Detecting Concepts for Medical Images: Contributions of the DeepLens Team at IUST to ImageCLEFmedical Caption 2025

Notebook for the ImageCLEFmedical Caption Lab at CLEF 2025

Amir Hossein Salimi Rudsari[1,*,†], Bahareh Kavousi Nejad[1,*,†], Malihe Hajihosseini[1,*] and Sauleh Eetemadi[1]

[1]Iran University of Science and Technology, Tehran, Iran

Abstract

The ImageCLEFmedical Caption Task 2025 challenge includes two subtasks: Concept Detection and Caption Prediction. This paper addresses the Concept Detection subtask, which involves automatically identifying relevant medical concepts from radiological images to support semantic image retrieval and clinical decision-making. We developed multiple deep learning models, including Convolutional Neural Networks (CNNs) combined with Feed-Forward Neural Networks (FFNNs), as well as a novel architecture named KClipMed, based on the K-Nearest Concept-Language-Image Pre-training (KCLIP) framework. KClipMed incorporates Top-k Concept Retrieval, a trainable Logit Temperature, and a Cross-Attention mechanism to enhance image-concept alignment. We also explored ensemble strategies, with the best-performing ensemble—combining EfficientNet and DenseNet—achieving an F1 score of 57.66% on the test set, placing second in the competition with a margin of 1.2% behind the top-ranked team.

Keywords

Concept Detection, Medical Image Analysis, CNN, FFNN, KClipMed, Top-k Retrieval, Cross-Attention, Ensemble

1. Introduction

The ImageCLEF [1] medical challenge is an integral part of ImageCLEF, an ongoing evaluation platform initiated in 2003 under the Conference and Labs of the Evaluation Forum (CLEF). The primary goal of ImageCLEF is to facilitate progress and innovation in methods used for indexing, retrieval, classification, and annotation of multimodal content, with a particular emphasis on biomedical applications. Over the years, ImageCLEFmedical has seen growing international engagement, considerably contributing to the advancements in medical image processing and automatic captioning techniques [2, 3].

For the 2025 ImageCLEFmedical Caption Task, two distinct subtasks were introduced: Concept Detection and Caption Prediction [4]. Concept Detection involves automatically assigning relevant medical concepts to biomedical images, which supports semantic searches and assists medical practitioners by offering preliminary structured insights [5]. The evaluation of system performance in this task was based on binary F1 scores, measuring the overlap between predicted and ground-truth concept sets [6]. Conversely, the Caption Prediction task focuses on generating detailed and accurate captions for medical images, thereby providing diagnostic information and aiding clinicians in their workflow.

Our group, DeepLens from Iran University of Science and Technology, participated specifically in the Concept Detection subtask for the 2025 edition. Our approach involved exploring three different

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

salimi_amirhosein@comp.iust.ac.ir (A. Salimi Rudsari); bahareh_kavousi@comp.iust.ac.ir (B. Kavousi Nejad); m_hajihosseini@comp.iust.ac.ir (M. Hajihosseini); sauleh@iust.ac.ir (S. Eetemadi)

https://www.sauleh.ir/ (S. Eetemadi)

0009-0009-8977-2800 (A. Salimi Rudsari); 0009-0004-7958-0227 (B. Kavousi Nejad); 0000-0002-0067-1184 (M. Hajihosseini); 0000-0003-1376-2023 (S. Eetemadi)

modeling strategies. Initially, we utilized several Convolutional Neural Networks (CNNs) paired with Feed-Forward Neural Networks (FFNNs) for classifying medical images into corresponding concepts.

In addition, we developed a novel model named KClipMed based on K-Nearest Concept-Language-Image Pre-training (KCLIP). The KClipMed architecture incorporates a Top-k Concept Retrieval mechanism, Cross-Attention, and a trainable Logit Temperature module, aiming to improve the alignment between visual features and medical concepts. This model was explored using two different image encoders: Swin Transformer and Vision Transformer (ViT).

Finally, we applied extensive ensemble approaches by aggregating predictions from all possible combinations of our proposed models using union-based aggregation. Through these comprehensive experimental strategies, we successfully secured second place in the Concept Detection subtask, closely following the leading team with only a 1.2% gap.

This paper outlines our methodological approaches, elaborates on the models we designed, discusses the experimental outcomes, and considers the implications of our findings for future developments in medical image analysis and concept-driven methodologies.

2. Related Work

A number of comprehensive reviews have summarized the role of deep learning in medical image analysis [7], laying the groundwork for tasks like concept detection. Concept detection in medical images aims to map each image to a set of relevant UMLS Concept Unique Identifiers (CUIs) for retrieval, annotation, or clinical support. Prior work in this area can be grouped into three main technical approaches that informed our system design.

2.1. Convolutional Neural Networks

Convolutional neural networks (CNNs) have long been the foundation for medical image classification and tagging [8]. DenseNet architectures [9], in particular, became popular in radiology after their success in thoracic disease detection, where their dense connectivity helped model subtle patterns in chest X-rays [10]. EfficientNet models [11] were introduced as a more parameter-efficient alternative that scales depth, width, and resolution in a balanced way, making them attractive for multi-label medical image analysis where computational efficiency matters [12]. Prior studies have shown that both DenseNet and EfficientNet backbones, fine-tuned on medical datasets[13, 14], provide strong baselines for multi-label tagging tasks such as concept detection in radiographs.

2.2. Transformer-Based Architectures and Concept-Aware Models

Transformer-based models have gained ground in medical image analysis because of their ability to model global relationships across image regions. Vision Transformers (ViTs) [15] partition images into patches and apply self-attention to learn contextual dependencies, while hierarchical variants like Swin Transformers [16] introduce shifted windows to efficiently capture both local and global structure. These architectures[17] have been successfully applied to radiographic image tagging, often outperforming conventional CNNs in settings where long-range dependencies between regions of interest are important.

Building on this foundation, concept-aware models [18] extend vision architectures by aligning image features with semantic representations of medical concepts. In our KClipMed models, we combine visual encoders (ViT or Swin Transformer backbones) with learnable embeddings representing individual CUIs. A top-K nearest retrieval mechanism and attention-based fusion allow the model to focus on semantically relevant concepts, enhancing its ability to discriminate between visually similar but semantically distinct classes.

Table 1
Dataset description for concept detection

Dataset	Training Set	Validation Set	Test Set
No. of images	80,091	17,277	19,267
No. of concept IDs (CUIs)	2,479	2,283	—
Max CUIs per image	28	26	—

2.3. Ensemble Strategies for Multi-Label Medical Image Tagging

Ensemble learning [19] has proven to be a simple and effective method for improving performance on multi-label tasks with class imbalance. By combining the outputs of models with different architectures — such as CNNs and Transformers — ensembles benefit from the complementary strengths of their components. Union-based aggregation strategies, in particular, are well suited to multi-label concept detection, as they improve recall by preserving any positive prediction from the constituent models without requiring complex calibration [20]. This approach helps stabilise performance on rare concepts and is computationally efficient.

Together, these prior works shaped the design of our system, which combines CNN baselines, Transformer encoders, concept-aware alignment, and union-based ensembles for robust medical concept detection.

3. Dataset

The data used in the 2025 ImageCLEFmedical Caption Task consists of curated medical images extracted from biomedical literature, accompanied by corresponding captions and manually controlled Unified Medical Language System (UMLS) terms as metadata. The dataset provided for the challenge aimed to facilitate advanced and diverse methodological approaches by offering a comprehensive collection of radiological images.

Specifically, for both training and development purposes, the Radiology Objects in COntext Version 2 (ROCOv2) dataset [21] was utilized. This dataset is an expanded and enhanced version of the original ROCO dataset, sourced primarily from biomedical articles within the PMC OpenAccess subset. The provided data was divided into three subsets: the training set comprising 80,091 images, the validation set including 17,277 images, and the test set consisting of 19,267 previously unseen images. The dataset statistics are summarized in Table 1.

For the Concept Detection subtask, the concepts were derived from a carefully filtered subset of the UMLS 2022 AB release, where filtering was applied based on semantic types and concept frequency to enhance the feasibility of recognizing relevant medical concepts from the provided images.

During data preprocessing, we observed that the original radiological images contained unnecessary white borders. Recognizing that these borders did not contribute valuable information and increased storage requirements, we systematically removed all white borders from the images. This preprocessing step effectively reduced the total size of the training dataset by approximately two gigabytes, thus facilitating more efficient data handling and model training [22].

Additionally, our analysis revealed that the training set includes **2,479 unique CUIs**, while the validation set includes **2,283 unique CUIs**. The most frequent concepts in the training data are listed in Table 2.

4. Methodology and Model Architectures

In this section, we present our approach to the multi-label concept detection challenge in ImageCLEF 2025. We developed and evaluated a wide range of deep learning architectures with the goal of identifying multiple medical concepts in each radiological image, even in the presence of challenges

Table 2
Top 10 most frequent UMLS concepts in the training data

UMLS CUI	UMLS Meaning	Frequency
C0040405	X-Ray Computed Tomography	27,790
C1306645	Plain X-ray	21,979
C0024485	Magnetic Resonance Imaging	12,666
C0041618	Ultrasonography	11,422
C0817096	Chest	10,226
C0002978	Angiogram	4,785
C0000726	Abdomen	4,280
C0037303	Bone Structure of Cranium	4,108
C0030797	Pelvis	3,661
C0023216	Lower Extremity	3,230

such as class imbalance, visual ambiguity, and overlapping semantics. Figure 1 provides an overview of our system design.

We grouped our models into three main categories: **CNN + FFNN**, **KClipMed**, and **Ensemble**. Each group focused on different strengths—CNN + FFNN models used powerful convolutional backbones to extract visual features; KClipMed models combined image and concept understanding through contrastive embeddings and attention mechanisms; and Ensemble models merged predictions from various architectures to improve reliability and overall performance. This strategy helped us make the most of each model type and build a balanced system that performs well across diverse and complex data. Since our models are based on pretrained architectures, we rely on transfer learning—a widely adopted strategy in medical imaging that has demonstrated significant benefits across domains [23].

4.1. CNN + FFNN Models

The CNN + FFNN group included EfficientNet and DenseNet architectures. All models were initialized with pretrained weights and fine-tuned with custom classification heads adapted for multi-label prediction. The original output layers were replaced with feedforward neural networks, using either GeM pooling or global average pooling for feature aggregation, followed by sigmoid activations to generate independent probability scores for each concept. Training was performed using the Binary Cross Entropy with Logits (BCEWithLogitsLoss) loss function and optimized with the Adam optimizer. Performance was evaluated based on micro-averaged F1 scores, and checkpoints were saved when improvements were observed.

4.1.1. DenseNet Models

We used DenseNet-121 [9] with an input resolution of 128×128 to reduce computational overhead while maintaining effective feature extraction. Its densely connected layers helped preserve gradient flow and capture fine-grained features. A simple linear classification head replaced the original output layer to support multi-label output, making it a solid and efficient baseline in our CNN lineup.

4.1.2. EfficientNet Models

We employed EfficientNet variants B0 to B3, selecting input resolutions from 224×224 to 300×300 based on model depth and hardware constraints. Each model used GeM pooling to aggregate spatial features, followed by a three-layer feedforward neural network with ReLU activations, dropout, and sigmoid outputs for multi-label classification. The output dimensionality of the feature extractor varied by model, with B3 producing 1536 features, B2 yielding 1408, and B1 outputting 1280. Training was conducted using the Adam optimizer with an initial learning rate of 1×10^{-3}. For B2 and B3, a `StepLR`

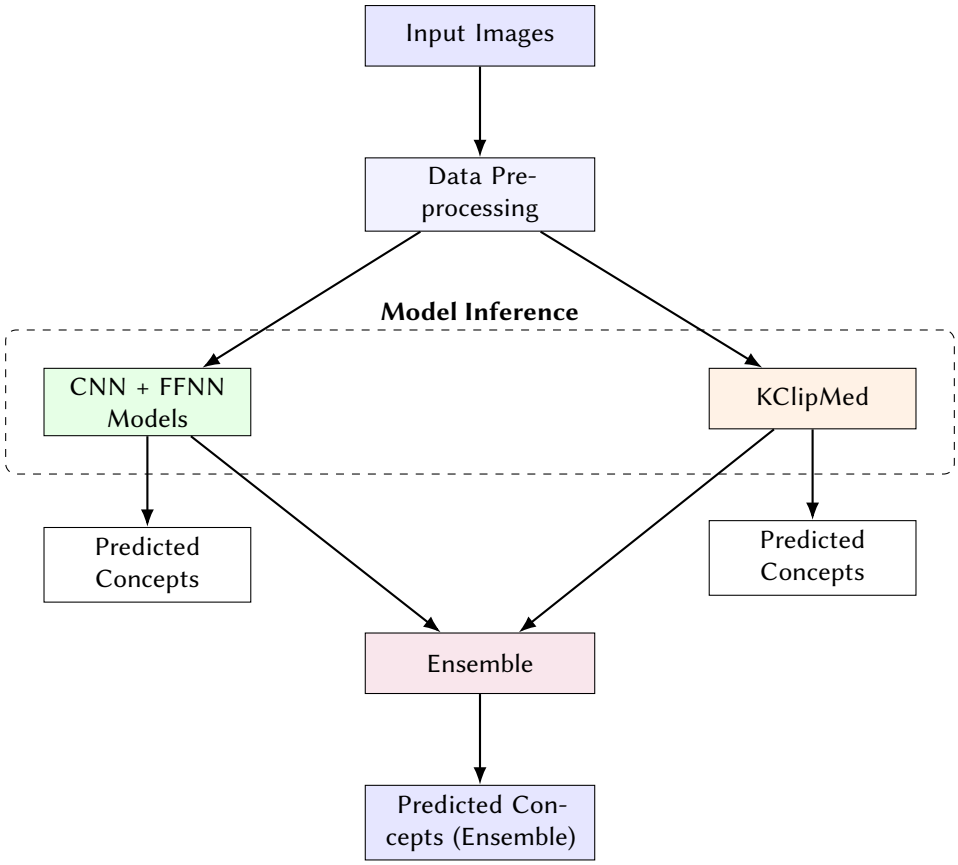

Figure 1: System Architecture for Concept Detection Task. Each model group (CNN and KClipMed) produces independent predictions. In parallel, an ensemble strategy combines outputs to generate a fused prediction.

scheduler reduced the learning rate every 5 epochs to stabilize convergence. All models were trained with BCEWithLogitsLoss, and in B3, label smoothing was also explored [24].

4.2. KClipMed Models

We incorporated two KClipMed models into our system to integrate semantic understanding into the image-based concept detection pipeline—one using a Swin Transformer backbone and another using a standard Vision Transformer (ViT). Both models follow a vision-language architecture that aligns image features with concept embeddings in a shared representation space, guided by contrastive learning principles.

In the Swin Transformer variant, image features are extracted using a Swin backbone. These features are projected into a fixed embedding space and compared to a table of learnable embeddings, each representing a distinct medical concept (CUI). For each image, the model selects the top-K most similar concept embeddings and combines them with the image representation using a multi-head attention mechanism. The final prediction logits are computed using a scaled dot-product between the attended image representation and the full set of concept embeddings. This mechanism allows the model to focus on semantically relevant concepts and enhance discriminative power across visually similar classes.

In the standard ViT-based KClipMed model, we used a ViT-Small backbone as the visual encoder. Instead of relying on precomputed class tokens, the model extracts token embeddings directly from image patches and uses a positional encoding scheme to maintain spatial coherence. The classification token from the ViT encoder is projected into the shared embedding space and processed similarly to the Swin-based model, including top-K concept selection and attention-based fusion. The use of a pure ViT backbone offers an alternative representation pathway and helps assess how transformer depth

and attention granularity affect performance in multi-label settings.

Both models were trained on images resized to 224×224 and optimized using BCEWithLogitsLoss and the AdamW optimizer with a learning rate of 2×10^{-4} and weight decay of 1×10^{-2}. The evaluation was conducted using both micro and macro F1 scores. The models were checkpointed using the best micro F1 score on the validation set. These KClipMed variants enriched the model ensemble by bringing semantic alignment capabilities and context-aware prediction into the concept detection pipeline, as illustrated in Figure 2 [25].

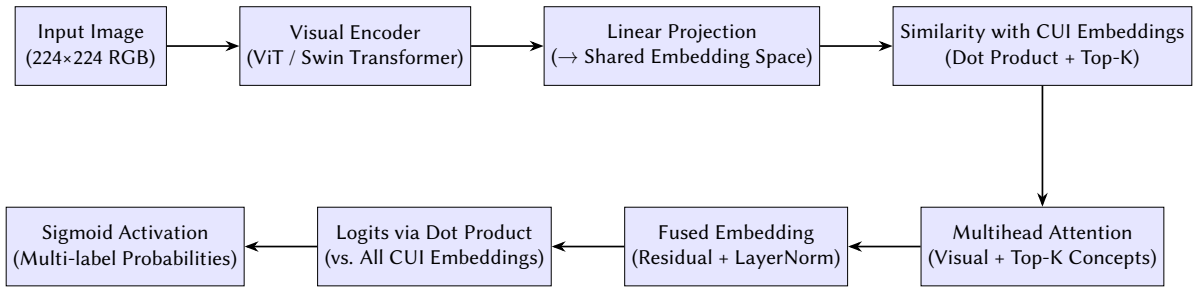

Figure 2: Architecture of the KClipMed model for multi-label concept detection.

Figure 2 outlines the complete data flow of KClipMed: from raw image input, through visual encoding using Swin or ViT, projection into a shared embedding space, retrieval of top-K concept embeddings, to multi-head attention fusion and final sigmoid-activated multi-label prediction. This pipeline is critical for aligning visual features with semantic cues and is our main methodological contribution.

We designed KClipMed to enhance semantic understanding by combining visual and concept embeddings. The use of a Top-k retrieval mechanism allows the model to focus on the most relevant concept candidates per image, improving precision. The trainable logit temperature enables the model to adapt similarity scores during training, refining decision boundaries. Additionally, the multihead module integrates context between image features and concept embeddings, helping the model differentiate between visually similar but semantically distinct concepts. These components were chosen based on their success in recent vision-language models and their interpretability for medical data alignment tasks.

4.2.1. Vision Transformer (ViT) Models

In addition to our KClipMed experiments, we implemented standalone Vision Transformer (ViT) models based on the ViT-B/16 architecture from the `torchvision` library. These models were designed to directly learn visual representations from image patches without convolutional priors. The default classification head was removed and replaced with a custom feedforward neural network (FFNN) that produced sigmoid-activated outputs for multi-label concept prediction.

Input images were resized to 224×224 and normalized using ImageNet statistics. The extracted class token embeddings were passed through a three-layer FFNN consisting of fully connected layers, ReLU activations, and dropout regularization. The models were trained using BCEWithLogitsLoss and optimized using Adam with a learning rate of 1×10^{-3}. Training and validation were monitored using micro and macro F1 scores.

We implemented two variants: a standard ViT model and a mixed-precision version. The mixed-precision model used PyTorch's `autocast` and `GradScaler` to reduce memory usage and accelerate training, especially beneficial for large batch sizes (e.g., 128). Training of the mixed-precision variant followed the same loop structure but wrapped forward and loss computations within a mixed-precision context. Both models performed well and contributed considerably to our ensemble pipeline, with the mixed-precision version offering a favorable trade-off between training speed and numerical stability.

These ViT-based models [26, 16] demonstrated the capacity of transformer-based architectures to extract meaningful features directly from pixel data in the absence of convolutional structure and served as a valuable addition to the diversity of our model ensemble.

4.3. Ensemble Models

In the final stage of model development, we explored ensemble strategies to combine predictions from multiple architectures. Ensembles included combinations of EfficientNet, DenseNet, and KClipMed models. We employed a logical OR (union) approach, where a concept was accepted if predicted by any model in the ensemble. This straightforward and effective strategy helped improve the system's robustness by leveraging the strengths of diverse architectures. Extensive experimentation with different model combinations confirmed that union-based ensembles considerably enhanced the overall consistency and performance of our concept detection system.

5. Experimental Results

This section details the technical environment in which the experiments were conducted, the metrics used for evaluation, and the results achieved. We also analyze model performance across various architectures and ensemble combinations to highlight trade-offs between complexity and accuracy.

5.1. Hardware and Software Setup

All experiments were executed on Google Colab using a single NVIDIA Tesla T4 GPU (15 GB VRAM). The main software stack is:

- **Python** 3.11.12
- **PyTorch** 2.6.0 +cu124
- `torchvision` 0.21.0 +cu124
- `pandas` 2.2.2
- `scikit-learn` 1.6.1

5.2. Evaluation Metrics

The official evaluation script of the ImageCLEFmedical Caption 2025 task [27] was employed. For each test image, the script computes a binary *F1-score* between the predicted and ground-truth concept sets using the `sklearn.metrics` implementation (default `binary` averaging) [19]. Two scores are reported:

1. **Primary F1**: computed on *all* concepts.
2. **Secondary F1 [1]**: computed only on the subset of manually annotated concepts, as required by the task.

Threshold tuning. Since model outputs are sigmoid-activated logits, we manually selected a global decision threshold for each run by testing several values (e.g., 0.3, 0.4, 0.5) on the validation set. The threshold yielding the best primary F1 score was chosen and applied to the final test predictions. This process was repeated independently for each model configuration. To determine optimal prediction thresholds, we performed grid search on the validation set, evaluating threshold values from 0.1 to 0.9 in 0.1 increments. The final thresholds for each model were selected based on the highest micro-F1 score. This manual tuning ensured fair comparisons between models and ensembles.

5.3. Results

In this section, we report both internal validation results and final test set performance.

Table 3
Micro-F1 scores of different model configurations on the validation set, sorted by performance.

Model Configuration	Micro-F1
EfficientNet-B1	0.5334
EfficientNet-B1 + EfficientNet-B0	0.5333
EfficientNet-B1 + KCLIP-ViT	0.5306
EfficientNet-B1 + DenseNet	0.5305
EfficientNet-B0	0.5294
EfficientNet-B1 + KCLIP-Swin	0.5294
EfficientNet-B0 + EfficientNet-B1 + KCLIP-ViT	0.5290
EfficientNet-B0 + DenseNet	0.5286
EfficientNet-B0 + EfficientNet-B1 + DenseNet	0.5284
EfficientNet-B0 + EfficientNet-B1 + KCLIP-ViT	0.5280
EfficientNet-B0 + KCLIP-ViT	0.5277
EfficientNet-B0 + KCLIP-Swin	0.5268
EfficientNet-B0 + KCLIP-ViT + KCLIP-Swin	0.5268
EfficientNet-B0 + DenseNet + KCLIP-ViT	0.5267
EfficientNet-B0 + DenseNet + KCLIP-Swin	0.5253
EfficientNet-B0 + EfficientNet-B1 + DenseNet + KCLIP-ViT	0.5245
EfficientNet-B0 + DenseNet + KCLIP-ViT + KCLIP-Swin	0.5244
EfficientNet-B1 + DenseNet + KCLIP-ViT	0.5232
EfficientNet-B1 + KCLIP-ViT + KCLIP-Swin	0.5223
DenseNet + KCLIP-ViT	0.5220
DenseNet + KCLIP-Swin	0.5214
EfficientNet-B0 + KCLIP-ViT + KCLIP-Swin	0.5205
EfficientNet-B1 + DenseNet + KCLIP-ViT + KCLIP-Swin	0.5202
EfficientNet-B0 + EfficientNet-B1 + DenseNet + KCLIP-ViT + KCLIP-Swin	0.5193
KCLIP-ViT + KCLIP-Swin	0.5169
KCLIP-ViT	0.5105
KCLIP-Swin	0.5073

5.3.1. Validation Set

To select the most effective model configurations, we evaluated a diverse set of architectures and ensembles on the validation set. Table 3 presents the resulting micro-F1 scores, sorted by performance. Simpler convolutional backbones (e.g., EfficientNet1) achieved solid baseline performance, while two-model ensembles such as *EffNet + DenseNet* and *EffNet1 + DenseNet* consistently outperformed single models. Ensembles incorporating vision–language encoders (KCLIP-ViT and KCLIP-Swin) often enhanced semantic understanding, but did not always yield gains in micro-F1, highlighting a trade-off between expressiveness and precision.

5.3.2. Test Set

Based on validation performance, we selected the top-performing models and submitted them to the official evaluation server. Table 4 shows the results on the hidden test set, reporting both the official **primary F1 score** (across all concepts) and **secondary F1 score** (restricted to manually annotated concepts) [1].

The best primary F1 score of **0.5766** was obtained by the *EffNet + DenseNet* ensemble (Submission ID: 1725). This configuration also showed good generalization with a secondary F1 of 0.9299. Interestingly, another submission using KCLIP-ViT achieved the highest secondary F1 score of **0.9306**, though with slightly lower primary performance.

Overall, ensembles of two or three well-tuned CNNs performed most consistently, whereas adding more diverse or pre-trained VLM modules did not always improve results. This suggests that while hybrid architectures offer potential, careful selection and thresholding remain critical for maximizing

Table 4
Performance of submitted runs on the test set. Best primary F1 is shown in bold.

Rank	Model / Configuration	Subm. ID	F1 (primary)	F1 (secondary)
1	EfficientNet-B0 + DenseNet	1725	**0.5766**	0.9299
2	EfficientNet-B1 + DenseNet	1704	0.5764	0.9231
3	EfficientNet-B0 + EfficientNet-B1	1678	0.5754	0.9156
4	EfficientNet-B0	1512	0.5752	0.9304
5	EfficientNet-B1	1677	0.5744	0.9225
6	EfficientNet-B0 + EfficientNet-B1 + DenseNet	1707	0.5739	0.9116
7	EfficientNet-B1 + KCLIP-ViT	1703	0.5725	0.9234
8	EfficientNet-B0 + KCLIP-ViT	1513	0.5724	**0.9306**
9	EfficientNet-B1 + KCLIP-Swin	1705	0.5720	0.9124
10	EfficientNet-B1 + DenseNet + KCLIP-ViT	1728	0.5715	0.9201
11	EfficientNet-B0 + KCLIP-Swin	1514	0.5711	0.9165
12	EfficientNet-B0 + EfficientNet-B1 + KCLIP-ViT	1706	0.5700	0.9119
13	EfficientNet-B0 + EfficientNet-B1 + KCLIP-Swin	1726	0.5691	0.9020

Table 5
Top submissions on the hidden test set from different participants in the 2025 ImageCLEFmedical task.

Rank	Participant	F1 (primary)	F1 (secondary)
1	auebnlpgroup	0.5888	0.9484
2	**deeplens**	**0.5766**	**0.9299**
4	mapan	0.5660	0.9298
5	oggyds312	0.5613	0.9104
6	ds4dh	0.5225	0.8672
7	oggysashimi	0.4543	0.7199
8	sakthiii	0.4003	0.9082
9	jaimage	0.3982	0.8329
10	ronghaopan	0.2398	0.5377
11	lekshmiscopevit	0.1494	0.2298

multi-label classification performance. The top submissions from various teams are shown in Table 5 [1].

Interestingly, the best results were achieved using a vanilla ensemble of EfficientNet and DenseNet, outperforming more complex transformer-based architectures. We attribute this to their complementary feature extraction styles: DenseNet's dense connectivity preserves low-level features, while EfficientNet's scaling strategy enhances semantic abstraction. Their combination likely provides a balanced representation that generalizes well across the medical domain, especially under class imbalance conditions.

6. Conclusion and Future Work

In this paper, we presented our system for the *ImageCLEFmedical 2025* concept detection task, focusing on multi-label classification of medical images using a diverse set of deep learning models. Our approach involved fine-tuning a variety of convolutional and transformer-based architectures, including EfficientNet, DenseNet, Vision Transformers (ViT), and KClipMed. To improve robustness and overall performance, we employed ensemble strategies that combined model outputs using a logical OR (union) operation. Our system was developed with careful attention to label encoding, image preprocessing, model diversity, and performance evaluation using micro and macro F1 scores.

Through extensive experimentation and fine-tuning, we found that combining models with different inductive biases and input resolutions considerably enhanced performance, especially in terms of

stability across validation subsets. KClipMed models contributed semantically-aware predictions, while ViTs offered competitive performance through patch-based representation learning.

For future work, we plan to extend our system by integrating interpretable AI techniques such as SHAP [28] to better understand model predictions and improve trustworthiness in clinical contexts. In addition, visual explanation methods like Grad-CAM [29] may help highlight image regions most responsible for specific predictions, increasing model transparency for clinicians. We are also exploring caption generation pipelines that incorporate longitudinal comparison and anatomy-wise controllability, inspired by recent advances in radiology reporting [30]. Furthermore, leveraging large-scale multilingual generative models [31] and vision-language alignment approaches [32] may enhance generalization to unseen medical concepts in zero-shot or few-shot scenarios.

Beyond the competition setting, we believe that automated concept detection systems like ours hold promising potential for practical use in medical education and research. For instance, they could assist radiology students in understanding complex imaging concepts through automatic annotation and feedback, or help researchers curate and label large-scale medical image datasets with minimal manual effort. In the long term, such systems could assist clinical decision-making workflows by surfacing relevant concepts during report generation or image review.

Ultimately, our goal is to build a concept detection system that is not only accurate but also transparent, scalable, and applicable across diverse medical imaging settings.

Code Availability

To support reproducibility and encourage further research, we have open-sourced our implementation, including all model definitions, training routines, evaluation scripts, and ensemble strategies. The codebase is available at: https://github.com/DeepLensIUST/ImageCLEF2025-Concept-Detection-DeepLens

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] H. Damm, T. M. G. Pakull, H. Becker, B. Bracke, B. Eryilmaz, L. Bloch, R. Brüngel, C. S. Schmidt, J. Rückert, O. Pelka, H. Schäfer, A. Idrissi-Yaghir, A. B. Abacha, A. G. S. de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2025 – Medical Concept Detection and Interpretable Caption Generation, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[2] B. Ionescu, H. Müller, R. Péteri, A. B. Abacha, V. Datla, S. A. Hasan, D. Demner-Fushman, S. Kozlovski, V. Liauchuk, Y. D. Cid, V. Kovalev, O. Pelka, C. M. Friedrich, A. García Seco de Herrera, V.-T. Ninh, T.-K. Le, L. Zhou, L. Piras, M. Riegler, P. Halvorsen, M.-T. Tran, M. Lux, C. Gurrin, D.-T. Dang-Nguyen, J. Chamberlain, A. Clark, A. Campello, D. Fichou, R. Berari, P. Brie, M. Dogariu, L. D. Ştefan, M. G. Constantin, Overview of the imageclef 2020: Multimedia retrieval in medical, lifelogging, nature, and internet applications, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2020, pp. 311–341.

[3] H. Müller, N. Michoux, D. Bandon, A. Geissbuhler, A review of content-based image retrieval systems in medical applications—clinical benefits and future directions, International Journal of Medical Informatics 73 (2004) 1–23. URL: https://www.sciencedirect.com/science/article/pii/S1386505603002119. doi:https://doi.org/10.1016/j.ijmedinf.2003.11.024.

[4] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of ImageCLEF 2025: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.

[5] S. Hu, Y. Gao, Z. Niu, Y. Jiang, L. Li, X. Xiao, M. Wang, E. F. Fang, W. Menpes-Smith, J. Xia, et al., Weakly supervised deep learning for covid-19 infection detection and classification from ct images, IEEE Access 8 (2020) 118869–118883. URL: https://doi.org/10.1109/ACCESS.2020.3005852. doi:10.1109/ACCESS.2020.3005852.

[6] D. Powers, Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness Correlation, Machine Learning: Science and Technology 2 (2008).

[7] J. Wang, H. Zhu, S.-H. Wang, Y.-D. Zhang, A review of deep learning on medical image analysis, Mobile Networks and Applications 26 (2021) 351–380. URL: https://doi.org/10.1007/s11036-020-01672-7. doi:10.1007/s11036-020-01672-7.

[8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. 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. URL: https://doi.org/10.1016/j.media.2017.07.005. doi:10.1016/j.media.2017.07.005.

[9] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely Connected Convolutional Networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269. doi:10.1109/CVPR35066.2017.

[10] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, in: NeurIPS Workshop on Machine Learning for Health, 2017.

[11] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 6105–6114. URL: http://proceedings.mlr.press/v97/tan19a.html.

[12] M. Shvetsova, D. Romanov, D. Dylov, Efficientnet-based convolutional neural network for automatic covid-19 detection from chest x-ray images, Frontiers in Medicine 8 (2021) 798055.

[13] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, C. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, A. Y. Ng, Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, Nature Scientific Data 6 (2019) 1–8.

[14] S. Gündel, S. Grbic, B. Georgescu, S. Liu, A. Maier, D. Comaniciu, Learning to recognize abnormalities in chest x-rays with location-aware dense networks, in: R. Vera-Rodriguez, J. Fierrez, A. Morales (Eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Springer International Publishing, Cham, 2019, pp. 757–765. URL: https://doi.org/10.1007/978-3-030-13469-3_88. doi:10.1007/978-3-030-13469-3_88.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16×16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR), 2021. URL: https://openreview.net/forum?id=YicbFdNTTy, presented at ICLR 2021.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9992–10002. doi:10.1109/ICCV48922.2021.00986.

[17] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, F. S. Khan, M. Shah, Transformers in medical imaging: A survey, Medical Image Analysis 88 (2023) 102802. URL: https://doi.org/10.1016/j.media.2023.102802. doi:10.1016/j.media.2023.102802.

[18] B. Boecking, N. Usuyama, S. Bannur, A. Sharafeldin, D. Castro, A. Schwaighofer, J. Valentin, O. Seneviratne, A. Nori, M. Kalra, et al., Making the most of text semantics to improve biomedical vision-language processing, in: Proceedings of the 2022 Conference on Health, Inference, and Learning, ACM, 2022, pp. 247–258. doi:10.1007/978-3-031-20059-5_1.

[19] G. Tsoumakas, I. Katakis, Multi-label classification: An overview, International Journal of Data Warehousing and Mining 3 (2007) 1–13. URL: https://doi.org/10.4018/jdwm.2007070101. doi:10.4018/jdwm.2007070101.

[20] D. H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259. URL: https://doi.org/10.1016/S0893-6080(05)80023-1. doi:10.1016/S0893-6080(05)80023-1.

[21] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. S. 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.

[22] H. Kauschke, K. Bogomasov, S. Conrad, Predicting Captions and Detecting Concepts for Medical Images: Contributions of the DBS-HHU Team to ImageCLEFmedical Caption 2024, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3740/paper-153.pdf.

[23] P. Kora, C. Ooi, O. Faust, R. U, A. Gudigar, W. Y. Chan, M. Kollati, K. Swaraja, P. Pławiak, U. Acharya, Transfer learning techniques for medical image analysis: A review, Biocybernetics and Biomedical Engineering 42 (2022) 79–107. doi:10.1016/j.bbe.2021.11.004.

[24] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14, MIT Press, Cambridge, MA, USA, 2014, p. 3320–3328. doi:10.5555/2969033.2969197.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, in: International Conference on Machine Learning, 2021. URL: https://proceedings.mlr.press/v139/radford21a/radford21a.pdf.

[26] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, D. Tao, A Survey on Vision Transformer, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2023) 87–110. doi:10.1109/TPAMI.2022.3152247.

[27] ImageCLEFmedical, ImageCLEFmedical Caption 2025 - Official Evaluation Script, 2025. URL: https://github.com/taubsity/clef-caption-evaluation.

[28] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.

[30] F. Dalla Serra, C. Wang, F. Deligianni, J. Dalton, A. O'Neil, Controllable chest X-ray report generation from longitudinal representations, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 4891–4904. URL: https://aclanthology.org/2023.findings-emnlp.325/. doi:10.18653/v1/2023.findings-emnlp.325.

[31] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru, S. Shleifer, P. S. Koura, V. Chaudhary, B. O'Horo, J. Wang, L. Zettlemoyer, Z. Kozareva, M. Diab, V. Stoyanov, X. Li, Few-shot learning with multilingual generative language models, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi,

United Arab Emirates, 2022, pp. 9019–9052. URL: https://aclanthology.org/2022.emnlp-main.616/. doi:10.18653/v1/2022.emnlp-main.616.

[32] T. Chen, Z.-G. Xu, X.-Y. Zheng, J. Wen, X.-Y. Liu, Z.-M. Li, H.-Y. Zhang, A survey on vision-language models for zero-shot image classification, Journal of Computer Science and Technology 38 (2023) 670–695. URL: https://doi.org/10.1007/s11390-023-4043-4. doi:10.1007/s11390-023-4043-4.