Uncertainty-Aware Biomedical Image Concept Detection via CNN Ensembles

Notebook for the Archimedes Unit at CLEF 2026

Spyridon Stavrou[3,*,†], Dimitris Boutzounis[3,*,†], Anna Chatzipapadopoulou[1,2,*,†], Foivos Charalampakos[1], Kalliopi V. Dalakleidi[1,2], Yannis Panagakis[2,3] and John Pavlopoulos[1,2]

[1]*Department of Informatics, Athens University of Economics and Business, 76, Patission Street, GR-104 34 Athens, Greece*
[2]*Archimedes Unit, Athena Research Center, 1, Artemidos Street, GR-151 25 Athens, Greece*
[3]*Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Panepistimiopolis, Zografou, GR-161 22 Athens, Greece*

Abstract

This article presents the methodology and results of the joint collaboration between the AUEB NLP Group and another group from NKUA representing the Archimedes Unit in the 10th edition of the ImageCLEFmedical Caption evaluation campaign. Our participation focused exclusively on the Concept Detection task, which involves the automatic association of biomedical images with relevant medical concepts. In this task, our system achieved 2nd place overall with a primary F_1 score of 0.5781 on the competition's final test set. Building upon our previous work, we experimented with the same family of models, combining image encoders based on Convolutional Neural Networks (CNNs) with Feed-Forward Neural Network (FFNN) classifiers. To improve robustness and generalization, we developed several ensemble strategies that combine predictions from multiple architectures. We further incorporated conformal prediction to provide more reliable and uncertainty-aware predictions. This allowed our system to better account for prediction confidence. Overall, our approach demonstrates that carefully designed CNN-based classifiers, ensemble methods, and conformal prediction can form a competitive pipeline for biomedical image concept detection.

Keywords

Natural Language Processing, Computer Vision, Biomedical Images, Convolutional Neural Networks, Multi-Label Classification, Conformal Prediction, Deep Learning

1. Introduction

ImageCLEF [1] is a long-running evaluation campaign organized in the context of the Conference and Labs of the Evaluation Forum (CLEF).[1] Since its launch in 2003, ImageCLEF has promoted the development and benchmarking of methods for annotating, indexing, retrieving, classifying, and generating multimodal data across a wide range of domains. Medical imaging has been a core part of the campaign since 2004, providing shared evaluation settings for clinically relevant visual and multimodal tasks.

A central component of the medical track is ImageCLEFmedical Caption [2], which focuses on automatic understanding of radiology images. The 2026 edition constitutes the 10th edition of the task and includes Concept Detection and Caption Prediction as standard subtasks, two newly introduced

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
spirosst16@gmail.com (S. Stavrou); dboutzounis2004@gmail.com (D. Boutzounis); ann.chatzipapadopoul@aueb.gr (A. Chatzipapadopoulou); phoebuschar@aueb.gr (F. Charalampakos); dalakleidi@aueb.gr (K. V. Dalakleidi); yannisp@di.uoa.gr (Y. Panagakis); annis@aueb.gr (J. Pavlopoulos)
www.linkedin.com/in/stavrou-spyridon (S. Stavrou); www.linkedin.com/in/dimitris-boutzounis (D. Boutzounis); https://www.linkedin.com/in/anna-chatzipapadopoulou/ (A. Chatzipapadopoulou); https://kvdalakleidi.wordpress.com/ (K. V. Dalakleidi); http://users.uoa.gr/~yannisp/ (Y. Panagakis); https://ipavlopoulos.github.io/ (J. Pavlopoulos)
0009-0008-0642-111X (S. Stavrou); 0009-0006-5778-9120 (D. Boutzounis); 0009-0005-1633-4966 (A. Chatzipapadopoulou); 0000-0001-6702-4398 (K. V. Dalakleidi); 0000-0003-0153-5210 (Y. Panagakis); 0000-0001-9188-7425 (J. Pavlopoulos)

[1]https://www.clef-initiative.eu/, Last accessed: 2026-05-19

synthetic-data subtasks, and an Explainability task. Although these subtasks are complementary, our participation this year focused exclusively on the standard **Concept Detection** task.

Concept Detection is formulated as a multi-label image classification problem. Given a radiology image from the ROCOv2 collection [3], the goal is to predict the set of relevant biomedical concepts, represented as Unified Medical Language System (UMLS)[2] Concept Unique Identifiers (CUIs). These concepts provide a structured semantic description of the image content, covering imaging modalities, anatomical structures, visual findings, and other clinically relevant entities. Beyond their direct use for image annotation, such concept sets can support downstream applications including medical image retrieval, report generation, and concept-aware captioning.

The task remains challenging due to the heterogeneity of radiology images, as well as because diseases can manifest with wide variability in size, texture, intensity, and appearance across imaging modalities and patient populations, while different diseases may also share highly similar visual characteristics. Furthermore, pathological regions are frequently small, diffuse, partially occluded, or embedded within normal anatomical variation, making them difficult to localize and distinguish from benign findings or imaging artifacts. The challenge is compounded by limited and noisy annotations, class imbalance due to the rarity of certain conditions, and the need for clinical context to interpret whether an observed feature truly indicates disease. Reliable concept prediction therefore requires models that are not only accurate, but also robust. In this work, we describe the joint collaboration between the AUEB NLP Group and another group from NKUA representing the Archimedes Unit in the ImageCLEFmedical Caption 2026 Concept Detection task. Building on our previous systems, we employ CNN-based image encoders combined with Feed-Forward Neural Network classifiers, and we investigate ensemble strategies, and conformal prediction to improve robustness and uncertainty-aware decision making. Our submitted system achieved 2nd place in the official Concept Detection ranking. All the code used for our experiments is available on Github.[3]

2. Data

The ImageCLEFmedical 2026 Concept Detection dataset comprises radiology images extracted from biomedical articles of the PubMed Central Open Access (PMC OA) subset[4], annotated with Unified Medical Language System (UMLS) [4] Concept Unique Identifiers (CUIs). It is based on an extended version of the ROCOv2 dataset [3], incorporating additional images and updated annotations.

The full dataset initially provided 97,364 training and 19,240 validation samples. This collection consists of radiology images, with each image mapped to a set of CUIs. Compared to last year's dataset, this represents an incremental increase of 21.6% for the training set and 11.4% for the validation set. Following standard machine learning practice, we restructured these partitions to maintain three distinct roles: a training set for model optimization, a validation set for calibration and early stopping, and a held-out development set for unbiased model selection and ablation studies. To this end, we isolated 7,005 images from the original training pool using a stratified re-splitting strategy that strictly preserves the multi-label class distribution across all partitions. Of the remaining pool, 90,359 images were allocated for training, while the official validation set of 19,240 samples was retained for calibration. From this point forward, our private held-out partition is referred to as the dev set.

2.1. Concept Detection

The Concept Detection task is framed as a multi-label classification problem requiring systems to predict 2,646 distinct biomedical concepts. The label vocabulary spans medical conditions, procedures, anatomical structures and imaging modalities such as computed tomography (CT), ultrasonography, magnetic resonance imaging (MRI), and positron emission tomography (PET/CT). Each concept is

[2]UMLS: https://www.nlm.nih.gov/research/umls/index.html, Last accessed: 2026-05-20
[3]https://github.com/dboutzounis/imageclef2026, Last accessed: 2026-05-26
[4]PMC Open Access: https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/, Last accessed: 2026-05-20

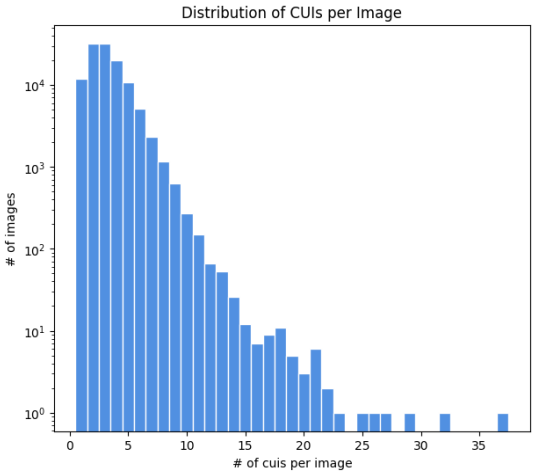
strictly represented by a standard UMLS CUI. A representative input image and its ground-truth concept annotations are shown in Figure 1.

CUI	UMLS Term
C1306645	Plain x-ray
C0817096	Chest
C0442872	Multiple cysts
Image ID	ImageCLEFmedical_Caption_2026_train_539
License	CC BY from Maepa et al.

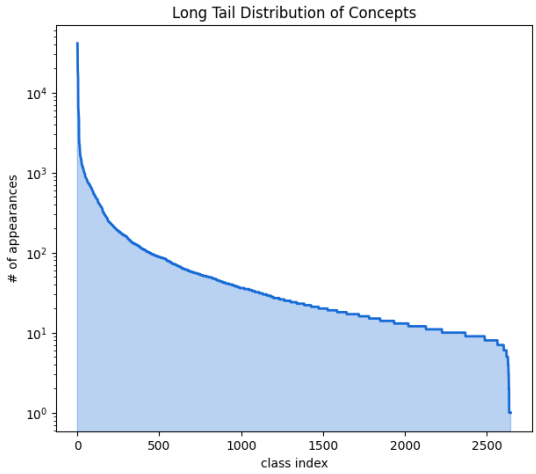
Figure 1: This figure, under CC BY from Maepa et al., presents an example from the ImageCLEFmedical 2026 dataset [3], illustrating the corresponding Concept Unique Identifiers (CUIs) and Unified Medical Language System (UMLS) terms.

The exploratory analysis of the annotations confirms the structural imbalance of the target label space, as illustrated in Figure 2. CUI frequencies follow a heavy long-tail distribution: the single most frequent concept appears in nearly 40,000 images, while by the 10th most common CUI occurrences drop to approximately 4,600. The highest-ranking concepts predominantly describe broad imaging modalities such as X-Ray Computed Tomography, Plain X-ray, MRI, and Ultrasonography, rather than specific pathologies (Table 1). Notably, this distribution is nearly identical to the previous year's, as the top 10 concepts remain completely unchanged, with the exception of a rank inversion between Abdomen (now 6th) and angiogram (now 7th). At the instance level, the majority of images are labeled with 2 to 5 CUIs, yielding an overall average of 3.22 concepts per image, consistent with last year's 3.20 average. While the minimum concept count of a single CUI occurs in 11,814 images (up from 10,018), a smaller fraction of highly complex cases extends up to 37 concepts per sample, a notable increase from the previous maximum of 28.

Data sparsity severely affects a large fraction of the concept set. As demonstrated in Table 2, several rare concepts occur only once within the entire dataset. These single-instance labels complicate the training process, as the system must learn to detect clinically relevant but highly isolated phenomena.



(a) Histogram showing the number of gold concepts per image.



(b) Visualization of the dataset's long-tail distribution.

Figure 2: Label distribution characteristics of the ImageCLEFmedical 2026 dataset [3].

Table 1
The ten most frequent concepts (CUIs) of the ImageCLEFmedical 2026 dataset [3], along with their corresponding UMLS terms, and the number of images they are associated with.

Most Common Concepts			
Rank	**CUI**	**UMLS Term**	**Images**
1	C0040405	X-Ray Computed Tomography	40,992
2	C1306645	Plain x-ray	31,818
3	C0024485	Magnetic Resonance Imaging	18,587
4	C0041618	Ultrasonography	17,143
5	C0817096	Chest	15,066
6	C0000726	Abdomen	6,537
7	C0002978	angiogram	6,055
8	C0037303	Bone structure of cranium	5,637
9	C0030797	Pelvis	5,342
10	C0023216	Lower Extremity	4,662

Table 2
Ten of the most rare concepts (CUIs) in the ImageCLEFmedical 2026 dataset [3], appearing only once, along with their corresponding UMLS terms.

Most Rare Concepts	
CUI	**UMLS Term**
C1962945	Radiographic imaging procedure
C1690005	MRI venography
C0243032	Magnetic Resonance Angiography
C1956110	Cone-Beam Computed Tomography
C0412650	Computed tomography of cervical spine
C0011906	Differential Diagnosis
C0202657	CT follow-up
C0203668	Radioisotope scan of bone
C0040395	tomography
C0598801	Diffusion weighted imaging

3. Methods

This section showcases the architecture, training pipelines, and ensembling strategies developed for the Concept Detection task.

3.1. Concept Detection

Following our established baselines from previous work [5, 6, 7, 8, 9, 10, 11, 12, 13], the core of our Concept Detection system relies on a diverse pool of neural image encoders. To promote model diversity, these backbones were trained with different seeds. Their resulting outputs were ensembled either through the Dual Threshold Aggregation strategy or through Soft Voting, both of which are described in later sections. In addition, a conformal regression layer was trained for application either before or after the ensembling stage.

3.1.1. CNN Backbone Architecture and Training

On every developed pipeline, a Convolutional Neural Network (CNN) backbone acts as the feature extractor, followed by a pooling mechanism and then paired with a classification head, which is a single linear layer with $|C|$ output neurons.

Prior to feature extraction, all input images undergo resizing and are subsequently normalized utilizing standard ImageNet statistics. The representative backbone architectures are optimized for multi-label concept detection by minimizing Binary Cross-Entropy with Logits loss function on our training set. Network parameters are updated via the AdamW [14] optimizer with an initial learning rate of 1e-4. The training phase is scheduled for a maximum of 30 epochs; however, an early stopping protocol with a patience of three epochs is strictly enforced to mitigate overfitting based on validation performance. Additionally, we employ a dynamic learning rate strategy via a Reduce-on-Plateau scheduler, which systematically reduces the learning rate (factor of 0.5) if the validation loss fails to improve after a single epoch.

For spatial feature aggregation, we replaced standard pooling operations with a learnable Generalized Mean (GeM) pooling layer [15]. The GeM layer computes the p-norm of the input feature maps, incorporating a clamping mechanism for numerical stability. Crucially, the exponent p is implemented as a fully trainable parameter. This parameterization allows the network to automatically learn the optimal pooling strategy for the target dataset by continuously interpolating between the behaviors of average pooling and max pooling, thereby enhancing the backbone's representational flexibility.

After spatial pooling and flattening, the extracted feature vector is processed by a sequential classifier comprising a dropout layer for regularization and a final linear projection mapped to the target number of classes. During inference, a sigmoid activation is applied to the network's output logits to generate class-specific probabilities. These probabilistic scores are subsequently binarized utilizing an optimized global threshold, which is calibrated based on the model's performance on the validation set.

3.1.2. Model Selection and Random Seeds

To build a diverse ensemble capable of capturing varied visual features, we selected a range of established CNN architectures implemented via the PyTorch framework. The model pool consists of **EfficientNet-B0** [16], **EfficientNet-V2-S** [17], **EfficientNet-V2-M** [17], **ConvNeXt-Tiny** [18], **DenseNet-121** [19], and **ResNet-50** [20]. All selected models were initialized using weights pre-trained on the ImageNet dataset, allowing the network to leverage rich, generalized feature representations before fine-tuning on our training set [21].

To introduce optimization diversity and reduce the risk of local minima, each specific architecture was trained independently using four distinct random seeds: 40, 41, 42, and 43 [22]. Varying these seeds ensures distinct stochastic initializations for the newly added classification layers across training epochs, ultimately enhancing the variance among the trained base models.

Following the training phase, we optimized the ensemble composition by executing a grid search to evaluate the performance of different model combinations. The outputs of the selected models were combined using the Dual Threshold Aggregation strategy. Through this grid search, we isolated the optimal subset of models and threshold parameters that maximized the samples-average F_1 score, which serves as the principal evaluation metric for the ImageCLEFmed Concept Detection task.

Rather than computing a single metric over the entire corpus at once, the samples-average approach evaluates the agreement between predictions and ground truth on a per-image basis. Let T denote the full collection of evaluated images. For any individual image $t \in T$, the model's predicted medical CUIs and the actual ground truth CUIs are structured as binary multi-hot vectors, denoted as p_t and g_t, respectively. A local score, $F_1^t(p_t, g_t)$, is first extracted specifically for this single sample:

$$F_1^t(p_t, g_t) = \frac{2|p_t \cap g_t|}{2|p_t \cap g_t| + |p_t \setminus g_t| + |g_t \setminus p_t|} \tag{1}$$

Where $|p_t \cap g_t|$ represents the correctly predicted concepts (True Positives), $|p_t \setminus g_t|$ represents the predicted concepts not in the ground truth (False Positives), and $|g_t \setminus p_t|$ represents the ground truth concepts that were not predicted (False Negatives). The global metric, F_1, is formulated by calculating the mean of these independent local F_1^t scores across all $|T|$ images in the dataset:

$$F_1 = \frac{1}{|T|} \sum_{t \in T} F_1^t(p_t, g_t) \tag{2}$$

Furthermore, the official evaluation framework distinguishes between two variations of this metric: a primary score that evaluates all the concepts, and a secondary score that strictly filters both p_t and g_t to a predefined subset of manually annotated concepts prior to calculating the aforementioned F_1^t score.

3.1.3. Concept Count Regressor and Conformal Prediction

Conformal prediction [23, 24] is a distribution-free framework for uncertainty quantification that can be applied on top of any predictive model. Instead of returning only a single point prediction, conformal prediction produces a prediction set or interval that is calibrated to contain the true target with a user-specified probability. A key advantage of this framework is that it does not require assumptions about the underlying data distribution; it only relies on the exchangeability of calibration and test samples.

In the inductive, or split conformal [25], setting, the available data are divided into a training set and a calibration set. The model is first trained on the training set, and the calibration set is then used to estimate the model's typical prediction error. Given a calibration set

$$\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{n}, \tag{3}$$

and a trained model f, we compute a nonconformity score for each calibration example. In the regression case, a common choice is the absolute residual:

$$R_i = |y_i - f(x_i)|, \qquad i = 1, \ldots, n. \tag{4}$$

The conformal uncertainty margin is then obtained as an empirical quantile of the calibration residuals:

$$q_\alpha = \text{Quantile}\left(\{R_i\}_{i=1}^{n}, \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right), \tag{5}$$

where α denotes the desired miscoverage level. For example, setting $\alpha = 0.1$ corresponds to a target coverage level of 90%.

For a new test example x_{test}, the split conformal prediction interval is therefore defined as

$$\Gamma_\alpha(x_{\text{test}}) = [f(x_{\text{test}}) - q_\alpha, f(x_{\text{test}}) + q_\alpha]\,. \tag{6}$$

Under the exchangeability assumption, this interval satisfies the finite-sample marginal coverage guarantee

$$\mathbb{P}\left(y_{\text{test}} \in \Gamma_\alpha(x_{\text{test}})\right) \geq 1 - \alpha. \tag{7}$$

In this work, conformal prediction is used as a calibration mechanism for the multi-label concept detection pipeline. Since each radiology image may be associated with a different number of biomedical concepts, the method provides a principled way to estimate a plausible range for the number of concepts that should be predicted. This range is then used to make the final predictions more robust by reducing extreme under-prediction and over-prediction cases.

To ensure our system predicts a realistic number of medical concepts per image, we introduced an auxiliary Concept Count Regressor paired with a conformal prediction framework [24].

This regressor relies on frozen features from a pre-trained backbone (e.g., frozen EfficientNet-V2-M with ImageNet weights) to preserve its learned feature extraction abilities. These features are pooled, flattened, and passed through a simple linear layer to predict the total number of concepts in the image using Mean Squared Error (MSE) loss. The regressor was trained exclusively on our training set, and we use early stopping during training to prevent the model from overfitting.

To provide statistical guarantees for our predictions, we apply **Split Conformal Regression** [26]. During the calibration phase, we first calculate the absolute residuals (errors) between the true concept count y_i and the predicted count $\hat{y}_i$ across our validation set:

$$R_i = |y_i - \hat{y}_i| \tag{8}$$

Next, we find the error quantile q required to achieve a desired confidence level $1 - \alpha$ (e.g., 90% confidence when $\alpha = 0.1$):

$$q = \text{Quantile}\left(R, \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right) \tag{9}$$

where n is the total number of calibration samples. As a result, for any new test image, we do not just get a single point estimate $\hat{y}$; we receive an interval: $[\hat{y} - q, \hat{y} + q]$.

Finally, we use these bounds to automatically correct our main model's predictions. We call this post-processing step **Bidirectional Conformal Rescue**:

- **Conformal Rescue (Lower Bound)**: If the base model predicts too few concepts (below the lower limit $\hat{y} - q$), we dynamically add the missing concepts by selecting the unpredicted labels with the highest probability logits until we reach the minimum bound.
- **Conformal Pruning (Upper Bound)**: If the base model predicts too many concepts (above the upper limit $\hat{y} + q$), we prune the surplus by dropping the predicted concepts with the lowest confidence scores until the total respects the maximum bound.

By integrating these statistical limits with the raw probability scores of our base models, we aim to constrain the cardinality of our predictions, helping to mitigate severe over- or under-prediction for any given image.

3.1.4. Ensemble Strategies

To improve prediction stability, our final framework employs a hierarchical ensemble strategy, integrating a diverse pool of standard backbones with a cross-validated sub-ensemble.

The primary voting pool consists of 18 independently trained models. This pool leverages the architectural variants selected during our intermediate grid search, cultivated across various random initialization seeds. During inference, each of these 18 models generates independent concept probabilities, which are binarized into distinct discrete predictions using their respective optimized global thresholds.

To further stabilize the predictions and provide a highly confident pipeline, we developed a secondary sub-ensemble using **Monte-Carlo Cross-Validation** (MCCV) [27]. Utilizing a Multilabel Stratified

Shuffle Split [28] to preserve the complex label distributions, we partitioned our training and validation data into five unique 80/20 splits. We trained a dedicated EfficientNet-V2-S backbone on each split, where each model's global threshold was specifically tuned using the 20% validation holdout from its respective split.

During aggregation, these five MCCV models operate under an intersection consensus rule: a concept is only output by this sub-ensemble if it is predicted by all five constituent models simultaneously. This unanimous voting requirement acts as a high-precision, conservative contributor to the final ensemble.

The modular nature of the Bidirectional Conformal Rescue mechanism allows for its integration at two distinct stages of the inference pipeline: either independently per base model or globally following the ensemble strategy. When implemented on a per-model basis, the statistically guaranteed lower and upper bounds are utilized to dynamically rescue or prune the discrete predictions of each autonomous network before their individual outputs are aggregated. Alternatively, when applied post-ensemble, the models first generate a combined set of predictions via the chosen voting or aggregation strategy, and this final consensus is subsequently constrained by the conformal limits.

To effectively combine the predictive representations learned by our diverse model pool, we investigated two distinct aggregation strategies during the ensemble phase:

- **Soft-Voting (Average Probabilities)**: This strategy preserves the continuous confidence signals of the constituent models by deferring binarization until the final stage. The raw, post-sigmoid probability distributions generated by each model are extracted and averaged to produce a unified mean probability score for every concept. Formally, given an ensemble of M models, the final discrete binary prediction $\hat{y}_c$ for a specific concept c is determined by:

$$\hat{y}_c = \mathbb{1}\left(\frac{1}{M}\sum_{m=1}^{M} p_{m,c} \geq \tau\right) \tag{10}$$

  where $p_{m,c}$ represents the probability predicted by model m for concept c, $\mathbb{1}(\cdot)$ is the indicator function, and τ represents the global decision threshold. Rather than relying on an arbitrary static value, τ is calibrated by maximizing the F_1 score on the validation set prior to inference. This approach allows the continuous confidence signals of all models to contribute jointly to the final decision, rather than being discarded by strict per-model binarization.
- **Dual Threshold Aggregation** [11]: Initially, each independent model applies its own uniquely optimized decision threshold to its continuous probability outputs, yielding a discrete binary vote (presence or absence) for every target concept. These individual binary votes are then summed across the entire ensemble pool to compute an accumulated vote count, V_c, for each specific concept. To formulate the preliminary consensus, a secondary ensemble-level voting threshold, L, is introduced. A concept is retained as a positive prediction only if $V_c \geq L$. The optimal value for L is empirically determined by sweeping through all possible vote requirements to maximize the evaluation metric.

4. Experiments, Submissions and Results

This section presents a comprehensive overview of our experimental results for this year's evaluation campaign [1] on the **Concept Detection** task. We comprehensively report on our official submissions, detailing the predictive performance achieved across our held-out dev set and the official unlabelled test set.

4.1. Concept Detection Submissions

Based on the results from our internal dev set, we selected our top five model setups for official submission. Each submission is built upon the 18 base models selected via grid search over the six CNN architectures trained across four random seeds, as described in Section 3.1.2. The standalone

performance and optimized global thresholds for all individual variants are detailed in Table 11. These models were combined using either the Dual-Threshold Aggregation or Soft-Voting strategies (Section 3.1.4). Some submissions further expanded this pool by incorporating the 5-split MCCV sub-ensemble of EfficientNet-V2-S models. Finally, the Bidirectional Conformal Rescue method (Section 3.1.3) was integrated flexibly, applied either to each individual model prior to ensembling or directly to the final aggregated output.

Consistent with the evaluation protocol of the task, our system's performance is quantified using the samples-average F_1 score described in equation 2, supplemented by a secondary F_1 score restricted to manually selected concepts like modality and anatomy.

Across both our dev set and the official test set, the ensemble strategies yielded a modest improvement over the standalone baseline architectures. As detailed in Table 11, the best individual models (e.g., ConvNeXt-Tiny) achieved a dev F_1 score of approximately 0.59. Our top ensemble configurations (Table 3) elevated this performance to approximately 0.60, representing a gain of roughly 1%. While incorporating both the MCCV models and the Bidirectional Conformal Rescue step yielded the highest test score (Run 983, Test F_1 = 0.5781), the performance gaps between the top ensemble variations are minimal, limited primarily to the fourth decimal place.

Table 3
Summary of our submissions to the ImageCLEFmedical 2026 Concept Detection task [2]. The table presents the scores of our systems on both our held-out dev set and the official test set. It also includes the rankings of these systems among all submissions from the 11 participating teams. Values in parentheses indicate the random seeds selected via grid search for each model. **MC**: Monte-Carlo, **EB0**: EfficientNet-B0 (seeds: 41, 42), **EV2S**: EfficientNet-V2-S (seeds: 40, 41, 42), **EV2M**: EfficientNet-V2-M (seeds: 40, 41, 42, 43), **D121**: DenseNet-121 (seeds: 40, 41, 42), **CN**: ConvNext-Tiny (seeds: 40, 41, 42, 43), **R50**: ResNet-50 (seeds: 41, 42), **BCR**: Bidirectional Conformal Rescue, **Dual-L**: number of L models used for thresholding in the ensemble, **SV**: soft-voting, averaging raw probability scores.

Individual Concept Detection Experiments					
Run ID	**Method**	**F1**		**Secondary F1**	**Rank**
		Dev	**Test**		
983	Dual-5(BCR(MC(EV2S)), BCR(EB0), BCR(EV2S), BCR(EV2M), BCR(D121), BCR(CN), BCR(R50))	0.6007	**0.5781**	0.9591	2
818	Dual-5(BCR(EB0), BCR(EV2S), BCR(EV2M), BCR(D121), BCR(CN), BCR(R50))	0.6007	0.5780	0.9590	4
984	Dual-6(BCR(MC(EV2S)), BCR(EB0), BCR(EV2S), BCR(EV2M), BCR(D121), BCR(CN), BCR(R50))	0.6005	0.5776	**0.9614**	5
931	Dual-5(MC(EV2S), EB0, EV2S, EV2M, D121, CN, R50)	0.6007	0.5771	0.9574	7
987	BCR(SV(MC(EV2S), EB0, EV2S, EV2M, D121, CN, R50))	**0.6008**	0.5764	0.9601	9

It is worth noting that while performance variations at the fourth decimal place may appear negligible from a strictly clinical or statistical perspective, such tiny optimizations are often the deciding factor in competitive benchmarking environments. This dynamic is clearly reflected in the final competition standings, where our team secured 2nd place overall. Our top-performing system achieved a primary F_1 score of **0.5781** and a secondary score of **0.9591**. These results highlight the highly competitive nature of the task, as our proposed framework trailed the first-place submission by a marginal difference of only 0.0009 on the primary evaluation metric (0.5790) and 0.0066 on the secondary metric (0.9657).

4.2. Impact of Bidirectional Conformal Rescue

To understand exactly how our cardinality constraints influence the final predictions, we conducted an ablation study on the Dual-Threshold ensemble strategy. As shown in Table 4, we compared the baseline ensemble (without any conformal adjustments) against our two proposed conformal strategies: applying the Bidirectional Conformal Rescue (BCR) to each model individually prior to voting, and applying it directly to the final aggregated output. We tracked the F_1 score on our dev set across varying voting thresholds from L=1 to L=9.

The data reveals that the baseline ensemble naturally reaches its peak performance at L=5, achieving an F_1 score of 0.6007. At this optimal threshold, the unadjusted models inherently predict a realistic number of concepts, meaning the conformal bounds rarely need to trigger.

However, the true value of the BCR mechanism becomes clear at the edges of the voting spectrum. When the threshold is very lenient (L=1, causing over-prediction) or too strict (L=8 and L=9, causing under-prediction), the baseline performance drops noticeably. In these scenarios, both BCR strategies act as a safety net, dynamically pruning excess concepts or rescuing missing ones to maintain a higher, more stable F_1 score.

Comparing the two placement strategies reveals distinct behaviors depending on the strictness of the ensemble:

- Mid-Range Thresholds (L=4 to L=7): Applying the conformal rescue to each model individually (BCR per Member) provides a slight edge. It ensures that only well-calibrated votes enter the final consensus pool, optimizing the mid-level agreements.
- Strict Thresholds (L=8 and L=9): As the voting rule becomes stricter, applying the rescue step after the ensemble vote (BCR post Ensemble) becomes the more effective strategy. Strict voting naturally discards many valid concepts because they fail to reach the high consensus requirement. Applying BCR as a post-processing step allows the system to intelligently rescue these discarded concepts using the combined confidence of the entire model pool.

Table 4
Ablation study of Bidirectional Conformal Rescue (BCR) strategies. The table reports the F_1 scores for an 18-model ensemble across varying Dual Threshold values (L) on our held-out dev set. We compare the baseline ensemble without conformal prediction (**No BCR**), applying BCR to each model individually prior to ensembling (**BCR per Member**), and applying BCR as a post-processing step on the final ensemble output (**BCR post Ensemble**). Bold values indicate the best performing strategy for each L threshold.

Ablation study of Bidirectional Conformal Rescue (BCR) strategies			
L	**No BCR**	**BCR per Member**	**BCR post Ensemble**
1	0.5626	0.5622	**0.5635**
2	0.5890	0.5889	**0.5890**
3	0.5969	0.5969	**0.5969**
4	0.5991	**0.5992**	0.5991
5	0.6007	**0.6007**	0.6007
6	0.6001	**0.6006**	0.6005
7	0.5991	**0.6002**	0.6000
8	0.5981	0.5990	**0.5993**
9	0.5968	0.5977	**0.5982**

**Where multiple strategies share identical values at four decimal places, bold formatting reflects differences at the fifth decimal place.*

This study confirms that integrating statistical cardinality limits successfully prevents the severe performance drops typically associated with rigid ensemble voting rules.

Furthermore, to verify that our calibration process successfully translates to unseen data, we evaluated the empirical coverage of the generated conformal intervals on the dev set, as detailed in Table 5. While the regressor was calibrated to guarantee a 90% target coverage level ($\alpha = 0.1$), the observed empirical

coverage safely exceeded this threshold. An analysis of the coverage errors revealed a strictly one-sided pattern: the system only failed when it underestimated the true complexity of an image. In every out-of-bound case, the Ground Truth (GT) concept count exceeded the predicted Upper Bound (UB). Notably, the true count never fell below the predicted Lower Bound (LB), meaning there were zero samples of extreme over-estimation.

Table 5
Observed Coverage Evaluation of Conformal Intervals. The table details the performance of our dynamically generated cardinality bounds evaluated on the held-out dev set that contains 7,005 images. The system safely exceeded the 90% target calibration coverage, with out-of-bound errors restricted entirely to target under-estimation. For clarity within the metric descriptions, GT denotes the Ground Truth concept count, while UB and LB represent the predicted Upper Bound and Lower Bound, respectively.

Conformal Prediction case	Count	Coverage
Correctly estimated (GT $\in$ [LB,UB])	6,756	96.45%
Under-estimated (GT $>$ UB)	249	3.55%
Over-estimated (GT $<$ LB)	0	0.00%

4.3. Evaluation Excluding High-Frequency Concepts

While the official F_1 score defined in Equation 2, evaluates overall performance, it is highly susceptible to inflation from the dataset's long-tail distribution (see Figure 2). As discussed previously, the ground truth is dominated by a few broad imaging modalities, which can boost the primary metric and mask a model's underlying performance on rarer, clinically specific pathologies. To address this baseline bias, we introduce the **Long-Tail** F_1 score. This custom metric applies a targeted filter prior to computing the local sample score $F_1^t(p_t, g_t)$. Specifically, it explicitly removes the four most frequent concepts (shown in Table 1)[5] from both the prediction vector p_t and the ground truth vector g_t. By isolating the long-tail distribution, this metric provides a more targeted assessment of the model's diagnostic utility on less common clinical findings.

Table 6
Summary of our submissions evaluated using the Long-Tail F_1 score. The table presents the Primary F_1 scores of our systems on our held-out dev set alongside the newly introduced Long-Tail F_1 scores. Values in parentheses indicate the random seeds selected via grid search for each model. **MC**: Monte-Carlo, **EB0**: EfficientNet-B0 (seeds: 41, 42), **EV2S**: EfficientNet-V2-S (seeds: 40, 41, 42), **EV2M**: EfficientNet-V2-M (seeds: 40, 41, 42, 43), **D121**: DenseNet-121 (seeds: 40, 41, 42), **CN**: ConvNext-Tiny (seeds: 40, 41, 42, 43), **R50**: ResNet-50 (seeds: 41, 42), **BCR**: Bidirectional Conformal Rescue, **Dual-L**: number of L models used for thresholding in the ensemble, **SV**: soft-voting, averaging raw probability scores.

Evaluation of Long-Tail Concept Detection		
Method	**Primary-F1**	**Long-Tail F1**
Dual-5(BCR(MC(EV2S)), BCR(EB0), BCR(EV2S), BCR(EV2M), BCR(D121), BCR(CN), BCR(R50))	0.6007	**0.2462**
Dual-5(BCR(EB0), BCR(EV2S), BCR(EV2M), BCR(D121), BCR(CN), BCR(R50))	0.6007	0.2460
Dual-6(BCR(MC(EV2S)), BCR(EB0), BCR(EV2S), BCR(EV2M), BCR(D121), BCR(CN), BCR(R50))	0.6005	0.2411
Dual-5(MC(EV2S), EB0, EV2S, EV2M, D121, CN, R50)	0.6007	0.2461
BCR(SV(MC(EV2S), EB0, EV2S, EV2M, D121, CN, R50))	**0.6008**	0.2411

[5]The number of frequent terms excluded depends on the tail and could be customized.

The results presented in Table 6 demonstrate the masking effect of dominant medical concepts on the Primary F_1 evaluation metric. Across all ensemble configurations, performance drops from roughly 0.60 on the Primary F_1 to approximately 0.24 on the Long-Tail F_1. This reduction illustrates how standard evaluation can be heavily influenced by the top-4 imaging modalities, potentially obscuring a model's underlying capabilities on rarer, clinically specific pathologies. Rather than replacing standard metrics, the Long-Tail F_1 score serves as a valuable complementary tool for assessing model robustness on less frequent classes. For instance, evaluating solely on the Primary F_1 would favor the Soft Voting ensemble (Run 987 in Table 3), which achieved the highest Primary score (0.6008) but performed the poorest on both the Long-Tail F_1 (0.2411) and the official test set (0.5764). In contrast, our Dual-5 thresholding strategy (Run 983 in Table 3) balanced these trade-offs more effectively, as it achieved the highest Long-Tail F_1 (0.2462) while also securing our best performance on the unseen test set (0.5781). Ultimately, incorporating long-tail evaluation provides a more nuanced understanding of a model's ability to handle a diverse clinical spectrum without overfitting to the most common concepts.

4.4. Error Analysis

To better understand the predictive behavior and limitations of our system, we conducted an error analysis using the dev set predictions from our highest-performing submitted model (Run ID 983 in Table 3). This subsection isolates class-level performance to identify both the primary strengths and the systematic vulnerabilities of the ensemble. We first examine the top-performing concepts alongside those that resulted in complete predictive failure. Following this, we investigate "hard-to-learn" classes/concepts that yielded persistently low accuracy despite high representation in the training data and utilize a bigram co-occurrence analysis to diagnose the statistical and semantic factors driving these specific performance deficits.

Table 7 details the top 15 medical concepts with the highest F_1 scores, based on the dev set predictions from our highest-performing submission. The results demonstrate that the ensemble performs extremely well on fundamental imaging modalities and major anatomical structures, consistent with their high representation in the training data. Interestingly, a small number of concepts with limited training support, such as Breast and Striated Muscle, also achieved above-average scores, suggesting that visually distinctive concepts may be learnable even from fewer examples, though this observation is based on a small number of classes.

In contrast to the clearly defined imaging types and specific body parts discussed earlier, Table 8 identifies the concepts that remained most challenging for our best-performing submitted model. Specifically, it lists the 15 concepts for which the model did not produce a correct prediction on the development set. Notably, this poor performance is not caused by a lack of training data. Concepts like "Neoplasms" (abnormal tissue growths) and "Brain" have a large number of training examples (1,944 and 1,394, respectively). Instead, the model struggles with terms that are too broad or visually inconsistent. The failures mainly involve wide disease categories (e.g., "Neoplasms", "Edema"), general body areas (e.g., "soft tissue", "Entire bony skeleton"), and vague descriptive words (e.g., "Enlarged", "Irregular"). Furthermore, the error breakdown for these classes shows a massive number of missed detections (False Negatives) and almost no false alarms (False Positives).

To further investigate the disconnect between data volume and predictive success, Table 9 isolates "hard-to-learn" concepts. These are classes that possess substantial training data (over 1,000 instances) yet still yield exceptionally low accuracy (an F_1 score below 0.30). This table reinforces the observation that simply providing more images does not automatically solve visual ambiguity. The list includes major organs (e.g., "Heart", "Liver"), distinct blood vessels (e.g., "Aorta"), and general physical characteristics (e.g., "Fluid behavior", "Cystic", "Nodule"). The poor performance across these highly supported classes suggests a problem with context and visual consistency. For instance, a major organ like the liver or heart can appear drastically different depending on the imaging technique used (such as an X-ray versus an MRI). Similarly, physical traits like "Fluid behavior" or a "Nodule" (a small lump) can occur anywhere in the body, lacking a fixed background or consistent shape. Consequently, the model struggles to learn a single, universal visual rule for these concepts, regardless of how many examples it processes.

Table 7
Best Performing Classes (Dev Support $\geq$ 5). This table presents the top 15 medical concepts ranked by their F_1 score, evaluated on our held-out dev set using the predictions from our best-performing submitted model (Run ID 983 in Table 3). For each concept, the table details the Concept Unique Identifier (CUI) and Concept Name, the number of instances present in the training set (Train Support) and the dev set (Dev Support), as well as the resulting F_1 score, False Positives (FP), and False Negatives (FN).

Top 15 Performing Medical Concepts						
CUI	**Concept Name**	**Train Support**	**Dev Support**	**F1**	**FP**	**FN**
C0041618	Ultrasonography	14,233	1,023	0.9966	6	1
C0040405	X-Ray Computed Tomography	34,055	2,448	0.9745	103	24
C1306645	Plain x-ray	26,531	1,907	0.9718	98	12
C0037303	Bone structure of cranium	4,715	339	0.9491	12	22
C0024485	Magnetic Resonance Imaging	15,475	1,113	0.9480	71	46
C0920367	Tomography, Optical Coherence	367	26	0.9057	3	2
C0002978	angiogram	5,387	387	0.8980	72	13
C0023216	Lower Extremity	3,911	281	0.8852	54	15
C1140618	Upper Extremity	1,848	133	0.8681	30	8
C0817096	Chest	12,559	903	0.8175	98	211
C0006141	Breast	172	12	0.8000	6	0
C0037949	Vertebral column	1,934	139	0.7692	45	24
C0032743	Positron-Emission Tomography	802	58	0.7246	30	8
C0226032	Anterior descending branch...*	705	51	0.6667	39	6
C1331262	Muscle, Striated	63	8	0.6154	1	4

**Anterior descending branch of left coronary artery*

Table 8
Worst Performing Classes (Dev Support $\geq$ 5). This table presents the 15 medical concepts with the lowest F_1 scores, evaluated on our held-out dev set using the predictions from our best-performing submitted model (Run ID 983 in Table 3). For each concept, the table details the Concept Unique Identifier (CUI) and Concept Name, the number of instances present in the training set (Train Support) and the dev set (Dev Support), as well as the resulting F_1 score, False Positives (FP), and False Negatives (FN).

Top 15 Worst Performing Medical Concepts						
CUI	**Concept Name**	**Train Support**	**Dev Support**	**F1**	**FP**	**FN**
C0027651	Neoplasms	1,944	140	0.0000	0	140
C0006104	Brain	1,394	100	0.0000	0	100
C0012359	Pathological Dilatation	1,297	93	0.0000	0	93
C1266909	Entire bony skeleton	1,282	92	0.0000	0	92
C0225317	soft tissue	1,241	89	0.0000	0	89
C0013604	Edema	1,129	81	0.0000	3	81
C0006663	Calcinosis	1,089	78	0.0000	0	78
C1261287	Stenosis	1,037	75	0.0000	3	75
C1510420	Cavitation	981	71	0.0000	0	71
C0025066	Mediastinum	980	70	0.0000	0	70
C0018827	Heart Ventricle	936	67	0.0000	3	67
C0087086	Thrombus	931	67	0.0000	0	67
C0442800	Enlarged	922	66	0.0000	0	66
C0205271	Irregular	901	65	0.0000	0	65
C0016169	pathologic fistula	829	60	0.0000	0	60

Table 9
Analysis of "Hard-to-Learn" Classes. This table presents medical concepts that achieved an F_1 score below 0.30 despite having substantial representation in the training dataset (Train Support > 1000). Evaluated on our held-out dev set using the predictions from our best-performing submitted model (Run ID 983 in Table 3), the table details for each concept the Concept Unique Identifier (CUI) and Concept Name, the number of instances present in the training set (Train Support) and the dev set (Dev Support), and the resulting F_1 score.

Top "Hard-to-Learn" Medical Concepts				
CUI	**Concept Name**	**Train Support**	**Dev Support**	**F1**
C0444611	Fluid behavior	2,224	160	0.0345
C0027651	Neoplasms	1,944	140	0.0000
C0205207	Cystic	1,761	127	0.0455
C0018787	Heart	1,577	113	0.1138
C0023884	Liver	1,467	105	0.1587
C0006104	Brain	1,394	100	0.0000
C0032227	Pleural effusion disorder	1,306	94	0.2778
C0012359	Pathological Dilatation	1,297	93	0.0000
C1266909	Entire bony skeleton	1,282	92	0.0000
C0028259	Nodule	1,262	91	0.0404
C0225317	soft tissue	1,241	89	0.0000
C0034052	Pulmonary artery structure	1,171	84	0.2321
C0013604	Edema	1,129	81	0.0000
C0006663	Calcinosis	1,089	78	0.0000
C0003483	Aorta	1,085	78	0.0482
C1261287	Stenosis	1,037	75	0.0000

To uncover exactly why these well-represented classes fail, Table 10 examines the context in which they appear. Specifically, it lists the top three other labels that most frequently accompany each "hard-to-learn" concept in the training data. First, it is important to note that many of these medical findings are inherently difficult to detect. For instance, terms like "Neoplasms" (abnormal growths), "Cystic," and "Fluid behavior" do not have a fixed size, shape, or clear border. They can appear anywhere in the body and often blend into surrounding healthy tissues, making them visually subtle even in a perfect scan.

The table reveals that this natural difficulty is severely worsened by how these images are captured. These challenging terms are heavily entangled with entirely different scanning methods, such as CT scans, MRIs, and Ultrasounds. Because a CT scan looks completely different from an MRI, the physical appearance of a "Neoplasm" changes drastically depending on how the image was taken. As a result, the model is forced to learn a concept that is already visually vague, while its background and texture constantly shift. This combination of ambiguous shapes and changing imaging methods prevents the system from locking onto a reliable pattern, explaining the near-zero accuracy despite thousands of training examples.

Table 10
Co-occurrence Analysis for "Hard-to-Learn" Classes. To diagnose the failure modes of the lowest-performing, highly-supported medical concepts, we analyzed their top bigram co-occurrences within the training set. Evaluated on our held-out dev set using the predictions from our best-performing submitted model (Run ID 983 in Table 3), the table presents the Target Concept and its F_1 score, alongside a comma-separated list of the top three most frequently co-occurring concepts (with their absolute occurrences in parentheses). *Abbreviations: CT = X-Ray Computed Tomography, MRI = Magnetic Resonance Imaging, US = Ultrasonography, Cranium = Bone structure of cranium, Path Dilatation = Pathological Dilatation.*

Top Co-occurring Concepts for "Hard-to-Learn" Classes		
Concept (CUI)	**F1**	**Top Co-occurring Concepts**
Fluid behavior (C0444611)	0.0345	CT (1015), MRI (579), US (330)
Neoplasms (C0027651)	0.0000	CT (933), MRI (572), US (183)
Cystic (C0205207)	0.0455	CT (767), US (475), MRI (413)
Heart (C0018787)	0.1138	Chest (496), Plain X-ray (454), US (330)
Liver (C0023884)	0.1587	CT (827), US (314), MRI (174)
Brain (C0006104)	0.0000	MRI (944), CT (404), Fluid behavior (95)
Pleural effusion (C0032227)	0.2778	Chest (830), Plain X-ray (652), CT (556)
Path Dilatation (C0012359)	0.0000	CT (529), Plain X-ray (297), Abdomen (289)
Entire bony skeleton (C1266909)	0.0000	Plain X-ray (557), CT (465), Cranium (218)
Nodule (C0028259)	0.0404	CT (741), US (272), Chest (234)
soft tissue (C0225317)	0.0000	CT (678), Plain X-ray (258), MRI (246)
Pulmonary artery (C0034052)	0.2321	CT (545), Angiogram (338), Chest (184)
Edema (C0013604)	0.0000	MRI (585), CT (390), soft tissue (87)
Aorta (C0003483)	0.0482	CT (426), US (387), Chest (143)
Calcinosis (C0006663)	0.0000	CT (594), Plain X-ray (274), US (139)
Stenosis (C1261287)	0.0000	Angiogram (440), CT (205), Plain X-ray (195)

4.5. Additional Concept Detection Experiments

4.5.1. Mondrian Conformal Regressor

Standard conformal prediction applies a single, global error margin across all images. However, the variance in medical concept counts can be highly heterogeneous. Specifically, dense, complex images inherently exhibit larger absolute prediction errors compared to simple, sparse images. Mondrian Conformal Prediction addresses this by dynamically scaling the statistical margin of error based on the predicted complexity of the specific image.

During calibration on the validation set, we partitioned the continuous count predictions into distinct operational categories, or bins, utilizing data-dependent thresholds. These bin boundaries were empirically defined by the 25th, 50th, and 75th percentiles of the validation prediction distribution.

For each resulting bin b, an independent conformal quantile q_b was calibrated on our validation set, to satisfy the targeted 90% confidence level ($\alpha = 0.1$). During inference, the system first predicts the raw concept count $\hat{y}$ and identifies its corresponding bin. It then applies the localized error margin rather than the global average, yielding a dynamically sized prediction interval: $[\hat{y} - q_b, \hat{y} + q_b]$. To ensure statistical robustness, the system incorporates a global fallback quantile, which acts as a fail-safe for any bins containing an insufficient number of calibration samples (e.g., $N < 10$).

Despite the theoretical advantages of localized, dynamic bounding, empirical evaluation (see tables 4, 12 for comparison) revealed that the Mondrian Conformal strategy yielded predictive performance identical to our standard, globally calibrated conformal approach. While the dynamic margins functioned as intended, the Mondrian binning process introduces additional structural complexity and inherently overfits the calibration to the specific data distribution of our validation set. Given that this added complexity did not translate into a tangible benefit in the final F_1 score, we opted not to include this configuration among our official competition submissions, favoring the simpler and equally effective global conformal bounds.

4.5.2. Meta Learner

For any given input image, we extracted the continuous probability vectors from all M constituent base models. These vectors were concatenated and flattened to form a high-dimensional feature representation of size $M \times C$, where C represents the total number of target concept classes. This vector served as the input to a trainable meta-learner, that consists of a single, fully connected linear layer. The meta-learner was tasked with mapping these combined features directly to a final, refined set of C concept logits, theoretically allowing the network to learn complex inter-model correlations and dynamically weight the most reliable architectures for specific concepts.

To prevent data leakage, our fine-tuned on training set models were frozen, and the meta-learner was trained exclusively using the extracted prediction features from our validation set. The layer was optimized over a brief training cycle utilizing a Binary Cross-Entropy with Logits loss function and an AdamW optimizer with weight decay. Final binary predictions were generated by applying a standard 0.5 threshold to the sigmoid-activated outputs of the meta-learner.

Despite the theoretical capacity of the meta-learner to adaptively correct individual base model biases, empirical evaluation on the dev set revealed suboptimal performance. The parameterized layer rapidly overfit to the specific prediction distributions and correlations present within the validation set. As a result, the model failed to generalize effectively to unseen data, yielding an F_1 score of 0.5888 noticeably lower than our unparameterized Soft-Voting and Dual-Threshold strategies.

4.6. Hardware Configuration

To accelerate the training phase, each model was optimized utilizing a single NVIDIA L4 GPU equipped with 24 GB of VRAM, supported by 54 GB of system RAM to ensure efficient data loading and batch processing.

5. Conclusions

Our participation in the ImageCLEFmedical Caption task provided an opportunity to explore innovative approaches in computer vision for biomedical image concept detection. Our approach demonstrated competitive performance in the Concept Detection task.

In the Concept Detection task, we achieved the 2$^{\text{nd}}$ place out of 11 participating groups. Building on our earlier work for the automatic association of biomedical images with relevant medical concepts, we explored the same group of models by combining image encoders that utilize Convolutional Neural Networks (CNNs) with Feed-Forward Neural Network (FFNN) classifiers. To enhance performance and generalization, we created various ensemble strategies that aggregate predictions from multiple architectures. We note that we disregard architectural complexity, opting solely for accuracy that is the objective of the challenge. In addition, we integrated conformal prediction to ensure more reliable and uncertainty-aware concept predictions. This enhancement enabled our system to incorporate explicit confidence intervals, providing a more statistically grounded basis for the generated concept sets. In summary, our approach illustrates that well-designed CNN-based classifiers, ensemble techniques and conformal prediction can establish a competitive pipeline for biomedical image concept detection.

In our future research, we aim to create a unified multitask model that can simultaneously tackle both Concept Detection and Caption Prediction tasks. This integrated framework will leverage reinforcement learning signals from both tagging and captioning metrics, allowing for more cohesive and interrelated outputs. Ultimately, we see these systems as foundational steps toward developing clinically valuable and robust multimodal AI tools in radiology and other fields.

Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. Thanks to the

developers of ACM consolidated LaTeX styles https://github.com/borisveytsman/acmart and to the developers of Elsevier updated LaTeX templates https://www.ctan.org/tex-archive/macros/latex/contrib/els-cas-templates.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Gemini in order to: Grammar and spelling check, Paraphrase and reword, and Improve writing style. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, M. Constantin, M. Dogariu, A. 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. 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. 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, 2026.

[2] 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, 2026. To appear.

[3] 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.

[4] O. Bodenreider, The unified medical language system (UMLS): Integrating biomedical terminology, Nucleic acids research 32 (2004) D267–70. doi:10.1093/nar/gkh061.

[5] V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption 2019, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, volume 2380 of *CEUR Workshop Proceedings*, Lugano, Switzerland, 2019.

[6] B. Karatzas, J. Pavlopoulos, V. Kougia, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption 2020, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, volume 2696 of *CEUR Workshop Proceedings*, Thessaloniki, Greece, 2020.

[7] F. Charalampakos, V. Karatzas, V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption Tasks 2021, in: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, volume 2936 of *CEUR Workshop Proceedings*, Bucharest, Romania, 2021, pp. 1184–1200.

[8] F. Charalampakos, G. Zachariadis, J. Pavlopoulos, V. Karatzas, C. Trakas, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmedical Caption 2022, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022, pp. 1355–1373.

[9] P. Kaliosis, G. Moschovis, F. Charalampakos, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group

at ImageCLEFmedical Caption 2023, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.
[10] M. Samprovalaki, A. Chatzipapadopoulou, G. Moschovis, F. Charalampakos, P. Kaliosis, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group in ImageCLEF medical 2024 (highlighted talk), in: Proceedings of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 2024.
[11] A. Chatzipapadopoulou, I. Pantelidis, F. Charalampakos, M. Samprovalaki, G. Moschovis, P. Kaliosis, K. V. Dalakleidi, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmedical Caption 2025, in: Proceedings of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain, 2025.
[12] A. Chatzipapadopoulou, Enhanced Biomedical Image Tagging, Bachelor's thesis, Athens University of Economics and Business, Department of Informatics, 2025. URL: http://nlp.cs.aueb.gr/theses/Bsc_Thesis_Chatzipapadopoulou.pdf.
[13] A. Chatzipapadopoulou, Ensemble Learning in Uncertainty Quantification for Multi-Label Prediction: From Theoretical Foundations to Medical Image Understanding, Master's thesis, Athens University of Economics and Business, Department of Informatics, 2025. URL: https://pyxida.aueb.gr/handle/123456789/12259. doi:10.26219/heal.aueb.9478, supervisors: John Pavlopoulos and Ion Androutsopoulos.
[14] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: ICLR, 2019.
[15] F. Radenović, G. Tolias, O. Chum, Fine-tuning CNN image retrieval with no human annotation, PubMed 41 (2019) 1655 – 1668. doi:10.1109/tpami.2018.2846566.
[16] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: ICML, 2019, pp. 6105 – 6114.
[17] M. Tan, Q. V. Le, Efficientnetv2: Smaller models and faster training, in: ICML, 2021, pp. 10096 – 10106.
[18] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: Computer Vision and Pattern Recognition, 2022, pp. 11966 – 11976. doi:10.1109/cvpr52688.2022.01167.
[19] G. Huang, Z. Liu, K. Q. Weinberger, Densely connected convolutional networks, in: Computer Vision and Pattern Recognition, 2016, pp. 2261 – 2269. doi:10.1109/cvpr.2017.243.
[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770 – 778. doi:10.1109/cvpr.2016.90.
[21] M. Raghu, C. Zhang, J. M. Kleinberg, S. Bengio, Transfusion: Understanding transfer learning for medical imaging, in: NeurIPS, 2019, pp. 3342 – 3352.
[22] S. Bethard, We need to talk about random seeds, ArXiv abs/2210.13393 (2022). URL: https://api.semanticscholar.org/CorpusID:252548394.
[23] G. Shafer, V. Vovk, A tutorial on conformal prediction, J. Mach. Learn. Res. 9 (2008) 371 – 421. doi:10.5555/1390681.1390693.
[24] A. N. Angelopoulos, S. Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification, CoRR abs/2107.07511 (2021). URL: https://arxiv.org/abs/2107.07511. arXiv:2107.07511.
[25] H. Papadopoulos, Inductive conformal prediction: Theory and application to neural networks, INTECH Open Access Publisher Rijeka, 2008. doi:10.5772/6078.
[26] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression, Journal of the American Statistical Association 113 (2017) 1094 – 1111. doi:10.1080/01621459.2017.1307116.
[27] G. Shan, Monte carlo cross-validation for a study with binary outcome and limited sample size, BMC Medical Informatics and Decision Making 22 (2022). doi:10.1186/s12911-022-02016-z.
[28] K. Sechidis, G. Tsoumakas, I. Vlahavas, On the stratification of multi-label data, in: D. Gunopulos, T. Hofmann, D. Malerba, M. Vazirgiannis (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 145–158. doi:10.1007/978-3-642-23808-6_10.

Appendix

Table 11
Benchmarking of Individual Base Models. This table details the standalone predictive performance of all constituent base models evaluated on our held-out dev set. For each Convolutional Neural Network (CNN) backbone architecture, we report the resulting F_1 score alongside its corresponding optimized global decision threshold. To evaluate initialization stability, these metrics are comprehensively reported across four distinct random weight seeds (40, 41, 42, and 43).

Cross-Seed Benchmarking of Base Architectures								
	Seed 40		Seed 41		Seed 42		Seed 43	
Base Model	**F1**	**Threshold**	**F1**	**Threshold**	**F1**	**Threshold**	**F1**	**Threshold**
ConvNeXt-Tiny	0.5926	0.41	0.5917	0.40	0.5886	0.41	0.5896	0.40
DenseNet-121	0.5862	0.46	0.5860	0.45	0.5841	0.45	0.5857	0.48
EfficientNet-B0	0.5865	0.42	0.5892	0.40	0.5877	0.38	0.5873	0.43
EfficientNet-V2-M	0.5905	0.42	0.5896	0.46	0.5906	0.46	0.5908	0.38
EfficientNet-V2-S	0.5889	0.39	0.5892	0.45	0.5901	0.42	0.5907	0.47
ResNet-50	0.5844	0.42	0.5848	0.40	0.5844	0.47	0.5855	0.47

Table 12
Performance of the Mondrian Conformal-First strategy. The table reports the F_1 scores on our held-out dev set for an 18-model ensemble across varying Dual Threshold values (L). We compare the baseline ensemble without conformal prediction (**No BCR**) against the dynamic, bin-based **Mondrian Conformal-First** approach applied prior to ensembling. The bold value indicates the peak performance threshold.

Evaluation of Mondrian Conformal-First strategy		
L	**No BCR**	**Mondrian Conformal-First**
1	0.5626	0.5622
2	0.5890	0.5889
3	0.5969	0.5969
4	0.5991	0.5992
5	0.6007	**0.6007**
6	0.6001	0.6006
7	0.5991	0.6002
8	0.5981	0.5990
9	0.5968	0.5977

**Where multiple strategies share identical values at four decimal places, bold formatting reflects differences at the fifth decimal place.*