Frequency-Aware Multi-Head Learning for Long-Tailed Medical Concept Detection in ImageCLEFmed Caption 2026

Bashar Alsaleh[1,*], Riad Sonbol[1,2] and Zeinab Baghdadi[1]

[1]Faculty of Artificial Intelligence Engineering, Syrian Private University, Damascus, Syria
[2]Department of Informatics, Higher Institute for Applied Sciences and Technology, Damascus, Syria

Abstract

This paper describes the submission of **team basharderar** to the ImageCLEFmed Caption 2026 Concept Detection Task. The task aims to predict clinically relevant Unified Medical Language System (UMLS) concepts from radiology images in a large-scale, sparse, and highly imbalanced multi-label setting. We explore a frequency-aware multi-head learning strategy built on a Swin Transformer visual backbone. The concept vocabulary is divided into frequency-based tiers, and each tier is assigned a dedicated prediction head. This design is used together with Asymmetric Loss, frequency-group loss weighting, exponential moving average, multilabel stratified validation, and post-hoc threshold calibration.

To address long-tailed evaluation more carefully, we report additional ablation and frequency-tier analyses comparing the proposed model with two single-head Swin-Base baselines: a standard Asymmetric Loss baseline and a class-weighted BCE baseline with post-hoc thresholds. The proposed frequency-aware model achieved a three-fold validation Micro-F1 of 0.5421 under per-head threshold calibration, which is comparable to the single-head ASL baseline and higher than the class-weighted single-head baseline under the global Micro-F1-oriented operating point. The frequency-tier analysis shows that aggregate Micro-F1 is strongly dominated by common concepts, while rare-concept recovery depends heavily on the selected operating point. On the official hidden test set, team basharderar achieved a primary score of 0.5725 and a secondary score of 0.9419, ranking 4th in the official task leaderboard. Overall, our results highlight the trade-off between aggregate Micro-F1 and rare-concept recovery in long-tailed medical concept detection, and suggest that frequency-aware calibration is an important component for analyzing such systems.

Keywords

medical concept detection, ImageCLEFmed Caption 2026, multi-label classification, long-tailed learning, frequency-aware learning, Swin Transformer, UMLS concepts, threshold calibration

1. Introduction

Medical image understanding is an important research direction in biomedical artificial intelligence because radiology images are often associated with clinically meaningful findings, anatomical structures, imaging modalities, and standardized medical concepts [1, 2, 3, 4]. In medical captioning systems, directly generating free-text descriptions from images can be challenging because generated text may be difficult to evaluate and may omit or misrepresent clinically relevant information. Concept-based prediction provides an interpretable intermediate representation by first identifying the medical concepts associated with an image before using them as semantic support for downstream caption generation.

The ImageCLEFmed Caption 2026 challenge addresses this direction through tasks related to medical concept detection and caption generation [1, 2]. In the Concept Detection Task, the goal is to predict UMLS-aligned concepts for each radiology image. This task is naturally formulated as a multi-label classification problem because a single image can be associated with multiple concepts simultaneously. Unlike single-label image classification, multi-label medical concept detection requires the model to make many binary decisions over a large concept vocabulary for each input image.

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany
*Corresponding author.
bashar_alsaleh@spu.edu.sy (B. Alsaleh); riadsonbol@spu.edu.sy (R. Sonbol); riad.sonbol@hiast.edu.sy (R. Sonbol); zeinabbaghdadi60@gmail.com (Z. Baghdadi)
0009-0007-8701-5588 (B. Alsaleh); 0000-0002-2785-2508 (R. Sonbol); 0009-0005-3261-0406 (Z. Baghdadi)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The challenge data are based on an updated and extended version of the ROCOv2 dataset [3]. In our development data, the task involved 116,604 radiology image samples and 2,646 unique UMLS concepts. Each image contained an average of approximately 3.22 concepts, which indicates a sparse target structure in which only a small fraction of the vocabulary is positive for each image. In addition to sparsity, the label distribution is strongly long-tailed: a small number of concepts occur frequently, whereas many concepts have limited training support.

Long-tailed label distributions are known to bias learning toward frequent labels and reduce performance on underrepresented classes [7, 8, 9, 10]. In a conventional single-head multi-label classifier, the optimization process can be dominated by frequent concepts because they contribute more stable and more frequent gradient signals during training. As a result, rare concepts may receive insufficient supervision, even though they can represent clinically specific and semantically important information.

This paper presents the system submitted by team basharderar to the ImageCLEFmed Caption 2026 Concept Detection Task. We explore a frequency-aware multi-head learning strategy based on a Swin Transformer visual encoder [12]. Rather than treating the full concept vocabulary as a single homogeneous output space, the proposed approach decomposes the label space into frequency tiers and assigns each tier a dedicated prediction head. In the revised analysis, we compare the proposed model with two single-head Swin-Base baselines and report both aggregate and frequency-tier metrics. This allows us to discuss the trade-off between global Micro-F1 and rare-concept recovery more objectively.

The remainder of this paper is organized as follows. Section 2 reviews related work on medical concept detection and long-tailed multi-label learning. Section 3 describes the task and dataset characteristics. Section 4 presents the proposed frequency-aware multi-head method. Section 5 reports the experimental setup, ablation results, frequency-tier analysis, rare-aware calibration, and official test-set result. Section 6 discusses the main observations and limitations. Section 7 concludes the paper.

2. Related Work

Medical concept detection is closely related to multi-label medical image classification, where the goal is to assign one or more clinically relevant labels to each image. In the context of ImageCLEFmed Caption, concept prediction is used as an intermediate semantic representation that can support medical caption generation and improve interpretability [5, 6, 2]. The use of UMLS concepts provides a standardized biomedical vocabulary, making the predicted labels more structured than free-text outputs [4].

Medical concept detection has been addressed in previous ImageCLEFmedical Caption campaigns, where the task has been formulated as the prediction of UMLS concepts from biomedical or radiology images. The 2024 task focused on caption prediction and concept detection using ROCOv2, while the 2025 edition extended the setting to medical concept detection, caption generation, and explainability [5, 6]. These prior campaigns highlight the continuing importance of concept prediction as a structured step toward medical image understanding and caption generation.

Large-scale medical image classification benchmarks such as ChestX-ray8 and CheXpert have shown the importance of multi-label learning in radiology, where a single image may be associated with several findings or observations [13, 14]. More broadly, deep learning has become a central methodology for medical image analysis, including classification, detection, segmentation, and related tasks [15]. Previous medical captioning and concept-based systems commonly relied on visual encoders to extract image-level representations and then predict medical concepts or generate textual descriptions.

Convolutional neural networks such as DenseNet have been widely used as strong visual backbones for image classification tasks because of their feature reuse and efficient gradient flow [16]. More recently, transformer-based visual backbones have become attractive for image understanding because attention mechanisms can model contextual relationships across image regions [12]. This is relevant for radiology images, where a visual concept may depend not only on a local pattern but also on its surrounding anatomical context.

A major challenge in medical concept detection is long-tailed multi-label imbalance. In such settings, a small number of labels occur frequently, while many labels appear only rarely. Standard binary

Table 1
Dataset statistics used in our experiments.

Statistic	Value
Development samples	116,604 images
Concept vocabulary	2,646 UMLS CUIs
Average concepts per image	3.22
Median concepts per image	3
Maximum concepts per image	37
Empty-label images	0
Target representation	Multi-hot vector
Task type	Multi-label classification
Image input size	224×224
Validation strategy	3-fold multilabel stratified validation

cross-entropy training can be dominated by easy negative labels and frequent positive labels, which can reduce sensitivity to rare concepts. Several loss-design strategies have been proposed to address imbalance, including focal-style losses and class-balanced learning objectives [7, 8]. Asymmetric Loss extends this direction for multi-label classification by applying different focusing behavior to positive and negative terms, making it suitable for sparse multi-label settings with many negative labels [11].

However, loss reweighting alone may not fully address the heterogeneity of a large long-tailed label space. When all concepts are predicted by a single output head, frequent and rare concepts share the same classifier space and thresholding behavior. This can make it difficult to select decision thresholds that work well across all frequency ranges. The present work explores a frequency-aware decomposition of the output space and analyzes how such a structure behaves compared with single-head ASL and class-weighted baselines.

3. Task and Dataset

The Concept Detection Task of ImageCLEFmed Caption 2026 aims to predict a set of medical concepts associated with each radiology image [2]. Each concept is represented using a Unified Medical Language System (UMLS) Concept Unique Identifier (CUI), which provides a standardized semantic representation of biomedical terminology [4]. Therefore, the task can be viewed as an intermediate semantic prediction stage for concept-based medical caption generation.

Formally, given an input radiology image x, the objective is to predict a binary vector $y \in \{0, 1\}^K$, where K is the number of unique medical concepts in the vocabulary. A value of 1 indicates that the corresponding concept is relevant to the image, while a value of 0 indicates that it is not assigned. Since each image can contain multiple concepts, the task is formulated as a multi-label classification problem rather than a single-label classification problem.

The released data include radiology images, concept annotations, and attribution metadata. During preprocessing, the concept annotations were parsed and converted into a fixed label vocabulary. Each image was then represented using a multi-hot target vector over the full concept space. This representation preserves the multi-label nature of the task and allows the model to learn independent relevance scores for all concepts.

A central characteristic of the dataset is the long-tailed label-frequency distribution. A small number of concepts occur very frequently, while many concepts have limited training support. Figure 1 illustrates this label-frequency imbalance across the concept vocabulary. This distribution is important because frequent concepts can dominate model optimization, whereas rare concepts may receive insufficient learning signal despite their potential clinical relevance.

In addition to label-frequency imbalance, the dataset also exhibits sparse label cardinality at the image level. Figure 2 shows the distribution of the number of concepts assigned to each image. Most images contain only a small number of positive concepts compared with the total vocabulary size.

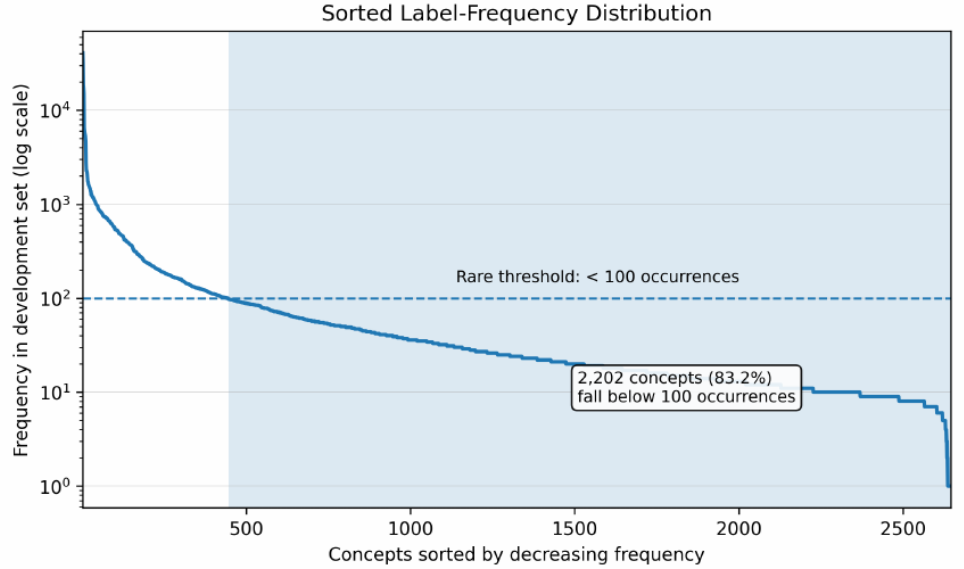

Figure 1: Sorted label-frequency distribution over the UMLS concept vocabulary. Concepts are ordered by decreasing frequency, and the dashed line marks the threshold of 100 occurrences used to define the rare-concept group. The plot shows that 83.2% of concepts have fewer than 100 occurrences.

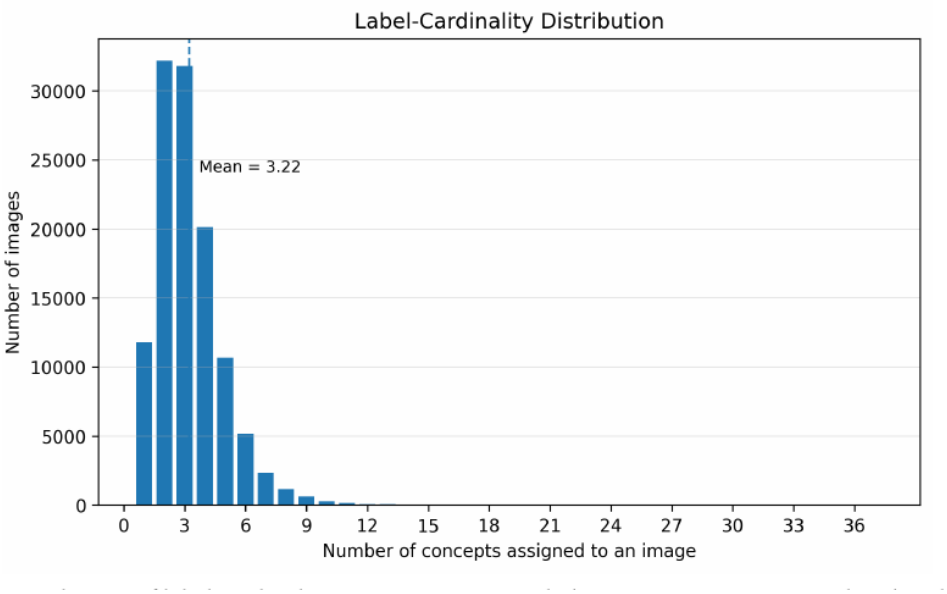

Figure 2: Distribution of label cardinality per image. Most radiology images are associated with only a small number of positive concepts compared with the full vocabulary size, highlighting the sparsity of the multi-label prediction problem.

This observation motivates careful calibration of the final decision thresholds, since predicting too few concepts reduces recall and semantic coverage, while predicting too many concepts increases false positives.

4. Proposed Method

Our proposed method is a frequency-aware multi-head learning framework designed for long-tailed medical concept detection. The main idea is to avoid treating the entire concept vocabulary as a single homogeneous output space. Instead, we decompose the label space into frequency-based tiers and assign a dedicated prediction head to each tier.

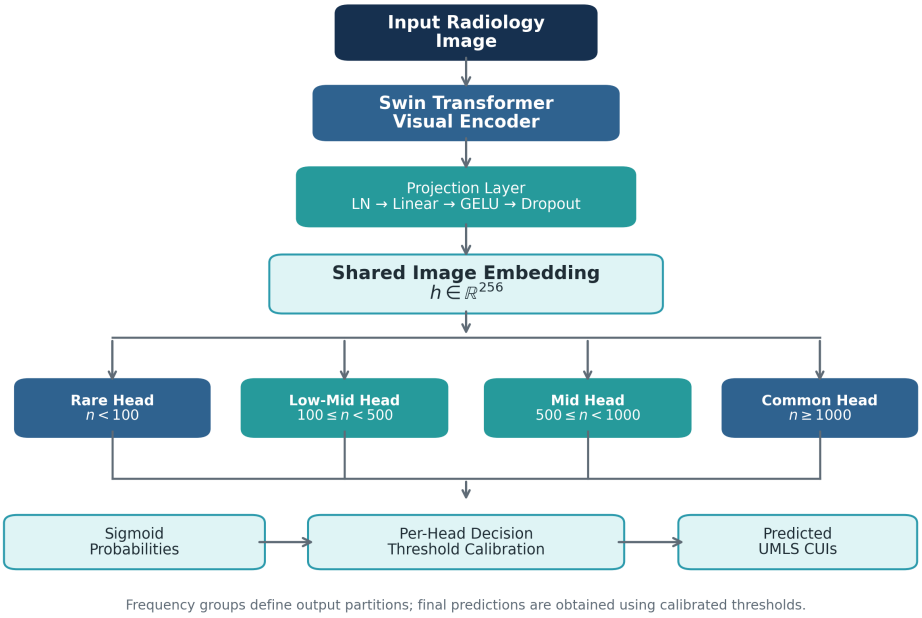

Figure 3: Overview of the proposed frequency-aware multi-head learning framework. The input radiology image is encoded using a Swin Transformer backbone, projected into a shared embedding space, and passed to four classification heads corresponding to different concept-frequency ranges. Per-head decision-threshold calibration is then applied to obtain the final predicted UMLS CUIs.

The overall architecture consists of five main components. First, the input radiology image is resized and normalized. Second, a Swin Transformer backbone extracts a visual representation from the image. Third, a projection block maps the backbone output into a compact shared embedding space. Fourth, four frequency-aware classification heads produce logits for different subsets of the UMLS concept vocabulary. Finally, sigmoid probabilities are converted into binary predictions using calibrated decision thresholds. Figure 3 illustrates the overall architecture.

4.1. Frequency-Aware Label-Space Decomposition

Based on the label-frequency analysis described in Section 3, we divide the concept vocabulary into four frequency tiers: rare concepts with fewer than 100 occurrences, low-mid-frequency concepts with 100 to fewer than 500 occurrences, mid-frequency concepts with 500 to fewer than 1000 occurrences, and common concepts with at least 1000 occurrences. These thresholds were selected as pragmatic frequency cutoffs to separate extremely low-support concepts from progressively more frequent concepts, rather than as statistically optimal partitions.

The resulting groups are not intended to be interpreted as perfectly homogeneous supervision groups. In particular, the rare head contains a large portion of the vocabulary, and concepts near a boundary may have similar frequencies despite being assigned to different groups. Nevertheless, the decomposition provides a simple and interpretable way to organize the long-tailed output space and to enable frequency-specific loss weighting, calibration, and analysis.

The rare group contains 2,202 concepts, corresponding to 83.2% of the full vocabulary. This confirms that most concepts have limited training support and that aggregate metrics such as Micro-F1 may be dominated by a small number of frequent concepts. Therefore, in addition to reporting global validation scores, we report performance separately for each frequency tier.

Table 2
Frequency groups used by the proposed multi-head model. The ranges are defined using concept occurrence counts in the development set.

Head	Frequency range	Concepts	Min	Max	Mean
Rare head	$n < 100$	2,202	1	98	27.67
Low-mid head	$100 \leq n < 500$	334	100	489	202.67
Mid head	$500 \leq n < 1000$	68	505	988	703.79
Common head	$n \geq 1000$	42	1,005	40,992	4,739.67

4.2. Swin Transformer Visual Encoder

The visual encoder of the proposed model is based on a Swin Transformer backbone [12]. We use `swin_base_patch4_window7_224` from the `timm` library and remove its original classification layer. In our implementation, the backbone is initialized using the publicly available `timm` weights pretrained on ImageNet-22K and fine-tuned on ImageNet-1K for the corresponding Swin-Base configuration. The role of the backbone is to extract a high-level visual representation from each radiology image.

The choice of a Swin Transformer is motivated by its shifted-window self-attention mechanism. Unlike purely convolutional models, transformer-based visual encoders can model contextual relationships between image regions. This is useful in radiology, where a medical concept may be associated with a local visual pattern but its interpretation can depend on the surrounding anatomical context.

4.3. Projection Layer and Shared Embedding Space

The output feature vector produced by the Swin backbone is passed through a projection block before being used by the classification heads. This block consists of layer normalization, a linear projection, GELU activation, and dropout. In our implementation, the projection dimension is set to 256.

The projection layer maps the backbone output into a compact shared embedding space that is used consistently by all frequency heads. Directly using a high-dimensional backbone feature vector for multiple heads may increase the number of classifier parameters and make optimization less stable. The 256-dimensional embedding provides a shared representation that is expressive enough for classification while remaining compact.

4.4. Multi-Head Classification Layer

After obtaining the shared image embedding, the model applies four prediction heads, one for each frequency tier. Each head predicts logits only for the concepts assigned to its corresponding frequency group, and the outputs are then mapped back to the full concept vocabulary.

Each head is implemented using a label-embedding formulation. For every concept in a head, a learnable vector and a bias term are defined. The logit of a concept is computed as the dot product between the shared image embedding and the corresponding concept vector, followed by the bias term.

It is important to clarify that this formulation is mathematically equivalent to applying a linear classifier to the shared embedding, with the rows of the classifier partitioned according to concept-frequency groups. Therefore, the contribution of the multi-head formulation should not be interpreted as adding extra nonlinear modeling capacity compared with a standard single-head linear classifier. Instead, its role is to explicitly organize the output space into frequency tiers, which enables group-specific loss weighting, threshold calibration, and frequency-tier analysis.

This distinction is important for interpreting the results. The proposed design is not claimed to solve long-tailed learning by architecture alone. Rather, it provides a structured framework for studying how different frequency groups behave under shared visual representations, weighted objectives, and calibrated decision thresholds.

4.5. Training Objective

The model is trained using Asymmetric Loss, which is commonly used for imbalanced multi-label classification [11]. In this task, each image contains only a small number of positive concepts compared with the full vocabulary size. As a result, the number of negative labels per image is much larger than the number of positive labels.

In our final configuration, Asymmetric Loss is used with $\gamma_{neg} = 2.0$, $\gamma_{pos} = 1.0$, clip $= 0.0$, and $\epsilon = 10^{-8}$. Since the clipping parameter was set to zero, the probability-shifting component of Asymmetric Loss was not used in our final system. Therefore, we interpret this loss mainly as an asymmetric focal-style objective based on different positive and negative focusing parameters, rather than attributing the results to asymmetric clipping. Evaluating non-zero clipping values is left for future work.

In addition, frequency-group loss weights are applied so that groups containing lower-frequency concepts contribute more strongly to the total loss. The total loss is computed as:

$$L = \lambda_1 L_{rare} + \lambda_2 L_{low-mid} + \lambda_3 L_{mid} + \lambda_4 L_{common},$$

where L_{rare}, $L_{low-mid}$, L_{mid}, and L_{common} are the losses of the four frequency heads. In our final configuration, the corresponding weights are 4.0, 2.0, 1.5, and 1.0.

4.6. Decision-Threshold Calibration

Since the task is multi-label, the model outputs one probability per concept. These probabilities must be converted into binary predictions using decision thresholds. A single global threshold is often suboptimal in long-tailed multi-label settings because rare and common concepts may have different score distributions.

For this reason, we calibrate decision thresholds using validation predictions. We evaluate global threshold calibration, per-head threshold calibration, and additional rare-aware calibration analysis. The per-head strategy allows the final decision boundary to adapt to the statistical behavior of each concept group, rather than forcing all concepts to use the same threshold.

4.7. Training and Validation Strategy

We evaluate the model using three-fold multilabel stratified validation. This strategy is more suitable than ordinary random splitting because it attempts to preserve the distribution of multiple labels across folds. For each fold, the model is trained on two parts of the data and validated on the remaining part.

Training uses AdamW optimization, mixed-precision training, gradient accumulation, and exponential moving average of model weights. We also used image-level weighted sampling as a heuristic to increase the probability of observing samples containing low-frequency concepts. However, because the task is multi-label, an image that contains a rare concept may also contain frequent concepts. Therefore, image-level resampling can also oversample co-occurring frequent labels. For this reason, weighted sampling is treated as a training heuristic rather than an isolated contribution, and the ablation results in Section 5 are used to provide a more careful comparison with single-head baselines.

After training, validation predictions are cached and used for post-hoc threshold calibration. We report both aggregate validation metrics and frequency-tier metrics to better characterize the trade-off between global Micro-F1 and rare-concept recovery.

5. Experiments and Results

This section describes the experimental configuration, internal validation results, frequency-tier analysis, rare-aware calibration analysis, and official test-set result obtained in the ImageCLEFmed Caption 2026 Concept Detection Task.

Table 3
Main training configuration of the final submitted model.

Component	Configuration
Visual backbone	Swin Base Patch4 Window7 224
Backbone pretraining	ImageNet-22K pretraining, ImageNet-1K fine-tuning
Input size	224×224
Projection dimension	256
Output design	Four frequency-aware heads
Loss function	Asymmetric Loss
ASL parameters	$\gamma_{neg} = 2.0$, $\gamma_{pos} = 1.0$, clip=0.0
Optimizer	AdamW
Learning rate	1×10^{-4}
Backbone learning rate	2×10^{-5}
Weight decay	1×10^{-4}
Batch size	32
Gradient accumulation	2
Effective batch size	64
EMA decay	0.999
Validation	3-fold multilabel stratified validation
Decision thresholding	Post-hoc per-head calibration

Table 4
Overall ablation results on the three-fold validation setting. The table compares the proposed frequency-aware model with two single-head Swin-Base baselines.

Method	Calibration	Micro-F1	Macro-F1$_{nz}$	Precision	Recall	Avg. pred.
Swin-Base + single-head + ASL	global threshold	0.5442	0.0075	–	–	–
Swin-Base + single-head + class-weighted BCE	global threshold	0.5250	0.0103	0.8404	0.3818	1.463
Frequency-aware multi-head Swin	per-head threshold	0.5421	0.0069	0.8629	0.3952	1.475

5.1. Experimental Setup

All images were resized to 224×224 pixels and normalized according to the input requirements of the Swin Transformer backbone. The final model used a Swin-Base visual encoder, a 256-dimensional projection layer, and four frequency-aware prediction heads. It was trained using AdamW optimization, Asymmetric Loss, frequency-group loss weighting, gradient accumulation, mixed-precision training, exponential moving average of model parameters, and post-hoc per-head threshold calibration.

To address ablation concerns, we additionally evaluated two single-head Swin-Base baselines. The first baseline uses a standard single-head classifier trained with Asymmetric Loss. The second baseline uses a standard single-head classifier trained with class-weighted binary cross-entropy and post-hoc threshold calibration. All models were evaluated using three-fold multilabel stratified validation, and results are reported using both aggregate metrics and frequency-tier metrics.

5.2. Overall Ablation Results

Table 4 reports the overall validation results of the proposed model and the two single-head Swin-Base baselines. The first baseline uses a standard single-head classifier trained with Asymmetric Loss, while the second baseline uses a standard single-head classifier trained with class-weighted binary cross-entropy and post-hoc threshold calibration. These baselines were included to better isolate the effect of the proposed frequency-aware formulation and to evaluate whether class weighting alone is sufficient to address the long-tailed label distribution.

The single-head ASL baseline achieved the highest aggregate validation Micro-F1, with a score of

Table 5
Frequency-tier validation breakdown. Values report group-level Micro-F1 for each concept-frequency tier. The rare, low-mid, mid, and common groups contain 2,202, 334, 68, and 42 concepts, respectively.

Method	Operating point	Rare	Low-mid	Mid	Common	Overall
Swin-Base + single-head + ASL	global threshold	0.0004	0.0318	0.0894	0.8039	0.5442
Swin-Base + single-head + class-weighted BCE	global threshold	0.0025	0.0499	0.1109	0.7780	0.5250
Swin-Base + single-head + class-weighted BCE	frequency-group thresholds	0.0668	0.1472	0.1831	0.7780	0.4217
Frequency-aware multi-head Swin	per-head threshold	0.0000	0.0225	0.0829	0.7977	0.5421

0.5442. The proposed frequency-aware multi-head model achieved a comparable Micro-F1 of 0.5421 under per-head threshold calibration. The class-weighted single-head baseline achieved a lower aggregate Micro-F1 of 0.5250 under the global Micro-F1-oriented operating point.

These results suggest that the proposed frequency-aware model does not uniformly outperform the strongest single-head ASL baseline in aggregate Micro-F1. However, it achieves comparable global performance while providing an explicit frequency-based structure for group-specific calibration and analysis. In contrast, class weighting alone did not improve the aggregate Micro-F1 compared with the proposed model under the global operating point.

5.3. Frequency-Tier Breakdown

Since aggregate Micro-F1 can be dominated by high-frequency concepts, Table 5 reports validation Micro-F1 separately for each frequency tier. This analysis is important because the task exhibits a strong long-tailed distribution, with 2,202 out of 2,646 concepts belonging to the rare group.

The single-head ASL baseline achieved strong performance on common concepts but almost no recovery of rare concepts, with a rare-group Micro-F1 of 0.0004. The class-weighted single-head baseline slightly improved rare-concept performance under the global threshold, reaching a rare-group Micro-F1 of 0.0025, but its aggregate Micro-F1 remained lower than the ASL baseline and the proposed model.

When frequency-group thresholds were applied to the class-weighted baseline, rare-label recovery improved substantially, reaching a rare-group Micro-F1 of 0.0668. However, this improvement came with a large decrease in overall Micro-F1, from 0.5250 under the global threshold to 0.4217 under frequency-group thresholds. This indicates that class weighting and frequency-specific thresholds can increase sensitivity to rare labels, but may also produce many additional false positives and reduce aggregate performance.

The proposed frequency-aware multi-head model achieved an overall Micro-F1 of 0.5421 under per-head threshold calibration, remaining close to the strongest ASL baseline and higher than the class-weighted single-head baseline under the global operating point. However, under this operating point, the rare-group Micro-F1 remained 0.0000. This finding shows that the proposed architecture alone does not solve rare-concept detection under the default micro-F1-oriented calibration. Instead, rare-label recovery remains highly sensitive to the selected operating point.

5.4. Rare-Aware Calibration Analysis

To further investigate whether the proposed rare head contains useful signal for underrepresented concepts, we performed an additional rare-aware calibration analysis. This analysis was applied only to the rare-concept group and is therefore not reported as a complete operating point for the full vocabulary. Its purpose is to evaluate whether lower rare-specific thresholds or top-K rare predictions can recover additional rare labels.

As shown in Table 6, the default threshold of 0.50 recovered no rare concepts. Reducing the rare-specific threshold to 0.20 increased rare-group Micro-F1 to 0.0236, with a recall of 0.0327 and 340 rare

Table 6
Rare-aware calibration analysis for the proposed frequency-aware model. The table reports metrics only for the rare-concept group.

Operating point	Rule	Precision	Recall	Micro-F1	Hit concepts	Coverage
Default threshold	$t = 0.50$	0.0000	0.0000	0.0000	0	0.00%
Rare-aware threshold	$t = 0.20$	0.0185	0.0327	0.0236	340	15.44%
Rare Top-K	$K = 2$	0.0133	0.0510	0.0211	374	16.98%

Table 7
Official ImageCLEFmed Caption 2026 Concept Detection leaderboard excerpt. The table reports both the primary and secondary official scores and uses the official corrected team names provided by the organizers where applicable.

Rank	Team	Primary score	Secondary score
1	dsgt_caption	0.5790	0.9657
2	archimedes	0.5781	0.9591
3	UIT_Worker	0.5725	0.9533
4	**basharderar (our results)**	**0.5725**	**0.9419**
5	AUEB/DMST	0.5721	0.9503

concepts receiving at least one true positive prediction. This rare-group Micro-F1 is approximately 60 times higher than the rare-group score of the single-head ASL baseline under its global threshold. A rare Top-K strategy with $K = 2$ achieved a rare-group Micro-F1 of 0.0211 and recovered 374 rare concepts.

These results indicate that the proposed model retains some discriminative signal for rare labels, but this signal is not recovered under the default micro-F1-oriented threshold. Therefore, rare-concept detection should be interpreted as an operating-point-dependent behavior rather than a uniform improvement produced by the architecture itself. The results also highlight the trade-off between rare-concept recovery and aggregate Micro-F1 in long-tailed medical concept detection.

5.5. Official Test-Set Result

The final submitted run was evaluated on the official hidden test set of the ImageCLEFmed Caption 2026 Concept Detection Task. Table 7 reports the official leaderboard excerpt provided for the top-performing teams, including both the primary and secondary scores.

Team basharderar (our results) achieved a primary score of 0.5725 and a secondary score of 0.9419, ranking 4th in the official task leaderboard. The primary-score difference between our submission and the highest-ranked submission was 0.0065. These official results show that the submitted system was competitive in the challenge setting, while the validation analyses above provide a more detailed interpretation of its behavior across frequency tiers.

6. Discussion

The experimental results show that long-tailed medical concept detection is strongly affected by the selected operating point. Although aggregate Micro-F1 is the primary validation metric, it is heavily influenced by high-frequency concepts. This is evident from the frequency-tier breakdown, where all evaluated models achieved substantially higher Micro-F1 on common concepts than on rare concepts.

The single-head ASL baseline remained a strong aggregate baseline, achieving the highest validation Micro-F1 among the internal validation experiments. The proposed frequency-aware multi-head model achieved comparable aggregate performance, but did not uniformly improve all frequency tiers. In particular, rare-concept recovery remained weak under the default per-head thresholding strategy.

This finding indicates that frequency-aware output decomposition alone is not sufficient to solve the rare-label problem under a micro-F1-oriented operating point.

The class-weighted single-head baseline provides an important comparison. Under the global threshold, it achieved lower aggregate Micro-F1 than the proposed model. However, when frequency-group thresholds were applied, it substantially improved rare-concept recovery, reaching a rare-group Micro-F1 of 0.0668. This improvement came at the cost of a large decrease in overall Micro-F1, indicating that class weighting and frequency-specific thresholding can increase sensitivity to rare labels but may also introduce many additional false positives. This result highlights the central trade-off in this task: improving rare-label recovery can conflict with maximizing aggregate Micro-F1.

The rare-aware calibration analysis of the proposed model further supports this interpretation. Although the proposed model did not recover rare concepts under the default threshold, lowering the rare-head threshold to 0.20 increased rare-group Micro-F1 to 0.0236 and recovered 340 rare concepts. This is approximately 60 times higher than the rare-group score of the single-head ASL baseline under its global threshold. This result suggests that the rare head retains some discriminative signal, but that this signal is not recovered under the default micro-F1-oriented threshold. Therefore, rare-label performance should be interpreted as strongly dependent on calibration and operating-point selection.

The multi-head design should also be interpreted carefully. Since each head applies a linear compatibility function to the same shared embedding, the label-embedding formulation is mathematically equivalent to a partitioned linear classifier. Therefore, the main contribution of the proposed structure is not additional nonlinear capacity, but rather the explicit organization of the output space into frequency tiers. Although the proposed frequency-aware architecture does not improve rare-concept detection under the default Micro-F1-oriented operating point, it enables frequency-specific calibration and reveals discriminative signal that remains inaccessible under a single global decision threshold. This suggests that the main benefit of the proposed design lies in providing a structured framework for frequency-aware optimization and analysis rather than automatically improving rare-label performance. In practical terms, separating the rare, low-mid, mid, and common concepts into different heads makes it possible to study and adjust the decision behavior of each frequency tier independently, instead of forcing all concepts to share one global thresholding behavior.

Several limitations remain. First, the frequency thresholds used to define the heads are pragmatic cutoffs rather than optimized partitions. Concepts near a boundary may have similar supervision levels despite being assigned to different heads, and the rare head contains a large majority of the vocabulary. Second, image-level weighted sampling is only a heuristic in the multi-label setting because rare concepts may co-occur with frequent concepts. Third, the final Asymmetric Loss configuration used a clipping value of zero, so the asymmetric probability-shifting mechanism was not used. Finally, the current model does not explicitly exploit UMLS ontology relations, concept co-occurrence structure, or external medical knowledge.

Future work should investigate adaptive thresholding, label-wise calibration, data-driven frequency partitions, ontology-aware concept modeling, and contrastive or label-aware objectives that may better support rare-concept learning without causing a large decrease in aggregate Micro-F1.

7. Conclusion

This paper presented the submission of team basharderar to the ImageCLEFmed Caption 2026 Concept Detection Task. We addressed the task as a large-scale, sparse, and long-tailed multi-label classification problem over UMLS medical concepts. The submitted system used a Swin Transformer visual encoder with a frequency-aware multi-head output structure, Asymmetric Loss, frequency-group loss weighting, exponential moving average, multilabel stratified validation, and post-hoc threshold calibration.

The revised validation analysis compared the proposed model with two single-head Swin-Base baselines. The proposed frequency-aware model achieved an overall validation Micro-F1 of 0.5421, which is comparable to the single-head ASL baseline and higher than the class-weighted single-head baseline under the global micro-F1-oriented operating point. The frequency-tier analysis showed

that rare-concept detection remains challenging and highly dependent on calibration. In particular, rare-aware thresholding revealed additional rare-label signal in the proposed model, improving rare-group Micro-F1 to 0.0236, while the class-weighted baseline achieved stronger rare recovery under frequency-group thresholds at the cost of substantially lower overall Micro-F1.

On the official hidden test set, team basharderar achieved a primary score of 0.5725 and a secondary score of 0.9419, ranking 4th in the official ImageCLEFmed Caption 2026 Concept Detection leaderboard. Overall, the results suggest that frequency-aware organization and calibration are useful for analyzing long-tailed medical concept detection systems, but rare-concept recovery remains an open challenge requiring more targeted calibration and learning strategies.

8. Code and Model Availability

The implementation code, project materials, and trained model checkpoints are publicly available through the project repository and its Releases section: https://github.com/BasharDerar/ImageCLEFmed-2026-Concept-Detection. The official dataset is not redistributed in the repository due to challenge usage restrictions. The released checkpoints correspond to the three validation folds of the final frequency-aware multi-head Swin Transformer model.

Acknowledgements

The authors would like to thank the organizers of ImageCLEFmed Caption 2026 for providing the dataset, evaluation platform, and benchmark environment.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT by OpenAI to support grammar and spelling checking, paraphrasing and rewording, improvement of writing style, formatting assistance, and reviewer-response preparation. The tool was also used to assist in organizing experimental results into tables and improving the clarity of selected manuscript sections. Generative AI was not used to train models, generate experimental results, create official challenge submissions, fabricate data, or replace the authors' scientific judgment. All AI-assisted content was reviewed, edited, and verified by the authors, who take full responsibility for the final content of the publication.

References

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

[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, and C. M. Friedrich. Overview of ImageCLEFmedical 2026 – Medical Concept Detection and Caption Generation with Synthetic

Data Extensions. In: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, September 21–24, 2026. To appear.

[3] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. Ben Abacha, A. García Seco de Herrera, H. Müller, P. Horn, F. Nensa, and 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–D270. doi:10.1093/nar/gkh061.

[5] J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, B. Bracke, H. Damm, T. M. G. Pakull, C. S. Schmidt, H. Müller, and C. M. Friedrich. Overview of ImageCLEFmedical 2024 – Caption Prediction and Concept Detection. In: CLEF 2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[6] 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. Ben Abacha, A. García Seco de Herrera, H. Müller, and 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.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal Loss for Dense Object Detection. In: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[8] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-Balanced Loss Based on Effective Number of Samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9268–9277.

[9] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis. Decoupling Representation and Classifier for Long-Tailed Recognition. In: International Conference on Learning Representations, 2020.

[10] A. K. Menon, S. Jayasumana, H. Jain, A. Veit, S. Kumar, A. Singh Rawat. Long-tail learning via logit adjustment. In: International Conference on Learning Representations, 2021.

[11] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor. Asymmetric Loss for Multi-Label Classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 82–91.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[13] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3462–3471.

[14] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 2019, pp. 590–597.

[15] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 2017, pp. 60–88.

[16] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708. doi:10.1109/CVPR.2017.243.