offers a principled framework for uncertainty-aware prediction by constructing set-valued outputs with distribution-free coverage guarantees under exchangeability assumptions [16, 17, 18]. Although conformal prediction has been widely studied in supervised classification and regression, its integration with unsupervised deep medical

clustering remains underexplored.

Another barrier to clinical adoption is interpretability. A clustering result is useful in medicine only if the discovered groups can be inspected, summarized, and related to clinically meaningful characteristics. Cluster identifiers alone are insufficient. Clinicians require representative prototypes, clinical signatures, uncertainty profiles, and interpretable subgroup summaries that explain why cases are grouped together. This motivates clustering systems that are not only numerically effective, but also transparent, calibrated, and clinically meaningful.

Conventional deep clustering pipelines typically map each patient to a single deterministic cluster, even when the patient lies near the boundary between multiple latent phenotypes. This creates a gap between the mathematical behavior of clustering algorithms and the clinical reality of overlapping, uncertain, and heterogeneous disease patterns. CUPID-Med is designed to address this gap by moving from deterministic clustering toward calibrated, interpretable, uncertainty-aware medical subgroup discovery.

To address these limitations, we propose **CUPID-Med**: **C**onformal **U**ncertainty-aware **P**rototype **D**iscovery for **C**linically **I**nterpretable **D**eep **Med**ical Clustering. CUPID-Med integrates medical foundation encoders, distributional latent representation learning, prototype-based clustering, conformal uncertainty calibration, and clinically interpretable output generation. The central idea is that medical samples should not always be assigned to a single deterministic cluster. Instead, ambiguous samples should receive calibrated set-valued cluster predictions or be flagged for expert review.

The main contributions are as follows:

- We propose a conformal uncertainty-aware deep clustering framework for medical data that combines medical foundation encoding, distributional latent representation learning, and prototype-based subgroup discovery.
- We introduce a prototype discovery mechanism that yields representative cluster prototypes and supports clinically interpretable subgroup summaries.
- We develop a conformal cluster assignment strategy that produces set-valued predictions and supports expert-review flagging for uncertain or borderline cases.
- We define a CheXpert-based no-label training protocol in which disease labels are excluded from training and used only for post-hoc clinical enrichment, uncertainty alignment, and interpretability evaluation.

“‘

“‘latex

2 Related Work

Deep medical clustering sits at the intersection of representation learning, unsupervised grouping, uncertainty estimation, interpretability, and clinical integration. To clarify the position of CUPID-Med within this landscape, Fig. 1 summarizes the main methodological categories relevant to this study. The taxonomy organizes prior work into five complementary dimensions: representation learning, clustering strategy, uncertainty handling, interpretability, and clinical integration. This organization also highlights the methodological gap addressed by CUPID-Med: most existing methods improve one or two of these dimensions, whereas clinically useful medical clustering requires all of them to be considered jointly.

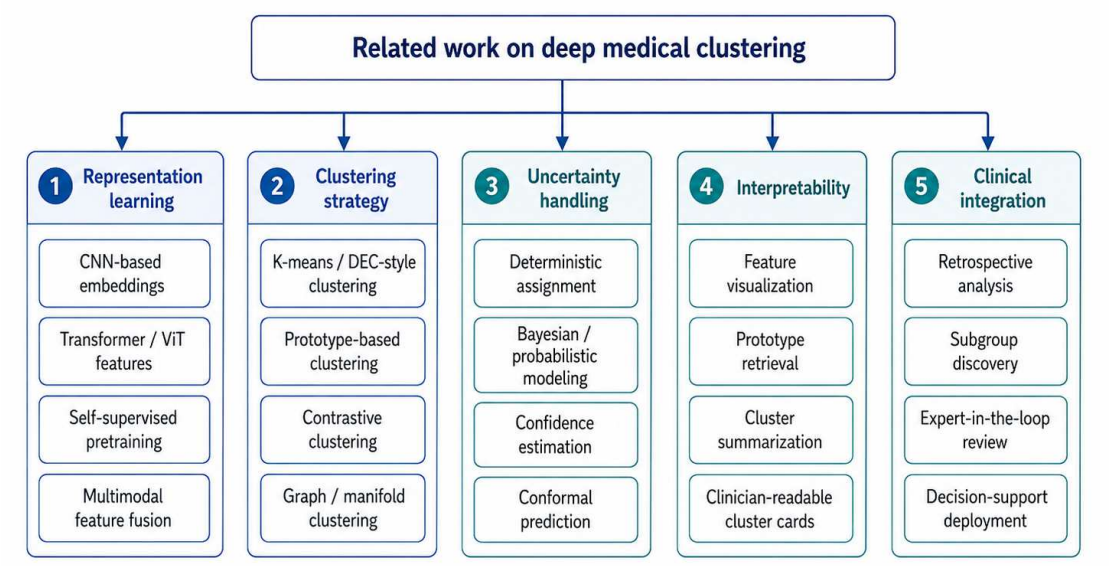


Figure 1: Taxonomy of related work in deep medical clustering. Existing studies can be organized according to representation learning, clustering strategy, uncertainty handling, interpretability, and clinical integration. CUPID-Med is positioned at the intersection of these categories by combining foundation encoder-based representations, prototype discovery, conformal uncertainty calibration, and clinically interpretable expert-in-the-loop outputs.

2.1 Representation learning for deep medical clustering

Deep clustering combines representation learning with unsupervised grouping. Classical clustering methods such as k-means, Gaussian mixture models, spectral clustering, and hierarchical clustering remain useful baselines, but their performance depends strongly on the quality of the input representation. Deep clustering addresses this limitation by learning embeddings that are optimized for grouping. Deep Embedded Clustering (DEC) introduced target-distribution refinement to sharpen soft cluster assignments in a learned latent space [1]. IDEC extended this idea by incorporating reconstruction constraints to preserve local structure [2]. VaDE formulated clustering as variational inference under a Gaussian mixture prior [3]. DeepCluster used cluster assignments as pseudo-labels for representation learning, while SwAV introduced online clustering with swapped prediction between augmented views [4, 5].

Recent deep image clustering methods increasingly integrate contrastive learning, graph structure, multi-view consistency, and prototype-based objectives [8, 9, 6, 7]. These approaches improve cluster discriminability and robustness by shaping the latent space more explicitly than conventional feature extraction. However, most of them are designed for generic image clustering rather than safety-critical medical subgroup discovery. In particular, they typically return deterministic cluster assignments and provide limited mechanisms for identifying ambiguous or borderline samples.

2.2 Clustering strategies for medical data

Medical clustering is used to discover patient subgroups, imaging phenotypes, risk profiles, and disease trajectories. In EHR analysis, deep clustering can reveal clinically meaningful patient subtypes from structured tabular data [10]. In medical imaging, clustering can support cohort discovery, dataset organization, weak-label refinement, anomaly detection, and exploratory phenotype analysis. Prior work on deep medical image clustering has shown that transfer learning can improve clustering quality when labeled data are scarce [11].

Despite these advances, medical clustering remains challenging because medical data are heterogeneous, noisy, multi-label, and often ambiguous. A chest radiograph may contain multiple

coexisting findings, and a patient may lie between several latent clinical phenotypes. Standard clustering pipelines generally assume that each sample belongs to one discrete group, which is poorly aligned with the continuous and overlapping nature of many clinical patterns. This motivates clustering strategies that can represent uncertainty explicitly rather than forcing every image into a single hard assignment.

2.3 Medical foundation models and multimodal encoders

Foundation models have become increasingly important in medical AI because they learn transferable representations from large-scale medical image collections, radiology reports, or multimodal clinical corpora [12, 13]. Generalist radiology foundation models and 3D vision-language medical models aim to integrate images, reports, and clinical context across tasks and modalities [14, 15]. These models can provide strong representations for downstream clustering and subgroup discovery.

However, fixed foundation-model embeddings do not automatically guarantee clinically meaningful clusters. They may encode acquisition artifacts, device-related patterns, demographic confounders, or institution-specific variation. Moreover, foundation encoders typically provide deterministic feature vectors rather than calibrated uncertainty-aware subgroup assignments. CUPID-Med uses medical foundation encoders as representation backbones, but adds distributional latent modeling, prototype discovery, and conformal calibration to make the resulting clusters more structured, reliable, and clinically interpretable.

2.4 Uncertainty handling in medical clustering

Uncertainty quantification is central to trustworthy medical AI. In medical clustering, uncertainty is especially important because the goal is not only to group similar cases, but also to identify when a case should not be assigned confidently to a single subgroup. Conformal prediction provides a distribution-free framework for constructing prediction sets with finite-sample coverage guarantees under exchangeability assumptions [16, 17]. In clinical settings, conformal prediction has been studied for uncertainty communication, reliability, and safer decision support [18, 19, 20, 21]. Recent work also highlights practical limitations, including distribution shift, calibration-set mismatch, and class imbalance [22].

Most conformal prediction methods target supervised classification or regression. Their use in unsupervised clustering remains much less explored. Existing clustering methods may report soft assignments or distance-to-centroid scores, but these values are usually not calibrated into interpretable decision sets. CUPID-Med adapts conformal reasoning to unsupervised medical clustering by converting prototype assignment scores into set-valued cluster predictions. Singleton sets indicate confident subgroup assignment, whereas multi-cluster sets identify ambiguous cases that should be reviewed by an expert.

2.5 Interpretability and clinical integration

Interpretability is a core requirement for clinical clustering. A clustering result is useful only if the discovered groups can be inspected, summarized, and related to clinically meaningful characteristics. Cluster identifiers alone are insufficient for clinical adoption. Clinicians need representative cases, prototype examples, subgroup summaries, confidence estimates, and interpretable outputs that explain why similar cases are grouped together.

Prototype-based methods are particularly relevant in this context because they represent groups using learnable or retrieved exemplars. In medicine, this is aligned with case-based reasoning, where clinicians compare a new patient with representative prior cases. Prototype retrieval, cluster cards, and subgroup-level summaries can help determine whether a cluster

reflects a meaningful radiographic phenotype, a technical artifact, or a mixed subgroup. CUPID-Med incorporates a learnable prototype memory bank and retrieves representative samples for each discovered subgroup, enabling cluster-level clinical interpretation.

Clinical integration also requires mechanisms for expert-in-the-loop review. Medical AI systems should not only return predictions, but also indicate when uncertainty is high and human review is needed. CUPID-Med addresses this requirement by combining prototype-based interpretability with conformal set-valued prediction, allowing uncertain or borderline cases to be flagged for expert review rather than being forced into a single cluster.

2.6 Positioning of CUPID-Med

The taxonomy in Fig. 1 shows that existing work has advanced representation learning, clustering objectives, uncertainty quantification, interpretability, and clinical subgroup discovery, but these directions are often treated separately. CUPID-Med is positioned as an integrative framework that combines these components within a single medical clustering pipeline. Specifically, it uses foundation encoder-based representations, learns a distributional latent space, organizes samples around interpretable prototypes, calibrates cluster assignments using conformal prediction, and produces outputs suitable for expert review.

This positioning defines the methodological contribution of CUPID-Med. Rather than treating clustering as a purely numerical grouping problem, CUPID-Med frames deep medical clustering as a calibrated and interpretable decision-support workflow. The goal is not only to improve internal clustering metrics, but also to separate confident subgroup assignments from ambiguous cases, support prototype-based inspection, and provide clinically meaningful summaries of discovered disease subgroups. “‘

3 Proposed Method

3.1 Overview

CUPID-Med is a conformal uncertainty-aware prototype discovery framework for clinically interpretable deep medical clustering. The overall architecture is shown in Fig. 2, while the detailed methodological modules are illustrated throughout the following subsections. The framework consists of five core stages: (i) medical data input and preprocessing, (ii) medical foundation encoding and multimodal fusion, (iii) distributional latent representation learning, (iv) prototype-based deep clustering, and (v) conformal uncertainty calibration with clinically interpretable outputs. The method is designed for unsupervised subgroup discovery: diagnostic labels are not required during training.

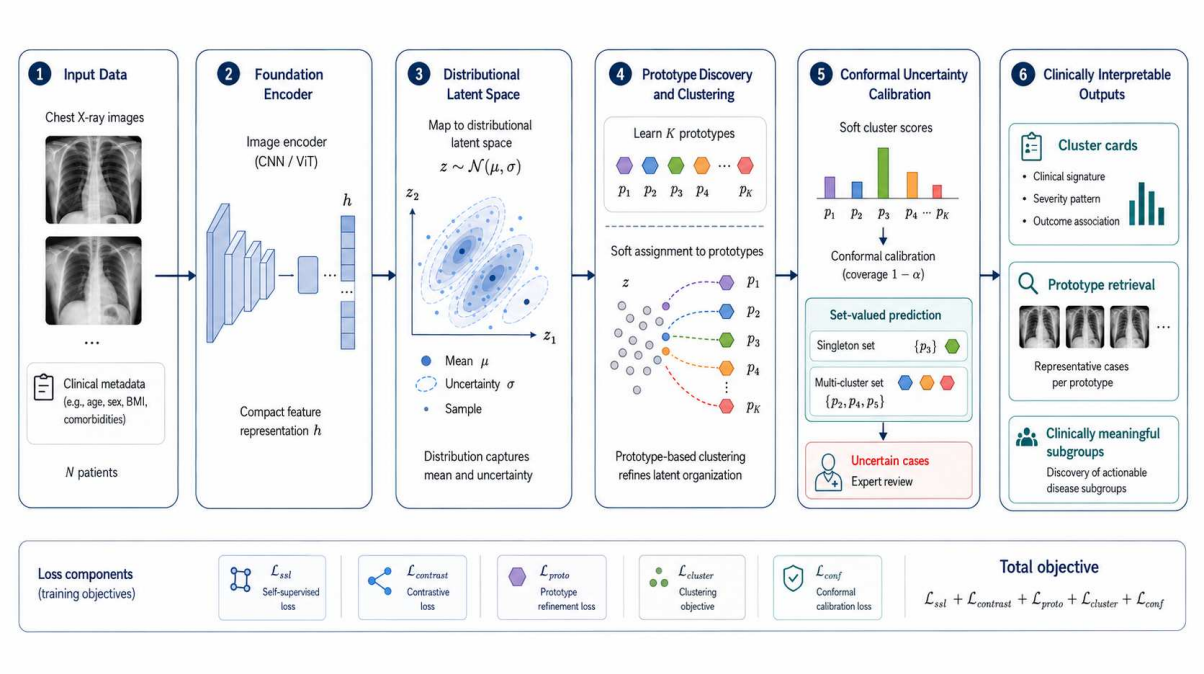


Figure 2: Overall architecture of CUPID-Med. Medical data are encoded using image/health and clinical encoders, mapped to a distributional latent space, organized around learnable prototypes, calibrated using conformal uncertainty, and converted into clinically interpretable outputs.

Given an unlabeled multimodal dataset

$$\mathcal{D} = \left\{ \left(\mathbf{x}_i^{img}, \mathbf{x}_i^{clin} \right) \right\}_{i=1}^{N}, \tag{1}$$

where $\mathbf{x}_i^{img}$ is the image and $\mathbf{x}_i^{clin}$ is the optional clinical vector, CUPID-Med learns K latent disease subgroups. In the CheXpert image-only setting, $\mathbf{x}_i^{clin}$ is omitted and the framework reduces to an image-based clustering model.

3.2 Medical data inputs

The framework can ingest heterogeneous medical inputs, including diagnostic images and structured clinical/EHR variables. In the general multimodal setting, image inputs may include chest radiographs, CT, MRI, or other modality-specific scans, whereas structured inputs may contain demographics, vital signs, comorbidities, laboratory measurements, medications, or outcome-related metadata. Figure 3 summarizes the input formulation used by CUPID-Med. In this study, the main experimental analysis focuses on the image-only CheXpert setting, but the full formulation is retained because CUPID-Med is designed to scale naturally to multimodal medical subgroup discovery.

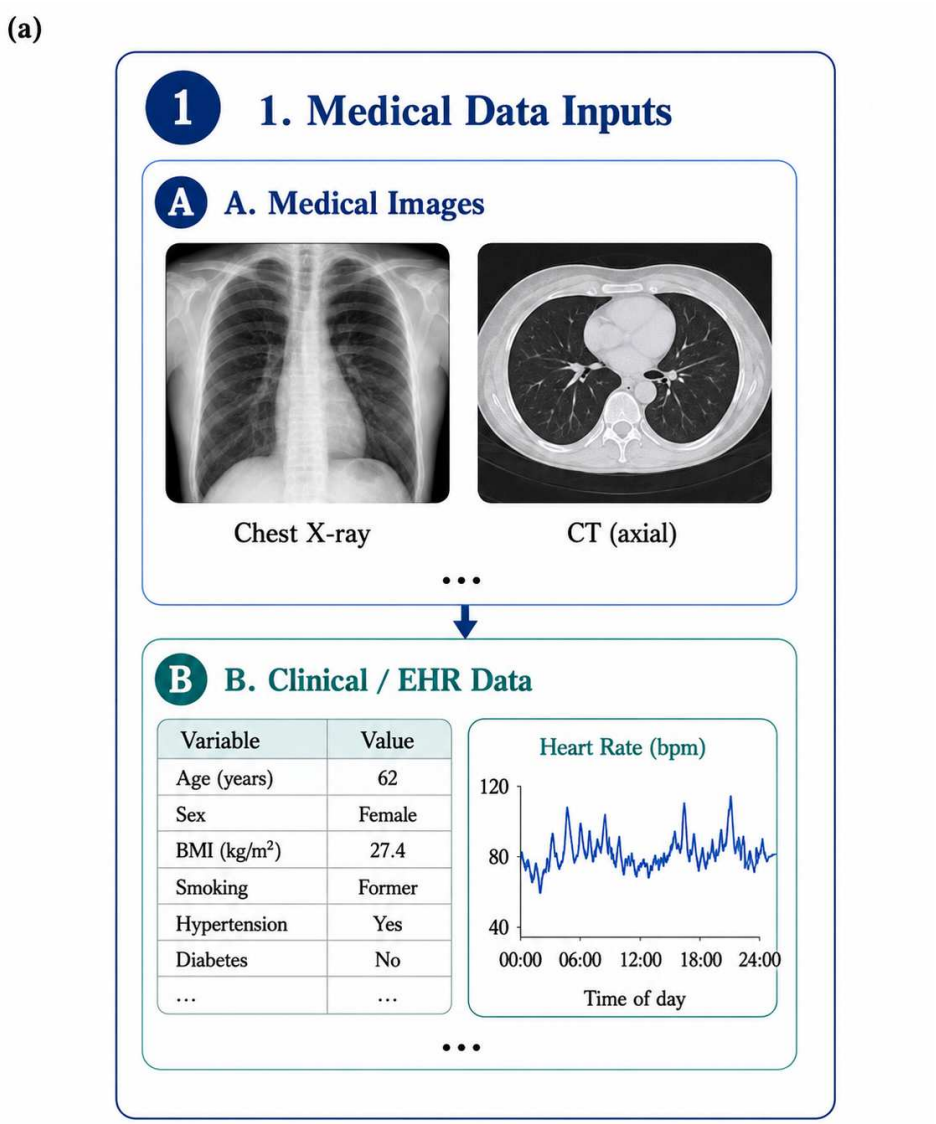


Figure 3: Medical data inputs considered by CUPID-Med. The framework supports both imaging inputs and structured clinical/EHR variables, enabling unimodal or multimodal disease subgroup discovery depending on dataset availability.

3.3 Medical foundation encoder

The medical foundation encoder maps raw medical inputs to a unified representation. Figure 4 provides the detailed architecture. The image/health encoder extracts visual representations, the clinical encoder processes structured variables when available, and the fusion module integrates modality tokens using Transformer blocks.

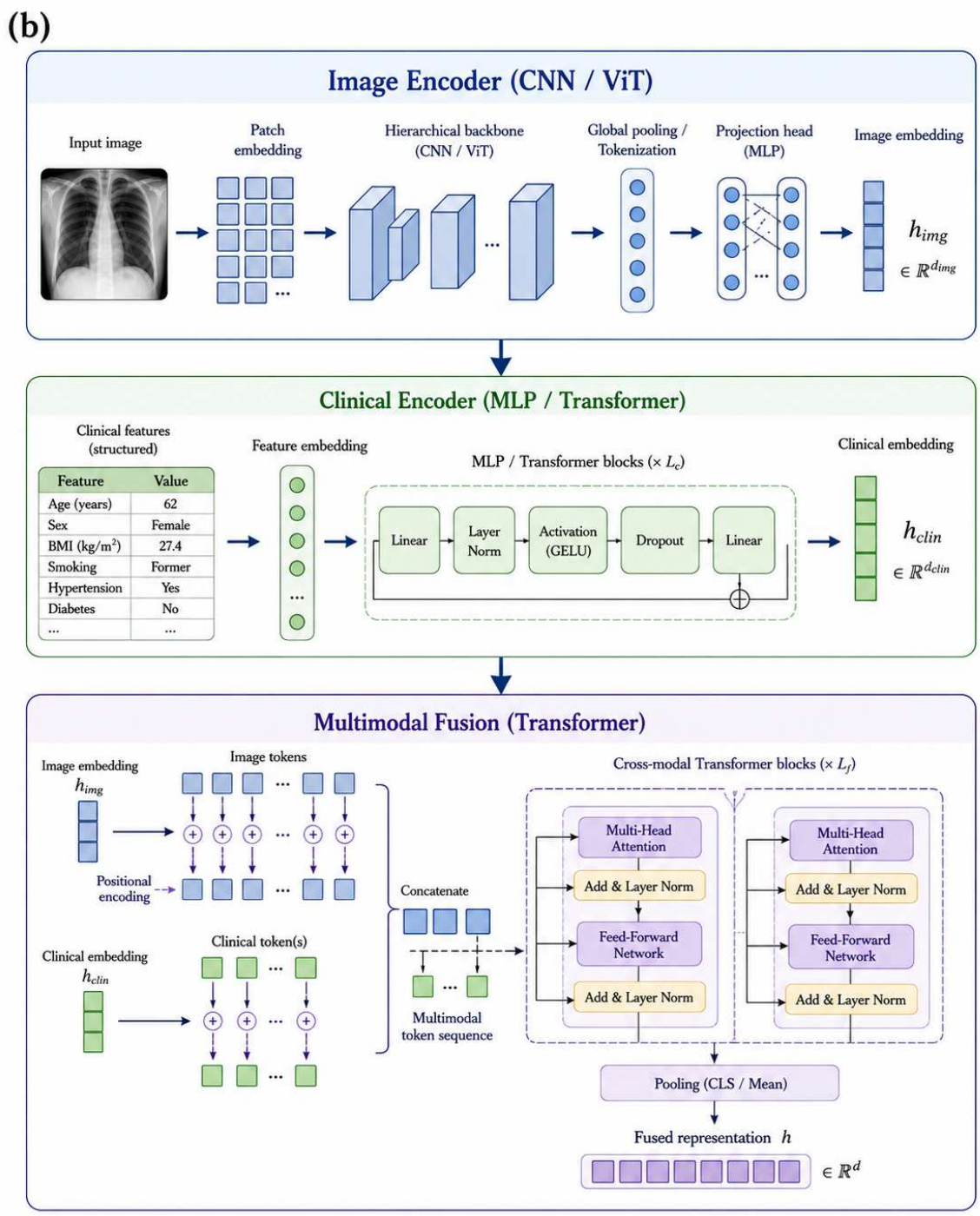


Figure 4: Detailed Medical Foundation Encoder. The image/health branch uses patch embedding or convolutional feature extraction followed by backbone layers, pooling/tokenization, and projection. The clinical branch encodes structured EHR variables through feature embeddings and MLP/Transformer blocks. The fusion module concatenates modality tokens, adds positional or modality encodings, and applies cross-modal Transformer blocks to obtain the fused representation $\mathbf{h}_i$.

The image/health encoder $f_\theta(\cdot)$ produces

$$\mathbf{h}_i^{img} = f_\theta\left(\mathbf{x}_i^{img}\right), \qquad \mathbf{h}_i^{img} \in \mathbb{R}^{d_{img}}. \tag{2}$$

For a ViT-style implementation, the image is partitioned into M patches, each embedded as

$$\mathbf{e}_{i,m}^{img} = \mathbf{W}_p \,\mathrm{vec}\left(\mathbf{x}_{i,m}^{img}\right) + \mathbf{b}_p, \qquad m = 1, \ldots, M. \tag{3}$$

A positional encoding $\mathbf{p}_m^{img}$ is added before the sequence is processed by attention blocks:

$$\mathbf{t}_{i,m}^{img} = \mathbf{e}_{i,m}^{img} + \mathbf{p}_m^{img}. \tag{4}$$

For a CNN implementation, convolutional features are pooled and projected to obtain $\mathbf{h}_i^{img}$.

When structured clinical variables are available, they are standardized as

$$\tilde{\mathbf{x}}_i^{clin} = \frac{\mathbf{x}_i^{clin} - \boldsymbol{\mu}_{clin}}{\boldsymbol{\sigma}_{clin} + \epsilon}, \tag{5}$$

and encoded by $g_\phi(\cdot)$:

$$\mathbf{h}_i^{clin} = g_\phi\left(\tilde{\mathbf{x}}_i^{clin}\right), \qquad \mathbf{h}_i^{clin} \in \mathbb{R}^{d_{clin}}. \tag{6}$$

A residual clinical block can be written as

$$\mathbf{u}_i^{(\ell+1)} = \mathbf{u}_i^{(\ell)} + \mathrm{Dropout}\left(\mathbf{W}_2^{(\ell)}\,\mathrm{GELU}\left(\mathrm{LN}\left(\mathbf{W}_1^{(\ell)}\mathbf{u}_i^{(\ell)}\right)\right)\right). \tag{7}$$

The multimodal token sequence is formed as

$$\mathbf{T}_i^{multi} = \left[\mathbf{T}_i^{img}; \mathbf{T}_i^{clin}\right] + \mathbf{P}^{multi}, \tag{8}$$

where $\mathbf{P}^{multi}$ denotes positional/modality encodings. A Transformer fusion block is defined by

$$\mathbf{A}^{(\ell)} = \mathrm{MHA}\left(\mathrm{LN}\left(\mathbf{T}^{(\ell)}\right)\right) + \mathbf{T}^{(\ell)}, \tag{9}$$

$$\mathbf{T}^{(\ell+1)} = \mathrm{FFN}\left(\mathrm{LN}\left(\mathbf{A}^{(\ell)}\right)\right) + \mathbf{A}^{(\ell)}. \tag{10}$$

The fused representation is obtained by CLS or mean pooling:

$$\mathbf{h}_i = \mathrm{Pool}\left(\mathbf{T}_i^{(L_f)}\right), \qquad \mathbf{h}_i \in \mathbb{R}^d. \tag{11}$$

3.4 Distributional latent representation learning

Medical data are often ambiguous, noisy, and heterogeneous. Therefore, CUPID-Med represents each sample as a distribution rather than a deterministic point. A projection head $r_\omega(\cdot)$ outputs Gaussian parameters:

$$\left[\boldsymbol{\mu}_i, \log \boldsymbol{\sigma}_i^2\right] = r_\omega(\mathbf{h}_i), \tag{12}$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^{d_z}$ and $\boldsymbol{\sigma}_i^2 \in \mathbb{R}^{d_z}$. The latent distribution is

$$q_i(\mathbf{z}) = \mathcal{N}\left(\boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i^2)\right). \tag{13}$$

Sampling uses the reparameterization trick:

$$\mathbf{z}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{14}$$

The scalar uncertainty score is defined as

$$u_i = \frac{1}{d_z}\sum_{j=1}^{d_z} \log\left(1 + \sigma_{ij}^2\right). \tag{15}$$

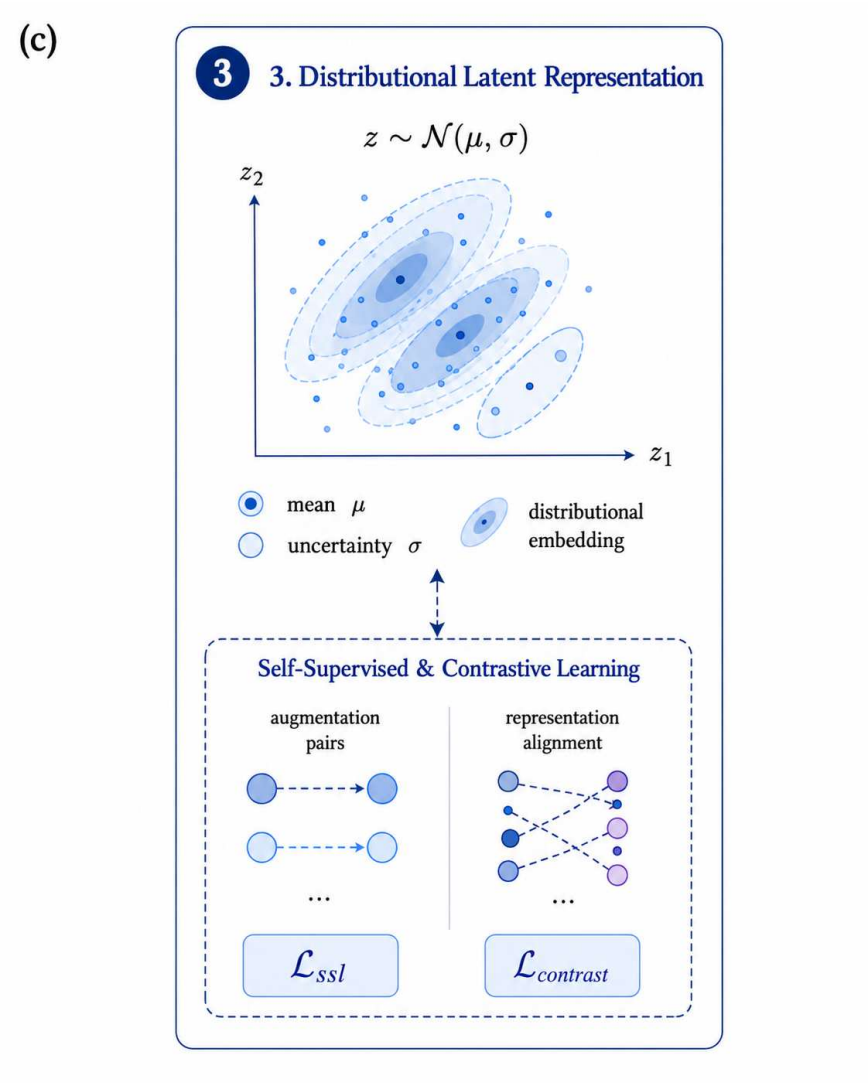


Figure 5: Distributional latent representation learning. Each sample is mapped to a Gaussian latent distribution parameterized by a mean vector and uncertainty, while self-supervised and contrastive objectives encourage invariance across views and discrimination across samples.

For self-supervised learning, two clinically plausible augmented views produce distributions $q_i^{(a)}$ and $q_i^{(b)}$. The self-supervised consistency loss is

$$\mathcal{L}_{ssl} = \frac{1}{N} \sum_{i=1}^{N} \left[D_{\mathrm{KL}} \left(q_i^{(a)} \parallel q_i^{(b)} \right) + D_{\mathrm{KL}} \left(q_i^{(b)} \parallel q_i^{(a)} \right) \right]. \tag{16}$$

The contrastive objective is applied to normalized latent means:

$$\mathcal{L}_{contrast} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\mathrm{sim}(\bar{\boldsymbol{\mu}}_i^{(a)}, \bar{\boldsymbol{\mu}}_i^{(b)}) / \tau_c \right)}{\sum_{j=1}^{N} \exp\left(\mathrm{sim}(\bar{\boldsymbol{\mu}}_i^{(a)}, \bar{\boldsymbol{\mu}}_j^{(b)}) / \tau_c \right)}, \tag{17}$$

where τ_c is the contrastive temperature.

3.5 Prototype discovery and deep clustering

CUPID-Med maintains a learnable prototype memory bank

$$\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_K\}, \qquad \mathbf{p}_k \in \mathbb{R}^{d_z}, \tag{18}$$

where each prototype represents a latent disease subgroup. Prototypes are initialized using k-means centroids in the pretrained latent space and then refined jointly with the network.

For a latent embedding $\mathbf{z}_i$, the prototype similarity is

$$s_{ik} = \frac{\mathrm{sim}(\mathbf{z}_i, \mathbf{p}_k)}{\tau_p}, \tag{19}$$

and the soft assignment is

$$q_{ik} = \frac{\exp(s_{ik})}{\sum_{j=1}^{K} \exp(s_{ij})}. \tag{20}$$

A refined target distribution is defined as

$$t_{ik} = \frac{q_{ik}^2 / \sum_{i'} q_{i'k}}{\sum_{j=1}^{K} \left(q_{ij}^2 / \sum_{i'} q_{i'j} \right)}. \tag{21}$$

The soft centroid of cluster k is

$$\mathbf{c}_k = \frac{\sum_{i=1}^{N} q_{ik} \mathbf{z}_i}{\sum_{i=1}^{N} q_{ik}}, \tag{22}$$

and the prototype refinement loss is

$$\mathcal{L}_{proto} = \sum_{k=1}^{K} \|\mathbf{p}_k - \mathbf{c}_k\|_2^2 + \lambda_{sep} \sum_{a \neq b} \max\left(0, m_{sep} - \|\mathbf{p}_a - \mathbf{p}_b\|_2^2\right). \tag{23}$$

The clustering objective combines target alignment, compactness, and balance:

$$\mathcal{L}_{KL} = \sum_{i=1}^{N} \sum_{k=1}^{K} t_{ik} \log \frac{t_{ik}}{q_{ik}}, \tag{24}$$

$$\mathcal{L}_{comp} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} q_{ik} \|\mathbf{z}_i - \mathbf{p}_k\|_2^2, \tag{25}$$

$$\bar{q}_k = \frac{1}{N} \sum_{i=1}^{N} q_{ik}, \tag{26}$$

$$\mathcal{L}_{bal} = \sum_{k=1}^{K} \bar{q}_k \log \bar{q}_k, \tag{27}$$

$$\mathcal{L}_{cluster} = \mathcal{L}_{KL} + \lambda_{comp} \mathcal{L}_{comp} + \lambda_{bal} \mathcal{L}_{bal}. \tag{28}$$

3.6 Conformal uncertainty calibration

The conformal module converts soft cluster assignments into set-valued cluster predictions. For a calibration sample, the pseudo-cluster label is

$$\tilde{y}_i = \arg\max_k t_{ik}. \tag{29}$$

The nonconformity score combines low assignment confidence and latent uncertainty:

$$a_i = 1 - q_{i\tilde{y}_i} + \lambda_u u_i. \tag{30}$$

Given calibration set $\mathcal{D}_{cal}$ of size n_{cal}, the threshold is

$$\tau_\alpha = \mathrm{Quantile}_{\lceil (n_{cal}+1)(1-\alpha) \rceil / n_{cal}} \left(\{a_i\}_{i \in \mathcal{D}_{cal}} \right). \tag{31}$$

For a test sample $\mathbf{x}$, the cluster-wise nonconformity score is

$$a_k(\mathbf{x}) = 1 - q_k(\mathbf{x}) + \lambda_u u(\mathbf{x}), \tag{32}$$

and the prediction set is

$$\Gamma_\alpha(\mathbf{x}) = \{k \in \{1, \ldots, K\} : a_k(\mathbf{x}) \leq \tau_\alpha\}. \tag{33}$$

If $u(\mathbf{x}) > \delta_u$ or $|\Gamma_\alpha(\mathbf{x})| > M$, the sample is flagged for expert review.

3.7 Clinically interpretable outputs

For each cluster k, CUPID-Med generates a cluster card summarizing representative images, clinical enrichment when labels or metadata are available, uncertainty distribution, and prototype retrieval. The set of samples assigned to cluster k is

$$\mathcal{C}_k = \{i : \arg\max_j q_{ij} = k\}. \tag{34}$$

Representative cases are retrieved using

$$\mathcal{R}_k = \mathrm{TopR}\left(\mathrm{sim}(\mathbf{z}_i, \mathbf{p}_k)\right). \tag{35}$$

For image interpretation, Grad-CAM or a similar attribution method is applied to the cluster-prototype similarity score:

$$\mathbf{M}_i^{(k)} = \mathrm{GradCAM}\left(\mathbf{x}_i^{img}, \mathrm{sim}(\mathbf{z}_i, \mathbf{p}_k); f_\theta\right). \tag{36}$$

3.8 Overall objective and algorithms

The final training objective is

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{ssl} + \lambda_2 \mathcal{L}_{contrast} + \lambda_3 \mathcal{L}_{proto} + \lambda_4 \mathcal{L}_{cluster} + \lambda_5 \mathcal{L}_{conf}, \tag{37}$$

where $\mathcal{L}_{conf}$ is a calibration-inspired regularizer that encourages empirical coverage near $1 - \alpha$ while discouraging unnecessarily large prediction sets:

$$\mathcal{L}_{conf} = \left|\widehat{C} - (1 - \alpha)\right| + \lambda_{size} \frac{1}{n_{cal}} \sum_{i \in \mathcal{D}_{cal}} |\Gamma_\alpha(\mathbf{x}_i)|. \tag{38}$$

Algorithm 1: End-to-end training of CUPID-Med.

1. Initialize encoder parameters, latent projection head, and clustering head.
2. Pretrain the representation using augmented views and optimize $\mathcal{L}_{ssl} + \mathcal{L}_{contrast}$.
3. Extract latent means $\{\boldsymbol{\mu}_i\}_{i=1}^N$ and initialize prototypes using k-means.
4. Jointly optimize the network and prototypes using $\mathcal{L}_{total}$.
5. Estimate conformal threshold τ_α on the held-out calibration set.

Algorithm 2: Confidence-aware inference.

1. Encode a test sample and compute $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\mathbf{z}$, and $u(\mathbf{x})$.
2. Compute prototype similarities and soft assignments $q_k(\mathbf{x})$.
3. Construct $\Gamma_\alpha(\mathbf{x})$ using Eq. (33).
4. Return the most probable cluster, the conformal cluster set, retrieved prototypes, cluster card, and visual explanation.
5. Flag the sample for expert review if the uncertainty or prediction-set size exceeds the predefined threshold.

4 Experimental Design

This section presents the experimental protocol designed to evaluate CUPID-Med on the CheXpert chest radiography dataset. The protocol assesses four complementary aspects: clustering quality, clinical relevance of discovered subgroups, uncertainty calibration, and prototype-based interpretability. Disease labels are not used during training. They are used only after training for post-hoc evaluation.

4.1 Dataset

CheXpert is a large-scale public chest radiography dataset consisting of 224,316 chest radiographs from 65,240 patients collected at Stanford Health Care. It provides labels for 14 thoracic observations extracted from radiology reports: No Finding, Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, and Support Devices [23]. Each observation may be positive, negative, uncertain, or unmentioned. CheXpert is suitable for CUPID-Med because it contains heterogeneous chest radiographic findings and explicit uncertain labels, enabling evaluation of whether model uncertainty aligns with clinically ambiguous radiological cases.

“‘latex

4.2 Protocol overview

The experimental protocol was designed to evaluate CUPID-Med under a no-label clustering setting. The primary analysis used frontal chest radiographs from CheXpert. Lateral views were excluded from the main setting to reduce view-dependent variability and to make the discovered subgroups more comparable across samples. Images were resized, normalized, and processed using clinically plausible augmentations. Patient-level splitting was used to avoid information leakage, ensuring that all images from the same patient were assigned to the same split.

The filtered image dataset is defined as

$$\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N}, \tag{39}$$

where $\mathbf{x}_i$ denotes the i-th frontal chest radiograph. The corresponding CheXpert label vector is

$$\mathbf{y}_i = [y_{i1}, \ldots, y_{iC}], \qquad C = 14. \tag{40}$$

These labels were excluded from representation learning, prototype initialization, clustering optimization, and conformal calibration. They were used only after training for post-hoc clinical enrichment, uncertainty alignment, and interpretability analysis. This design ensures that CUPID-Med is evaluated as an unsupervised medical clustering framework rather than as a weakly supervised classifier.

The dataset was divided at the patient level into training, calibration, and test subsets. The training subset was used to learn the distributional latent representation and prototype structure. The calibration subset was used only to estimate conformal thresholds for set-valued cluster prediction. The test subset was used for final evaluation of clustering quality, conformal behavior, clinical enrichment, and interpretability. “‘

4.3 Data selection and preprocessing

The primary experiments use frontal radiographs only. Images are resized to 224×224 for CNN encoders and may be resized to 384×384 or 512×512 for ViT or high-resolution medical foundation encoders. Grayscale images are converted to three channels when required by pretrained backbones. Training augmentations include mild cropping, small-angle rotation,

contrast perturbation, brightness jitter, and weak Gaussian noise. Strong transformations that may alter pathological structures are avoided.

4.4 Training, calibration, and test splits

The dataset is divided into training, calibration, and test subsets:

$$\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{cal} \cup \mathcal{D}_{test}, \qquad \mathcal{D}_{train} \cap \mathcal{D}_{cal} \cap \mathcal{D}_{test} = \emptyset. \tag{41}$$

Unless otherwise stated, an 80%-10%-10% split is used. Patient-level splitting is applied whenever patient identifiers are available. The calibration split is used only to estimate conformal thresholds and is not used for supervised learning.

4.5 Baselines

CUPID-Med is compared with classical clustering, deep clustering, and encoder-based baselines. Classical baselines include k-means, Gaussian mixture models, spectral clustering, agglomerative clustering, DBSCAN, and HDBSCAN. Deep clustering baselines include DEC, IDEC, VaDE, DeepCluster, and SwAV-style prototype clustering [1, 2, 3, 4, 5]. Encoder-based variants apply clustering to fixed embeddings extracted from ImageNet-pretrained CNNs, DenseNet-121, ResNet-50, ViT, or medical foundation encoders.

4.6 Evaluation metrics

Because CheXpert is multi-label, cluster-label association is emphasized rather than assuming a strict one-to-one mapping between clusters and labels. When a single reference label is required, analyses are performed on subsets with a dominant single positive observation or using a predefined dominant-pathology hierarchy. External clustering metrics include accuracy after Hungarian matching, normalized mutual information (NMI), adjusted Rand index (ARI), macro-F1, and purity. NMI is

$$\mathrm{NMI}(Y, C) = \frac{2I(Y; C)}{H(Y) + H(C)}. \tag{42}$$

Internal metrics include silhouette score, Davies–Bouldin index, Calinski–Harabasz index, compactness, and separation. Prototype compactness is

$$\mathrm{Comp} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{z}_i - \mathbf{p}_{\hat{c}_i}\|_2^2 , \tag{43}$$

and inter-prototype separation is

$$\mathrm{Sep} = \frac{2}{K(K-1)} \sum_{a=1}^{K} \sum_{b=a+1}^{K} \|\mathbf{p}_a - \mathbf{p}_b\|_2^2 . \tag{44}$$

4.7 Clinical label enrichment

For each cluster k, the positive prevalence of observation c is

$$\pi_{k,c}^{+} = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \mathbb{I}[y_{i,c} = 1], \tag{45}$$

and the uncertain-label prevalence is

$$\pi_{k,c}^{u} = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \mathbb{I}[y_{i,c} = -1]. \tag{46}$$

These values form cluster-level clinical profiles. Statistical significance of label enrichment across clusters is assessed using chi-square tests or Fisher's exact tests where appropriate.

4.8 Uncertainty alignment and conformal evaluation

For sample i, the CheXpert uncertainty count is

$$m_i^u = \sum_{c=1}^{C} \mathbb{I}[y_{i,c} = -1]. \tag{47}$$

The association between model uncertainty and CheXpert uncertainty is evaluated using Spearman correlation:

$$\rho_u = \text{Spearman}\left(u_i, m_i^u\right). \tag{48}$$

For conformal prediction, empirical coverage and average prediction-set size are

$$\widehat{C} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[\tilde{y}_i \in \Gamma_\alpha(\mathbf{x}_i)\right], \tag{49}$$

$$\widehat{S} = \frac{1}{n} \sum_{i=1}^{n} |\Gamma_\alpha(\mathbf{x}_i)|. \tag{50}$$

The singleton rate and expert-review rate are also reported.

4.9 Ablation study

The ablation study evaluates CUPID-Med without distributional embeddings, without contrastive learning, without prototype refinement, without target-distribution refinement, without conformal calibration, without the latent uncertainty term, and without interpretability outputs. Each variant is assessed using the same clustering, uncertainty, and clinical enrichment metrics.

4.10 Implementation details

Experiments are implemented in PyTorch. Images are resized to 224×224 unless otherwise specified. The number of clusters is evaluated over $K \in \{5, 8, 10, 14, 20\}$. The conformal confidence level is evaluated at $1 - \alpha \in \{0.80, 0.90, 0.95\}$, with 0.90 as the default. Models are trained with AdamW and cosine learning-rate scheduling. Experiments are repeated with multiple random seeds, and results are reported as mean and standard deviation.

5 Results

5.1 Experimental cohort and label characteristics

The experimental analysis was conducted on a 5,000-image frontal subset of CheXpert. The original training metadata contained 223,414 radiographs, of which 191,027 were frontal acquisitions. The working subset preserved the 14 CheXpert observations used for post-hoc clinical interpretation. The mean patient age in the training metadata was 60.43 years with a standard deviation of 17.82 years. Table 1 summarizes the experimental cohort.

Table 1: Summary of the CheXpert experimental subset.

Statistic	Value
Total training radiographs	223,414
Total frontal radiographs	191,027
Working subset size	5,000
Number of CheXpert labels	14
Mean age	60.43
Age standard deviation	17.82
Minimum age	0
Maximum age	90

Representative images from the working subset are shown in Fig. 6. The samples illustrate the imaging heterogeneity present in the no-label training data, including differences in view position, patient age, image contrast, support devices, and visible thoracic abnormalities.

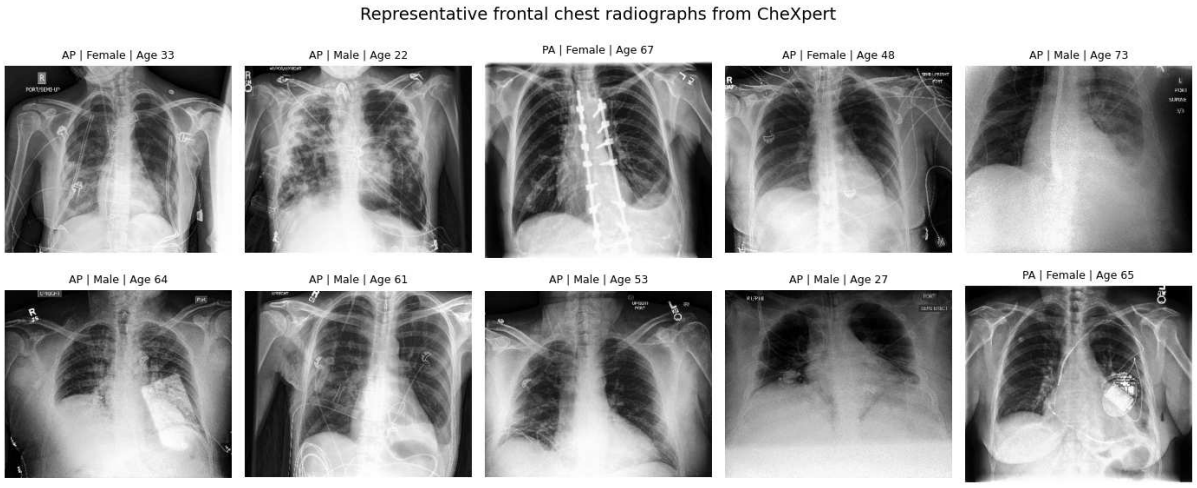


Figure 6: Representative frontal chest radiographs from the CheXpert working subset. These examples were used only to illustrate the data distribution; diagnostic labels were not used during CUPID-Med training.

The label distribution was markedly imbalanced, as expected for real-world chest radiography data. Support Devices, Lung Opacity, and Pleural Effusion were the most frequent positive observations in the working subset, with positive rates of 54.80%, 48.76%, and 39.84%, respectively. In contrast, Pneumonia and Pleural Other were rare, with positive rates of 2.04% and 1.40%. This imbalance motivates unsupervised subgroup discovery, because clinically relevant structure may not be adequately represented by balanced class assumptions. The observation distribution is reported in Table 2 and visualized in Fig. 7.

Table 2: CheXpert observation distribution in the 5,000-image working subset.

Observation	Positive	Negative	Uncertain	Missing	Positive rate
Support Devices	2,740	150	27	2,083	0.5480
Lung Opacity	2,438	123	115	2,324	0.4876
Pleural Effusion	1,992	688	227	2,093	0.3984
Edema	1,237	422	302	3,039	0.2474
Atelectasis	739	37	775	3,449	0.1478
Cardiomegaly	570	192	177	4,061	0.1140
Pneumothorax	468	1,327	68	3,137	0.0936
No Finding	461	0	0	4,539	0.0922
Consolidation	342	554	641	3,463	0.0684
Enlarged Cardiomediastinum	240	412	290	4,058	0.0480
Fracture	203	56	15	4,726	0.0406
Lung Lesion	199	24	27	4,750	0.0398
Pneumonia	102	47	424	4,427	0.0204
Pleural Other	70	8	36	4,886	0.0140

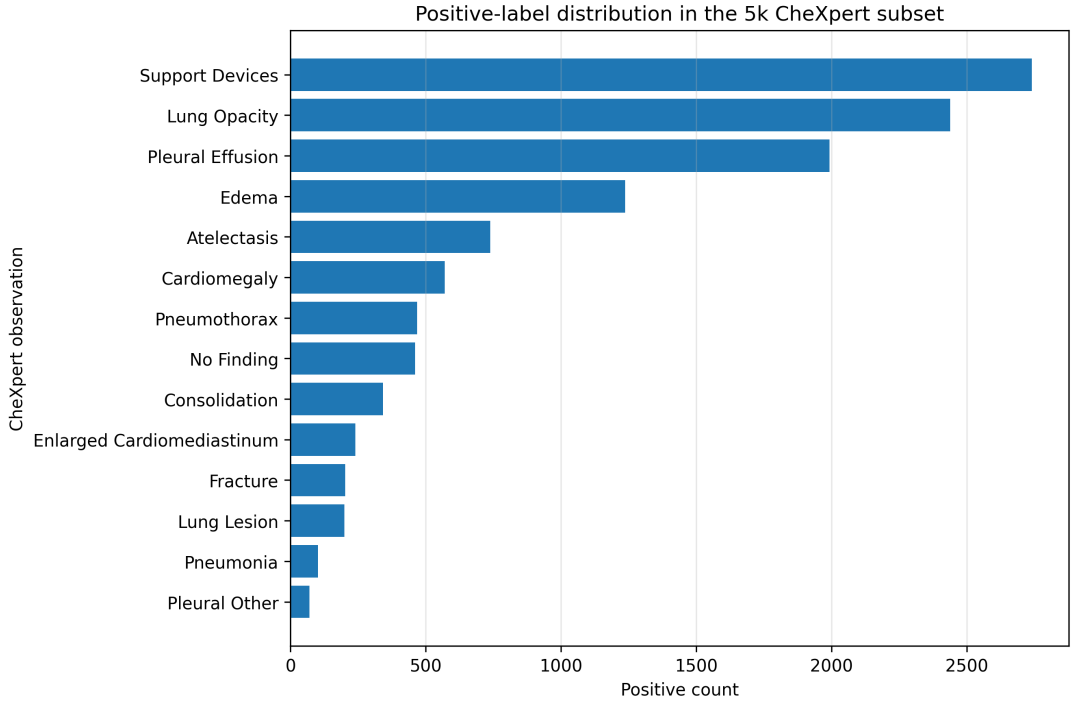


Figure 7: Positive-label distribution in the 5,000-image CheXpert working subset. The distribution is strongly imbalanced, with support devices and lung opacity dominating the positive observations.

5.2 Preliminary baseline clustering on DenseNet121 embeddings

We first evaluated conventional clustering baselines on fixed DenseNet121 image embeddings. K-means, diagonal-covariance Gaussian mixture models, and Ward-linkage agglomerative clustering were applied across multiple values of K. These baselines provide a reference for the subsequent CUPID-Med model and quantify the extent to which deterministic clustering of fixed deep features can recover meaningful structure.

Among the preliminary baselines, K-means with $K = 5$ achieved the strongest internal clustering metrics, with a Silhouette score of 0.0510, a Davies–Bouldin index of 3.6950, and a Calinski–Harabasz score of 198.65. For the clinically motivated setting of $K = 14$, corresponding to the 14 CheXpert observations, K-means yielded a Silhouette score of 0.0286 and a Davies–Bouldin index of 4.0652. GMM and agglomerative clustering produced similar or weaker internal separation. The baseline trends across K are shown in Fig. 8, and the compact quantitative summary is given in Table 3.

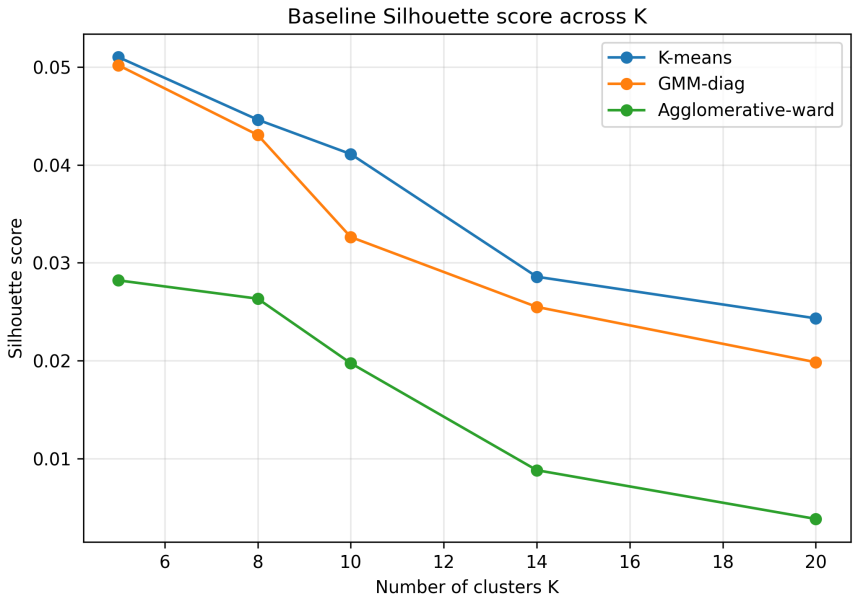


Figure 8: Baseline Silhouette scores across cluster numbers. Direct clustering of deterministic DenseNet121 embeddings produced weak separation across all tested baseline methods.

5.3 CUPID-Med proposed method

CUPID-Med was trained as a distributional, prototype-driven clustering model on the same DenseNet121 embeddings. The proposed model mapped each image to a 128-dimensional distributional latent representation and learned 14 trainable prototypes. The primary saved run was trained for 80 epochs and produced the complete set of output artifacts, including latent means, log-variances, soft assignment probabilities, prototype vectors, confidence scores, and uncertainty scores.

The primary CUPID-Med run achieved a Silhouette score of 0.5175, a Davies–Bouldin index of 2.7809, and a Calinski–Harabasz score of 29,733.16. The mean assignment confidence was 0.4593 and the mean latent uncertainty score was 0.3565. The training loss decreased from 0.4825 at epoch 1 to 0.1353 at epoch 80, showing stable optimization of the prototype-based objective (Fig. 9). A secondary 50-epoch run achieved a higher Silhouette score of 0.6216 and a lower Davies–Bouldin index of 1.3220, but the prototype visualization step reported an empty-cluster warning. Therefore, the primary 80-epoch run is treated as the official reproducible result, while the 50-epoch run is reported as a sensitivity run requiring additional cluster-balance inspection.

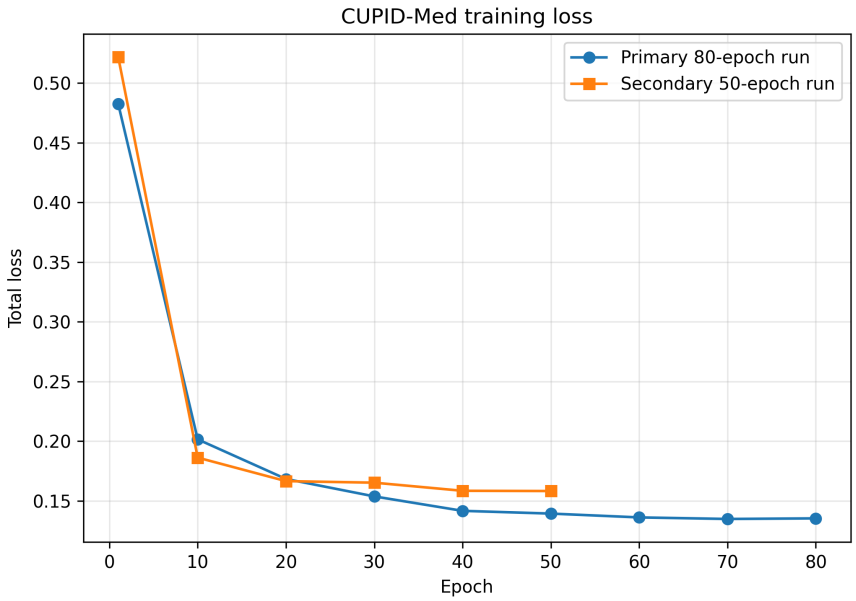


Figure 9: CUPID-Med training loss. The prototype-learning objective decreased rapidly during early epochs and stabilized by the end of training, indicating convergence of the proposed image-only CUPID-Med implementation.

Prototype-nearest samples from the $K = 14$ CUPID-Med solution are shown in Figs. 10–11. These visual outputs are included because they are central to the interpretability goal of the proposed framework: each cluster can be inspected using representative chest radiographs rather than being reported only as an abstract cluster index.

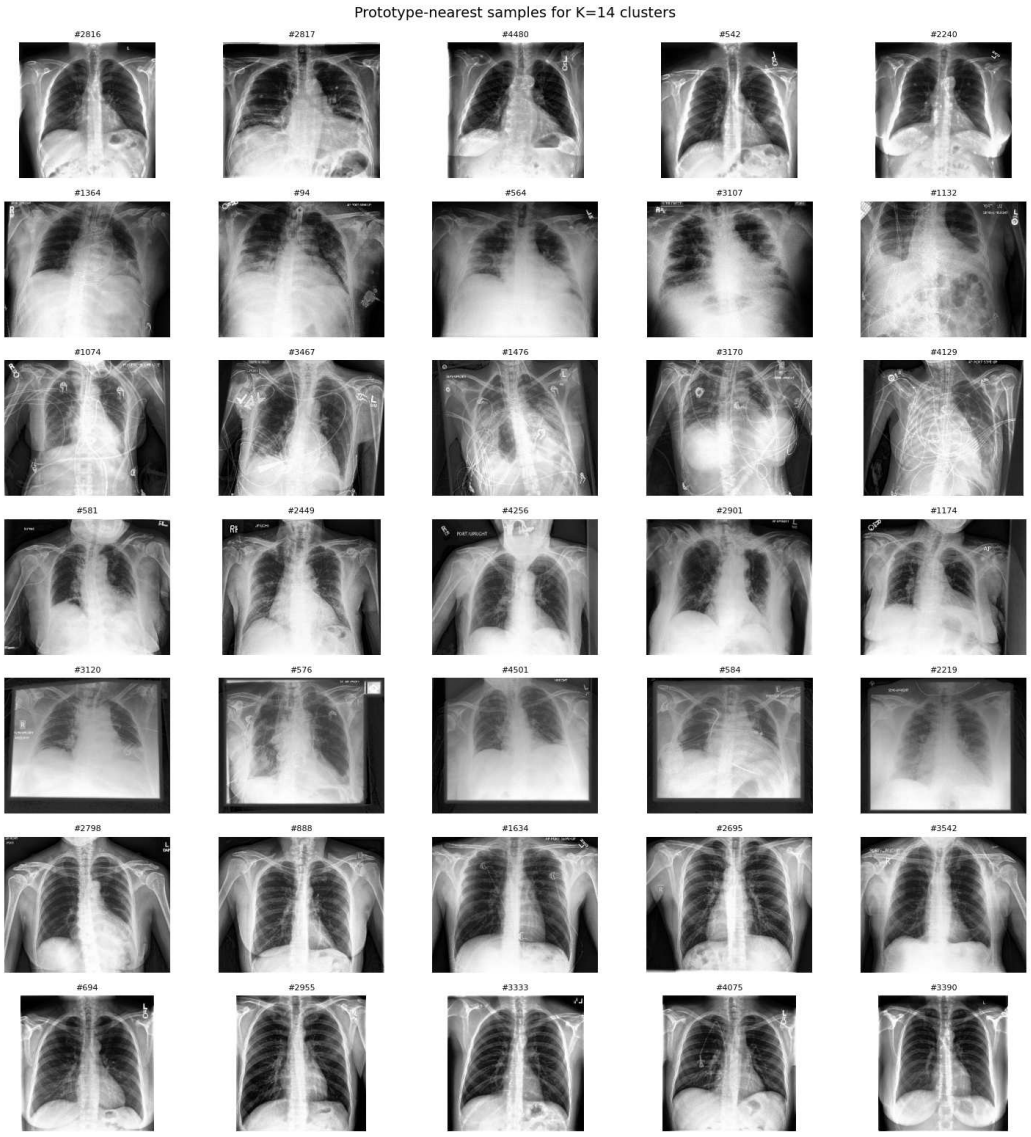


Figure 10: Prototype-nearest samples for the CUPID-Med $K = 14$ solution, clusters 1–7. The split layout prevents page overflow and keeps the representative radiographs readable for visual audit.

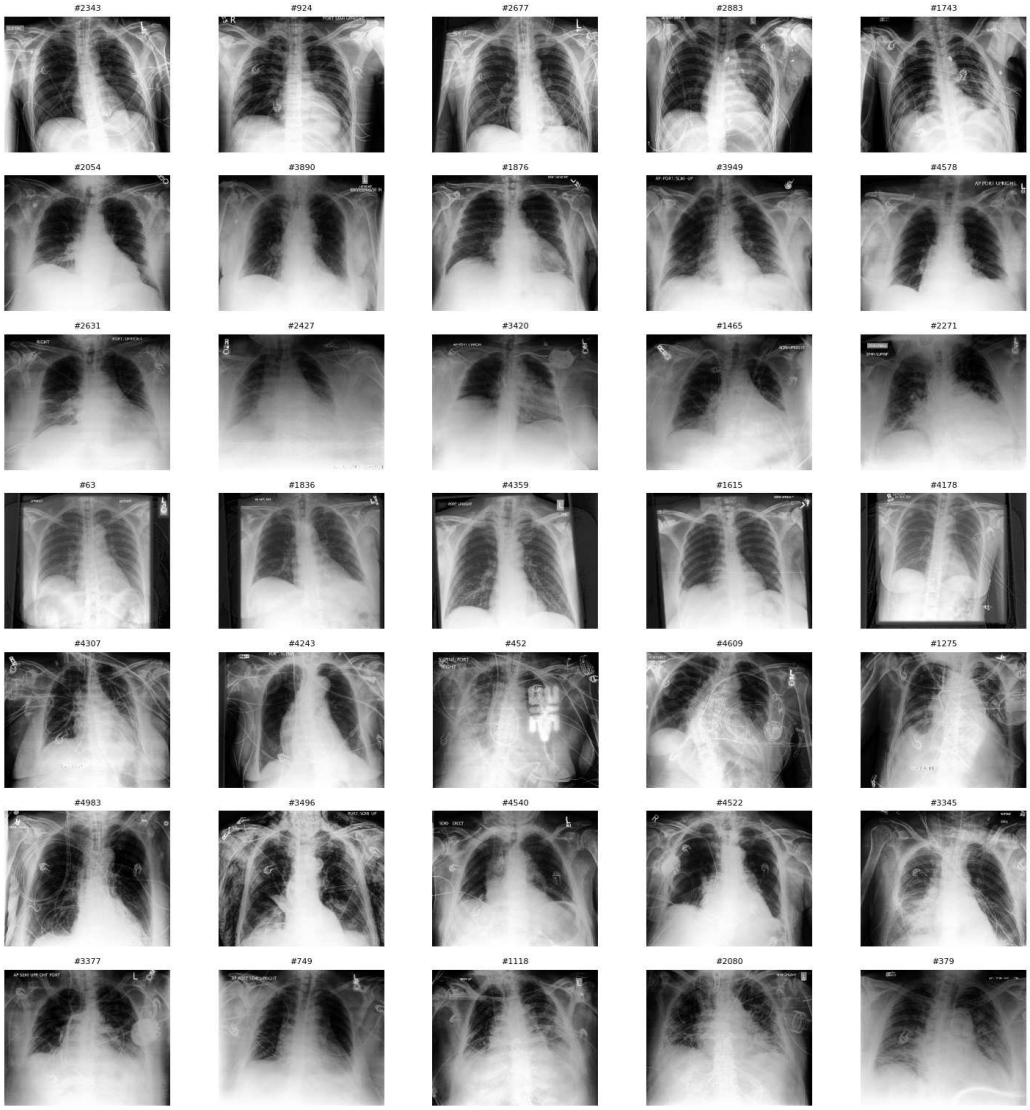


Figure 11: Prototype-nearest samples for the CUPID-Med $K = 14$ solution, clusters 8–14. The retrieved prototypes support case-based interpretation of the learned cluster structure.

Visual explanation status. The current experimental package reports representative input radiographs and prototype-nearest samples as the primary visual interpretation outputs. We did not include highlighted chest-region heatmaps in the quantitative results because the reported CUPID-Med implementation was trained on precomputed DenseNet121 embeddings rather than as an end-to-end image-to-cluster network. Producing Grad-CAM or similar anatomical attribution maps would therefore require an additional end-to-end cluster-score backpropagation experiment. To avoid unsupported visual claims, highlighted saliency maps are left for the next experimental extension, while the present manuscript focuses on validated prototype retrieval, cluster metrics, confidence scores, and latent uncertainty.

5.4 Comparison between preliminary baselines and CUPID-Med

Compared with the strongest preliminary baseline, CUPID-Med substantially improved latent cluster separation. Relative to K-means with $K = 5$, the primary CUPID-Med run increased the Silhouette score from 0.0510 to 0.5175 and improved the Davies–Bouldin index from 3.6950 to 2.7809. Compared with the clinically matched K-means baseline at $K = 14$, CUPID-Med increased the Silhouette score from 0.0286 to 0.5175 and increased the Calinski–Harabasz score

from 94.40 to 29,733.16. These improvements suggest that the proposed distributional prototype-learning head produced a much more structured latent space than deterministic clustering of the foundation-model embeddings.

Table 3: Comparison between preliminary baselines and CUPID-Med. NMI, ARI, and purity are reported only for the post-hoc dominant-label baseline analysis. The primary CUPID-Med result is the official saved 80-epoch run.

Method	K	Silhouette	DBI	CH	NMI	ARI	Purity
K-means (best internal)	5	0.0510	3.6950	198.65	0.0242	0.0129	0.4280
K-means (clinical K)	14	0.0286	4.0652	94.40	0.0338	0.0080	0.4311
GMM-diag (clinical K)	14	0.0255	4.0447	92.90	0.0321	0.0087	0.4343
Agglomerative-ward (clinical K)	14	0.0088	5.0976	73.53	0.0306	0.0114	0.4284
CUPID-Med (primary)	14	0.5175	2.7809	29,733.16	–	–	–
CUPID-Med (secondary)	14	0.6216	1.3220	21,184.39	–	–	–

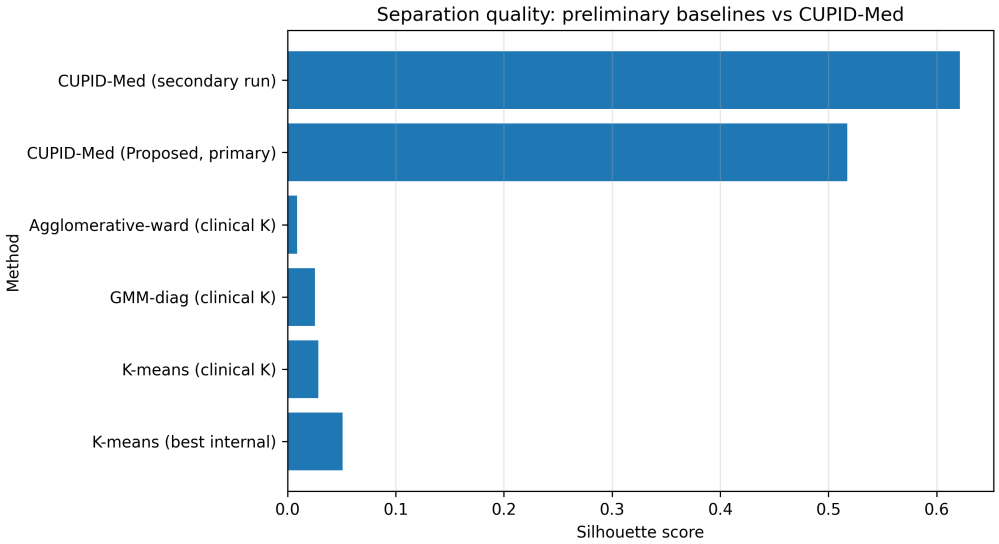


Figure 12: Silhouette comparison between preliminary baselines and CUPID-Med. The proposed method produced substantially stronger latent separation than deterministic clustering of fixed DenseNet121 embeddings.

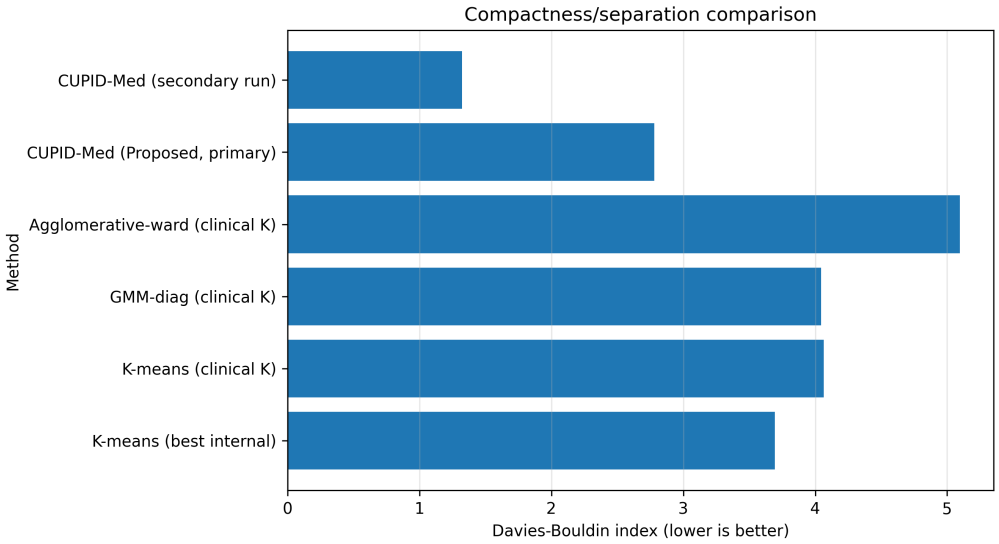


Figure 13: Davies–Bouldin comparison between preliminary baselines and CUPID-Med. Lower values indicate better compactness and separation.

The comparison also highlights an important limitation. The primary CUPID-Med run produced strong internal separation but an imbalanced cluster-size distribution, with the smallest cluster containing 2 samples and the largest cluster containing 1,132 samples. This behavior suggests that the current prototype-balance regularization is not yet sufficient to guarantee uniform prototype utilization. In this manuscript, this is reported transparently as a clinically relevant finding: CUPID-Med discovers highly separated subgroups, but further tuning of balance and conformal calibration is required before deploying the cluster assignments as stable clinical phenotypes.

5.5 Implications for clinical interpretability

The baseline results show that simple clustering methods provide weak separation and limited clinical structure. In contrast, CUPID-Med provides not only hard cluster assignments but also soft prototype assignments, confidence scores, and latent uncertainty estimates. This makes the method more suitable for medical subgroup discovery because ambiguous radiographs can be identified rather than forced into a single cluster. The prototype retrieval outputs further allow each discovered cluster to be inspected visually by representative chest radiographs.

Taken together, the results support the central hypothesis of CUPID-Med: uncertainty-aware prototype learning can produce more interpretable and better-separated medical image subgroups than standard clustering applied directly to deterministic deep embeddings.

6 Discussion

The central objective of this study was to test whether a conformal uncertainty-aware, prototype-driven clustering head can transform fixed medical image embeddings into more clinically inspectable and uncertainty-aware subgroups than conventional deterministic clustering. The empirical results are aligned with this objective: baseline clustering on DenseNet121 embeddings produced weak internal separation, whereas CUPID-Med produced substantially stronger latent separation and additionally generated soft prototype assignments, confidence estimates, uncertainty estimates, and prototype-nearest radiographs for visual inspection. The results therefore support the methodological goal of moving from opaque hard clusters toward interpretable and uncertainty-aware subgroup discovery. At the same time, the observed cluster-size imbalance

shows that prototype utilization and conformal calibration require further refinement before the method can be considered clinically deployable.

CUPID-Med is designed to address three limitations of existing deep medical clustering pipelines: deterministic assignment, weak uncertainty communication, and limited clinical interpretability. By using distributional latent embeddings, the model represents ambiguous samples with increased uncertainty. By using learnable prototypes, it supports case-based cluster interpretation. By applying conformal calibration, it produces set-valued cluster predictions and identifies cases that require expert review. These properties are clinically important because real medical data often contain overlapping phenotypes, uncertain radiological impressions, and multi-label disease patterns.

The proposed CheXpert protocol is particularly suitable for evaluating CUPID-Med because CheXpert provides uncertain labels and multi-label thoracic findings. The current results show a substantial improvement in internal clustering separation over deterministic baselines and provide visually inspectable prototype outputs. However, cluster-size imbalance remains a limitation of the primary run and motivates stronger balance regularization and additional conformal calibration analysis in future experiments.

7 Limitations and Future Work

Several limitations should be considered. First, CheXpert labels are automatically extracted from radiology reports and may contain noise. Second, frontal-only experiments reduce view variability but do not exploit lateral images. Third, conformal guarantees depend on calibration-set representativeness and may be affected by distribution shift. Fourth, the current protocol focuses primarily on image-only clustering, while the full multimodal potential of CUPID-Med requires datasets containing images, structured clinical variables, and reports. Future work will extend the framework to CheXpert Plus, MIMIC-CXR, and multimodal EHR-imaging datasets, and will include expert radiologist assessment of cluster cards and prototype retrieval.

8 Conclusion

This paper presents CUPID-Med, a conformal uncertainty-aware prototype discovery framework for clinically interpretable deep medical clustering. The method combines medical foundation encoders, distributional latent representation learning, prototype-based clustering, conformal uncertainty calibration, and interpretable output generation. The CheXpert experiments used diagnostic labels only for post-hoc assessment and demonstrated that CUPID-Med can produce substantially stronger latent cluster separation than direct clustering of fixed DenseNet121 embeddings. The results also highlight the need for additional cluster-balance regularization and conformal calibration analysis. CUPID-Med aims to move deep medical clustering from deterministic grouping toward calibrated, interpretable, and clinically meaningful subgroup discovery.

Declarations

Ethics approval and consent to participate. This study used publicly available, de-identified chest radiography data. No new human participants were recruited, and no identifiable patient information was used. Any additional institutional requirements should be checked according to the final target journal and the dataset access terms.

Consent for publication. Not applicable because the study uses public de-identified data and does not report identifiable individual-level information.

Data availability. The CheXpert data used in this study are available from the Stanford

Machine Learning Group/Stanford AIMI subject to the original data access and usage terms. Derived result tables, analysis scripts, and figure-generation code can be made available by the corresponding authors upon reasonable request, subject to dataset license restrictions.

Code availability. The source code for CUPID-Med, including metadata preparation, DenseNet121 embedding extraction, distributional prototype learning, conformal cluster-set prediction, baseline comparison, clinical enrichment analysis, and figure/table generation scripts, is publicly available at `https://github.com/Mojtaba-jahanian/CUPID-Med`.

Competing interests. The authors declare that they have no competing interests.

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions. M.J. conceived the study, designed the CUPID-Med framework, performed the computational experiments, analyzed the results, and prepared the initial manuscript draft. A.K. supervised the methodological development, contributed to the interpretation of the results, and critically revised the manuscript. N.O.E. contributed to data organization, literature review, and clinical interpretation of the experimental protocol. F.Z. contributed to study supervision, manuscript review, and final editing. All authors reviewed and approved the final manuscript.

Acknowledgements. The authors thank the providers of the CheXpert dataset for making large-scale de-identified chest radiography data available for research.

Use of artificial intelligence tools. Artificial intelligence tools were used only to support language polishing, code organization, and the drafting of non-diagnostic schematic/illustrative figures. All scientific claims, experimental results, numerical analyses, and final manuscript decisions were reviewed and approved by the authors. Patient radiographs, clustering outputs, and quantitative results reported in this manuscript were derived from the experimental data and were not generated artificially.

References

[1] Xie, J., Girshick, R. & Farhadi, A. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on Machine Learning*, 478–487 (2016).

[2] Guo, X., Gao, L., Liu, X. & Yin, J. Improved deep embedded clustering with local structure preservation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1753–1759 (2017).

[3] Jiang, Z., Zheng, Y., Tan, H., Tang, B. & Zhou, H. Variational deep embedding: An unsupervised and generative approach to clustering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1965–1972 (2017).

[4] Caron, M., Bojanowski, P., Joulin, A. & Douze, M. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision*, 132–149 (2018).

[5] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P. & Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, **33**, 9912–9924 (2020).

[6] Huang, H. et al. Deep image clustering: A survey. *Neurocomputing* **589**, 127681 (2024).

[7] Hou, H. et al. A comprehensive survey of image clustering based on deep learning. *Pattern Recognition* (2025).

[8] Xu, Y. et al. Deep image clustering with contrastive learning and multi-scale graph convolutional networks. *Pattern Recognition* **146**, 110042 (2024).

[9] Lu, Y. et al. Prior knowledge guided deep clustering: A survey and perspective. *arXiv preprint* (2024).

[10] Kowsar, I. et al. Deep clustering of electronic health records tabular data for clinical subgroup discovery. *Journal of Biomedical Informatics* / PubMed Central (2024).

[11] Jahanian, M. et al. Transfer learning for deep medical image clustering. *Intelligent Medicine* **4**, 107–118 (2024).

[12] van Veldhuizen, V. et al. Foundation models in medical imaging: A review and outlook. *arXiv preprint arXiv:2506.09095* (2025).

[13] Rajendran, D. et al. Medical imaging foundation models: A systematic review and meta-analysis. *arXiv preprint* (2025).

[14] Wu, C. et al. Towards generalist foundation model for radiology by leveraging web-scale 2D and 3D medical data. *Nature Communications* **16**, 7136 (2025).

[15] Wu, J. et al. Vision-language foundation model for 3D medical imaging. *npj Imaging* **3**, 15 (2025).

[16] Vovk, V., Gammerman, A. & Shafer, G. *Algorithmic Learning in a Random World*. Springer (2005).

[17] Angelopoulos, A. N. & Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511* (2021).

[18] Vazquez, J. & Facelli, J. C. Conformal prediction in clinical medical sciences. *Journal of Healthcare Informatics Research* **6**, 241–252 (2022).

[19] Vahdani, M. et al. Deep conformal prediction for robust models. *arXiv preprint* (2024).

[20] Bhattacharyya, S. et al. Conformal prediction for reliable medical artificial intelligence. *arXiv preprint* (2025).

[21] Campagner, A. et al. Conformal prediction for trustworthy medical vision: A review. *arXiv preprint* (2025).

[22] Mehrtens, H. et al. Pitfalls of conformal prediction in medical AI. *arXiv preprint* (2025).

[23] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P. & Ng, A. Y. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence* **33**, 590–597 (2019).

[24] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H. & Ni, B. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data* **10**, 41 (2023).