UIT_Worker at ImageCLEFmedical 2026: DenseNet-121 Full-Fit Medical Concept Detection with Asymmetric Loss

Notebook for the DS-MED Lab (UIT) at CLEF 2026

Thai Q. To[1,2], Dinh C. Pham[1,2], Nhan H. Bui[1,2] and Thien B. Nguyen-Tat[1,2,*]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

This paper describes the proposed UIT_Worker system for the ImageCLEFmedical 2026 Concept Detection task. The task formulates medical concept detection as a large-scale multi-label classification problem over 2,646 Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) using 116,604 annotated development images. Our final system employs a DenseNet-121 backbone initialized with ImageNet weights and is trained end-to-end using a full-fit strategy on all development images to maximize exposure to rare concepts. To address the severe long-tail label distribution, we optimize the model with Asymmetric Loss using $\gamma^- = 4.0$, $\gamma^+ = 1.0$, and clipping parameter $c = 0.05$. Training is performed for six epochs using AdamW, cosine annealing, and PyTorch Automatic Mixed Precision with GradScaler, while inference uses a fixed decision threshold of 0.60. The submitted single-model system achieved an official primary F1-score of 0.5725 and a secondary F1-score of 0.9533, ranking third in the ImageCLEFmedical 2026 Concept Detection task. These results indicate that a carefully optimized ImageNet-initialized DenseNet-121 with imbalance-aware learning provides an effective and reliable solution for extreme long-tail medical concept detection.

Keywords

ImageCLEFmedical 2026, Medical Concept Detection, Multi-label Classification, DenseNet-121, Asymmetric Loss, Long-tail Learning, Automatic Mixed Precision, UMLS, Medical Image Analysis

1. Introduction

Medical images are a central source of diagnostic and scientific evidence in modern healthcare. Radiological figures, clinical scans, and biomedical images support diagnosis, medical education, research communication, and case-based retrieval. As the volume of medical imaging data continues to increase, automatic systems that can assign standardized semantic concepts to images are increasingly important for indexing, retrieval, downstream caption generation, and clinical decision support. Automated medical image annotation is therefore not merely a classification exercise [1]; it is a foundation for connecting visual biomedical evidence with structured medical knowledge.

ImageCLEF has provided a long-running international benchmark for multimedia retrieval and medical image understanding [2]. The ImageCLEFmedical Concept Detection task asks participants to predict Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) from medical images [3]. Given an input image, a system must output all relevant concepts rather than selecting a single class. This makes the task inherently multi-label: one image may simultaneously contain modality information, anatomical regions, imaging views, procedures, devices, or visible abnormalities. The task is also difficult because the concepts are sparse, visually heterogeneous, and strongly imbalanced.

The 2026 development data used in our experiments contains 116,604 images and 2,646 unique UMLS CUI labels [3]. Although the label vocabulary is large, each image is associated with only a small subset of the full vocabulary. Our parsing of the official annotations gives an average of 3.2206 labels per

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany
*Corresponding author.
24521598@gm.uit.edu.vn (T. Q. To); 24520306@gm.uit.edu.vn (D. C. Pham); 24521226@gm.uit.edu.vn (N. H. Bui); thiennntb@uit.edu.vn (T. B. Nguyen-Tat)
0009-0008-6058-5254 (T. Q. To); 0009-0009-3247-8352 (D. C. Pham); 0009-0007-6605-6874 (N. H. Bui); 0000-0002-4809-7126 (T. B. Nguyen-Tat)

image, meaning that for each training sample the overwhelming majority of the 2,646 label positions are negative. This creates an extreme long-tail setting: frequent modality and anatomy concepts appear repeatedly, while many specific or clinically valuable concepts are observed only a small number of times. Under a standard binary cross-entropy (BCE) objective, the abundance of easy negative labels can dominate the gradient and suppress the learning signal from rare positives [5, 10].

Our team, UIT_Worker, addressed the task with a deliberately simple but carefully optimized convolutional pipeline. Instead of relying on a complex multi-stage architecture or an aggressive ensemble, we focused on three practical principles. First, we used DenseNet-121 [4], a compact convolutional network with dense feature reuse and strong historical performance in medical image classification. Second, we used standard ImageNet initialization rather than RadImageNet or other medical-domain pretraining, because our empirical trials showed that ImageNet initialization combined with stable full fine-tuning was more robust for this task. Third, we optimized the network with Asymmetric Loss (ASL) [5], which is designed for imbalanced multi-label learning and explicitly reduces the contribution of easy negative labels.

A key design choice of our final submission is the full-fit strategy. During exploratory development, the official split of 97,364 training images and 19,240 validation images was useful for understanding the dataset and selecting a reproducible inference threshold. For the final model, however, we trained on all 116,604 labeled development images. This choice was motivated by the long-tail distribution: withholding nearly twenty thousand images would also remove valuable examples of rare concepts. In the final competitive setting, we therefore prioritized maximum rare-label coverage over validation-based checkpoint selection.

Our submitted system achieved a primary F1-score of 0.5725 and a secondary F1-score of 0.9533 on the official hidden test set, ranking third in the ImageCLEFmedical 2026 Concept Detection task. The result indicates that a single DenseNet-121 model, when trained with the right loss function, thresholding rule, and full-data fitting strategy, remains highly competitive for large-scale medical concept detection. The main contributions of this paper are as follows:

- We present a reproducible DenseNet-121 pipeline for ImageCLEFmedical 2026 Concept Detection over 2,646 UMLS concepts.
- We describe a final full-fit training strategy that uses all 116,604 labeled development images to improve exposure to rare long-tail concepts.
- We apply Asymmetric Loss with $\gamma^- = 4.0$, $\gamma^+ = 1.0$, and clipping parameter $c = 0.05$ to address extreme imbalance in the multi-label target vectors.
- We combine AdamW, cosine annealing, and Automatic Mixed Precision with GradScaler to train efficiently under GPU memory constraints.
- We report an official primary F1-score of 0.5725 and analyze why the final ImageNet-initialized full-fit model outperformed both RadImageNet initialization and an ensemble submission.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 describes the task data and label representation. Section 4 presents preprocessing and augmentation. Section 5 details the proposed model and training strategy. Section 6 reports the official results and submission analysis. Section 7 concludes the paper and outlines future work.

2. Related Work

2.1. Medical concept detection

Medical concept detection can be viewed as multi-label visual recognition over a biomedical vocabulary. Unlike standard single-label classification, the objective is to identify all semantic concepts that are visually or contextually associated with an image. In ImageCLEFmedical Caption and Concept Detection tasks, these labels are commonly represented as UMLS CUIs, which provide a standardized terminology

for biomedical entities. Such structured labels are valuable because they can support image retrieval, caption generation, and downstream clinical or educational applications [3].

The task introduces challenges that are more severe than those found in many general-domain image classification benchmarks. Medical figures are heterogeneous across modalities, resolutions, contrast patterns, acquisition settings, and visual layouts. Some images contain embedded markers or multi-panel arrangements, while others show subtle anatomy or pathology. Moreover, the label distribution is heavily imbalanced: a small number of concepts occur very frequently, whereas many labels are rare. This combination of visual diversity and long-tail supervision has motivated approaches based on strong convolutional backbones, imbalance-aware loss functions, calibrated thresholds, and, in some cases, label-dependency modeling [9].

2.2. Dense convolutional networks for medical imaging

Convolutional neural networks remain strong baselines for medical image analysis. DenseNet introduced dense connectivity, where each layer receives feature maps from all preceding layers and passes its own feature maps to all subsequent layers within a dense block [4]. This design improves gradient flow, encourages feature reuse, and keeps the parameter count relatively efficient. These properties are particularly relevant in medical imaging, where discriminative evidence may occur at multiple scales and where over-parameterized models can overfit rare or noisy labels.

DenseNet-121 has been widely adopted in medical image classification. Prior large-scale chest X-ray studies, including ChestX-ray8 [6] and the CheXNeXt system [7], demonstrated the effectiveness of DenseNet-based architectures for thoracic disease recognition and contributed to their widespread use in medical imaging applications. For our official submission, we used DenseNet-121 initialized with ImageNet weights and replaced the original classification head with a 2,646-output linear classifier, corresponding to the complete set of UMLS CUIs observed in the development annotations.

2.3. Learning under long-tail multi-label imbalance

In multi-label concept detection, each image has only a few positive labels and thousands of negative labels. Standard binary cross-entropy treats all positive and negative entries symmetrically. As a result, training can be dominated by easy negatives, especially in large label spaces. Focal Loss was introduced to reduce the influence of easy examples in dense detection [10], but multi-label classification requires a more asymmetric treatment because the positive and negative terms play different roles.

Asymmetric Loss was proposed to improve multi-label classification under imbalance by applying different focusing strengths to positive and negative samples [5]. Its negative focusing and clipping mechanism suppress easy negative labels while preserving useful gradients for positive labels and hard negatives. This makes ASL well suited to ImageCLEFmedical concept detection, where rare concepts can otherwise be overwhelmed by the background of absent labels.

2.4. Pretraining and ensemble considerations

Medical-domain pretraining is often attractive because it can provide representations closer to radiological imagery than ImageNet, and prior medical-imaging studies have shown that preprocessing and initialization choices can strongly affect downstream performance [8]. However, it is not guaranteed to improve every downstream task. In our experiments, RadImageNet initialization did not outperform the standard ImageNet-initialized DenseNet-121 under the same practical learning-rate regime. The medical-pretrained model also appeared less stable during fine-tuning, which we interpret as behavior consistent with catastrophic forgetting: useful pretrained features may be rapidly overwritten when the downstream task is broad, heterogeneous, and heavily long-tailed.

We also explored an ensemble that incorporated predictions from additional architectures, including ConvNeXt and EfficientNet variants. Although ensembles often improve robustness, our ensemble submission reached 0.5712 F1, slightly below the final single DenseNet-121 full-fit model. We therefore treat this result as an empirical observation rather than proof that ensembling is generally ineffective.

A plausible explanation is correlated error: models trained on the same imbalanced label distribution may agree on frequent concepts while jointly missing difficult rare labels. This observation shaped our final emphasis on a single, well-calibrated, imbalance-aware model rather than a heavier but less discriminative ensemble.

3. Dataset and Label Representation

We used the official ImageCLEFmedical 2026 development data for the Concept Detection task. The challenge uses an updated and extended version of the ROCOv2 dataset, comprising 116,604 medical images with CUI annotations [3, 11]. The provided split contains 97,364 training images and 19,240 validation images. For each image, the annotation field stores zero or more CUI strings separated by semicolons. After parsing all annotations, removing empty entries, and normalizing whitespace, we obtained 2,646 unique UMLS concepts.

Table 1
Summary of the ImageCLEFmedical 2026 development data used in our experiments.

Property	Value
Total development images	116,604
Official training images	97,364
Official validation images	19,240
Unique CUI labels	2,646
Average labels per image	3.2206
Median labels per image	3
Maximum labels per image	37
Positive label density	0.1218%
Concepts appearing at most 10 times	420
Final fitting images	116,604
Input resolution	224×224

To make the visual heterogeneity of the dataset concrete, Figure 1 shows six representative validation images from IDs `valid_18294–valid_18299`. These examples illustrate different modalities, anatomical regions, and annotation styles. They are used only for dataset illustration; the model is trained to predict CUI sets rather than captions.

Each image is represented by a sparse multi-hot vector. Let K denote the number of unique concepts; in our experiments, $K = 2646$. For image i, the target vector is

$$\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iK}] \in \{0, 1\}^K, \tag{1}$$

where $y_{ik} = 1$ if concept k is assigned to image i, and $y_{ik} = 0$ otherwise. The average of 3.2206 labels per image confirms that the target matrix is extremely sparse. This corresponds to a positive label density of only 0.1218% across the full label vector. Figure 2 visualizes this sparsity and long-tail behavior: most images contain only a few labels, concept frequencies span several orders of magnitude, and 420 concepts appear at most ten times in the development set. This sparsity is the central reason for using ASL rather than standard BCE.

The official split was used during development to inspect the label distribution, monitor training behavior, and decide a simple global inference threshold. For the final submitted model, we trained on the union of the official training and validation sets. This full-fit setting increases the number of examples available for rare concepts and was more aligned with the final hidden-test objective.

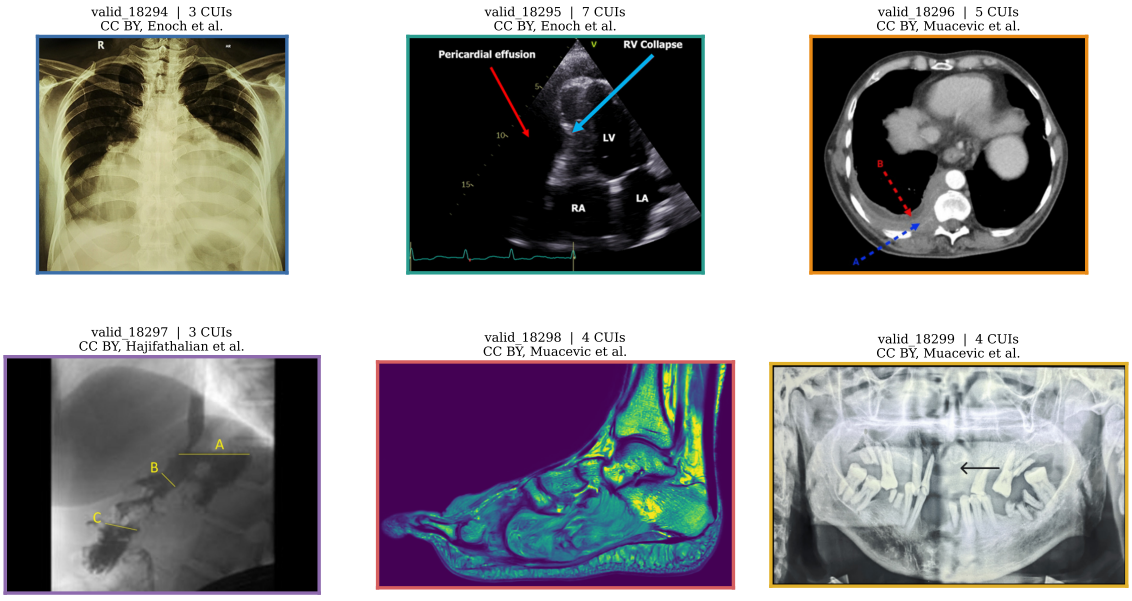

Figure 1: Representative validation samples from the ImageCLEFmedical 2026 development set. The examples illustrate the visual heterogeneity of the task, including different imaging modalities, anatomical regions, annotation styles, and numbers of associated UMLS CUI labels. Images are reproduced from the ImageCLEFmedical 2026 development dataset, which is based on an updated and extended version of the ROCOv2 dataset. Image attributions: valid_18294 (Enoch et al., CC BY); valid_18295 (Enoch et al., CC BY); valid_18296 (Muacevic et al., CC BY); valid_18297 (Hajifathalian et al., CC BY); valid_18298 (Muacevic et al., CC BY); valid_18299 (Muacevic et al., CC BY).

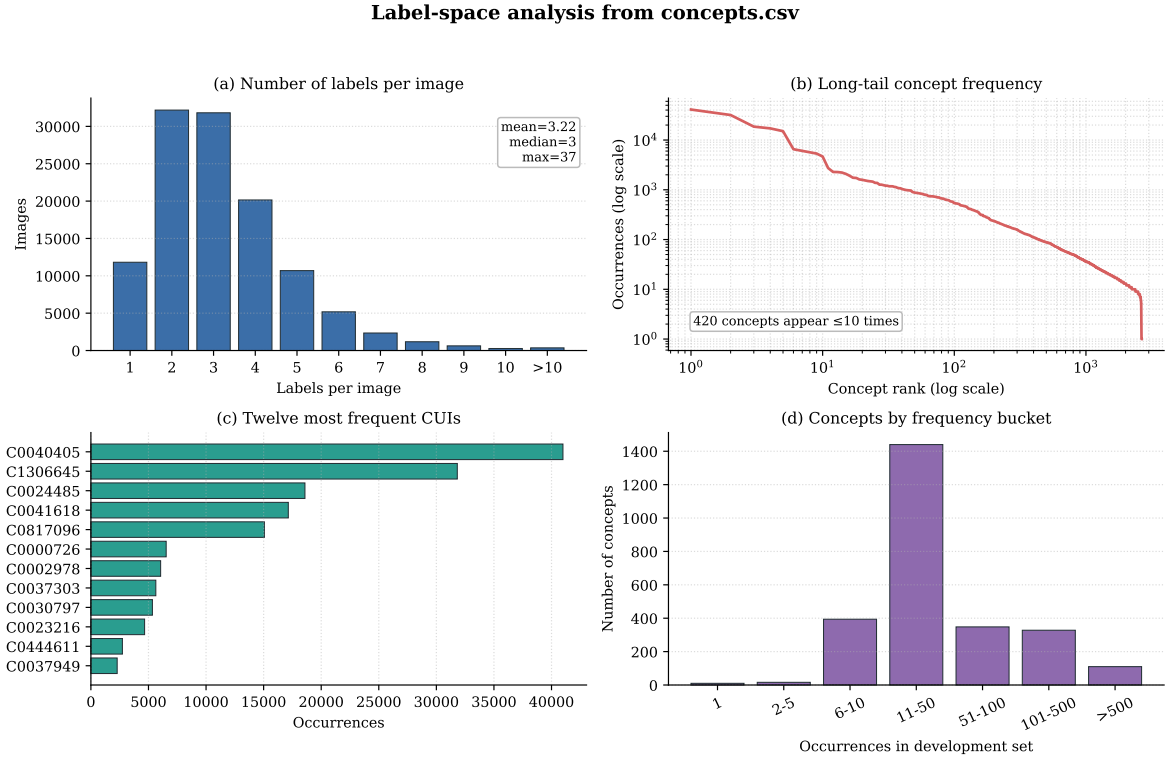

Figure 2: Label-space analysis from `concepts.csv`: number of labels per image, long-tail concept frequency, frequent CUIs, and concept-frequency buckets.

4. Preprocessing and Data Augmentation

4.1. Image preprocessing

All images were loaded from disk and converted to RGB format. Since the DenseNet-121 backbone was initialized with ImageNet weights, every image was resized to 224×224 pixels and normalized with the standard ImageNet channel-wise statistics. The mean vector was $(0.485, 0.456, 0.406)$ and the standard deviation vector was $(0.229, 0.224, 0.225)$.

The resizing operation standardizes heterogeneous image dimensions and enables efficient mini-batch training. Although resizing may compress high-resolution details, it provides a practical balance between computational cost and feature quality for the large development set. The use of ImageNet normalization is also consistent with the pretrained backbone and avoids a mismatch between pretraining and fine-tuning input statistics.

4.2. Training augmentation

During training, we used lightweight data augmentation implemented with Albumentations [12]. The augmentation pipeline included random shift, scale, and rotation with probability 0.5, as well as random brightness and contrast adjustment with probability 0.5. These transformations simulate moderate variation in image positioning, scanner contrast, and acquisition conditions while avoiding overly aggressive changes that could distort subtle medical findings. During inference, only resizing and normalization were applied.

5. Method

5.1. Model architecture

Our final model is based on DenseNet-121 [4]. We used the implementation from the `timm` library [13] and initialized the backbone with standard ImageNet-pretrained weights [14]. We did not use RadImageNet weights in the final submission. The original classification head was replaced with a task-specific linear layer that outputs one logit for each of the 2,646 UMLS concepts.

Let $\mathbf{h} \in \mathbb{R}^{1024}$ be the feature vector produced by the DenseNet-121 backbone after global pooling. The classifier computes

$$\mathbf{z} = W\mathbf{h} + \mathbf{b}, \tag{2}$$

where $W \in \mathbb{R}^{2646 \times 1024}$ and $\mathbf{z} \in \mathbb{R}^{2646}$ is the logit vector. The predicted probability for concept k is obtained independently using the sigmoid function:

$$\hat{p}_k = \sigma(z_k) = \frac{1}{1 + e^{-z_k}}. \tag{3}$$

This formulation allows the model to assign multiple concepts to the same image. A dropout rate of 0.2 was used before the final classifier for regularization. Figure 3 summarizes the final DenseNet-121 classifier used in our submission.

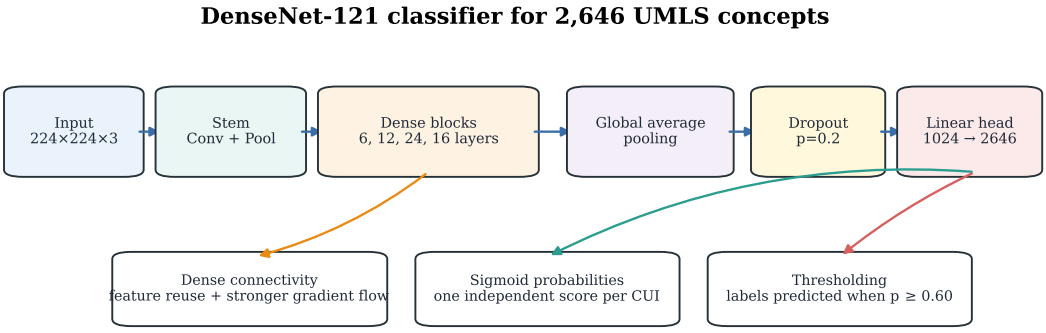

Figure 3: DenseNet-121 classifier for multi-label medical concept detection. The ImageNet-initialized backbone produces a 1024-dimensional feature vector, and the task-specific linear head outputs 2,646 independent sigmoid probabilities corresponding to UMLS concepts.

5.2. Asymmetric Loss

A major difficulty in this task is the imbalance between positive and negative labels. For each image, only a few of the 2,646 labels are positive, while most labels are absent. To reduce the dominance of easy negatives, we optimized the network with Asymmetric Loss [5].

For each class probability $p_k = \sigma(z_k)$, ASL applies asymmetric focusing to positive and negative terms. In our implementation, the negative probability is adjusted with a small clipping parameter c to reduce the contribution of very easy negatives. We used

$$\tilde{p}_k^- = \min(1 - p_k + c, 1), \tag{4}$$

with clipping parameter $c = 0.05$. The loss for a multi-label target vector is then expressed as

$$\mathcal{L}_{ASL} = -\frac{1}{K} \sum_{k=1}^{K} \left[y_k (1 - p_k)^{\gamma^+} \log(p_k) + (1 - y_k)(1 - \tilde{p}_k^-)^{\gamma^-} \log(\tilde{p}_k^-) \right]. \tag{5}$$

We set $\gamma^- = 4.0$ and $\gamma^+ = 1.0$. The larger negative focusing parameter strongly down-weights easy negative entries, allowing the optimization process to focus more on rare positive labels and hard negatives. Figure 4 illustrates the focusing behavior used in our submission. This behavior is particularly important in our 2,646-label setting, where the negative label mass is overwhelming.

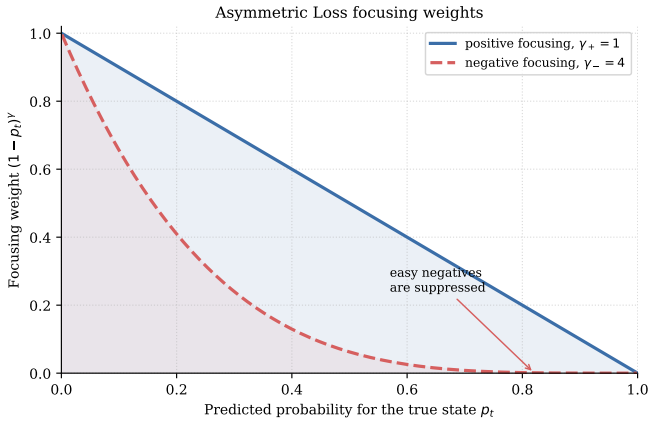

Figure 4: Asymmetric Loss focusing weights used in our submission. The larger negative focusing parameter, $\gamma^- = 4.0$, suppresses easy negative labels more aggressively than the positive focusing term $\gamma^+ = 1.0$.

5.3. Full-fit training strategy

Our final submission used a full-fit strategy. During analysis, the official split of 97,364 training images and 19,240 validation images was useful for estimating the label distribution and checking the behavior of the thresholding rule. For the final model, however, we merged these subsets and trained on all 116,604 labeled images. The motivation is straightforward: in an extreme long-tail dataset, the validation subset contains many rare-label examples that are valuable for representation learning. Excluding them from final training reduces the model's exposure to precisely the concepts that are hardest to learn.

This strategy differs from a conventional validation-based research protocol and should be interpreted as the final competition-fitting procedure rather than as a replacement for controlled model selection. We did not select the final checkpoint by validation F1. Instead, we used a fixed training recipe and selected the final checkpoint based on training stability and loss behavior. Across the 6-epoch schedule, the training loss remained stable, and the final hidden-test score indicates that this full-data fitting strategy was beneficial for the official evaluation setting.

5.4. Optimization and memory-efficient training

We trained the model with AdamW [15], using an initial learning rate of 2×10^{-4} and a weight decay of 10^{-4}. A cosine annealing schedule [16] decayed the learning rate to a minimum of 10^{-6} over 6 epochs:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_0 - \eta_{min}) \left(1 + \cos\left(\frac{\pi t}{T_{max}}\right)\right), \tag{6}$$

where $T_{max} = 6$. The batch size was 32.

To fit the large training set efficiently within GPU memory limits, we used PyTorch Automatic Mixed Precision (AMP) with `GradScaler` [17]. AMP executes many operations in half precision to reduce memory consumption and improve throughput, while `GradScaler` prevents numerical underflow by scaling losses before backpropagation. This data-engineering component was essential for processing the full 116,604-image dataset without VRAM overflow.

Table 2
Training configuration used for the final UIT_Worker submission.

Component	Setting
Backbone	DenseNet-121
Initialization	ImageNet pretrained
Medical-domain pretraining	Not used
Number of output labels	2,646
Input size	224×224
Dropout	0.2
Optimizer	AdamW
Learning rate	2×10^{-4}
Weight decay	10^{-4}
Epochs	6
Batch size	32
Scheduler	Cosine annealing
Minimum learning rate	10^{-6}
Loss function	Asymmetric Loss
ASL parameters	$\gamma^- = 4.0$, $\gamma^+ = 1.0$, clipping $c = 0.05$
Precision	AMP with GradScaler
Final fitting data	All 116,604 development images
Inference threshold	0.60

5.5. Inference and submission generation

At inference time, each test image was processed with the deterministic transform pipeline: RGB conversion, resizing to 224×224, and ImageNet normalization. The model produced a 2,646-dimensional logit vector. We applied the sigmoid function to obtain per-concept probabilities and used a fixed threshold $\tau = 0.60$:

$$\hat{y}_k = \begin{cases} 1, & \text{if } \hat{p}_k \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The binary predictions were decoded back into CUI strings using the fitted label mapping. For each test image, the predicted CUIs were joined by semicolons to match the required submission format. The final inference run generated predictions for 15,249 test images, and the submitted file contained the required `ID` and `CUIs` columns.

Figure 5 summarizes the complete end-to-end pipeline, from image and CUI loading to fixed-threshold submission generation.

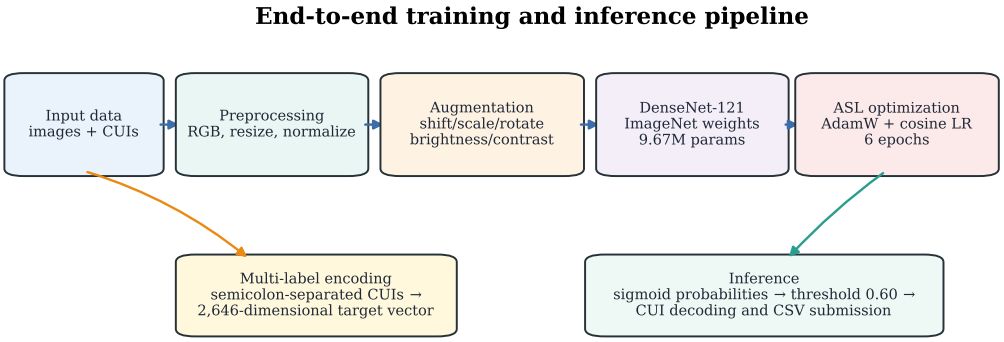

Figure 5: Overview of the proposed DenseNet-121 full-fit concept detection pipeline. The system loads images and CUI labels, applies RGB conversion, resizing, ImageNet normalization, lightweight augmentation during training, DenseNet-121 feature learning, ASL optimization, and fixed-threshold CUI decoding.

6. Results and Analysis

6.1. Official results

The official leaderboard snapshot is shown in Table 3. Our submitted UIT_Worker entry achieved a primary F1-score of 0.5725 and a secondary F1-score of 0.9533 on the hidden test set, ranking third in the ImageCLEFmedical 2026 Concept Detection task. The margin to the top-ranked entry was small, indicating that the proposed single-model DenseNet-121 full-fit system was competitive with the strongest submissions.

Table 3
Official leaderboard snapshot for the ImageCLEFmedical 2026 Concept Detection task.

Rank	Owner	Primary F1	Secondary F1
1	dsgt_caption	0.5790	0.9657
2	archimedes	0.5781	0.9591
3	UIT_Worker	0.5725	0.9533
4	basharderar	0.5725	0.9419
5	AUEB/DMST	0.5721	0.9503

The result demonstrates that a carefully trained convolutional baseline remains competitive for large-scale medical concept detection. DenseNet-121 provides efficient and reusable visual features, ASL directly addresses the sparse multi-label target structure, and the threshold of 0.60 provides a conservative precision-recall balance suitable for the imbalanced label space.

6.2. Empirical observations and submission analysis

Table 4 summarizes the main empirical observations that guided the final submission. The final ImageNet-initialized DenseNet-121 trained with full-fit ASL reached the best score among our tested strategies. In contrast, medical-domain initialization with RadImageNet did not improve the final system, and the ensemble strategy was slightly worse than the single model.

Table 4
Summary of key empirical observations from our submission development.

Strategy	F1	Observation
DenseNet-121 + ImageNet + ASL + full-fit	**0.5725**	Best official configuration; stable under the final 116,604-image full-fit setting.
Ensemble with ConvNeXt/EfficientNet variants	0.5712	Slightly lower score; models tended to reinforce similar mistakes on hard rare labels.
RadImageNet initialization	0.5314	Did not outperform ImageNet initialization; fine-tuning was less stable with the standard learning-rate recipe.

The first important observation is that ImageNet pretraining was sufficient and more reliable than RadImageNet in our final setting. Although RadImageNet is closer to the target domain, the optimization behavior was more fragile. With the learning rate used in our main recipe, the model appeared to lose useful pretrained structure quickly during full fine-tuning. We interpret this behavior as consistent with catastrophic forgetting, while acknowledging that a dedicated learning-rate and freezing-stage study would be required to prove the mechanism conclusively.

The second observation is that ensembling was not automatically beneficial. Our ensemble combined multiple architectures, including ConvNeXt and EfficientNet variants, but the score decreased to 0.5712. We hypothesize that the issue was not model capacity but correlated error. Because the models were trained on the same sparse and imbalanced labels, they may have agreed on common concepts while jointly failing on difficult rare CUIs. Averaging such predictions can amplify frequent-label confidence without recovering missed rare concepts; we refer to this informally as an echo-chamber-like effect.

The third observation concerns the effectiveness of the proposed full-fit training strategy. To validate our design choice, we compared the final full-fit model with the same DenseNet-121 configuration trained only on the official training split while reserving the validation split for model selection. The comparison is summarized in Table 5.

Table 5
Comparison between the standard train/validation split and the proposed full-fit training strategy.

Training strategy	Primary F1	Secondary F1
Standard train/validation split	0.5669	0.9413
Full-fit (116,604 development images)	**0.5725**	**0.9533**

Training on the complete development set consistently improved both evaluation metrics. Specifically, compared with the standard train/validation split, the primary F1-score increased from 0.5669 to 0.5725, while the secondary F1-score improved from 0.9413 to 0.9533. Although the absolute gain is modest, the results provide empirical evidence supporting our hypothesis that exposing the model to all available labeled images, including additional rare-label examples contained in the validation split, improves

generalization on the hidden test set. This experiment therefore justifies the full-fit strategy adopted for the final submission.

6.3. Error characteristics and limitations

Despite the strong official ranking, the system has several limitations. First, it uses a single global threshold. A threshold of 0.60 worked well overall, but frequent and rare concepts can have very different calibration profiles. Class-wise thresholds could improve recall for rare labels while controlling false positives for frequent labels. Second, the model predicts each CUI independently and does not explicitly model label dependencies. In medical images, concepts such as modality, anatomy, and procedure are often correlated. Third, the system uses only image-level supervision and does not exploit textual metadata, captions, or external knowledge graphs during prediction. Fourth, our submission analysis is not a complete controlled ablation: we did not report BCE versus ASL, threshold sweeps, augmentation ablations, or rare-versus-frequent label breakdowns, which would provide a deeper explanation of the final score.

These limitations suggest that future improvements should focus not merely on larger backbones, but on better calibration, rare-label specialization, and dependency-aware prediction. The failure of our naive ensemble also suggests that future ensembles should be diversity-aware, rather than simply combining several high-capacity models trained with similar objectives.

7. Conclusion and Future Work

We presented the UIT_Worker submission to the ImageCLEFmedical 2026 Concept Detection task. The system formulates medical concept detection as large-scale multi-label classification over 2,646 UMLS CUIs. Our final model uses DenseNet-121 initialized with standard ImageNet weights, trained end-to-end on all 116,604 development images with a full-fit strategy. The model is optimized with Asymmetric Loss using $\gamma^- = 4.0$, $\gamma^+ = 1.0$, and clipping parameter $c = 0.05$, together with AdamW, cosine annealing for 6 epochs, and AMP with GradScaler. A fixed threshold of 0.60 converts sigmoid probabilities into final CUI predictions.

The official result, 0.5725 primary F1 and 0.9533 secondary F1, ranked third in the ImageCLEFmedical 2026 Concept Detection task. The results suggest that: under extreme long-tail supervision, a carefully trained single convolutional model with an imbalance-aware objective can be more reliable than more complex alternatives that are not explicitly calibrated for rare labels. In our experiments, ImageNet initialization with full fitting was stronger than RadImageNet initialization, and a naive ensemble slightly reduced performance, likely because of correlated errors on difficult labels.

Future work will focus on three directions. First, we plan to investigate class-wise threshold calibration to better balance rare-label recall and frequent-label precision. Second, we will explore label-dependency modeling through graph-based classifiers or transformer decoders over CUI embeddings. Third, we will revisit ensembles using diversity-aware training, per-concept specialist models, or calibrated late fusion, rather than simple averaging of models that share the same error patterns.

Code Availability

The implementation of the proposed DenseNet-121 full-fit system is publicly available at: https://github.com/nhanbui246/ImageCLEFmedical-2026-UIT-Worker.

Acknowledgments

This work was conducted at the University of Information Technology, Vietnam National University, Ho Chi Minh City. The authors thank the ImageCLEFmedical 2026 organizers for providing the dataset, evaluation platform, and benchmark task.

This research was funded by University of Information Technology-Vietnam National University HoChiMinh City under grant number CS4-2026-01.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: grammar and spelling check, paraphrase and reword, language polishing, and translation between Vietnamese and English. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88. doi:10.1016/j.media.2017.07.005.

[2] B. Ionescu, H. Müller, D.-C. Stanciu, A. Radu, R.-G. Bolborici, M. Negru, A.-F. Ene, V.-M. Vasilescu, A.-A. Nicolae, L.-D. Ştefan, M.-G. Constantin, M. Dogariu, A.-G. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryılmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W.-w. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C.-N. Florea, M. Ivanovici, Overview of ImageCLEF 2026: Multimodal Challenges in Medicine, Science, Agritech, and Security, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science (LNCS), Jena, Germany, September 21–24, 2026.

[3] H. Damm, T. M. G. Pakull, L. Reinartz, B. Bracke, B. Eryılmaz, P. Nath, R. Brüngel, C. S. Schmidt, H. Schäfer, A. Ben Abacha, A. García Seco de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2026 – Medical Concept Detection and Caption Generation with Synthetic Data Extensions, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, September 21–24, 2026. To appear.

[4] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708. doi:10.1109/CVPR.2017.243.

[5] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 82–91. doi:10.1109/ICCV48922.2021.00015

[6] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. M. Summers, 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, pp. 2097–2106. doi:10.1109/CVPR.2017.369.

[7] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, A. Y. Ng, Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists, PLOS Medicine 15 (2018) e1002686. doi:10.1371/journal.pmed.1002686.

[8] T. B. Nguyen-Tat, T. Q. Hung, P. T. Nam, V. M. Ngo, Evaluating pre-processing and deep learning methods in medical imaging: Combined effectiveness across multiple modalities, Alexandria Engineering Journal 119 (2025) 558–586. doi:10.1016/j.aej.2025.01.090.

[9] G.-P. Bui-Hoang, M.-H. Dinh-Doan, V.-M. Luong, T. B. Nguyen-Tat, UIT-Oggy at ImageCLEFmedi-

cal 2025 caption: CSRA-enhanced concept detection and BLIP-driven vision-language captioning, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988. doi:10.1109/ICCV.2017.324.
[11] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. Ben Abacha, A. García Seco de Herrera, H. Müller, P. Horn, F. Nensa, C. M. Friedrich, ROCOv2: Radiology objects in context version 2, an updated multimodal image dataset, Scientific Data 11 (2024). doi:10.1038/s41597-024-03496-6.
[12] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A. A. Kalinin, Albumentations: Fast and flexible image augmentations, Information 11 (2020) 125. doi:10.3390/info11020125.
[13] R. Wightman, PyTorch Image Models, 2019. URL: https://github.com/huggingface/pytorch-image-models.
[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.
[15] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations (ICLR), 2019.
[16] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, in: International Conference on Learning Representations (ICLR), 2017.
[17] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., PyTorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, volume 32, 2019.
[18] O. Bodenreider, The Unified Medical Language System (UMLS): Integrating biomedical terminology, Nucleic Acids Research 32 (2004) D267–D270. doi:10.1093/nar/gkh061.