ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

School of Information Sciences and Technology
Department of Informatics
Athens, Greece

Master Thesis
in
Computer Science

Ensemble Learning in Uncertainty Quantification for Multi-Label Prediction: From Theoretical Foundations to Medical Image Understanding

Anna Chatzipapadopoulou

Supervisors: Assistant Prof. John Pavlopoulos
Department of Informatics
Athens University of Economics and Business

Prof. Ion Androutsopoulos
Department of Informatics
Athens University of Economics and Business

September 2025

Anna Chatzipapadopoulou

Ensemble Learning in Uncertainty Quantification for Multi-Label Prediction: From Theoretical Foundations to Medical Image Understanding

September 2025

Supervisors: Assistant Prof. John Pavlopoulos, Prof. Ion Androutsopoulos

Athens University of Economics and Business

School of Information Sciences and Technology

Department of Informatics

Information Processing Laboratory

Athens, Greece

Abstract

Reliable prediction in high-stakes domains requires methods that can handle complex outputs and quantify their own uncertainty. Multi-label classification (MLC) provides the framework to address tasks where multiple labels may be simultaneously relevant, while conformal prediction (CP) provides principled guarantees on the reliability of its uncertainty estimates. This thesis investigates ensemble strategies that combine these two approaches, and in a second part applies multi-label classification methods, including ensemble-based techniques, to the task of medical concept detection. In the first part, we present a theoretical and empirical study of conformal ensembles, combining the formal coverage guarantees of CP with the robustness and diversity benefits of homogeneous and heterogeneous ensembles. We propose an ensemble conformal prediction (ECP) framework for multilabel classification, in which individually conformalized models are aggregated using standard strategies such as majority voting, probability averaging, and F1-weighted fusion. We adapt existing theoretical results to analyze coverage properties under these ensembles, and evaluate their performance across benchmark datasets. Results demonstrate that conformal ensembles consistently improve macro-F1 while maintaining valid coverage, and at the same time produce more compact and informative prediction sets compared to single-model or post-hoc conformal baselines. In the second part, we address the task of multi-label medical image concept detection, examining a range of architectures and strategies, including ensemble-based methods, as part of our participation in the ImageCLEFmedical Caption 2025 challenge. Our approach employs CNN–FFNN architectures with various backbone encoders, per-label threshold optimization to address extreme label imbalance, and diverse ensemble aggregation strategies, including union, intersection, and consensus-driven methods. Experiments on the ImageCLEFmedical dataset show that these ensembles achieved highly competitive performance in concept detection, ranking first in the 2025 competition.

Περίληψη

Η αξιόπιστη πολυετικετική πρόβλεψη είναι κρίσιμη σε εφαρμογές υψηλής σημασίας, όπου τα μοντέλα πρέπει όχι μόνο να επιτυγχάνουν υψηλή ακρίβεια αλλά και να παρέχουν αξιόπιστες εκτιμήσεις αβεβαιότητας. Στην εργασία αυτή μελετώνται στρατηγικές ensemble learning για πολυετικετική ταξινόμηση, με έμφαση στη μέθοδο conformal prediction (CP), και εξετάζεται η εφαρμογή τους στον εντοπισμό ιατρικών εννοιών. Στο πρώτο μέρος παρουσιάζεται θεωρητική και πειραματική μελέτη των conformal ensembles, που συνδυάζουν τις στατιστικές εγγυήσεις κάλυψης του CP με την ανθεκτικότητα και ποικιλία ομοιογενών και ετερογενών συνόλων μοντέλων. Αναλύονται διαφορετικές στρατηγικές συνδυασμού (πλειοψηφία, μέσος όρος πιθανοτήτων, F1-weighted fusion) και αξιολογούνται τόσο ως προς την κάλυψη όσο και την ακρίβεια σε benchmark σύνολα δεδομένων. Τα αποτελέσματα δείχνουν ότι οι conformal ensembles βελτιώνουν συστηματικά την απόδοση και τη βαθμονόμηση, παρέχοντας πιο συμπαγή και κατατοπιστικά σύνολα προβλέψεων σε σχέση με μεμονωμένα μοντέλα ή μεταγενέστερες προσεγγίσεις CP. Στο δεύτερο μέρος, οι παραπάνω ιδέες εφαρμόζονται στον εντοπισμό ιατρικών εννοιών σε εικόνες, στο πλαίσιο της συμμετοχής μας στον διαγωνισμό ImageCLEFmedical Caption 2025. Χρησιμοποιούνται CNN–FFNN αρχιτεκτονικές με διαφορετικούς backbone encoders, βελτιστοποίηση κατωφλίων ανά ετικέτα για την αντιμετώπιση ακραίας ανισορροπίας, και διάφορες στρατηγικές συνδυασμού (ένωση, τομή, consensus). Τα πειράματα στο σύνολο δεδομένων ImageCLEFmedical έδειξαν ότι τα προτεινόμενα ensembles πέτυχαν κορυφαία απόδοση, κατακτώντας την πρώτη θέση στον διαγωνισμό του 2025. Συνολικά, η εργασία αναδεικνύει πώς ο συνδυασμός της στατιστικής εγκυρότητας των conformal ensembles με τις απαιτήσεις του ιατρικού εντοπισμού εννοιών μπορεί να οδηγήσει σε πιο αξιόπιστες και ακριβείς πολυετικετικές προβλέψεις.

Acknowledgements

I would like to express my sincere appreciation to my supervisors, John Pavlopoulos and Ion Androutsopoulos, for granting me the opportunity to work under their guidance and for their steadfast support throughout the development of this thesis. Their persistent guidance, constructive feedback, and encouragement have significantly shaped my research direction and expanded my academic perspective.

I am also deeply grateful to our biomedical team members, Postdoctoral Researcher Kalliopi Dalakleidi, PhD candidates Foivos Charalampakos, Georgios Moschovis, and Panagiotis Kaliosis, Master's student Marina Samprovalaki, and BSc Ippokratis Pantelidis, whose insightful discussions, technical advice, and practical guidance were invaluable at every stage of this work. Their expertise and collaborative spirit greatly enriched both the theoretical and applied aspects of this project.

Finally, I wish to express my heartfelt thanks to my family and friends for their unwavering support, patience, and belief in me. Their love and understanding have been my foundation and source of strength throughout this journey.

Contents

1 Introduction

Medical imaging is a cornerstone of modern healthcare, essential for diagnosis, treatment planning, and patient monitoring across virtually all medical specialties. The rapid evolution of imaging technologies—spanning X-rays, MRIs, PET/CT scans, and ultrasounds—has led to a surge in both the volume and complexity of imaging data [Naj22]. This exponential growth presents a pressing challenge: developing automated systems capable of interpreting medical images accurately and efficiently, to alleviate the burden on radiologists without compromising diagnostic precision.

A key task in this space is medical concept detection—automatically identifying and tagging anatomical structures, pathological findings, and imaging modalities within medical images. This task falls under multi-label classification (MLC), where each image may correspond to multiple, often interdependent, labels. Complicating matters further is the severe class imbalance typical of medical datasets, where rare but critical findings are vastly outnumbered by more common cases.

Designing robust multi-label classification (MLC) systems for medical imaging entails several core challenges. First, models must represent dependencies among labels, whereby the presence of one condition can alter the likelihood of others. Second, they must address severe class imbalance so that rare but clinically important findings are not missed. Finally, in safety-critical settings such as healthcare, systems should couple high predictive accuracy with well-calibrated uncertainty. When the evidence is weak or ambiguous, a robust model should modulate its output—by flagging low confidence, deferring, or presenting a compact set of plausible labels with coverage guarantees (e.g., via conformal prediction)—so that clinicians can adjudicate among a narrow, clinically meaningful set of alternatives. This behavior preserves utility on difficult cases, supports clinician oversight, and reduces the risk of high-stakes errors by aligning the model's expressed confidence with the strength of the underlying evidence.

This thesis addresses these challenges by bridging theoretical advances in uncertainty-aware machine learning with practical applications in medical image analysis. The first part focuses on enhancing conformal prediction (CP) for MLC through ensemble learning. CP is a model-agnostic framework that constructs prediction sets with formal coverage guarantees under minimal assumptions, making it well-suited for high-stakes applications [VGS05]. However, standard CP often yields overly conservative sets and fails to capture label dependencies. To address this, we develop novel ensemble conformal methods with

new coverage bounds and aggregation strategies. Empirical evaluations on MS-COCO, Yeast, and Emotions datasets show our methods consistently outperform single-model and post-hoc calibrated baselines, achieving higher F1-scores, valid coverage, and more compact predictions.

The second part of this thesis focuses on the application of these methods to medical concept detection, using the ImageCLEFmedical Caption challenges of 2024 and 2025 as a real-world case study. In those competitions, our team developed a high-performing system based on deep convolutional neural networks (CNNs) for feature extraction, combined with feed-forward neural network (FFNN) classifiers and a suite of ensemble techniques. We introduced per-label threshold optimization using coordinate ascent and designed custom aggregation strategies—such as dual-threshold and partial intersection methods—to improve robustness. These systems achieved top-tier results in the competitions, ranking 2nd in 2024 and 1st in 2025 in the Concept Detection task. In this thesis, we extend our previous work by integrating ensemble conformal prediction into the medical concept detection pipeline. This allows us to assess whether the statistical rigor and coverage guarantees of CP translate effectively to clinical multi-label tasks. In doing so, we not only build upon our competition-tested architectures but also explore the trade-offs between predictive performance and uncertainty calibration in real-world, imbalanced, and high-dimensional medical datasets. The broader significance of this research lies in demonstrating that theoretical advances in uncertainty-aware learning can be successfully applied to practical, safety-critical problems. Our methods offer a general framework for combining accuracy, robustness, and calibrated uncertainty in MLC, with particular relevance to healthcare applications. The integration of CP into deep learning-based medical imaging opens promising avenues for future work, including context-aware prediction sets, uncertainty-guided explainability, and multimodal AI systems that account for clinical risk in their outputs.

1.1 Motivation and Problem Statement

The motivation for this thesis arises from the need to develop reliable, uncertainty-aware methods for multi-label classification that are both theoretically grounded and practically applicable in high-stakes domains. While the first part of this work advances ensemble-based conformal prediction methods and validates them across diverse benchmark datasets, the second part focuses on large-scale medical concept detection, where robustness, calibration, and domain-specific optimization are crucial. By applying the CP ensemble methodology developed in the general MLC setting to the medical domain—alongside a broader deep learning framework incorporating CNN–FFNN architectures, per-label threshold optimization, and specialized ensemble strategies—this thesis bridges methodological innovation with real-world deployment. The central problem addressed is how to design

and adapt ensemble-based prediction systems that achieve high accuracy, maintain valid and informative uncertainty estimates, and remain effective under the severe imbalance, label dependencies, and operational constraints characteristic of clinical applications.

1.2 Thesis Structure

This thesis is organized into the following chapters:

Chapter 2: Related Work

This chapter reviews the literature on multi-label classification, conformal prediction, and ensemble learning, with a focus on their applications in medical imaging and uncertainty quantification. It also discusses key benchmarks, methodological trends, and the limitations of current approaches that motivate the contributions of this thesis.

Chapter 3: Methodology

This chapter details the methodology developed in this thesis. The first part introduces ensemble-based conformal prediction methods for multi-label classification, including theoretical justifications and aggregation strategies. The second part describes the medical concept detection framework used in the ImageCLEFmedical competitions and explains how conformal prediction is integrated into this pipeline.

Chapter 4: Data

This chapter presents the datasets used in both the theoretical and applied parts of the thesis. It includes benchmark MLC datasets (MS-COCO, Yeast, Emotions) as well as the ImageCLEFmedical Caption 2024 and 2025 datasets. Dataset characteristics, preprocessing steps, label distributions, and data splits are described in detail.

Chapter 5: Experiments and Evaluation

This chapter presents the experimental setup and results. It includes evaluations of the proposed conformal ensemble methods on benchmark datasets and their application to medical concept detection. Performance is assessed using accuracy, F1-score, coverage, prediction set size, and robustness under class imbalance. Comparisons with baselines are also included.

Chapter 6: Conclusions and Future Work

The final chapter summarizes the key contributions of the thesis and reflects on its implications for uncertainty-aware machine learning in medical imaging. It also outlines directions for future research, including adaptive and multimodal conformal methods, clinical integration, and broader deployment in high-stakes domains.

2 Background and Related Work

2.1 Multi-Label Classification (MLC)

Multi-label classification (MLC) is a supervised learning paradigm in which each instance may be associated with multiple labels simultaneously, in contrast to single-label classification where exactly one class is assigned [TKV10]. This setting arises naturally in many domains: a photograph may contain both a *person* and a *bicycle*, a news article may belong to both *politics* and *economy*, and a chest X-ray may exhibit *cardiomegaly* and *pleural effusion*. Such cases highlight that co-occurring labels are not incidental but often encode meaningful dependencies, especially in medical imaging where the presence of one finding can increase or decrease the likelihood of another.

Formally, let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the input feature space and $\mathcal{L} = \{\lambda_1, \dots, \lambda_L\}$ the finite label set. The output space is $\mathcal{Y} = \{0,1\}^L$, where $y_\ell = 1$ indicates the presence of label λ_ℓ. A dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ consists of feature vectors $\mathbf{x}_i \in \mathcal{X}$ and binary label vectors $\mathbf{y}_i \in \mathcal{Y}$. The goal is to learn a function $f : \mathcal{X} \to [0,1]^L$ producing label-wise scores, which are converted into predictions via a thresholding rule

$$\widehat{y}_\ell(\mathbf{x}) = \begin{cases} 1 & \text{if } f_\ell(\mathbf{x}) \geq \tau_\ell, \\ 0 & \text{otherwise,} \end{cases}$$

with thresholds τ_ℓ set globally or per label.

The complexity of MLC stems from three main aspects. First, **label imbalance**: most datasets exhibit a few highly frequent "head" labels alongside many rare "tail" ones. Second, **label dependencies**: labels often co-occur (e.g., *pneumonia* with *lung opacity*) or are mutually exclusive. Third, **scalability**: the exponential size of the label space complicates both modeling and evaluation, motivating the use of metrics such as micro- and macro-averaged F_1 instead of simple accuracy.

Traditional approaches fall into two categories. *Problem transformation* methods reduce MLC to standard classification tasks, e.g., Binary Relevance (BR), which trains one classifier per label, Classifier Chains (CC), which model dependencies by propagating predictions across labels [Rea+11], and Label Powerset (LP), which treats each unique label combination

as a class. In contrast, *algorithm adaptation* methods extend existing algorithms directly to the multi-label setting, such as ML-kNN [ZZ07] or ranking-based SVMs [EW01].

More recently, deep learning has become dominant for MLC, as neural architectures can learn shared feature representations and capture complex inter-label relationships. Approaches leveraging attention mechanisms, graph neural networks, and transfer learning have shown strong results, while cost-sensitive learning addresses the challenges of imbalance [Liu+21; Wan+17].

In summary, MLC provides a flexible framework for modeling real-world problems with overlapping categories, but its challenges—imbalance, dependencies, and scale—require methods that go beyond naive extensions of single-label classification. These challenges motivate the uncertainty-aware and ensemble-based methods studied in this thesis.

2.1.1 Descriptive Properties of Multi-Label Data

The statistical characteristics of multi-label datasets strongly influence both model design and evaluation. In particular, label cardinality and density, class imbalance, and inter-label dependencies are central descriptors that affect predictive performance and the choice of appropriate metrics [TKV10; ZZ14].

Label Cardinality and Density. Two standard measures quantify the extent of multi-labelness in a dataset. The *label cardinality* is the average number of labels per instance:

$$\text{Cardinality}(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}|Y_i|,$$

while the *label density* normalizes this value by the number of labels L:

$$\text{Density}(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i|}{L}.$$

High cardinality and density imply that models must output multiple active labels per example, whereas low values indicate sparse assignments where false positives are more critical.

Class Imbalance and Long-Tail Distributions. Label frequencies in real-world MLC corpora often follow a long-tail distribution [HG09]. A few labels occur frequently, while many appear rarely or only once. This can be expressed through the marginal probability

$$p(\lambda_\ell) = \frac{1}{n} \sum_{i=1}^{n} y_{i\ell},$$

which is typically highly skewed. In medical imaging, frequent descriptors such as *chest* or *plain X-ray* dominate, whereas rare but clinically important labels like *pulmonary embolism* may occur in less than 1% of cases. Such imbalance biases models toward common classes and motivates remedies such as re-sampling, cost-sensitive learning, or threshold calibration.

Label Correlations. Labels in MLC are rarely independent, but often exhibit strong dependencies [Rea+11]. These can be positive (e.g., *pneumonia* with *lung opacity*), negative (e.g., *normal finding* versus pathology), or conditional. Correlations are commonly quantified via a co-occurrence matrix

$$C_{\ell m} = \sum_{i=1}^{n} \mathbb{I}\{y_{i\ell} = 1 \ \wedge \ y_{im} = 1\},$$

or normalized indices such as the Jaccard similarity

$$J(\lambda_\ell, \lambda_m) = \frac{C_{\ell m}}{C_{\ell\cdot} + C_{\cdot m} - C_{\ell m}}.$$

Exploiting these dependencies through methods like Classifier Chains [Rea+11] or graph-based models has been shown to significantly improve predictive performance.

Illustrative Analyses. To analyze these properties, researchers typically inspect: (i) label frequency histograms to expose long-tail imbalance, (ii) co-occurrence heatmaps to highlight correlations, and (iii) distributions of label cardinality to gauge per-instance label load. Such visualizations are provided for this thesis' datasets in Chapter 4, grounding the above concepts in empirical evidence.

2.1.2 Challenges in Multi-Label Classification

While multi-label classification (MLC) enables the modeling of complex, multi-faceted phenomena, it also introduces several challenges absent or less pronounced in single-label settings [ZZ14; TKV10].

Inter-label dependencies. Labels often exhibit structured co-occurrence patterns that can be positive (e.g., *pneumonia* with *lung opacity*), negative (e.g., *normal finding* vs. pathology), or conditional on features or other labels. Simple methods such as Binary Relevance treat labels independently, thereby discarding valuable relational information. Methods that explicitly model dependencies—such as Classifier Chains [Rea+11] or graph-based approaches [Hua+20]—tend to achieve superior performance.

Class imbalance. Real-world datasets typically follow a long-tail distribution where a few "head" labels dominate and many "tail" labels are rare [HG09]. This imbalance is especially problematic in medical imaging, where rare but critical conditions may occur in less than 1% of cases. Without corrective strategies, models achieve poor recall for rare labels. Remedies include resampling, cost-sensitive loss weighting, and threshold calibration [Cha+15].

Scalability. As the number of labels L grows, the label space $\{0,1\}^L$ expands exponentially, making exhaustive modeling infeasible. Even training one classifier per label can be computationally prohibitive for large L. To address this, dimensionality-reduction strategies such as low-rank label embeddings [TL12], sparse output coding [Hsu+09], and hierarchical partitioning [CGZ06] have been proposed.

Evaluation complexity. No single metric captures all aspects of performance in MLC. Metrics such as Hamming loss, subset accuracy, and micro-/macro-averaged F1 emphasize different trade-offs [ZZ14]. In high-stakes domains, metric choice is application-driven: medical imaging often prioritizes recall for rare labels, whereas recommendation systems may emphasize precision to preserve user trust.

2.1.3 Classical Approaches to Multi-Label Classification

Before the rise of deep learning, research on multi-label classification (MLC) primarily extended single-label paradigms to handle multiple outputs. Classical methods are commonly divided into *problem transformation* techniques, which reduce MLC to simpler subproblems, and *algorithm adaptation* approaches, which directly modify existing learners to produce multi-label outputs [ZZ14; TKV10].

Problem Transformation Methods. These approaches reformulate the MLC task into one or more single-label problems, enabling the reuse of established algorithms:

- **Binary Relevance (BR):** Trains L independent binary classifiers, one per label. BR is simple and scalable but ignores label dependencies.

- **Classifier Chains (CC)** [Rea+11]: Extends BR by feeding predictions of earlier labels as features into subsequent classifiers, thus modeling conditional dependencies. Ensembles of CC with randomized label orders mitigate error propagation.

- **Label Powerset (LP):** Treats each observed label combination as a single multi-class label, capturing dependencies exactly but suffering from sparsity and scalability issues when L is large. Variants such as Pruned LP address this by discarding rare combinations.

- **Random k-Labelsets (RAkEL)** [TV07]: Decomposes the label set into smaller random subsets of size k and applies LP locally. By aggregating predictions across multiple such subsets, RAkEL balances dependency modeling with tractability.

Algorithm Adaptation Methods. Instead of decomposing the task, these approaches extend learning algorithms to handle multiple labels directly:

- **ML-kNN** [ZZ07]: An adaptation of k-nearest neighbors that estimates per-label posterior probabilities from neighbor counts. Effective on small to medium datasets, though computationally expensive at inference.

- **Rank-SVM** [EW01]: Formulates MLC as a ranking task by learning pairwise label orderings, useful when ranked predictions are required but costly due to $O(L^2)$ comparisons.

- **Neural networks with sigmoid outputs:** Early neural MLC models replaced the softmax layer with L sigmoids trained via binary cross-entropy. These leverage shared representations but do not inherently capture inter-label dependencies, often requiring additional mechanisms (e.g., attention, graph-based models).

- **Other adaptations:** Historical proposals include BRkNN (a hybrid of BR and kNN), multi-label decision trees and ensembles, and probabilistic graphical models (Bayesian networks, Markov random fields). These explicitly model dependencies but face scalability challenges as label spaces grow.

Overall, classical approaches laid the foundation for MLC by highlighting the trade-off between scalability, ability to model dependencies, and robustness to imbalance. Their limitations motivated the development of modern deep learning–based methods that integrate representation learning with structured output modeling.

2.1.4 Evaluation Metrics in Multi-Label Classification

Evaluating multi-label classification (MLC) systems is inherently more complex than single-label classification, as no single metric fully captures all performance aspects. The choice of metric depends on the application, the relative cost of errors, and whether outputs are binary label sets, ranked lists, or calibrated probabilities [ZZ14; TKV10]. Broadly, metrics fall into three categories: set-based, label-based, and ranking-based.

Set-based metrics. Set-based measures directly compare the predicted label set $\hat{\mathbf{y}}$ with the ground truth $\mathbf{y}$. The *Hamming loss* quantifies the fraction of misclassified label–instance pairs:

$$\text{HammingLoss} = \frac{1}{NL} \sum_{i=1}^{N} \sum_{\ell=1}^{L} \mathbb{I}[y_{i\ell} \neq \hat{y}_{i\ell}], \tag{2.1}$$

where N is the number of instances and L the number of labels. It is simple and scalable but does not account for label dependencies. *Subset accuracy* (exact match ratio) metric requires that the predicted label set matches the true label set exactly:

$$\text{SubsetAccuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\mathbf{y}_i = \hat{\mathbf{y}}_i]. \tag{2.2}$$

While intuitive, it is overly strict in high-cardinality problems, as a single error invalidates the prediction.

Label-based metrics. Label-based metrics evaluate performance per label and then aggregate results using macro- or micro-averaging. Precision, recall, and the F1-score are most common:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2.3}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.4}$$

Macro-averaging weighs all labels equally, highlighting performance on rare labels, whereas micro-averaging favors frequent labels by aggregating across all decisions. The choice reflects domain priorities: e.g., macro-F1 is crucial in medical tasks where rare conditions are clinically important.

Ranking-based metrics. When models output scores rather than binary predictions, ranking-based metrics provide finer evaluation [SS00; ZZ14]. *Average precision* (AP) computes the area under the precision–recall curve for each label, then averages across labels. *Coverage error* measures the average number of top-ranked labels required to cover

all true labels. Other metrics include one-error, ranking loss, and nDCG, particularly relevant when the order of labels matters.

Trade-offs and considerations. Metrics emphasize different error types: minimizing Hamming loss may bias toward conservative predictions, while maximizing macro-F1 often boosts recall at the expense of more false positives. In healthcare, recall for rare but critical findings may outweigh precision, while in domains with costly verification, precision may dominate. Thus, comprehensive evaluation typically reports multiple complementary metrics, ensuring a balanced assessment of model behavior.

2.1.5 Modern Trends in Multi-Label Classification

Recent advances in deep learning have significantly reshaped the landscape of multi-label classification (MLC), offering powerful solutions to long-standing challenges such as modeling inter-label dependencies, addressing label imbalance, and learning from high-dimensional data [ZZ14; Wan+17].

Deep learning for shared feature representation. Deep neural networks (DNNs) are now the dominant paradigm for MLC across domains, due to their ability to learn hierarchical feature representations directly from raw inputs. Convolutional neural networks (CNNs) for images, recurrent architectures for sequential data, and transformer-based encoders for text have been successfully adapted by replacing the softmax layer with L independent sigmoid outputs [Wan+17; Liu+17]. Shared hidden layers enable joint representation learning, reducing reliance on handcrafted features and improving generalization, particularly when labels are correlated.

Attention mechanisms for label dependency modeling. To explicitly model dependencies between labels and input features, attention mechanisms have become increasingly popular. Label-wise attention networks assign distinct attention weights for each label, allowing fine-grained feature–label interactions [You+19]. More broadly, transformer architectures leverage self-attention to capture both local and global dependencies, and in some variants, inter-label relations as well [Vas+17; Che+19]. These mechanisms enable models to dynamically focus on the most relevant evidence for each label.

Graph-based methods for label relationships. Many MLC tasks involve structured label relationships that can be naturally represented as graphs. Graph neural networks (GNNs) exploit this structure by propagating information among label nodes, allowing correlated or hierarchical labels to inform each other [Che+19; Wu+20]. For example, in

biomedical imaging, anatomical labels and associated pathologies can be modeled jointly, improving recall for rare findings. Integrating GNNs with deep encoders provides a unified way to learn both input representations and structured label interactions.

Cost-sensitive learning. Severe label imbalance remains a pervasive issue in MLC. Cost-sensitive learning strategies address this by reweighting losses, resampling training data, or optimizing per-label thresholds [Cha+15]. Such approaches are particularly critical in high-stakes domains like medicine, where false negatives for rare conditions can have disproportionately severe consequences. By prioritizing rare or clinically important labels, cost-sensitive methods mitigate bias toward frequent labels.

Pretraining and transfer learning. Pretraining on large-scale datasets followed by task-specific fine-tuning has become a standard practice in MLC. In vision, CNNs pretrained on ImageNet serve as feature extractors for multi-label tagging in specialized domains [He+16], while in NLP, pretrained transformers such as BERT and its successors are fine-tuned for multi-label text classification [Dev+18]. Transfer learning also enables cross-domain adaptation, where knowledge learned in general-purpose domains (e.g., natural images) improves performance in resource-scarce applications (e.g., medical imaging).

2.1.6 Multi-Label Classification in Safety-Critical Domains

While multi-label classification (MLC) is widely applied across domains, its deployment in *safety-critical* settings—such as healthcare, autonomous driving, environmental monitoring, and industrial process control—introduces requirements that extend beyond predictive accuracy. In such contexts, erroneous predictions may have severe or irreversible consequences, making the *quantification and communication of uncertainty* as important as predictive correctness [BBK19; Ova+19].

The role of uncertainty quantification. Uncertainty quantification enables risk-aware decision-making by estimating the confidence associated with each prediction. In medical imaging, diagnostic support systems must indicate when predictions are uncertain, prompting closer review by clinicians [Jia+12]. In autonomous driving, perception modules should flag low-confidence detections so that the control system can trigger conservative fallback strategies [FRD21]. Such mechanisms are essential not only for human oversight but also for integration into broader pipelines that are constrained by regulatory and safety standards.

Common pitfalls in conventional MLC models. Standard deep learning–based MLC models are often poorly calibrated, exhibiting systematic overconfidence in their predictions [Guo+17]. Miscalibration is particularly dangerous in safety-critical applications where probability thresholds guide decisions: underestimating uncertainty for rare but critical labels (e.g., *pneumothorax* in medical imaging) can lead to missed detections, while overestimating uncertainty for common labels may trigger unnecessary interventions. Moreover, conventional metrics such as F1-score or Hamming loss do not assess calibration quality, allowing unreliable models to appear deceptively strong under traditional evaluations.

Motivation for ensemble and conformal prediction approaches. These limitations motivate methods that provide not only accurate predictions but also reliable uncertainty estimates. Ensemble methods, which aggregate outputs from multiple diverse models, are well-established techniques for improving both accuracy and calibration in classification tasks [LPB17]. Conformal prediction (CP) offers a complementary, model-agnostic framework that produces prediction sets with formal, distribution-free coverage guarantees [VGS05; AB21]. In safety-critical MLC applications, combining ensembling with CP holds particular promise: ensembles reduce overconfidence and capture epistemic uncertainty, while CP ensures rigorous coverage control. This thesis builds upon these principles, first exploring ensemble-based CP for general MLC tasks and subsequently adapting these techniques to biomedical concept detection in radiological imaging.

2.2 Uncertainty Quantification in MLC

Machine learning models, particularly those based on deep neural networks, have achieved remarkable success across a wide range of tasks. However, their predictions are often accompanied by an implicit assumption of certainty, even in cases where the model has little evidence to support its output. In many real-world applications—especially safety-critical domains such as healthcare, autonomous driving, or legal decision-making—this overconfidence poses a significant risk. A model that cannot accurately indicate when it is uncertain may produce highly confident yet incorrect predictions, potentially leading to costly or even life-threatening consequences.

Uncertainty quantification (UQ) seeks to address this limitation by providing reliable measures of predictive confidence. Instead of returning only point predictions or label scores, a well-calibrated model can communicate the degree of uncertainty associated with each prediction, enabling risk-aware decision-making. In medical image analysis, for example, an automated concept detection system that flags uncertain predictions can alert clinicians to review specific findings more closely, improving patient safety and diagnostic reliability. This is particularly relevant in multi-label classification (MLC)

settings, where multiple interdependent labels are predicted simultaneously and where the cost of misclassification varies significantly between labels.

By equipping machine learning models with principled uncertainty estimates, we can bridge the gap between high predictive accuracy and real-world trustworthiness. This thesis focuses on methods that not only improve accuracy in MLC but also provide robust, interpretable measures of uncertainty with statistical guarantees, forming a foundation for the integration of AI systems into safety-critical workflows.

2.2.1 Types of Uncertainty

In supervised learning, uncertainty is generally categorized into two main types: *aleatoric* and *epistemic* uncertainty [DD09]. These correspond to different sources of predictive uncertainty and have distinct implications for model design and interpretation.

Aleatoric uncertainty. Aleatoric uncertainty arises from inherent variability in the data generation process. This type of uncertainty is irreducible: no matter how much data is collected, the noise present in the measurements, labeling process, or underlying phenomena cannot be eliminated. For example, in medical imaging, aleatoric uncertainty may stem from low image quality, motion artifacts, or ambiguous visual patterns that even human experts find difficult to interpret. In MLC, this may manifest when certain visual cues are shared between multiple labels, making it inherently unclear which labels apply to a given instance.

Epistemic uncertainty. Epistemic uncertainty, also known as model uncertainty, reflects a lack of knowledge about the true mapping from inputs to outputs. It is caused by limited, sparse, or biased training data, as well as model misspecification. Epistemic uncertainty is reducible: collecting more representative data or improving the model architecture can reduce it. In medical imaging, epistemic uncertainty might be high for rare conditions that appear in only a few training examples. In MLC, epistemic uncertainty is particularly important for rare labels or unusual label combinations that the model has not encountered frequently.

Interplay in multi-label settings. In practice, both types of uncertainty often coexist. For example, in a radiology image containing an unusual combination of findings, aleatoric uncertainty might arise from ambiguous image features, while epistemic uncertainty may result from the model's lack of prior exposure to similar cases. Disentangling these uncertainties can be valuable for decision-making: high aleatoric uncertainty might indicate an

inherently ambiguous case, whereas high epistemic uncertainty may suggest that further data collection or model retraining could improve performance.

Understanding the nature of uncertainty in a given prediction is essential for designing effective uncertainty quantification strategies and for interpreting model outputs responsibly, particularly in high-stakes MLC tasks.

2.2.2 Calibration and Overconfidence

While predictive accuracy is a primary performance metric in machine learning, in many applications it is equally important that a model's predicted probabilities accurately reflect the true likelihood of correctness. This property is known as *calibration* [Guo+17]. A perfectly calibrated classifier is one in which, for all predictions assigned a confidence score of p, the proportion of correct predictions is also p. For example, among all predictions to which the model assigns a confidence of 0.8, approximately 80% should be correct.

Overconfidence in modern neural networks. Empirical studies have shown that modern deep neural networks, despite achieving high accuracy, often suffer from severe miscalibration, typically manifesting as overconfidence [Guo+17]. This means that the model's predicted probabilities are systematically higher than the actual observed accuracy, especially on out-of-distribution examples or rare classes. Overconfidence is particularly problematic in safety-critical domains, where a high-probability incorrect prediction may lead to decisions with severe consequences. In multi-label classification (MLC), overconfidence can manifest for individual labels or propagate through correlated label structures, compounding the risk of erroneous predictions.

Reliability diagrams and calibration metrics. Calibration can be evaluated visually and quantitatively. Reliability diagrams (see Figure 2.1) plot the observed accuracy against predicted confidence scores, allowing deviations from the diagonal (perfect calibration) to be observed. They provide an intuitive visual assessment of how well a model's predicted probabilities correspond to actual correctness. In the diagram, each point corresponds to a bin of predictions grouped by confidence, with the x-coordinate representing the average predicted confidence and the y-coordinate representing the empirical accuracy within that bin. Deviations below the diagonal indicate overconfidence, while deviations above indicate underconfidence. The accompanying histogram shows the proportion of predictions falling into each confidence bin, illustrating the distribution of the model's confidence scores. Quantitatively, calibration is often measured using metrics such as the *Expected Calibration Error* (ECE) and the *Maximum Calibration Error* (MCE). ECE, for instance, is computed by partitioning predictions into M bins based on their predicted

probability and then calculating the weighted average of the absolute differences between accuracy and confidence in each bin:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{2.5}$$

where B_m is the set of samples in bin m, $\mathrm{acc}(B_m)$ is the empirical accuracy, $\mathrm{conf}(B_m)$ is the mean predicted confidence, and n is the total number of samples.

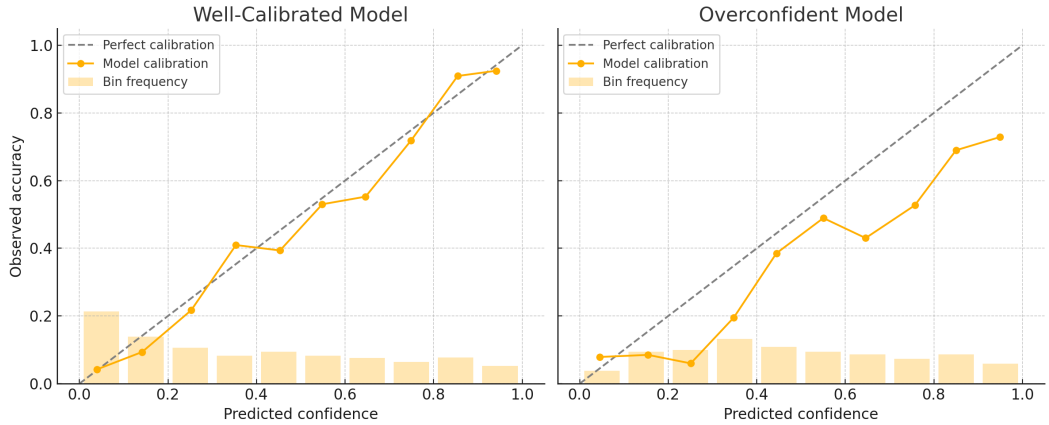


Fig. 2.1: Reliability diagrams for two models. (*Left*) A well-calibrated model, where most points lie close to the diagonal, indicating predicted probabilities align with actual observed accuracies. (*Right*) An overconfident model, where points fall below the diagonal, showing that predicted probabilities are systematically higher than true accuracies. The bars indicate the proportion of predictions in each confidence bin. Reliability diagrams provide a visual means of assessing calibration quality, with the dashed line representing perfect calibration.

Consequences of poor calibration. In safety-critical MLC tasks, poor calibration can result in two failure modes: (i) *false assurance*, where incorrect predictions are assigned high confidence and thus go unchallenged by human reviewers, and (ii) *false alarm*, where correct predictions are assigned low confidence, leading to unnecessary interventions or reviews. Both cases degrade the overall effectiveness of a human–AI collaboration and can erode trust in the system.

Motivation for improved uncertainty estimation. Addressing miscalibration is therefore essential for reliable uncertainty quantification. Techniques such as post-hoc calibration (e.g., temperature scaling, isotonic regression), Bayesian modeling, and ensemble methods have all been shown to improve calibration. In this thesis, we place particular emphasis on ensemble methods and conformal prediction, as these approaches not only enhance calibration but also provide formal guarantees on predictive uncertainty—an especially valuable property in safety-critical multi-label applications.

2.2.3 Approaches to Uncertainty Quantification

Several methodologies have been developed to quantify uncertainty in machine learning models, ranging from fully Bayesian treatments to computationally efficient post-hoc calibration techniques. This subsection reviews four widely used families of approaches, with a focus on their conceptual underpinnings, practical considerations, and relevance to multi-label classification (MLC).

Bayesian Neural Networks. Bayesian neural networks (BNNs) [Nea12] extend standard neural networks by placing probability distributions over their weights rather than learning fixed point estimates. This allows the model to capture epistemic uncertainty through a posterior distribution over parameters, from which predictions are obtained by marginalizing over the weight distribution. In practice, the true posterior is intractable and must be approximated using methods such as variational inference [**graves2011practical**; **blundell2015weight**] or Markov chain Monte Carlo (MCMC) sampling. While BNNs provide a theoretically principled framework for uncertainty estimation, they are computationally expensive to train and scale poorly to very deep architectures, limiting their adoption in large-scale MLC problems.

Monte Carlo Dropout. Monte Carlo (MC) Dropout [GG16] offers a computationally inexpensive approximation to Bayesian inference. The method leverages dropout, a regularization technique, at inference time: multiple stochastic forward passes are performed with dropout activated, producing a distribution of predictions for each input. The mean of these predictions serves as the final output, while their variance provides an estimate of epistemic uncertainty. MC Dropout is simple to implement in existing architectures, incurs minimal modifications to training, and has been successfully applied in a variety of domains. However, its uncertainty estimates can be sensitive to the chosen dropout rate and may not match the fidelity of more explicit Bayesian approaches.

Ensemble Methods. Ensemble methods combine the outputs of multiple independently trained models to improve predictive performance and obtain uncertainty estimates from the diversity of predictions [LPB17]. In the simplest case, bootstrap ensembles train each model on a different resampled subset of the training data, while *snapshot ensembles* [Hua+17a] capture multiple network states from a single training run using cyclical learning rates. In MLC, ensembles can help mitigate label imbalance effects and improve calibration by averaging over multiple diverse decision boundaries. The predictive variance across ensemble members naturally reflects epistemic uncertainty, making this approach highly compatible with conformal prediction frameworks discussed later in this thesis.

Post-hoc Calibration. Post-hoc calibration methods adjust a trained model's output scores to improve the alignment between predicted probabilities and observed accuracies, without altering the underlying model parameters. Popular techniques include *Platt scaling* [**platt1999probabilistic**], which fits a logistic regression model to the outputs; *temperature scaling* [Guo+17], which divides logits by a learned scalar temperature parameter; and *isotonic regression* [**zadrozny2002transforming**], a non-parametric method that fits a monotonic function mapping raw scores to calibrated probabilities. These methods are computationally inexpensive and easy to integrate, but they only address calibration and do not inherently improve predictive accuracy or capture epistemic uncertainty.

In summary, while Bayesian approaches offer the most principled uncertainty estimates, their high computational cost often limits practical use. MC Dropout and ensemble methods provide more scalable alternatives, with ensembles typically offering superior calibration and robustness in MLC settings. Post-hoc calibration serves as a lightweight complement to these methods, improving trustworthiness without retraining.

2.3 Conformal Prediction

Conformal Prediction (CP) is a framework for producing prediction sets that are guaranteed, under mild assumptions, to contain the true label with a user-specified probability [VGS05]. Unlike conventional classifiers, which output a single predicted label or a probability distribution over labels, CP returns a set of plausible labels whose size depends on the model's confidence in its prediction. This approach offers a rigorous, model-agnostic method for uncertainty quantification, making it particularly attractive in high-stakes decision-making domains such as medical diagnosis, autonomous driving, and financial risk assessment.

The fundamental guarantee of CP is *coverage validity*. Formally, for a given significance level $\alpha \in (0, 1)$, a CP predictor constructs a prediction set $\Gamma_\alpha(\mathbf{x})$ such that:

$$\mathbb{P}(y \in \Gamma_\alpha(\mathbf{x})) \geq 1 - \alpha, \tag{2.6}$$

where the probability is taken over the joint distribution of the training and test data. This means that, on average, at least $(1 - \alpha) \times 100\%$ of the prediction sets produced will contain the correct label. Crucially, this guarantee holds *without any assumptions* on the correctness of the underlying model, provided that the data are *exchangeable*.

CP methods differ in how they partition the available data for model training and calibration. The most common variants are:

- **Transductive CP:** The original formulation, in which the model is retrained for each test instance with and without candidate labels to compute nonconformity scores. While offering exact validity, this approach is computationally infeasible for modern large-scale problems.

- **Inductive CP:** A more practical variant that splits the data into a *proper training set* for model fitting and a *calibration set* for computing nonconformity scores. This reduces computation but retains validity under exchangeability.

- **Split Conformal Prediction:** A simplification of inductive CP where the split between training and calibration is fixed, avoiding cross-validation or retraining, making it well-suited for large datasets.

- **Mondrian CP:** A label-conditional extension that calibrates prediction sets separately for different categories (e.g., per class or per subgroup), allowing coverage guarantees to hold *conditionally* within each group. This is particularly useful in imbalanced classification, where global calibration may hide systematic under-coverage for rare labels.

In the context of this thesis, CP provides a principled framework for converting probabilistic predictions into well-calibrated, set-valued outputs with formal guarantees. Its model-agnostic nature allows it to be applied to deep neural networks, ensemble methods, and other architectures without altering the underlying learning algorithm, making it an ideal choice for integrating with the multi-label classification systems developed in this work.

2.3.1 Mathematical Framework of Conformal Prediction

Conformal Prediction (CP) is a distribution-free framework for uncertainty quantification that outputs *prediction sets* rather than point estimates, with the guarantee that the true label is contained in the set with user-specified probability $1 - \alpha$ [VGS05; SV08; AB21]. Unlike Bayesian or parametric methods, CP requires only that data be *exchangeable*, making its validity robust to model misspecification.

Problem Setup. Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the input space and $\mathcal{Y}$ the label space (e.g. $\{1, \ldots, K\}$ in classification). Given training data

$$\{Z_i\}_{i=1}^n = \{(X_i, Y_i)\}_{i=1}^n \sim P_{XY},$$

and a new input X_{n+1}, CP constructs a prediction set $\hat{C}_\alpha(X_{n+1}) \subseteq \mathcal{Y}$ such that

$$\mathbb{P}\Big(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\Big) \geq 1 - \alpha, \tag{2.7}$$

where $\alpha \in (0, 1)$ is the miscoverage rate.

Core Components.

1. **Nonconformity Measure.** A function $A : \mathcal{Z}^m \times \mathcal{Z} \to \mathbb{R}$ quantifies how "atypical" an example $z = (x, y)$ is relative to a reference set.

   - Regression: $A(S, (x, y)) = |y - \hat{f}(x)|$.
   - Classification: $A(S, (x, y)) = 1 - \hat{p}_y(x)$, where $\hat{p}_y(x)$ is the model-predicted probability for y.

2. **Data Splitting.** The dataset is partitioned into $\mathcal{D}_{\text{train}}$ (to fit the model f) and $\mathcal{D}_{\text{cal}}$ (for calibration).

3. **Calibration Scores.** For each calibration point $Z_i = (X_i, Y_i)$,

$$S_i = A(\mathcal{D}_{\text{train}}, Z_i). \tag{2.8}$$

4. **Test Scores.** For a new input X_{n+1} and candidate label $y \in \mathcal{Y}$,

$$S_{n+1}(y) = A(\mathcal{D}_{\text{train}}, (X_{n+1}, y)). \tag{2.9}$$

5. **Quantile Threshold.** The $(1 - \alpha)$ empirical quantile of calibration scores is

$$Q_{1-\alpha} = \inf \left\{ s : \frac{1}{n_{\text{cal}} + 1} \sum_{i=1}^{n_{\text{cal}}} \mathbb{I}\{S_i \leq s\} \geq 1 - \alpha \right\}. \tag{2.10}$$

6. **Prediction Set.**

$$\hat{C}_\alpha(X_{n+1}) = \{y \in \mathcal{Y} : S_{n+1}(y) \leq Q_{1-\alpha}\}. \tag{2.11}$$

Validity Guarantee.

Theorem 2.3.1 (Finite-sample validity [VGS05]). *If $\{Z_1, \ldots, Z_{n+1}\}$ are exchangeable, then Eq. (2.7) holds for any nonconformity function A.*

Proof. Exchangeability ensures that the rank of $S_{n+1}(Y_{n+1})$ among $\{S_1, \ldots, S_{n_{\text{cal}}}, S_{n+1}(Y_{n+1})\}$ is uniformly distributed. The quantile $Q_{1-\alpha}$ guarantees inclusion with probability at least $1 - \alpha$. □

Remarks.

- CP requires only exchangeability, a weaker assumption than i.i.d. sampling.
- Prediction sets adapt to model confidence: smaller under confident predictions, larger under uncertainty.
- In large label spaces, computing $S_{n+1}(y)$ for every y can be expensive, motivating approximations such as inductive CP and efficient nonconformity functions [VGS05; AB21].

2.3.2 Types of Conformal Prediction for Classification

Conformal Prediction (CP) admits several variants depending on how nonconformity scores are grouped and calibrated. In classification, the three most widely used approaches are *Standard CP*, *Mondrian CP*, and *Adaptive CP*.

Standard Conformal Prediction. Standard CP treats all calibration examples as a single pool. Nonconformity scores are computed as

$$S_i = A(\mathcal{D}_{\text{train}}, (x_i, y_i)),$$

and for a new input x, test scores $S_{n+1}(y)$ are obtained for each $y \in \mathcal{Y}$. The prediction set is

$$\hat{C}_\alpha(x) = \{y \in \mathcal{Y} : S_{n+1}(y) \leq Q_{1-\alpha}\}, \tag{2.12}$$

where $Q_{1-\alpha}$ is the $(1 - \alpha)$ quantile of calibration scores. This method is model-agnostic and simple but can produce overly conservative sets when class imbalance is severe, as majority-class scores dominate the quantile.

Mondrian Conformal Prediction (Label-Conditional). Mondrian CP [VGS05] mitigates imbalance by conditioning calibration on the true label. For each class y, a separate quantile $Q^{(y)}_{1-\alpha}$ is computed from calibration points with label y. The prediction set is then

$$\hat{C}^{\text{Mondrian}}_\alpha(x) = \{y \in \mathcal{Y} : S_{n+1}(y) \leq Q^{(y)}_{1-\alpha}\}. \tag{2.13}$$

This ensures *marginal coverage per class*, making it particularly useful in imbalanced domains such as biomedical imaging, where rare labels must not be systematically undercovered.

Adaptive Conformal Prediction. Adaptive CP [**romano2019conformalized**] produces *instance-dependent* sets by adjusting the effective confidence level according to input difficulty. A function $g(\cdot)$ maps test scores to an adaptive miscoverage rate α_{eff}, yielding smaller sets for confident cases and larger ones for ambiguous inputs. While marginal validity is preserved under mild conditions, designing $g(\cdot)$ requires care to avoid label-dependent bias.

Comparison.

- **Standard CP:** simplest, broadly applicable, but inefficient under imbalance.
- **Mondrian CP:** class-conditional coverage, effective for imbalanced or safety-critical problems.
- **Adaptive CP:** more informative, input-dependent sets, but adds complexity.

The choice depends on task requirements: standard CP suffices for balanced problems, while Mondrian CP is preferable in imbalanced or critical domains, and adaptive CP is best suited for applications demanding highly informative, variable-sized prediction sets.

2.3.3 Conformal Prediction for Multi-Label Classification

Extending Conformal Prediction (CP) to multi-label classification (MLC) introduces unique challenges due to the structured, high-dimensional nature of the label space. Given an instance $\mathbf{x} \in \mathcal{X}$ with binary label vector $\mathbf{y} \in \{0,1\}^L$, the goal is to construct a prediction set $\hat{C}_\alpha(\mathbf{x}) \subseteq \{1,\dots,L\}$ that contains the true label set with probability at least $1-\alpha$:

$$\mathbb{P}\left(\mathbf{y} \subseteq \hat{C}_\alpha(\mathbf{x})\right) \geq 1-\alpha,$$

while keeping $\hat{C}_\alpha(\mathbf{x})$ as small as possible.

Independent Per-Label CP. A standard approach applies CP independently to each label ℓ, analogous to the Binary Relevance (BR) strategy. For each ℓ, nonconformity scores $S^{(\ell)}$ are calibrated, yielding thresholds $Q^{(\ell)}_{1-\alpha}$ and per-label predictions:

$$\hat{C}^{(\ell)}_{\alpha}(\mathbf{x}) = \begin{cases} \{\ell\}, & S^{(\ell)}_{n+1} \leq Q^{(\ell)}_{1-\alpha}, \\ \emptyset, & \text{otherwise.} \end{cases}$$

The final set is the union $\hat{C}_{\alpha}(\mathbf{x}) = \bigcup_{\ell} \hat{C}^{(\ell)}_{\alpha}(\mathbf{x})$. This ensures marginal validity for each label [VGS05; Lei+18], but ignores inter-label dependencies, often leading to overly conservative (large) prediction sets.

Accounting for Label Dependencies. Real-world MLC tasks exhibit strong correlations (positive, negative, or conditional) between labels—for example, *pneumonia* co-occurring with *lung opacity*. Independent CP fails to exploit these relationships, forcing coverage guarantees that hedge against all plausible label combinations. To address this, several extensions have been proposed:

- **Tree-structured CP.** Cauchois et al. [CGD21a] introduced tree-structured classifiers with conformal scoring, achieving efficient confidence sets by hierarchically partitioning the label space.
- **Hierarchical CP with multiple testing.** Tyagi and Guo [TG24] formulated MLC as a multiple hypothesis testing problem, using split-conformal p-values with hierarchical corrections (e.g., Bonferroni) to guarantee family-wise error control.
- **Correlation-aware nonconformity.** Katsios and Papadopoulos [KP24] proposed Mahalanobis-distance-based nonconformity measures, capturing correlations between classifier errors across labels and yielding more efficient prediction sets.

Coverage in Multi-Label CP. While per-label CP guarantees

$$\mathbb{P}(y_{\ell} \in \hat{C}^{(\ell)}_{\alpha}(x)) \geq 1 - \alpha, \quad \forall \ell,$$

it does not ensure joint coverage of the entire label vector. Achieving vector-level coverage requires structured CP frameworks—such as hierarchical testing [TG24] or tree-based approaches [CGD21a]—which explicitly model dependencies and control family-wise error.

2.3.4 Metrics for Evaluating Conformal Predictors

The evaluation of Conformal Prediction (CP) centers on two criteria: *validity* and *efficiency* [VGS05; AB21]. Validity ensures that prediction sets achieve the desired coverage probability (typically $1 - \alpha$), while efficiency reflects how small and informative these sets are. Both aspects are critical: trivially valid sets may include all labels, but are uninformative; conversely, very compact sets may sacrifice coverage.

We distinguish between single-label and multi-label classification.

Validity Metrics. **Single-label classification.** In the single-label setting, each input x has a true label $y \in \mathcal{Y}$. The CP guarantee ensures marginal coverage:

$$\mathbb{P}\Big(y \in \hat{C}_\alpha(x)\Big) \geq 1 - \alpha, \tag{2.14}$$

under the assumption of exchangeability. Empirically, coverage is estimated as

$$\widehat{\text{Coverage}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\Big\{y_i \in \hat{C}_\alpha(x_i)\Big\}. \tag{2.15}$$

Multi-label classification. For MLC, each input x is associated with a binary label vector $\mathbf{y} \in \{0, 1\}^L$. Two common coverage notions are used:

- **Marginal coverage (per label):** The probability that each true label ℓ is included in the prediction set:

$$\frac{1}{L} \sum_{\ell=1}^{L} \mathbb{P}\Big(\ell \in \hat{C}_\alpha(x) \;\Big|\; y_\ell = 1\Big) \geq 1 - \alpha. \tag{2.16}$$

- **Empirical coverage (per instance):** The proportion of true labels captured per instance, averaged across the dataset:

$$\widehat{\text{Coverage}} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\hat{Y}(x_i) \cap Y_i|}{\max(1, |Y_i|)}, \tag{2.17}$$

  where $\hat{Y}(x_i)$ is the predicted set and Y_i the ground-truth set.

Efficiency Metrics. The standard efficiency measure is the *average set size*:

$$\widehat{\text{Size}} = \frac{1}{n} \sum_{i=1}^{n} |\hat{C}_\alpha(x_i)|, \tag{2.18}$$

which reflects informativeness (smaller is better). In MLC, it corresponds to the average number of labels predicted per instance.

Composite Metrics (Single-Label). Beyond coverage and set size, additional criteria capture residual uncertainty in single-label CP [VGS05]:

- **Observed Unconfidence (OU):** The average maximum p-value assigned to incorrect labels:

$$\text{OU} = \frac{1}{n} \sum_{i=1}^{n} \max_{y \neq y_i} p_y(x_i).$$

- **Observed Fuzziness (OF):** The average total p-value mass of incorrect labels:

$$\text{OF} = \frac{1}{n} \sum_{i=1}^{n} \sum_{y \neq y_i} p_y(x_i).$$

Lower values of OU and OF indicate sharper, less ambiguous predictions.

Trade-off Between Validity and Efficiency. Prediction sets can trivially satisfy validity by including all labels, but such sets lack utility. Conversely, aggressively minimizing set size can lead to undercoverage. The core challenge in CP is to balance *validity* (coverage guarantees) with *efficiency* (compactness), a trade-off that becomes especially demanding in high-dimensional or multi-label problems [AB21].

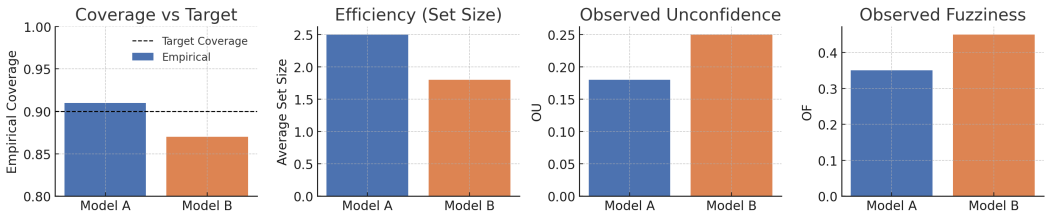


Fig. 2.2: Illustration of CP evaluation metrics for two predictors (A and B). **Left:** Empirical coverage compared to the target ($1 - \alpha = 0.90$). **Center-left:** Average set size (efficiency). **Center-right:** Observed Unconfidence (OU). **Right:** Observed Fuzziness (OF). Model A is conservative (larger sets, higher coverage), while Model B is sharper but less reliable.

2.3.5 Why Combine Conformal Prediction with Ensembles

Although Conformal Prediction (CP) provides distribution-free coverage guarantees, its efficiency depends heavily on the quality and stability of the underlying model. Ensemble learning, which aggregates multiple predictors, naturally complements CP by improving calibration, reducing variance, and enriching dependency modeling [Die00; LPB17; AB21].

Calibration Stability. The reliability of CP hinges on well-calibrated nonconformity scores. Single models, particularly deep neural networks, often exhibit miscalibration and overconfidence. Ensembles mitigate this by averaging predictions, producing smoother probability estimates and more stable p-values, which lead to tighter prediction sets [Guo+17; LPB17].

Variance Reduction. Prediction-set size in CP is sensitive to variance in nonconformity scores. Ensembles reduce such variance through diversity in initialization, architectures, or data resampling, thereby improving efficiency without compromising finite-sample validity [Die00].

Dependency Modeling. In multi-label classification, inter-label correlations are crucial. Heterogeneous ensembles or ensembles trained on different label subsets can capture complementary dependency structures. When combined with CP, these ensembles yield prediction sets that remain valid while being more informative [Rea+11; TV07].

2.4 Ensemble Learning

Ensemble learning combines multiple predictive models (*base learners*) to form a single aggregated predictor. The central intuition is that while individual models may suffer from bias, variance, or limited representational capacity, an ensemble can leverage their complementary strengths to improve generalization—an idea closely related to the "wisdom of the crowd" principle [Die00; Kun14].

Theoretical support comes from the bias–variance decomposition: if base learners are diverse and their errors are not perfectly correlated, aggregation reduces variance without substantially increasing bias. This is particularly advantageous in high-variance models such as deep neural networks, where different random seeds or training subsets can yield markedly different solutions.

Beyond accuracy, ensembles often enhance *calibration* and *uncertainty quantification*. By averaging outputs across models, ensembles tend to produce smoother probability estimates and reduce overconfidence, leading to more reliable predictive uncertainty [LPB17; Guo+17]. Such properties make them especially valuable in safety-critical domains like medical imaging, where both accuracy and trustworthy uncertainty estimates are essential.

Ensembles are widely applied, from decision-tree methods such as Random Forests to deep ensembles in computer vision and NLP. In multi-label classification (MLC), they mitigate class imbalance, improve predictions for rare labels, and better capture inter-label dependencies. When combined with Conformal Prediction (CP), ensembles further stabilize coverage, reduce overly conservative prediction sets, and incorporate model diversity into nonconformity scoring—a central theme explored in this thesis.

2.4.1 Formal Framework and Notation

Let $\mathcal{H}$ denote a hypothesis space of predictive models mapping from the input space $\mathcal{X} \subseteq \mathbb{R}^d$ to the output space $\mathcal{Y}$. In classification, a model $h \in \mathcal{H}$ may output either: (i) a predicted class label $\hat{y} \in \mathcal{Y}$, or (ii) a probability vector $\hat{\mathbf{p}}(x) \in [0,1]^{|\mathcal{Y}|}$ with $\sum_{y \in \mathcal{Y}} \hat{p}_y(x) = 1$ (single-label case) or $\hat{\mathbf{p}}(x) \in [0,1]^L$ in the multi-label case.

An *ensemble* consists of M base learners

$$\mathcal{E} = \{h_1, h_2, \ldots, h_M\}, \quad h_m \in \mathcal{H},$$

trained on the same data distribution but differing in initialization, data sampling (e.g., bootstrapping), feature subsets, architectures, or hyperparameters. The ensemble prediction function aggregates their outputs:

$$F(x) = \mathcal{A}(h_1(x), h_2(x), \ldots, h_M(x)), \tag{2.19}$$

where $\mathcal{A}$ is an aggregation rule such as:

- **Majority voting** for hard labels:

$$\hat{y}_{\text{ens}}(x) = \arg\max_{y \in \mathcal{Y}} \sum_{m=1}^{M} \mathbb{I}\{h_m(x) = y\}.$$

- **Probability averaging** for soft outputs:

$$\hat{\mathbf{p}}_{\text{ens}}(x) = \tfrac{1}{M} \sum_{m=1}^{M} \hat{\mathbf{p}}_m(x),$$

which is especially useful when coupled with probabilistic frameworks such as Conformal Prediction, since it preserves calibration properties.

Bias–Variance Perspective. For a true label Y and predictor $\hat{f}(x)$, the expected squared error decomposes as

$$\mathbb{E}\Big[(\hat{f}(x) - Y)^2\Big] = \underbrace{(\mathbb{E}[\hat{f}(x)] - Y)^2}_{\text{Bias term}} + \underbrace{\mathbb{E}\Big[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2\Big]}_{\text{Variance}} + \sigma^2.$$

where σ^2 is irreducible noise. Ensembles reduce the variance term by averaging across diverse learners, typically without substantially increasing bias [Die00].

Relation to Uncertainty Quantification. The distribution of ensemble outputs $\{\hat{\mathbf{p}}_m(x)\}_{m=1}^M$ captures *epistemic uncertainty* (due to model variability), while their mean reflects consensus. This dual perspective makes ensembles particularly valuable when combined with Conformal Prediction, where model diversity stabilizes nonconformity scores and yields more efficient prediction sets.

2.4.2 Categories of Ensemble Methods

Ensemble methods differ in how they induce diversity among base learners and aggregate predictions. Diversity is essential: if learners produce identical outputs, ensembling provides no benefit [Die00; KW03]. The main families are summarized below.

Bagging. Bootstrap aggregating (bagging) [Bre96] trains each learner on a bootstrap sample of the data, reducing variance by averaging predictions. It is particularly effective for unstable learners such as decision trees; Random Forests extend this by randomizing feature selection. Bagging is less effective for already low-variance models.

Boosting. Boosting builds learners sequentially, reweighting training examples so that later models focus on harder cases [FS97; SS99]. Gradient boosting frameworks (e.g., XGBoost, LightGBM) are highly competitive on tabular data, though boosting can overfit in noisy datasets.

Stacking. Stacked generalization [Wol92] combines diverse base learners by training a meta-learner on their outputs, usually via out-of-fold predictions to prevent overfitting. It flexibly leverages heterogeneous models but increases computational cost.

Random Subspaces. The random subspace method [Ho98] trains learners on random feature subsets, which is effective in high-dimensional domains such as text or images. Often combined with bagging (as in Random Forests), it may degrade accuracy if key features are excluded.

Snapshot Ensembles. Snapshot ensembles [Hua+17a] create multiple "snapshots" of a single network by cycling the learning rate and saving models at different local minima. They provide diverse predictors at nearly the cost of training one model, though with lower diversity than fully independent training.

Deep Ensembles. Deep ensembles [LPB17] train multiple neural networks independently with different seeds and data shuffles. They improve accuracy and calibration, making them popular for uncertainty estimation, but they are computationally expensive when M is large.

Hybrid Methods. Hybrid ensembles combine mechanisms (e.g., deep bagging, feature-randomized boosting, stacked subspace models), often achieving further gains but at the cost of greater complexity in design and tuning.

2.4.3 Ensemble Diversity and Its Quantification

The effectiveness of an ensemble depends critically on the *diversity* among its base learners: if models make correlated errors, aggregation offers little benefit over a single predictor [Die00; KW03]. Diversity reduces the covariance of errors across learners, thereby lowering the variance component of ensemble error [HS90].

Sources of Diversity. Diversity can be induced through (i) data sampling (e.g., bagging), (ii) feature sampling (random subspaces), (iii) model heterogeneity (stacking different algorithms), (iv) randomization in training (seeds, shuffling), and (v) hyperparameter variation. These mechanisms encourage complementary decision boundaries and reduce correlated errors.

Quantifying Diversity. Several measures capture pairwise diversity among classifiers. The **Q-statistic** [KW03] is defined as

$$Q_{ij} = \frac{N_{11}N_{00} - N_{01}N_{10}}{N_{11}N_{00} + N_{01}N_{10}},$$

where N_{ab} counts instances where classifier i and j are correct ($a = 1$) or incorrect ($a = 0$). Values near 1 indicate identical predictions (no diversity), 0 independence, and negative

values complementary errors (high diversity). Other metrics include the **disagreement measure**,

$$\mathrm{Dis}_{ij} = \frac{N_{01} + N_{10}}{N_{11} + N_{00} + N_{01} + N_{10}},$$

and correlation coefficients between correctness indicators. Ensemble-level diversity is typically averaged across all pairs.

Balancing Accuracy and Diversity. High diversity alone is not sufficient: base learners must also be accurate. Effective ensembles strike a trade-off between individual accuracy and disagreement [KW03]. Excessive diversity obtained by weakening base learners reduces overall performance, while low diversity yields redundancy.

2.4.4 Ensemble Aggregation Strategies in Classification

Once base learners are trained, their predictions must be combined into a single, final output, and the choice of aggregation strategy can strongly influence ensemble performance, particularly in multi-label classification (MLC) and uncertainty quantification settings [Die00; Kun14].

Two of the most common aggregation approaches are **majority voting** and **probability averaging**. In majority voting, each base classifier casts one vote for its predicted class, and the class with the most votes is chosen; in MLC, this is applied per label, with each model voting for presence or absence. Majority voting is simple, interpretable, and robust when base learners are diverse and have comparable accuracy. Probability averaging, by contrast, combines the probability distributions of base learners (or per-label probabilities in MLC) and averages them; labels are then predicted positive if their mean probability exceeds a threshold. This approach produces smoother, better-calibrated outputs, which is particularly advantageous in settings that integrate ensembles with conformal prediction [LPB17].

In many applications, base learners differ in predictive performance or calibration quality, motivating **weighted aggregation**. Weighted averaging assigns larger influence to stronger models:

$$\hat{p}(y|x) = \frac{\sum_{m=1}^{M} w_m \, \hat{p}_m(y|x)}{\sum_{m=1}^{M} w_m},$$

where w_m is the weight of model m, often derived from validation metrics such as accuracy, F1-score, or calibration error, or learned automatically via a meta-model as in stacking [Wol92].

In MLC, using a single threshold (e.g., $\tau = 0.5$) for binarizing probabilities can be suboptimal due to class imbalance and heterogeneous calibration across labels. **Per-label threshold optimization** therefore assigns distinct thresholds τ_ℓ for each label ℓ, tuned on a validation set to maximize metrics such as macro-F1 [TKV10]. More advanced methods, such as coordinate ascent, jointly optimize thresholds across labels to avoid locally suboptimal solutions that arise from treating labels independently.

When ensembles are applied for **uncertainty estimation**, aggregation must also capture not only the mean prediction but also the dispersion across models. Variance of predicted probabilities can serve as a measure of epistemic uncertainty, while the entropy of the averaged probability distribution reflects overall predictive uncertainty [LPB17]. These signals are complementary: variance highlights disagreement among models, whereas entropy measures uncertainty in the ensemble consensus. In conformal prediction, aggregated probabilities are used to compute nonconformity scores, and variance or entropy estimates can inform adaptive prediction set construction [AB21]. By combining both central tendency and dispersion across base learners, ensembles provide richer and more reliable uncertainty quantification than single classifiers.

2.5 Ensembles for Multi-Label Classification

Ensemble methods are particularly valuable in multi-label classification (MLC) because they can simultaneously address multiple challenges inherent to the task, including class imbalance, rare label prediction, and modeling of label dependencies.

2.5.1 Addressing Label Imbalance and Rare Labels

In MLC datasets, label frequencies often follow a long-tail distribution: a few labels occur frequently while many are rare but potentially critical. Ensembles mitigate this by:

- **Model diversity**: Different base learners may capture different parts of the label space, increasing the chance of correctly predicting rare labels.
- **Resampling strategies**: Combining bagging with label-aware sampling (e.g., stratified sampling per label) ensures that minority labels appear more frequently in training subsets.
- **Cost-sensitive aggregation**: Weights in the ensemble can be adjusted to favor models performing better on rare labels.

Empirically, ensemble averaging tends to smooth over extreme probability predictions, improving recall for rare labels while controlling false positives.

2.5.2 Leveraging Label Dependencies

Many labels in MLC are correlated — either positively (e.g., *pneumonia* and *lung opacity*) or negatively (e.g., *fracture* and *normal finding*). Ensembles can model these dependencies in several ways:

- **Classifier Chains (CC) ensembles**: Each chain models sequential label dependencies; aggregating multiple random chains reduces sensitivity to label order.
- **Label Powerset (LP) ensembles**: Each base learner models joint label combinations; aggregating over different subsets of labels reduces the combinatorial explosion.
- **Graph-based ensembles**: Base learners incorporate graph neural networks (GNNs) or attention layers to propagate information between labels; ensemble diversity arises from different graph structures or message-passing depths.

2.5.3 Examples of MLC-Specific Ensemble Architectures

Several ensemble designs are tailored for multi-label tasks:

- **RAkEL (Random k-Labelsets)** [TV11]: Builds multiple LP classifiers, each trained on a random subset of k labels; predictions are aggregated per label via majority voting.
- **Ensemble Classifier Chains (ECC)** [Rea+11]: Trains multiple CC models with different random label orders, improving robustness to order sensitivity.
- **Hybrid neural-symbolic ensembles**: Combines deep feature extractors (e.g., CNNs, Transformers) with symbolic multi-label learners (e.g., ML-kNN), aggregating predictions to benefit from both representation learning and explicit label dependency modeling.
- **Conformal-ensemble hybrids**: Integrates conformal prediction into each base model of an MLC ensemble, allowing aggregation not only of predictions but also of calibrated uncertainty estimates.

2.6 Related Work

2.6.1 Ensembling via resampling: ACP and CCP

Early work on CP ensembling aimed to recover efficiency lost by split CP while retaining practicality. *Aggregated conformal prediction (ACP)* averages (or otherwise combines) per-split conformal outputs across many resamples; *cross-conformal prediction (CCP)* mirrors cross-validation and combines fold-wise CP results [CEN14; Vov15]. Empirically, both improve informational efficiency (smaller sets) relative to a single split. However, theory shows that naive p-value averaging can be conservative (over-coverage) unless additional stability conditions hold for the score/model; calibration analyses and refined definitions quantify when ACP/CCP are near-valid and when they inflate set sizes [Lin+17]. Practical takeaways: (i) ensembling over splits reduces variance of set size; (ii) mean aggregation may be overly conservative; (iii) median or more robust combiners can mitigate instability [Lin+17].

2.6.2 Aggregating across models: voting, scores, and sets

Beyond resampling the same model, several works study how to aggregate predictions from multiple, potentially heterogeneous models. Cherubin [Che19] analyzed majority voting of conformal classifiers, deriving finite-sample guarantees for label-wise coverage in classification and showing how independence assumptions across models shape the bounds. Other approaches move beyond voting to score-based aggregation. Rivera et al. [OPT25] formalize multi-score aggregation with coverage guarantees and demonstrate improved efficiency in both classification and predict-then-optimize tasks. Luo and Zhou [LZ25] extend this idea by learning weights over multiple nonconformity scores, with the aim of minimizing expected set size while retaining coverage. Gasparin and Ramdas [GR24] instead propose an online scheme where model-wise sets are combined with time-varying weights updated according to performance, maintaining marginal coverage over time while adapting to nonstationary settings. Finally, Yang and Kuchibhotla [YK25] study the problem of selecting or aggregating from a family of conformal regions to yield the smallest-width valid region, giving algorithms with either approximate coverage and exact minimality, or finite-sample coverage with near-minimal width. Collectively, these works move the field from bagging p-values to explicitly learning how to combine multiple sources of conformal evidence, whether through voting, weighted score aggregation, or adaptive region selection.

2.6.3 Specialized ensemble CP constructions

In high dimensions, *Random Projection Ensemble CP* (RPECP) first ensembles random projections and base classifiers, then conformalizes and finally uses a designed voting rule to output both a point label and a calibrated set. The pipeline explicitly targets improved statistical efficiency (fewer false labels in the set) while honoring coverage [Qia+24]. This is conceptually related to random-projection ensemble classification, but adapted to produce conformal sets rather than hard labels.

2.6.4 Multilabel and multiclass conformal prediction

For multiclass and multilabel settings, CP must control set size growth while preserving coverage. Foundational results construct valid confidence sets and propose tree-structured classifiers to address label interactions and avoid exponentially large sets [CGD21b]. In multilabel specifically, Papadopoulos introduced *cross-conformal multilabel* predictors that treat labelsets as structured outputs and combine foldwise CP to provide calibrated confidence over label subsets [Pap14; Pap22]. More recently, Tyagi and Guo propose a *tree-based multilabel CP* that frames labelset selection as hierarchical multiple testing on split-conformal p-values, controlling family-wise error while yielding compact, dependency-aware prediction sets [TG23]. These strands differ in how they encode label dependence (label-powerset vs. hierarchical trees), how they calibrate (split vs. cross-conformal), and whether they aggregate over models/scores or over structured label hypotheses.

2.6.5 Medical Image Classification

Medical image classification is a specialized branch of computer vision focused on diagnosing diseases, identifying abnormalities, and supporting clinical decision-making from medical imaging data. Unlike generic image classification, which benefits from large-scale benchmarks such as ImageNet [Den+09], medical applications face challenges of limited annotated datasets, strict regulatory standards, and the demand for interpretability in high-stakes settings [KPA20]. With the advent of deep learning, convolutional neural networks (CNNs) have become the standard approach, replacing hand-engineered features with end-to-end representation learning. A landmark example is CheXNet [Raj+17], a DenseNet [Hua+17b]-based system fine-tuned for pneumonia detection in chest X-rays, which demonstrated the effectiveness of transfer learning from natural image pretraining. Subsequent work extended this paradigm to architectures such as ResNet [He+16], EfficientNet [TL19], and Vision Transformers (ViTs) [Dos+20], often combined into ensembles to enhance robustness and calibration. In parallel, retrieval-based approaches such as k-NN classifiers leveraging CNN encoders also proved competitive in ImageCLEFmedical, assigning concepts by propagating labels from visually similar training images [KPA19;

Cha+21]. Although later surpassed by CNN–FFNN classifiers, these methods highlighted the usefulness of instance-based reasoning in medical concept detection. In recent years, ensemble CNN–FFNN systems have consistently secured top positions, including second place in the 2024 ImageCLEFmedical concept detection task, confirming the competitiveness of approaches developed by AUEB's NLP group [Pel+19; Pel+20; PBG+21; RBG+22; Kal+23b]. Beyond 2D classification, three-dimensional CNNs such as V-Net [MNA16] have been developed to process volumetric data (CT, MRI), proving particularly effective for tumor and organ segmentation, while generative approaches like GANs [Shi+18] and diffusion models [DN21; Kha+22] have been employed for data augmentation, addressing the scarcity of annotated training examples. Finally, interpretability remains a central requirement in clinical deployment, with techniques such as Grad-CAM [Sel+20] providing saliency heatmaps that highlight the regions most influential to a model's prediction. Overall, medical image classification has evolved into a domain where transfer learning, architectural innovation, ensemble modeling, retrieval-based reasoning, and uncertainty-aware interpretability intersect to meet the dual demands of predictive performance and clinical reliability, as also reflected in prior work from AUEB's NLP group and related research contributions [KPA19; KPA20; Cha+21; Cha+22; Cha25].

3 Methodology

In this chapter, we present the methodological framework adopted in this thesis. The methodology is divided into two main parts, reflecting the two distinct but complementary contributions of this work. The first part focuses on ensemble methods for multilabel classification with conformal prediction, where we explore different model architectures, ensemble strategies, and uncertainty quantification mechanisms. The second part describes the approach developed for the *ImageCLEFmedical 2025* challenge, where we participated in the *Concept Detection* task using convolutional neural networks, feed-forward classifiers, and ensemble aggregation techniques. Together, these methodologies form the basis of our experimental investigations presented in the following chapter.

3.1 Part I: Multilabel Ensemble Methods

In the first part of this Chapter, we focus on the task of **multilabel classification (MLC)** with an emphasis on ensemble learning and uncertainty quantification through conformal prediction. The objective is to design methods that can handle inputs associated with multiple labels simultaneously, while also providing rigorous statistical guarantees on prediction reliability. This part therefore introduces the general problem formulation, describes the base classifiers employed, and develops ensemble conformal prediction strategies together with their theoretical properties. Experimental evaluation of these methods is presented later in Chapter 5.

3.1.1 Problem Formulation

We consider the task of multilabel classification (MLC), where each input $x \in \mathbb{R}^d$ can be associated with multiple labels drawn from a label set of size L. Formally, let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the input space and $\mathcal{Y} = \{0, 1\}^L$ the label space, where each $y \in \mathcal{Y}$ is a binary vector indicating the presence or absence of each label. Given a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, the objective is to learn a function $f : \mathcal{X} \to [0, 1]^L$ that produces per-label confidence scores. These scores can then be thresholded to yield a predicted label vector $\hat{y} \in \{0, 1\}^L$ for each input x.

In addition to standard evaluation criteria for MLC, such as macro-F1 we also assess the quality of prediction sets through metrics including empirical coverage, marginal coverage,

and average set size. Together, these measures provide a comprehensive assessment of both predictive performance and the reliability of uncertainty quantification. The methodology presented in this thesis is therefore designed to balance predictive accuracy, calibration, and interpretability in multilabel classification tasks.

3.1.2 Base Classifiers

To establish a diverse foundation for multilabel classification, we employ a range of base learners that span linear, shallow, and deep neural architectures. These models are used both for standalone evaluation and as components of ensemble strategies. By incorporating architectures of varying complexity and inductive biases, we are able to assess not only their individual performance but also the benefits of diversity when aggregated in ensembles. The following subsections describe the models considered:

Logistic Regression (LR). A simple yet effective linear classifier, logistic regression is trained independently for each label under the binary relevance paradigm. Each model minimizes the logistic loss and outputs probabilities that reflect the likelihood of a label being present. Despite its simplicity, LR provides a strong baseline and is often competitive in multilabel tasks, particularly for labels with clear linear separability.

Stochastic Gradient Descent (SGD). We also consider linear classifiers trained with stochastic gradient descent, which are well-suited for high-dimensional settings due to their efficiency in online updates. To ensure reliable probability estimates, these classifiers are calibrated using Platt scaling (via `CalibratedClassifierCV`), thereby producing outputs suitable for uncertainty quantification.

Multilayer Perceptron (MLP). The MLP represents a shallow neural architecture with a single hidden layer equipped with ReLU activations. Training is performed using binary cross-entropy loss across labels. Although lightweight compared to more advanced deep architectures, MLPs can capture non-linear dependencies between features and serve as an important bridge between linear and deep models.

Recurrent Neural Network (RNN). For sequential modeling, we employ a unidirectional LSTM network applied to fixed-length CLIP embeddings. The recurrent layer captures temporal or structural patterns in the embeddings, and its output is passed through a dense layer with sigmoid activation to yield per-label probabilities. This design allows the model to exploit contextual dependencies in the feature space beyond static feedforward processing.

Transformer Encoder. To capture long-range dependencies and richer contextual information, we include a two-layer transformer encoder with multi-head self-attention. The model processes CLIP embeddings as tokens and aggregates contextualized representations before passing them through a shared linear classifier that outputs label-specific confidence scores. Transformers have demonstrated strong performance in multilabel tasks due to their ability to model complex relationships across input features.

MLP-Mixer. Finally, we consider the MLP-Mixer, a lightweight alternative to attention-based models. This architecture applies token mixing and channel mixing operations, combined with layer normalization, to the CLIP embeddings. The resulting representations are fed into a sigmoid-activated output layer for multilabel probability estimation. The MLP-Mixer balances expressiveness with computational efficiency, making it a suitable candidate for ensemble inclusion.

All classifiers are trained independently for each label, following the binary relevance strategy to ensure scalability across large label spaces. Each model produces calibrated probability estimates that can be directly thresholded in standard classification settings or transformed into nonconformity scores for conformal prediction. When used in ensembles, their outputs are combined through label- and model-specific aggregation schemes, allowing us to construct prediction sets that jointly emphasize accuracy, diversity, and theoretical coverage guarantees.

3.1.3 Ensemble Learning

Ensemble learning constitutes a central component of our methodology, as it offers a principled way to enhance predictive robustness and mitigate the limitations of individual models. By combining multiple classifiers, ensembles can reduce variance, improve calibration, and provide more reliable uncertainty estimates. In this work, ensembles are considered both as independent baselines and as integral elements of the conformal prediction framework. We investigate three broad categories of ensembles: homogeneous, heterogeneous, and stacked approaches.

Homogeneous Ensembles. In the homogeneous setting, we train M independent instances of the same base classifier (e.g., logistic regression or MLP), each on bootstrap-resampled versions of the training data. The intuition behind this approach is that resampling introduces diversity in the decision boundaries of otherwise identical models, thereby reducing the risk of overfitting to specific idiosyncrasies of the data. At inference time, the predictions of the M models are combined using one of the following aggregation schemes:

- **Majority Voting (MV):** Binary predictions from each model are aggregated by a simple majority rule, with the most frequently predicted label assignment chosen as the final decision.

- **Probability Averaging (PA):** Rather than relying on hard decisions, the probabilistic outputs of the models are averaged across instances. The aggregated probabilities are then thresholded (e.g., at 0.5) to form the final prediction.

This strategy is particularly effective for reducing variance and improving calibration in settings where base models are sensitive to random initialization or stochastic data sampling.

Heterogeneous Ensembles. We also construct ensembles composed of diverse model architectures, combining both linear learners (LR, SGD) and non-linear deep models (MLP, RNN, Transformer, MLP-Mixer). Each classifier is trained per label under the binary relevance assumption, and their outputs are aggregated at inference time. We evaluate the following schemes:

- **Majority Voting (MV):** Binary predictions from all models are combined via unweighted majority.

- **Probability Averaging (PA):** Probabilistic outputs from heterogeneous models are averaged and thresholded to determine final label assignments.

- **F1-Weighted Voting:** To introduce label-specific adaptivity, predictions are weighted by each model's F1 score on a held-out validation set. This ensures that models demonstrating stronger performance on a particular label exert greater influence on the ensemble's final decision.

By leveraging architectural diversity, heterogeneous ensembles aim to capture a wider range of representational patterns, thereby improving both predictive accuracy and robustness.

Stacked Ensembles. To further enhance predictive quality, we implement a stacked ensemble approach. In this framework, the predictions of multiple base classifiers (e.g., LR, SGD, MLP) on a calibration set are used as input features to train a meta-classifier. We employ logistic regression as the meta-learner, trained separately for each label to output calibrated probabilities. At test time, the meta-classifier aggregates the predictions of the base models into a refined probability estimate, which is then post-processed using conformal prediction thresholds to yield the final prediction set. Unlike majority voting or

averaging, stacking allows the model to learn data-driven weighting schemes, effectively discovering how to best combine the strengths of different base learners. Moreover, by incorporating conformal calibration at the meta-level, this approach preserves formal coverage guarantees while benefiting from the flexibility of learned aggregation.

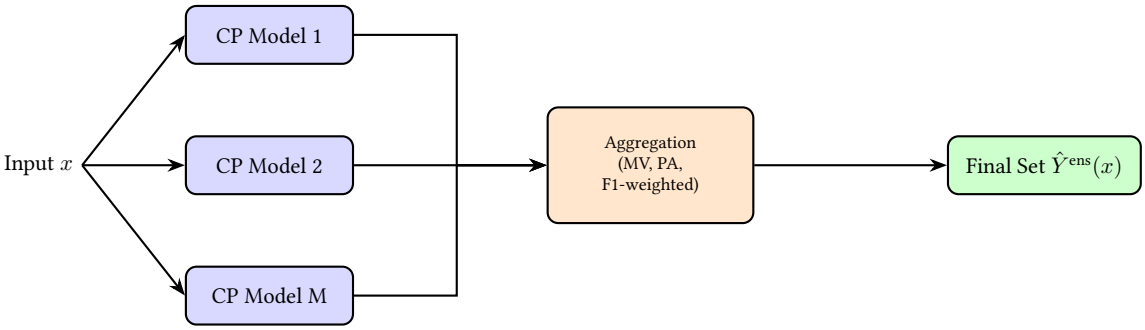


Fig. 3.1: Conceptual diagram of Ensemble Conformal Prediction (ECP). Each input is processed by multiple CP-calibrated base models. Their outputs are aggregated (via majority voting, probability averaging, or weighted voting), yielding the final prediction set.

3.1.4 Conformal Prediction

An important component of this work is the integration of conformal prediction (CP) into the multilabel classification setting. Conformal prediction provides a model-agnostic framework for generating uncertainty-aware outputs with formal statistical guarantees. Unlike standard classifiers that return point predictions, CP produces *prediction sets* that, with high probability, contain the true labels. This makes CP particularly suitable for applications where reliability and transparency are as critical as raw predictive accuracy.

In this thesis, we employ the **Mondrian conformal prediction** framework, which adapts the classical CP approach to the multilabel scenario by calibrating predictions separately for each label. The central objective is to construct prediction sets that maximize the inclusion of true labels while minimizing the inclusion of spurious ones, under a user-defined miscoverage rate α (e.g., $\alpha = 0.1$ corresponds to 90% target coverage).

Nonconformity Scores and Calibration For each label j, we train an independent probabilistic classifier f_j, such as logistic regression, MLP, or SGD. To quantify the uncertainty of the model's predictions, we compute *nonconformity scores* on a separate calibration set. Specifically, for a calibration instance x_i with $y_i^{(j)} = 1$, the nonconformity score is defined as:

$$s_i^{(j)} = 1 - f_j(x_i),$$

where $f_j(x_i)$ denotes the predicted probability of label j for input x_i. The intuition is that lower predicted probabilities for true positive labels indicate higher model uncertainty, and hence greater nonconformity.

From the distribution of nonconformity scores, we derive a label-specific threshold q_j using the $(1-\alpha)$ quantile:

$$q_j = \text{Quantile}_{1-\alpha}(\{s_i^{(j)}\}_{i:y_i^{(j)}=1}).$$

This calibration step ensures that, with probability at least $1-\alpha$, the true label will be included in the final prediction set.

Prediction Sets At inference time, given a new instance x, the prediction set $\hat{Y}(x)$ is constructed by including each label j whose nonconformity score falls below the calibrated threshold:

$$\hat{Y}(x) = \{j : 1 - f_j(x) \leq q_j\}.$$

In this way, conformal prediction transforms probabilistic outputs into uncertainty-aware prediction sets that adapt dynamically to the model's confidence on each label.

Evaluation Metrics The performance of conformal prediction is evaluated not only in terms of classification accuracy but also in terms of the statistical validity and efficiency of its prediction sets. We adopt the following metrics:

- **Empirical Coverage:** The proportion of true labels captured by the prediction set, averaged across all validation instances:

  $$\text{Coverage} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{Y}(x_i) \cap Y_i|}{\max(1, |Y_i|)}.$$

  This measures whether the prediction sets achieve the intended coverage level.

- **Average Set Size:** The mean number of labels predicted per instance:

  $$\text{Set Size} = \frac{1}{N}\sum_{i=1}^{N}|\hat{Y}(x_i)|.$$

  Smaller sets indicate more efficient and informative predictions.

- **Marginal Coverage:** The probability that an individual true label is included in the prediction set, averaged across all labels:

  $$\text{Marginal Coverage} = \frac{1}{L}\sum_{j=1}^{L}\mathbb{P}(j \in \hat{Y}(x) \mid j \in Y).$$

  This provides a per-label reliability measure.

- **Macro-F1 Score:** The harmonic mean of precision and recall computed across all labels, providing a standard measure of multilabel predictive accuracy.

By combining these metrics, we evaluate CP along three key dimensions: (1) predictive accuracy, (2) statistical validity of coverage guarantees, and (3) efficiency of the resulting prediction sets. This multi-faceted evaluation ensures that the models are not only accurate but also trustworthy and interpretable.

3.1.5 Ensemble Conformal Prediction: Theoretical Guarantees

A central goal of this thesis is to examine how the formal coverage guarantees of conformal prediction extend when combined with ensemble learning. While conformal prediction provides finite-sample marginal coverage guarantees at the level of individual models, the situation becomes more complex once predictions are aggregated across multiple calibrated predictors. In this section, we develop theoretical bounds for ensemble conformal prediction (ECP) for *majority voting*. Our analysis builds on existing results for single-label conformal ensembles [Che19], extending them to the multilabel setting and introducing novel aggregation rules.

Setup and Assumptions Let $\mathcal{D} = \{(x_i, Y_i)\}_{i=1}^n$ denote a multilabel dataset with input features $x_i \in \mathbb{R}^d$ and output label sets $Y_i \subseteq \mathcal{L} = \{1, \ldots, L\}$. Data points are assumed i.i.d. according to an unknown distribution $\mathcal{P}$. For each label ℓ, we train M independent base conformal models (e.g., via bootstrap resampling or random initialization). Model $m \in \{1, \ldots, M\}$ produces a prediction set $\hat{Y}^{(m)}(x)$, calibrated so that marginal coverage holds:

$$\mathbb{P}_{(x,Y)\sim\mathcal{P}}\Big(\ell \in \hat{Y}^{(m)}(x) \;\Big|\; \ell \in Y\Big) \;\geq\; 1 - \alpha.$$

The ensemble prediction set $\hat{Y}^{\text{ens}}(x)$ is then formed by aggregating $\hat{Y}^{(1)}(x), \ldots, \hat{Y}^{(M)}(x)$ via a chosen rule (e.g., majority vote). We seek to understand how the marginal coverage guarantees at the base level translate into guarantees for the ensemble.

Majority Voting We first consider the case where base conformal predictors vote on the inclusion of each label. Let $X_m^{(\ell)} \in \{0, 1\}$ indicate whether label ℓ is included in the prediction set of model m. The ensemble includes label ℓ if at least k models agree, for some threshold $k \leq M$. The following lemma adapts the analysis of Cherubin [Che19] to the multilabel case.

Lemma 3.1.1 (Majority-vote lower bounds, cf. [Che19]). *Assume independence across models and let each model satisfy* $\mathbb{P}(X_m^{(\ell)} = 1) \geq 1 - \alpha$. *Then, for any threshold* $k \in \{1, \ldots, M\}$,

$$\mathbb{P}\left(\sum_{m=1}^{M} X_m^{(\ell)} \geq k\right) \geq \sum_{r=k}^{M} \binom{M}{r} (1-\alpha)^r \alpha^{M-r}.$$

In particular, unanimity voting ($k = M$) *yields*

$$\mathbb{P}\left(\sum_{m=1}^{M} X_m^{(\ell)} \geq M\right) \geq (1-\alpha)^M.$$

Interpretation. The right-hand side corresponds to the tail probability of a Binomial$(M, 1-\alpha)$ distribution, providing a valid lower bound on ensemble coverage. Intuitively, if each base model covers a true label with probability at least $1 - \alpha$, then requiring a majority of models to agree leads to a probability of coverage that is at least the binomial tail. - For unanimity ($k = M$), coverage is at least $(1-\alpha)^M$. - For simple majority ($k = \lceil M/2 \rceil$), the binomial bound is looser but coverage is usually higher in practice because base models often perform better than the nominal level ($p_m > 1 - \alpha$).

This demonstrates that majority-vote ensembles cannot systematically under-cover relative to the baseline guarantees, and often improve coverage through aggregation.

Theorem 3.1.2 (Unanimity ensemble coverage). *Under the assumptions of Lemma 3.1.1, a unanimity-voting ensemble satisfies*

$$(1-\alpha)^M \leq \mathbb{P}\Big(\ell \in \hat{Y}^{ens}(x) \,\Big|\, \ell \in Y\Big) \leq 1.$$

Discussion. These results describe idealized extremes. The conservative lower bound arises from assuming complete independence across models, while the upper bound reflects the trivial case where every base predictor always includes the label. In practice, base learners trained on overlapping data or with similar architectures exhibit correlated errors. This reduces diversity, pushing empirical coverage values between the theoretical bounds. Consequently, the bounds should be interpreted as providing intuition for best- and worst-case behaviour, rather than exact predictions of performance.

Theoretical Summary and Implications The analysis above yields the following theorem.

Theorem 3.1.3 (Ensemble coverage bounds under independence)**.** *Assume independence across models. For unanimity voting ($k = M$),*

$$(1-\alpha)^M \;\le\; \mathbb{P}\Big(\ell \in \hat{Y}^{ens}(x) \;\Big|\; \ell \in Y\Big) \;\le\; 1.$$

For intermediate thresholds $k < M$, coverage is lower-bounded by the corresponding binomial tail (Lemma 3.1.1).

Proof Sketch. The lower bound follows by applying the binomial tail probability to the event that at least k of the M independent models include the true label. The upper bound is trivial, since the inclusion probability is bounded above by one.

Practical Implications. These theoretical results establish that ensemble conformal predictors inherit and often strengthen the coverage guarantees of their base components. However, the extent of improvement depends critically on the degree of independence among models. High diversity (e.g., via heterogeneous architectures or resampling) pushes performance closer to the upper end of the bounds, while correlated ensembles behave more conservatively. Thus, the theory provides guiding principles for the design of ensemble conformal systems, but empirical evaluation remains essential.

Empirical Validation In Chapter 5, we validate these observations on multilabel benchmarks. Taken together, the theory and experiments demonstrate that ensemble conformal prediction offers a principled pathway to combine the robustness of ensembles with the statistical guarantees of conformal prediction.

3.2 Baselines and Proposed Methods

In order to systematically evaluate the effectiveness of ensemble conformal prediction (ECP), we compare our proposed approaches against a diverse set of baseline methods. The baselines serve two purposes: (i) to establish the performance of conventional multilabel classifiers with and without conformal calibration, and (ii) to isolate the effect of applying conformal prediction either before or after ensembling. We then introduce our proposed methods, which fully integrate conformal calibration into the ensemble learning process. This organization allows us to assess the incremental benefits of ECP relative to existing strategies.

3.2.1 Baseline Methods

We consider three categories of baselines, each representing a progressively stronger integration of conformal prediction into the multilabel classification pipeline:

(a) Standard Classification. The simplest baseline corresponds to conventional multilabel classification without conformal calibration. Each base model is trained independently per label under the binary relevance framework, and predictions are obtained by thresholding label probabilities at a fixed cutoff (typically 0.5). While this approach provides a useful performance reference, it lacks any mechanism for quantifying predictive uncertainty or guaranteeing statistical coverage.

(b) Single-Model Conformal Prediction. In this setting, conformal prediction is applied to an individual classifier. Each model produces calibrated thresholds on a held-out calibration set, yielding label-wise prediction sets that satisfy marginal coverage guarantees. This baseline establishes the benefits of CP when applied to a single model, but does not exploit the robustness advantages of ensembling.

(c) Post-Hoc Conformal Ensembles. Here, ensembles of classifiers are first constructed using either homogeneous or heterogeneous base learners. Conformal calibration is then applied *after* aggregation, for example by calibrating averaged probabilities or majority-vote outputs. This setting isolates the role of CP as a post-processing step, without directly integrating it into the training of ensemble members. While effective, this strategy risks diluting model diversity, as calibration is performed on already aggregated outputs.

3.2.2 Proposed Methods: Ensemble Conformal Prediction

Our proposed methods differ fundamentally from the above in that conformal prediction is integrated directly into the ensemble framework. Each base model is independently calibrated, ensuring that label-wise coverage guarantees are preserved at the individual model level. Aggregation is then applied to the already calibrated outputs, leveraging both model diversity and the statistical validity of CP. This design allows us to retain the benefits of ensembling— variance reduction, robustness, and adaptivity—while also providing formal coverage guarantees.

(a) Conformal Ensembles. We evaluate both homogeneous and heterogeneous ensembles of CP models. Each ensemble aggregates predictions through majority voting,

probability averaging, or F1-weighted voting, where weights are determined by per-label validation performance. This approach combines the predictive strengths of diverse base models with conformal calibration, producing prediction sets that are both uncertainty-aware and robust to model variability.

(b) Conformal Stacking (StackECP). To further exploit model complementarity, we propose a stacked ensemble in which the outputs of multiple base CP models are used as input features for a logistic regression meta-classifier. The meta-classifier is trained on a calibration set and itself undergoes conformal calibration. At inference time, StackECP generates calibrated prediction sets by combining information from all base learners in a data-driven manner. Unlike simple voting or averaging, stacking allows the ensemble to learn optimal aggregation strategies, while conformal calibration ensures that statistical validity is preserved.

Tab. 3.1: Comparison of baseline and proposed methods.

Method Category	Description	Role of Conformal Prediction (CP)
	Baselines	
Standard Classification	Independent multilabel models with fixed probability thresholding (e.g., 0.5).	No CP; only point predictions without coverage guarantees.
Single-Model CP	Individual classifier calibrated using CP on a held-out calibration set.	CP applied at the single-model level; provides label-wise coverage guarantees but no ensemble robustness.
Post-Hoc Conformal Ensembles	Ensemble (homogeneous or heterogeneous) built first, then calibrated after aggregation.	CP applied after ensembling; coverage guarantees may be diluted due to calibration on aggregated outputs.
	Proposed Methods	
Conformal Ensembles	Each base model independently calibrated with CP; predictions combined via voting, averaging, or F1-weighted rules.	CP integrated before ensembling; preserves individual guarantees and leverages diversity in aggregation.
Stacked Conformal Ensembles (StackECP)	Meta-classifier trained on outputs of multiple CP-calibrated models, then calibrated itself.	CP applied at both base and meta levels; ensures coverage while learning optimal aggregation strategies.

Algorithm 1: Ensemble Conformal Prediction (ECP)

Input: Training set $\mathcal{D}_{\text{train}}$, calibration set $\mathcal{D}_{\text{cal}}$, base learners $\{f^{(m)}\}_{m=1}^{M}$, target miscoverage rate α, aggregation rule $\mathcal{A}$ (e.g., majority vote, probability averaging, F1-weighted).

Output: Prediction set $\hat{Y}^{\text{ens}}(x)$ for a new input x.

Step 1: Train Base Models.

for $m = 1$ **to** M **do**

  Train base learner $f^{(m)}$ on $\mathcal{D}_{\text{train}}$. ;

Step 2: Calibrate Each Model with CP.

for *each label* $\ell \in \{1, \ldots, L\}$ *and model* m **do**

  Compute nonconformity scores $s_\ell^{(m)}(x_i) = 1 - f_\ell^{(m)}(x_i)$ for $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$ with $y_i^{(\ell)} = 1$. ;

  Set threshold $\tau_\ell^{(m)} = \text{Quantile}_{1-\alpha}(\{s_\ell^{(m)}(x_i)\})$. ;

Step 3: Predict with Calibrated Models.

For a new input x, each base model m outputs a label-wise conformal set:

$$\hat{Y}^{(m)}(x) = \{\ell : 1 - f_\ell^{(m)}(x) \leq \tau_\ell^{(m)}\}.$$

Step 4: Aggregate Predictions.

Combine $\{\hat{Y}^{(m)}(x)\}_{m=1}^{M}$ using aggregation rule $\mathcal{A}$:

$$\hat{Y}^{\text{ens}}(x) = \mathcal{A}\left(\hat{Y}^{(1)}(x), \ldots, \hat{Y}^{(M)}(x)\right).$$

return $\hat{Y}^{ens}(x)$

3.3 Part II: ImageCLEF Concept Detection

The second part of this Chapter is dedicated to the **ImageCLEF concept detection task**, where the goal is to identify and localize semantic concepts within medical images. Unlike Part I, which addressed the more general problem of multilabel ensemble classification, this part focuses on a concrete applied benchmark. The task requires designing models capable of detecting clinically relevant concepts with high accuracy, while addressing challenges such as class imbalance, and the need for robust evaluation protocols. In this part, we outline the specific methodology adopted for ImageCLEF, followed by experimental results and analysis in the context of the competition framework.

3.3.1 CNN-FFNN

The first system we developed is based on a **CNN encoder coupled with a Feed-Forward Neural Network (FFNN)** classifier as illustrated in Figure 3.2. This architecture follows prior work by the AUEB NLP Group [KPA20; Cha+21; Cha+22; Kal+23a; Sam+24; Cha25; Cha+25], while extending it with stronger backbones and systematic ensembling strategies.

Feature Extraction with CNN Backbones. The backbone CNN is responsible for transforming raw input images into a rich, high-dimensional representation. We employ three ImageNet-pretrained architectures of varying complexity: *EfficientNetB0*, *DenseNet121*, and *ConvNeXt-Tiny*. Each input image is resized to $224 \times 224 \times 3$ and normalized according to the preprocessing scheme of the selected backbone. Feature maps are extracted from the final convolutional block of the network and aggregated into a fixed-size embedding via **Generalized Mean (GeM) pooling** [RTC19].

GeM Pooling. GeM pooling is a parametric and differentiable pooling operation that generalizes traditional max and average pooling. Given a spatial activation map X_k corresponding to the k-th feature channel, GeM computes:

$$f_k^{(g)} = \left(\frac{1}{|X_k|} \sum_{x \in X_k} x^{p_k} \right)^{1/p_k},$$

where $|X_k|$ is the number of spatial elements and p_k is a learnable pooling parameter. - When $p_k = 1$, GeM reduces to average pooling. - As $p_k \to \infty$, GeM approximates max pooling.

Since p_k is optimized via backpropagation, the model can adaptively interpolate between these pooling behaviors, effectively learning how much to emphasize high activations versus distributed evidence across the feature map. This flexibility makes GeM particularly suitable for tasks like medical concept detection, where both localized strong signals and diffuse contextual evidence can be informative. Empirically, GeM pooling has been shown to yield more discriminative embeddings than non-trainable pooling operations.

Classification via FFNN. The pooled embedding is fed into a lightweight FFNN, which in our final implementation consists of only an output layer with $|\mathcal{C}|$ neurons, where $\mathcal{C}$ is the set of medical concepts. Each neuron employs a sigmoid activation, producing an independent probability estimate for each concept. A global threshold τ is applied across all concepts, with the value selected via grid search on the validation set to maximize the

F_1-score, the primary competition metric. This formulation treats concept detection as a multi-label classification task where concepts are modeled independently but trained jointly.

Training Objective and Optimization. The model is trained using the **binary cross-entropy (BCE) loss**, computed independently per concept and summed across all labels:

$$\mathcal{L} = \sum_{c \in \mathcal{C}} \left[-y_c \log(\hat{y}_c) - (1 - y_c) \log(1 - \hat{y}_c) \right],$$

where $y_c \in \{0, 1\}$ denotes ground truth for concept c, and $\hat{y}_c \in (0, 1)$ is the predicted probability. Optimization is performed with the Adam optimizer [KB17], using a learning rate of 10^{-3}. A learning rate scheduler reduces the learning rate upon stagnation of validation loss (patience: 1 epoch). Early stopping (patience: 3 epochs) is used to prevent overfitting. Each model is trained for up to 20 epochs with a batch size of 16.

Regularization and Practical Design Choices. Dropout and stochastic weight averaging were considered but ultimately disabled, as preliminary experiments did not indicate consistent improvements. In the final setup, both embedding-level dropout and FFNN hidden layers were removed.

Ensembling Variants. To reduce variance and increase robustness, we trained multiple instances of this CNN-FFNN model using different backbones and random seeds. Predictions were combined using two strategies:

- Union: A concept is predicted if at least one model assigns it above-threshold probability (favoring recall).
- Intersection: A concept is predicted only if all models agree (favoring precision).

These ensembling strategies offer complementary behaviors, providing useful baselines for the more sophisticated aggregation methods presented later in Subsection 3.3.2.

3.3.2 Ensemble Strategies

To enhance robustness and predictive reliability in multilabel concept recognition, we designed a set of ensemble strategies that integrate predictions from models trained with diverse CNN backbones and data splits. These ensembles operate at two complementary levels: (*i*) the *model level*, by introducing architectural diversity across base learners, and

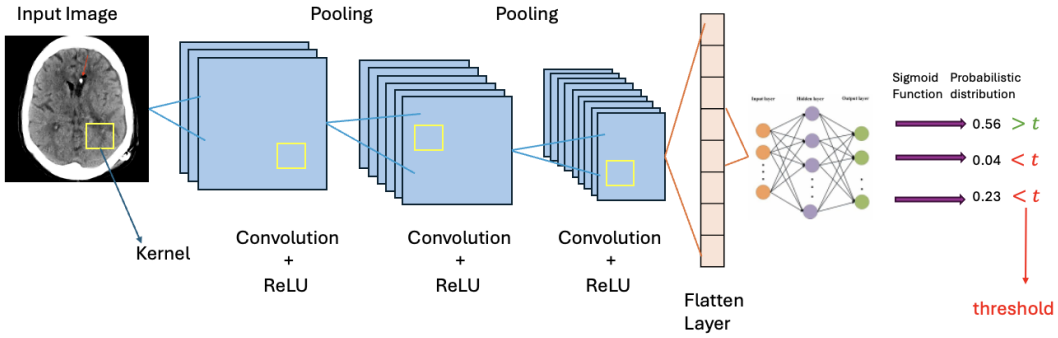


Fig. 3.2: The system uses a CNN for feature extraction and an FFNN for classification, with GeM pooling to generate image embeddings. Concepts are predicted using sigmoid probabilities, with a threshold t applied uniformly. This figure is reproduced from our previous work [Cha25].

(*ii*) the *prediction level*, by aggregating outputs through alternative consensus mechanisms. This two-level design allows the system to capitalize on complementary strengths of individual models, while mitigating their weaknesses through aggregation.

Our experimental ensembles comprised models trained with three distinct encoders: **EfficientNet-B0** [TL19], **DenseNet-121** [Hua+17b], and **ConvNeXt-Tiny** [Liu+22]. To further increase diversity within the EfficientNet-B0 family, we employed a Monte-Carlo cross-validation strategy. Specifically, five models were trained on different train–validation splits of the dataset, with a consistent development set across all splits. During inference, their outputs were aggregated using the *intersection rule*, such that only concepts predicted by all five models were retained. This conservative aggregation ensured high-confidence predictions, albeit at the cost of reduced recall. The resulting intersection set was subsequently merged (via *union*) with predictions from an additional EfficientNet-B0 trained on the entire training+validation set, as well as with outputs from DenseNet-121 and ConvNeXt-Tiny. This hybrid scheme combined the high precision of the intersection ensemble with the broader coverage of union-based integration, aiming to achieve a balanced trade-off between recall and precision.

Beyond these baseline operations, we designed two more refined aggregation strategies to provide adaptive concept inclusion:

Dual Threshold Aggregation. The goal of this strategy is to explicitly balance precision and recall by differentiating between highly confident predictions and those supported by partial consensus. Let $V_{i,j}$ denote the number of models that assign concept j to image i, and let M denote the total number of models in the ensemble. We first identify a *core set* of concepts that achieve full consensus:

$$\text{core}_{i,j} = \begin{cases} 1, & \text{if } V_{i,j} = M \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

While the core set guarantees maximal reliability, restricting predictions exclusively to this set often yields overly conservative outputs. To address this, we introduce a *border set*, which includes concepts predicted by at least L models, where $L < M$:

$$\text{border}_{i,j} = \begin{cases} 1, & \text{if } L \leq V_{i,j} < M \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

The final prediction is obtained as the union of the two sets:

$$\hat{P}_{i,j} = \text{core}_{i,j} \cup \text{border}_{i,j} \tag{3.3}$$

This dual-threshold formulation guarantees that concepts with unanimous agreement are always preserved, while allowing additional concepts to be included when supported by a sufficiently strong majority. The parameter L serves as a tunable consensus threshold, controlling the trade-off between precision and recall.

Partial Intersection Aggregation. The second strategy extends the intersection rule with a fallback mechanism, designed to prevent empty prediction sets in cases of model disagreement. Formally, the *core set* is again defined as:

$$\text{core}_{i,j} = \begin{cases} 1, & \text{if } V_{i,j} = M \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

If the core set is non-empty (i.e., $\sum_j \text{core}_{i,j} > 0$), predictions are restricted exclusively to this set:

$$\hat{P}_{i,j} = \text{core}_{i,j} \tag{3.5}$$

However, if the intersection is empty ($\sum_j \text{core}_{i,j} = 0$), the strategy reverts to a relaxed majority rule, including all concepts predicted by at least L models ($L \in \{2, 3\}$ in practice):

$$\hat{P}_{i,j} = \begin{cases} 1, & \text{if } V_{i,j} \geq L \\ 0, & \text{otherwise} \end{cases} \quad \text{for } \sum_j \text{core}_{i,j} = 0 \tag{3.6}$$

This hierarchical design ensures that predictions default to the most conservative (precision-oriented) aggregation when possible, while maintaining recall through fallback inclusion when full consensus is absent.

Overall, these ensemble strategies reflect different points along the precision–recall spectrum: *union* favors coverage, *intersection* enforces strict agreement, while *dual threshold* and *partial intersection* introduce tunable or hierarchical consensus rules. By combining these methods, we sought to systematically explore how structural diversity (across backbones and splits) and aggregation diversity (across consensus rules) interact to influence multilabel prediction performance.

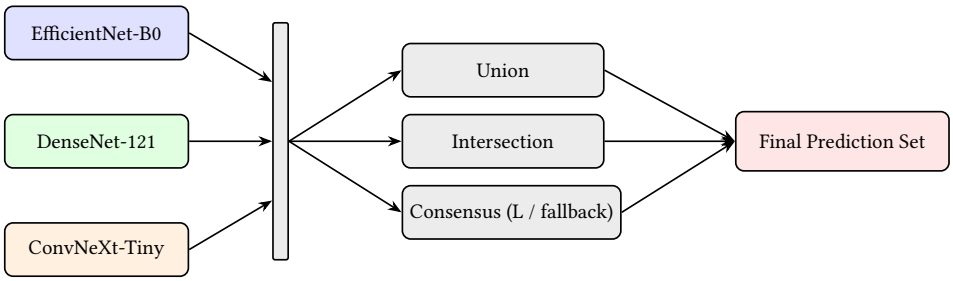


Fig. 3.3: Ensemble aggregation overview. Predictions from the three CNN backbones are combined by *Union*, *Intersection*, or a *Consensus* rule (covers dual-threshold and partial-intersection variants) to form the final concept set.

3.3.3 Ultrasonography Specific Experiments

Beyond the models officially submitted to the Concept Detection task, we performed a set of additional experiments aimed at improving classification performance on **ultrasonography** images. Preliminary error analysis revealed that our models systematically underperformed on this modality compared to others such as X-ray and MRI. To mitigate this gap, we explored targeted fine-tuning strategies designed to incorporate modality-specific information while preserving generalization across the full label space.

Two-Phase Fine-Tuning. Our first approach employed a sequential fine-tuning procedure. In the initial phase, the model was trained on a subset of the training data excluding

all ultrasonography images (**Dataset 1A**). This allowed the model to learn concept representations from modalities with higher predictive stability. Once trained, we preserved both the learned weights and the mapping between output neurons and their associated concept labels.

In the second phase, we expanded the model's output layer to cover the full set of 2,479 concepts, since **Dataset 1B** (containing only ultrasonography images) introduced additional labels not present in Dataset 1A. For overlapping labels, weights learned during the first phase were retained, ensuring continuity of knowledge. Fine-tuning on Dataset 1B thus enabled the model to specialize on ultrasonography while maintaining alignment with the broader label space.

Modality-Specific Masking. Building on the two-phase procedure, we designed a more refined training scheme that explicitly accounts for modality-label associations. The training split was again partitioned into two subsets: one excluding ultrasonography images (**Dataset 2A**) and one containing only ultrasonography images (**Dataset 2B**). A unified label vocabulary was defined as the union of concepts across both subsets, and the model's output layer was structured accordingly.

During training, irrelevant labels were dynamically masked depending on the active dataset. For instance, when training on Dataset 2A, labels exclusive to ultrasonography were masked, and vice versa when training on Dataset 2B. This masking ensured that the model focused only on labels meaningful for the modality under consideration, thereby facilitating knowledge transfer across modalities while preserving specialization where needed.

Results. Although both strategies improved interpretability and appeared promising for enhancing modality-specific adaptation, neither approach surpassed the overall performance of our primary models (see Tables 5.6 and 3.2). Consequently, these models were not included in the final submission.

Overall, these supplementary experiments demonstrate the potential and limitations of modality-aware fine-tuning in large-scale multilabel concept detection. While targeted strategies can reduce domain-specific error, they also highlight the trade-off between specialization and generalization in heterogeneous medical imaging tasks.

Tab. 3.2: Performance of exploratory models evaluated on our held-out development (private test) set. These models were not submitted to the official test set.

Model	F1 (dev)	F1 (val)
CNN + FFN (baseline)	0.5872	**0.5891**
Fine-Tuned → Ultrasonography	**0.5891**	0.5774
Masking (1)	0.5868	0.5773
Masking (2)	0.5806	—

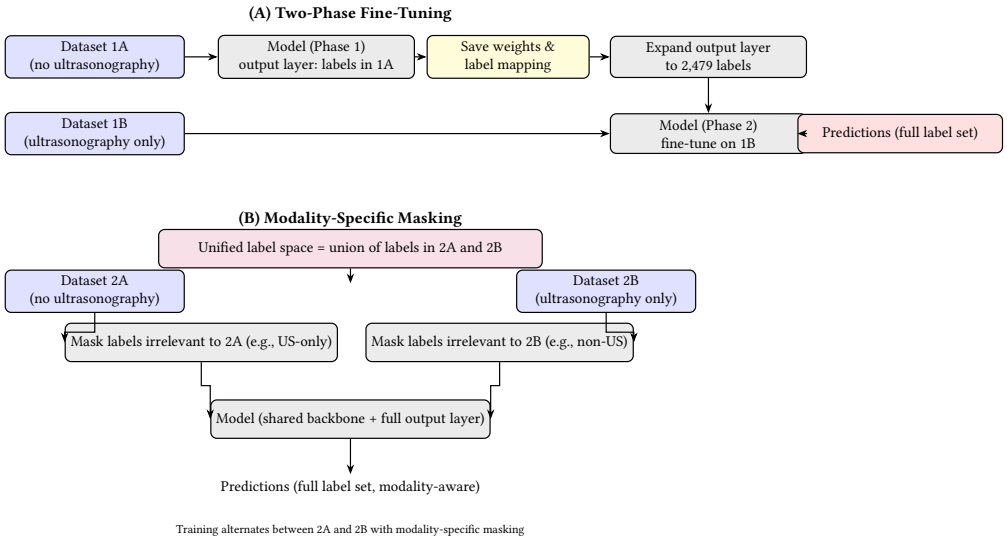


Fig. 3.4: Schematic of additional concept-detection experiments. (A) *Two-Phase Fine-Tuning*: train without ultrasonography (Dataset 1A), save weights and label mapping, expand the output layer to the full label set, then fine-tune on ultrasonography-only data (Dataset 1B). (B) *Modality-Specific Masking*: define a unified label space; when training on non-ultrasonography data (2A), mask ultrasonography-only labels; when training on ultrasonography data (2B), mask non-ultrasonography labels.

4 Data

A central component of this thesis is the empirical evaluation of methods for multi-label classification and medical concept detection. The choice of datasets is therefore critical: they must reflect both the methodological challenges of multi-label learning in general and the domain-specific requirements of medical imaging. This chapter introduces the datasets used in the two main parts of the thesis, highlighting their scale, complexity, and relevance to the research questions addressed.

For the conformal prediction experiments, we make use of several established multi-label benchmarks from diverse domains, including computer vision, biology, and audio analysis. These datasets are widely adopted in the literature because they exhibit challenging characteristics such as label imbalance, high cardinality, long-tail distributions, and complex inter-label dependencies. Evaluating models across such heterogeneous sources ensures that the findings are not tied to a single application area but instead generalize across different forms of structured multi-label data. In particular, these datasets provide a rigorous testbed for assessing both predictive performance and the validity and efficiency of uncertainty quantification methods such as conformal prediction.

For the medical concept detection experiments, we employ the **ImageCLEFmedical 2025** dataset, which constitutes the official benchmark for the annual Concept Detection task. This dataset consists of a large-scale collection of radiological images annotated with UMLS concepts, spanning imaging modalities, anatomical regions, and pathological findings. The dataset is notable for its scale, label diversity, and clinical relevance, as well as for the substantial class imbalance that mirrors real-world medical data. Its use within the ImageCLEF competition further ensures comparability with state-of-the-art systems developed by other research groups worldwide. Within this thesis, the dataset serves as the foundation for evaluating CNN–FFNN architectures, ensemble strategies, and threshold optimization methods in a realistic, safety-critical domain.

Taken together, the datasets used in this work provide a comprehensive empirical basis for analysis. The general-purpose multi-label benchmarks enable the study of conformal predictors and ensemble methods under controlled and diverse conditions, while the ImageCLEFmedical corpus grounds the research in a clinically meaningful application. This combination allows us to investigate both methodological contributions and their practical implications in medical imaging.

4.1 Datasets for Conformal Prediction

To evaluate the effectiveness and generality of the proposed ensemble conformal prediction (ECP) framework, three established multilabel classification benchmarks from distinct domains are employed: computer vision, bioinformatics, and music information retrieval. The selection of these datasets is motivated by their complementary characteristics in terms of sample size, label space cardinality, label density (i.e., the average number of positive labels per instance), and degree of label co-occurrence. Together, they provide a heterogeneous and challenging testbed for assessing predictive accuracy, calibration quality, and robustness of uncertainty quantification.

MS-COCO (Vision). The Microsoft Common Objects in Context (MS-COCO) dataset [Lin+14] is a large-scale benchmark widely used in computer vision. In its multilabel formulation, each image may contain multiple object categories (e.g., `person`, `car`, `dog`), yielding a label space of 80 classes. The dataset exhibits both high label cardinality and substantial class imbalance: frequent categories such as `person` dominate, while many categories are relatively rare. Moreover, strong co-occurrence patterns (e.g., `person` + `bicycle`) add further complexity. To reduce computational overhead and ensure comparability, all experiments are conducted on pre-extracted CLIP ViT-B/32 embeddings [Rad+21] rather than raw pixels. This choice emphasizes label prediction and conformal calibration rather than representation learning, while providing efficiency and reproducibility.

Yeast (Biology). The Yeast dataset [EW01] is a canonical benchmark in multilabel bioinformatics, originally proposed for predicting protein functional classes. It contains 2,417 instances, each described by a set of expression-based and sequence-derived features, annotated with 14 functional labels. Label density is moderate, with most genes associated with a small subset of functions. Unlike COCO, the Yeast dataset operates on structured biological features rather than raw images, introducing different inductive biases and noise sources. It is therefore valuable for testing the robustness of ECP across feature types and domains.

Emotions (Audio/Music). The Emotions dataset [Tro+08] is a music tagging corpus where short audio tracks are annotated with multiple emotion-related labels such as `happy`, `sad`, or `relaxing`. It contains 593 instances and 6 labels, making it small in both sample size and label space. The dataset is relatively balanced, unlike COCO or Yeast, and primarily challenges models in low-data regimes. From an uncertainty quantification perspective, Emotions is useful to evaluate whether conformal predictors remain reliable under small-sample conditions and perceptual label associations.

Comparative Overview. Table 4.1 summarizes key statistics of the three datasets. COCO represents a large-scale, high-cardinality, imbalanced vision benchmark; Yeast provides a medium-scale, structured biological dataset with moderate density; and Emotions serves as a small-scale, perceptual dataset with balanced labels. This diversity allows a systematic evaluation of the proposed methods across settings that vary in scale, domain, and statistical properties.

Tab. 4.1: Summary of multilabel datasets used for conformal prediction experiments. Density denotes the average number of positive labels per instance.

Dataset	Domain	Samples	Labels	Avg. Density
MS-COCO	Vision (Images)	123,287	80	2.90
Yeast	Biology (Genes)	2,417	14	4.24
Emotions	Music (Audio)	593	6	1.87

In summary, the three benchmarks complement one another: COCO stresses scalability and handling of long-tailed distributions; Yeast probes nonlinear feature–label mappings in biological domains; and Emotions evaluates robustness in small-sample, perceptual tasks. Their combined use enables a comprehensive evaluation of ensemble conformal prediction methods across heterogeneous application areas.

4.2 Datasets for ImageCLEFmed Concept Detection

The second part of the experimental study is conducted on the dataset provided for the *ImageCLEFmedical 2025 Caption Task*, which is an extended version of the ROCOv2 corpus [Rüc+24]. This dataset is built from radiology figures extracted from biomedical publications in the PubMed Central Open Access (PMC OA) repository,[1] and constitutes the foundation for all three subtasks of the challenge: Concept Detection, Caption Prediction, and Explainability. The present work focuses exclusively on the Concept Detection task.

Each image in the dataset is paired with a diagnostic caption and annotated with a set of medical concepts, expressed as UMLS Concept Unique Identifiers (CUIs). In total, the full collection comprises 97,368 annotated images, divided by the organizers into 80,091 training samples and 17,277 validation samples. The associated concept vocabulary is extensive, consisting of 2,479 distinct CUIs covering anatomical structures, imaging modalities, diagnostic findings, and pathological conditions. The task is inherently challenging because of (i) the large and heterogeneous label space, (ii) the highly multi-label nature of the problem, and (iii) strong class imbalance, with a long-tailed frequency distribution ranging

[1] PMC Open Access: `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`, last accessed: 2025-05-20

from highly prevalent modalities such as *X-ray Computed Tomography* to singleton concepts occurring only once in the corpus.

For model development and hyperparameter tuning, the official training and validation sets were merged and then repartitioned using stratified sampling into three disjoint subsets: training (75%), validation (10%), and development (15%). Stratification was carried out with respect to both concept frequency distributions and caption lengths in order to preserve the statistical properties of the dataset across splits. This resulted in 73,027 images for training, 9,736 for validation, and 14,605 for development. All internal experiments reported in this thesis are based on these splits, while final results were computed on the hidden official test set.

The official test set for ImageCLEFmed 2025 consists of 19,267 previously unseen radiology images, derived from ROCOv2 [Rüc+24]. As the gold-standard annotations for this split are withheld, final evaluation is conducted by the challenge organizers. This guarantees a fair comparison across all participating systems and ensures that reported scores reflect true generalization performance.

Tab. 4.2: Summary of datasets used in this chapter. Counts refer to the portions used in the experiments (see text). Avg. labels denotes the average number of positive labels per instance.

Dataset	Domain	Samples	Labels	Avg. labels	Features / Notes
MS-COCO (2014)	Vision	≈123k (train+val)	80	≈2.9	CLIP ViT-B/32 embeddings (512-d); multi-object co-occurrence; severe class imbalance
Yeast	Biology	2,417	14	≈4.2	Tabular features; z-score normalization; canonical benchmark for protein function prediction
Emotions	Audio/Music	593	6	≈1.8	Tabular features; z-score normalization; perceptual/emotional tagging
ImageCLEFmed 2025	Medical Imaging	97,368 (train+val)	2,479	3.20	Radiology images with UMLS CUIs; long-tail distribution; CNN inputs resized to 224×224 with ImageNet-style preprocessing

4.3 Concept Detection

The Concept Detection task is formally defined as a large-scale multilabel classification problem in which each radiology image must be annotated with a subset of clinically relevant biomedical concepts. The label space consists of 2,479 distinct concepts, each uniquely identified by a Concept Unique Identifier (CUI) in the Unified Medical Language System (UMLS) ontology [Bod04]. These concepts span a wide semantic spectrum, encompass-

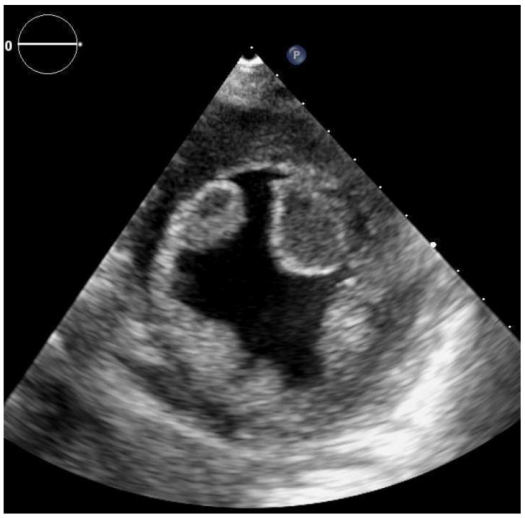
ing imaging modalities (e.g., *X-ray Computed Tomography*, *Magnetic Resonance Imaging*, *Ultrasonography*, *PET/CT*), anatomical entities (e.g., *chest*, *pelvis*), imaging protocols (e.g., *angiogram*, *CT follow-up*), and fine-grained diagnostic findings. By grounding each label in UMLS, the task ensures consistency with established biomedical ontologies and facilitates interoperability with external knowledge bases.

From a machine learning perspective, this setting presents multiple sources of complexity. First, the size of the label space is substantially larger than in most general-purpose multilabel benchmarks (e.g., COCO with 80 labels), which makes the prediction problem high-dimensional and exacerbates the risk of label sparsity. Second, the label distribution follows a pronounced long-tail pattern, with a small number of concepts such as *X-ray Computed Tomography* and *chest* appearing in tens of thousands of images, while the majority of concepts occur only a handful of times, and many are singletons. This imbalance creates strong asymmetries in the availability of training data across labels, challenging models to learn both frequent and rare concepts simultaneously. Third, the task exhibits high label density and co-occurrence: on average, each image is annotated with more than three concepts, often spanning multiple semantic categories (e.g., modality + anatomy + clinical finding). Consequently, the task requires capturing not only marginal label probabilities but also complex dependencies between labels.

The clinical relevance of the task further amplifies its importance. Accurately detecting concepts such as imaging modality or anatomical structure provides the foundation for downstream tasks, including automatic caption generation, clinical decision support, and content-based retrieval in large medical archives. Errors in concept detection are not uniformly problematic: failing to recognize frequent modality concepts may degrade generalizability, while misclassifying rare diagnostic findings risks overlooking information of potentially high clinical value. Hence, reliable uncertainty quantification and balanced predictive performance are both critical for the deployment of such systems in practice.

Figure 4.1 illustrates a concrete example from the ImageCLEFmedical 2025 dataset. The radiology image, depicting an ultrasonography examination, is annotated with three concepts: *Ultrasonography*, *Left ventricular structure*, and *Structure of papillary muscle*. The table shows their corresponding CUIs and UMLS terms, emphasizing the structured, ontology-linked nature of the labels. Such annotations exemplify the multi-faceted character of the task: each image often involves the interaction of modality, anatomy, and clinical focus, making the prediction problem substantially richer than standard object recognition benchmarks.

Distributional Characteristics A closer inspection of the concept frequency distribution reveals one of the central challenges of the Concept Detection task: its pronounced long-tail structure. As shown in Figure 4.2b, a small number of very frequent concepts



CUI	UMLS Term
C0041618	Ultrasonography
C0225897	Left ventricular structure
C0030352	Structure of papillary muscle
ID: ImageCLEFmedical_Caption_2025_train_41494	
CC BY [Magdás et al. (2021)]	

Fig. 4.1: This figure, under CC BY from Magdás et al. (2021), presents an example from the ImageCLEFmedical 2025 dataset [Rüc+24], illustrating the corresponding Concept Unique Identifiers (CUIs) and Unified Medical Language System (UMLS) terms.

dominate the dataset, whereas the majority of concepts occur only rarely. For example, the most common concept, *X-ray Computed Tomography*, is present in more than 34,000 images, while hundreds of concepts appear fewer than ten times across the entire collection, and a substantial number are observed only once (singletons). This heavy-tailed distribution implies that any predictive system must operate effectively across drastically different data regimes: learning reliable classifiers for high-frequency concepts, while also generalizing from extremely limited evidence for rare ones.

Table 4.3 lists the ten most frequent concepts, which are primarily broad imaging modalities or anatomical descriptors such as *X-ray*, *MRI*, and *chest*. These high-frequency categories provide coarse-grained information about the imaging study and are well-represented in training data, making them relatively easier for models to learn. However, a system that performs well only on such frequent categories risks overfitting to generic information while neglecting the fine-grained and clinically specific labels that are often crucial for diagnostic utility.

At the opposite extreme, Table 4.4 shows examples of singleton concepts, each observed in only a single image. These include highly specialized imaging protocols (*Diffusion Weighted Imaging*, *MRI Venography*), rare anatomical structures (*Structure of adductor canal*), and narrow diagnostic procedures (*Root canal post*). Such categories highlight the inherent data

Tab. 4.3: The ten most frequent concepts (CUIs) in the ImageCLEFmedical 2025 dataset [Rüc+24], along with their UMLS terms and frequency counts.

Most Common Concepts			
Rank	**CUI**	**UMLS Term**	**Images**
1	C0040405	X-Ray Computed Tomography	34,055
2	C1306645	Plain x-ray	26,531
3	C0024485	Magnetic Resonance Imaging	15,475
4	C0041618	Ultrasonography	14,237
5	C0817096	Chest	12,559
6	C0002978	Angiogram	5,387
7	C0000726	Abdomen	5,300
8	C0037303	Bone structure of cranium	4,715
9	C0030797	Pelvis	4,449
10	C0023216	Lower Extremity	3,911

sparsity faced in this benchmark: for many concepts, there is effectively no opportunity to learn robust visual representations from the available training data. In practice, this means that models must either rely on transfer learning from related concepts or risk ignoring these rare but clinically significant categories.

Tab. 4.4: Twelve example singleton concepts (CUIs) from the ImageCLEFmedical 2025 dataset [Rüc+24], each appearing in only one image.

CUI	**UMLS Term**
C0598801	Diffusion weighted imaging
C0202657	CT follow-up
C1956110	Cone-Beam Computed Tomography
C0011906	Differential Diagnosis
C0040395	Tomography
C1690005	MRI venography
C0243032	Magnetic Resonance Angiography
C0183062	Root canal post
C0203668	Radioisotope scan of bone
C0412650	Computed tomography of cervical spine
C1962945	Radiographic imaging procedure
C0225273	Structure of adductor canal

This skewed distributional profile creates a dual evaluation challenge: models must be judged not only on their ability to achieve high overall F_1 scores, which are naturally dominated by frequent labels, but also on their robustness in recognizing rare concepts. From an application standpoint, the difficulty of rare concepts is particularly consequential, since these often correspond to subtle diagnostic findings or less common imaging procedures that may hold disproportionately high clinical value. As a result, the long-tail distribution in ImageCLEFmed 2025 represents not just a technical obstacle but also a proxy for the real-world heterogeneity of medical imaging data.

Label Density and Co-occurrence Beyond the imbalance in individual concept frequencies, the dataset also exhibits substantial variation in *label density*, that is, the number of concepts assigned to each image. As shown in Figure 4.2a, the majority of radiology figures are annotated with two to four concepts, but the range extends from single-label cases (10,018 images) up to highly complex cases with 28 distinct concepts. On average, images are annotated with approximately 3.20 concepts each, indicating that the dataset systematically captures the multi-faceted nature of clinical imaging: modality, anatomy, and findings frequently co-occur and must all be correctly identified for complete semantic coverage.

This variability has direct implications for model design. From a learning perspective, low-density cases resemble traditional single- or few-label classification problems, where standard discriminative models can often perform well. In contrast, high-density cases require models to capture more intricate interdependencies among labels, since many concepts co-occur in structured patterns. For example, modality–anatomy pairs such as *chest + plain X-ray* or *abdomen + CT* are frequent and provide relatively strong cues. However, there also exist rare and clinically specific co-occurrence patterns (e.g., *angiogram + pelvis + vascular graft*), which are sparsely represented and therefore more difficult for purely data-driven systems to learn reliably.

The importance of label co-occurrence is amplified in multilabel conformal or ensemble frameworks, where predictive sets must balance individual label accuracy with calibration across combinations of labels. Poorly modeling dependencies can lead to inflated prediction sets or missed concepts, undermining the clinical interpretability of outputs. This challenge mirrors real-world diagnostic practice: a radiology image rarely conveys isolated information, but rather a constellation of complementary cues that must be integrated to produce a meaningful interpretation.

Taken together, these observations highlight that the Concept Detection task is not only defined by its large and heterogeneous label space, but also by the complex structural patterns in which labels appear. Models are therefore evaluated not just on their ability to recognize frequent and visually distinctive concepts, but also on their capacity to handle variable label density and to generalize across diverse co-occurrence structures. This combination of long-tail imbalance and structured dependencies makes ImageCLEFmedical 2025 a particularly challenging benchmark for multilabel learning and uncertainty-aware prediction.

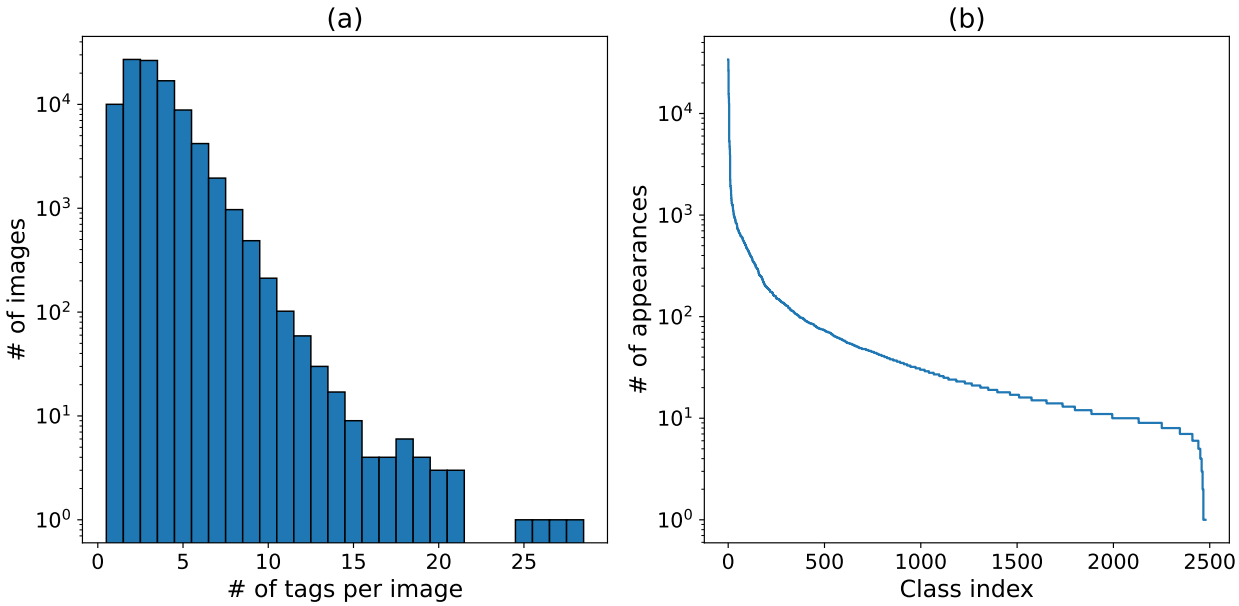


Fig. 4.2: Distributional statistics of the ImageCLEFmedical 2025 concept detection dataset [Rüc+24]. (a) Histogram of the number of concepts per image, illustrating variation in label density. (b) Long-tail distribution of concept frequencies, with a small number of very frequent labels and many rare ones.

5 Experimental Analysis

This Chapter presents the experimental evaluation of the methods proposed in Chapter 3. Our goal is twofold: first, to assess the effectiveness of ensemble conformal prediction (ECP) in the context of multilabel classification, focusing on predictive accuracy, uncertainty quantification, and statistical coverage guarantees; and second, to evaluate the performance of convolutional neural network (CNN) based systems on the *ImageCLEFmed Concept Detection* task. The experiments are organized as follows. We begin by describing the datasets used in both settings, highlighting their characteristics and the preprocessing steps applied. We then provide a detailed account of the evaluation protocols, including the metrics employed to measure both predictive performance and uncertainty calibration. Finally, we present and analyze the results of our experiments, comparing baseline approaches with the proposed methods, and discussing both strengths and limitations.

5.1 Training Setup for Conformal Prediction

All experiments are conducted under the **binary relevance** (BR) framework, in which each label is modeled as an independent binary classification problem. This formulation is widely adopted in multilabel learning due to its scalability and modularity: it allows heterogeneous classifiers to be trained per label and permits straightforward integration with conformal calibration procedures. Although BR ignores label dependencies by construction, this limitation can be partially mitigated through ensemble aggregation strategies, as explored in Section 3.1.3. To ensure that results are robust rather than artifacts of specific random initializations, each experiment is repeated under multiple random seeds, and results are averaged across runs.

Data Preparation. For the COCO dataset, image-level features are extracted using the `CLIP ViT-B/32` model [Rad+21], which is pretrained on image–text pairs from a large corpus. Using CLIP embeddings instead of raw pixels serves two purposes: (i) it reduces computational overhead, enabling efficient experimentation across multiple models and ensembles, and (ii) it provides a semantically rich representation that captures both visual and textual context, improving generalization in downstream multilabel classification. All images are processed using CLIP's standard pipeline (resize, center crop, normalization), resulting in 512-dimensional feature vectors. In contrast, the Yeast and Emotions datasets

are already available in tabular form with numerical features, which are standardized using z-score normalization to improve convergence during training. No additional feature engineering or dimensionality reduction is performed, in order to maintain comparability with prior work.

Dataset Splits. To obtain unbiased evaluation while enabling calibration, each dataset is divided into three disjoint subsets: training, calibration, and validation. This is achieved using a two-stage stratified sampling strategy that preserves the empirical distribution of labels. Specifically, 60% of the data are allocated to training, while the remaining 40% are evenly split into calibration and validation subsets (20% each). The calibration set is used exclusively to compute label-wise nonconformity thresholds, while the validation set is reserved for model selection and hyperparameter tuning. For ensemble experiments, bootstrap resampling of the training set is applied to promote diversity among base learners, following standard bagging principles.

Model Architectures. A broad spectrum of classifiers is evaluated, spanning linear, shallow, and deep architectures. Linear models include logistic regression (LR) with L2 regularization and stochastic gradient descent (SGD) trained with logistic loss. The latter is calibrated using Platt scaling via `CalibratedClassifierCV`, ensuring probabilistic outputs that are compatible with conformal prediction. Neural models are implemented in PyTorch and include:

- **MLP:** one hidden layer with 256 ReLU units, followed by a sigmoid output layer for multilabel prediction.
- **RNN:** a unidirectional LSTM with hidden size 256, applied to CLIP embeddings (reshaped into sequences of length 1), followed by a dense sigmoid classifier.
- **Transformer encoder:** two stacked layers, each with 4 self-attention heads, projecting into contextualized representations that feed into a shared linear classifier.
- **MLP-Mixer:** alternating token-mixing and channel-mixing feedforward layers with GELU activations and layer normalization, reflecting recent advances in vision architectures that eschew explicit attention mechanisms.

Training Protocol. All neural models are trained using the Adam optimizer [KB17] with a fixed learning rate of 10^{-3} and binary cross-entropy loss. Training is conducted for 10 epochs with batch size 64 on COCO and 128 for the smaller Yeast and Emotions datasets. Mini-batches are shuffled at each epoch to improve generalization. To ensure

reproducibility, random seeds are fixed across Python, NumPy, and PyTorch (including CUDA backends). Training is executed on a single NVIDIA RTX 2080 Ti GPU.

Conformal Prediction Setup. For each label j, nonconformity scores are computed on the calibration set, and thresholds q_j are determined at the $(1 - \alpha)$ quantile. Unless stated otherwise, we set $\alpha = 0.1$, corresponding to a 90% target coverage rate. This choice reflects a common trade-off between reliability (coverage) and informational efficiency (set size). In ablation studies (Section 5.2.2), the effect of varying α is investigated.

Ensembles. Both homogeneous and heterogeneous ensembles are explored, with ensemble sizes $M = 3$ or $M = 5$. Homogeneous ensembles use bootstrap-resampled training subsets, while heterogeneous ensembles combine base learners of different architectures. Aggregation strategies include majority voting, probability averaging, and F1-weighted voting, the latter assigning label-specific weights proportional to validation performance. Stacked ensembles are also implemented, where predictions from base models on the calibration set are used as features to train a logistic regression meta-classifier, which is subsequently calibrated via CP.

Evaluation Protocol. Performance is reported using multiple complementary metrics: macro-F1 (to capture balanced predictive performance across labels), exact match accuracy (a stringent criterion requiring all labels to be correct), empirical coverage, marginal coverage, and average prediction set size. Statistical significance is assessed using paired Wilcoxon signed-rank tests applied to results from five independent random seeds (42, 100, 2021, 7, 999). All results are reported as mean $\pm$ standard deviation across seeds. This evaluation framework ensures both robustness and statistical reliability of the observed differences. To demonstrate this, Tables 5.3 and 5.2 report mean and standard deviation values across multiple runs for representative methods on all three datasets.

5.2 Results

Roadmap. This section presents the empirical evaluation of the proposed **Ensemble Conformal Prediction (ECP)** framework. The analysis is structured around three main components. First, a direct comparison is made between single-model CP, post-hoc conformal ensembles, and ECP across three benchmark datasets (**Emotions**, **Yeast**, **COCO**), highlighting their trade-offs in terms of empirical coverage, marginal coverage, average prediction set size, and macro-F1 performance. Second, the runtime overhead and statistical robustness of ECP relative to baselines are examined. Finally, targeted ablation studies are presented to isolate the contributions of key design factors, including ensemble

size, aggregation strategy, model diversity, and the specified miscoverage rate α. Together, these analyses provide a comprehensive view of how ECP balances the competing goals of predictive accuracy, compactness, and statistical validity in multilabel classification.

Table 5.1 summarizes the comparative performance of all evaluated methods. The table is organized into four blocks: (i) non-conformal baselines (standard binary relevance and simple ensembles), (ii) single-model conformal predictors, (iii) post-hoc conformal ensembles, and (iv) our proposed ECP variants. Performance is reported on the three datasets spanning different domains and degrees of complexity, thereby providing a heterogeneous testbed for evaluation. Complementing the numerical results, Figure 5.1 plots the trade-off between average set size and predictive performance. Taken together, the table and figure provide both quantitative and visual evidence of the advantages of ECP.

Interpretation of Figure 5.1. The scatter plots illustrate the Pareto frontier between compactness and predictive strength. Single-model CP methods (triangles) typically guarantee coverage but at the expense of larger prediction sets, while post-hoc ensembles (squares) improve efficiency only marginally. In contrast, ECP methods (circles) shift the frontier upward and leftward, simultaneously reducing set size and improving macro-F1. This effect is most visible on the COCO dataset, where the label space is large and highly imbalanced, but the same qualitative pattern is observed across all three benchmarks.

Overall Performance. Across datasets, Ensemble Conformal Prediction (ECP) consistently improves upon both single-model conformal predictors and post-hoc conformal ensembles. The key strength of ECP is that it preserves *valid coverage guarantees* while producing more compact and informative prediction sets, thereby avoiding the over-conservativeness that often plagues single-model CP. In particular, whereas individual conformal models (e.g., SGD on Emotions, or MLP on COCO) frequently generate very large prediction sets with limited discriminative power, ECP leverages aggregation to balance coverage with predictive sharpness. The result is systematically higher macro-F1 without sacrificing calibration. Importantly, these improvements are consistent across all three benchmarks despite their differences in scale, domain, and label distribution.

Emotions. The Emotions dataset is the smallest of the three (six labels, few hundred samples), which makes it a challenging low-data regime. Nevertheless, the results show that aggregation can still extract meaningful signal:

- **Best performance:** the *Stacked Heterogeneous ECP* achieves the top macro-F1 of 0.6596, improving over logistic regression with CP (0.6467). The improvement may seem small in absolute terms, but in this low-data setting even marginal gains

are significant, showing that stacking can exploit complementary patterns among diverse base models.

- **Compact sets:** homogeneous MLP ensembles with weighted averaging also perform strongly (F1 = 0.6467) while producing the most compact sets (3.17 labels per instance). This highlights the efficiency of probabilistic aggregation relative to naïve voting.

- **Over-conservativeness:** CP with SGD attains perfect coverage (EC = 1.0, MC = 1.0), but does so by predicting all possible labels (average set size = 6.00), which yields the lowest F1 (0.4692). This illustrates a central limitation of single-model CP: coverage is maintained, but usefulness is lost.

Takeaway for Emotions: even in small datasets where base models are individually weak, ECP prevents trivial over-coverage and achieves a better balance between reliability and informativeness.

Yeast. The Yeast dataset (14 labels, moderate size) reveals clearer contrasts between aggregation strategies:

- **Top accuracy:** both the homogeneous MLP-WA and the heterogeneous MV ensembles reach the best macro-F1 of 0.4710. The heterogeneous version does so with slightly more compact sets (10.10 vs. 10.38), whereas the homogeneous ensemble achieves the highest coverage (EC = 0.9130, MC = 0.8932).

- **Single-model CP:** performance ranges between 0.4531–0.4682 F1, producing consistently large sets (10.5–10.6) with only modest accuracy, showing again that ensembling is essential for efficiency.

- **Post-hoc CP ensembles:** provide some improvement over single CP (e.g., Het. Ensemble–CP (1) reaches 0.4625 F1), but fall short of ECP, confirming that conformalizing before aggregation is more effective than conformalizing after.

Takeaway for Yeast: with its noisy biological features and moderate label size, ECP demonstrates robustness—not only maintaining coverage near the 90% target but also squeezing out extra predictive power compared to both single-model and post-hoc baselines.

COCO. The COCO dataset, with 80 labels and extreme imbalance, is the most demanding benchmark. Here the advantages of ECP are particularly evident:

- **Highest F1:** the *Heterogeneous ECP with Weighted Voting* attains the top macro-F1 of 0.5745 with a compact set size of 7.22. This shows that weighting models by label-specific validation F1 yields superior error correction in high-cardinality problems.

- **Smallest sets:** the *Homogeneous LR–MV ensemble* achieves the smallest prediction sets (7.16) while maintaining competitive F1 = 0.5674. This demonstrates that even simple majority voting can effectively counteract the conservativeness of single CP in large label spaces.

- **Single-model CP:** produces sets of size 8–9 with F1 around 0.51–0.54, confirming that individual predictors struggle with COCO's imbalance.

- **Post-hoc CP ensembles:** improve modestly (F1 = 0.5482, set size 7.83), but again fall behind ECP, emphasizing the benefit of calibrating models individually before aggregation.

Takeaway for COCO: ensemble diversity and weighted aggregation are crucial in large-scale, imbalanced MLC. They deliver both stronger accuracy and tighter sets than any single model can achieve, while still respecting coverage guarantees.

Key Takeaways. The combined evidence across all datasets leads to three overarching conclusions:

1. **ECP improves both calibration and predictive accuracy**, overcoming the trade-off that limits single-model CP. Coverage is preserved, but set size and informativeness are substantially improved.

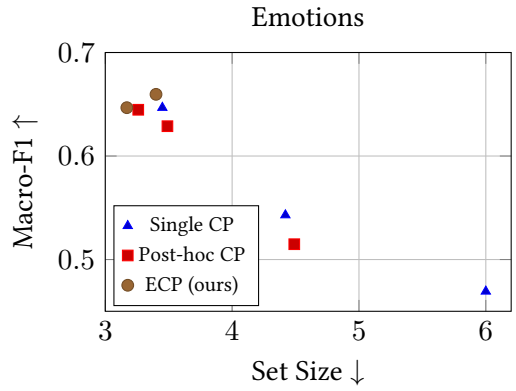
2. **Model diversity pays off.** Heterogeneous ensembles generally outperform homogeneous ones, especially on COCO, because they exploit complementary strengths and reduce correlated errors.

3. **Aggregation strategy matters.** Weighted averaging and stacked ensembles consistently deliver the best trade-offs, while simple majority voting remains surprisingly competitive in less complex domains.

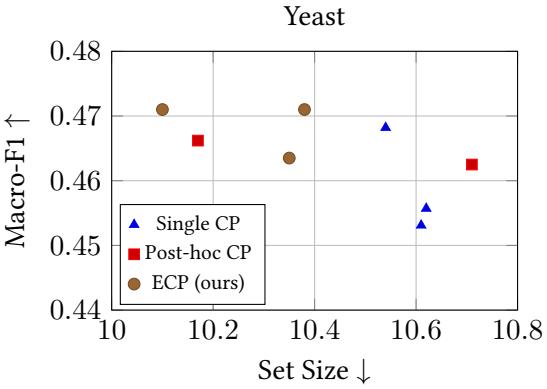
Overall, these findings establish ECP as a principled and effective framework for uncertainty-aware multi-label classification. It extends the theoretical guarantees of conformal prediction into practice, showing that when combined with ensemble learning, coverage need not come at the cost of accuracy or usability.

Tab. 5.1: Performance comparison across Emotions, Yeast, and COCO datasets. EC = Empirical Coverage, MC = Marginal Coverage, Set Size = average predicted labels per instance, F1 = Macro-F1. Best CP-based results per dataset are **bold**.

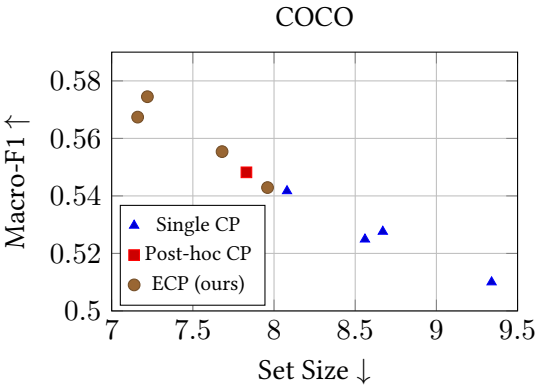
Dataset	Method	EC	MC	Set Size ↓	F1 ↑
		Emotions			
Non-Conformal Baselines	BR (LR)	–	–	–	0.6146
	Ensemble (LR)	–	–	–	0.3491
Single-Model CP	CP (LR)	0.8908	0.8765	3.45	0.6467
	CP (MLP)	0.9034	0.8930	4.42	0.5429
	CP (SGD)	1.0000	1.0000	6.00	0.4692
Post-hoc CP Ensembles	Het. Ensemble–CP	0.9663	0.8944	3.49	0.6288
	Multi-MLP–CP	0.8904	0.8776	4.49	0.5148
	Stacked Het.–CP	0.8867	0.8776	3.26	0.6447
ECP (ours)	Hom. CP (MLP–WA)	0.8717	0.8629	**3.17**	0.6467
	Stacked Het. CP	0.8992	0.8833	3.40	**0.6596**
		Yeast			
Non-Conformal Baselines	BR (LR)	–	–	–	0.3497
	Ensemble (LR)	–	–	–	0.3491
Single-Model CP	CP (LR)	0.8916	0.8826	10.61	0.4531
	CP (MLP)	0.9029	0.8929	10.54	0.4682
	CP (SGD)	0.8976	0.8708	10.62	0.4557
Post-hoc CP Ensembles	Het. Ensemble–CP (1)	0.9037	0.8929	10.71	0.4625
	Het. Ensemble–CP (2)	0.8742	0.8703	10.17	0.4662
ECP (ours)	Hom. CP (MLP–WA)	**0.9130**	**0.8932**	10.38	**0.4710**
	Hom. CP (LR–MV)	0.9089	0.8729	10.35	0.4635
	Het. CP (MV)	0.9100	0.8703	**10.10**	**0.4710**
		COCO			
Non-Conformal Baselines	BR (LR)	–	–	–	0.6981
	CLIP-RNN	–	–	–	0.7002
	Ensemble (LR)	–	–	–	0.6964
	Label Bagging (10)	–	–	–	0.4828
	Label Bagging (40)	–	–	–	0.6940
Single-Model CP	CP (LR)	0.9103	0.8998	8.56	0.5249
	CP (MLP)	0.9092	0.8992	9.34	0.5100
	CP (SGD)	0.9111	0.9009	8.67	0.5276
	CP (RNN)	0.9115	0.9006	8.08	0.5417
Post-hoc CP Ensembles	Het. Ensemble–CP	0.9045	0.8917	7.83	0.5482
ECP (ours)	Hom. CP (LR–MV)	0.8962	0.8965	**7.16**	0.5674
	Hom. CP (LR–WA)	0.9042	0.8887	7.96	0.5429
	Het. CP (MV)	0.9064	0.8870	7.68	0.5554
	Het. CP (WV)	0.8871	0.8577	7.22	**0.5745**



(a) Emotions



(b) Yeast



(c) COCO

Fig. 5.1: Trade-off between prediction set size and Macro-F1 for conformal methods across datasets. Single-model CP (triangles), post-hoc CP ensembles (squares), and our ECP methods (circles).

Reliability Across Runs. Tables 5.3 and 5.2 provide a complementary view of performance stability across random seeds. The extended metrics table (Table 5.3) shows that ECP consistently delivers more compact prediction sets with higher macro-F1 than single-model CP, while maintaining comparable or improved coverage. For instance, on Emotions, the stacked ensemble raises macro-F1 from 0.554 to 0.647 while reducing the average set size (3.24 vs. 3.97). Similarly, on Yeast, the heterogeneous ensemble maintains stable coverage while producing slightly smaller sets and marginally higher macro-F1.

The aggregate macro-F1 results in Table 5.2 confirm that these gains are statistically reliable across all three benchmarks. On COCO, the most challenging dataset, ECP improves from 0.542 to 0.558 macro-F1 with very low variance, showing that the improvements are systematic rather than random fluctuations. On Emotions, the benefit is even larger (+0.093 macro-F1), while Yeast demonstrates smaller but consistent gains.

Taken together, these results strengthen the conclusion that ensemble conformal prediction not only improves accuracy and efficiency, but also yields stable and reproducible performance across runs, enhancing the reliability of uncertainty-aware multilabel classification.

Tab. 5.2: Macro-F1 (mean ± std) across five runs.

Dataset	Method	Macro-F1
COCO	Single CP (MLP)	0.542 ± 0.001
	ECP (ours)	**0.558 ± 0.003**
Emotions	Single CP (MLP)	0.554 ± 0.015
	ECP (Stacked Ensemble)	**0.647 ± 0.024**
Yeast	Single CP (MLP)	0.466 ± 0.007
	ECP (Het. Ensemble)	**0.467 ± 0.004**

Tab. 5.3: Extended reliability metrics (mean ± std) across five runs for Emotions and Yeast.

Dataset	Method	Coverage	MC	Set Size	F1
Emotions	Single CP (MLP)	0.8641 ± 0.0255	0.8617 ± 0.0231	3.97 ± 0.23	0.5542 ± 0.0151
	ECP (Stacked)	0.8641 ± 0.0419	0.8678 ± 0.0345	**3.24** ± 0.30	**0.6467** ± 0.0238
Yeast	Single CP (MLP)	0.9055 ± 0.0063	0.9025 ± 0.0093	10.55 ± 0.14	0.4661 ± 0.0071
	ECP (Het.)	0.9103 ± 0.0082	0.8851 ± 0.0183	**10.36** ± 0.25	**0.4667** ± 0.0035

5.2.1 Runtime and Computational Analysis

Overview. All experiments were conducted on a workstation with an NVIDIA GeForce RTX 2080 Ti GPU (11GB VRAM), 64GB RAM, and Ubuntu 20.04. Runtime was measured for the calibration and inference phases on the COCO dataset, which serves as the most computationally demanding benchmark among those considered in this thesis. A single

conformal predictor required on average ∼16 minutes to complete calibration and inference, while an ensemble of five models required ∼22 minutes. This corresponds to an overhead of approximately 38%. The increase in runtime scales nearly linearly with the ensemble size, reflecting the fact that each model must be calibrated independently. Importantly, this additional cost remains practical for configurations of three to five models—the range most frequently considered in our evaluation. Furthermore, the independence of ensemble members means that wall-clock time can be reduced considerably through parallelization.

Runtime Scaling. To illustrate the relationship between ensemble size and runtime, Figure 5.2 presents the measured values for $M \in \{1, 5\}$ and interpolated estimates for intermediate ensemble sizes. The plot confirms that runtime grows linearly with the number of ensemble members. Notably, moving from one to five models increases runtime by only six minutes on average, while providing measurable improvements in coverage and predictive robustness (see Section 5.2). This trade-off underscores the practical feasibility of adopting ECP in real-world applications where both efficiency and uncertainty calibration are important.

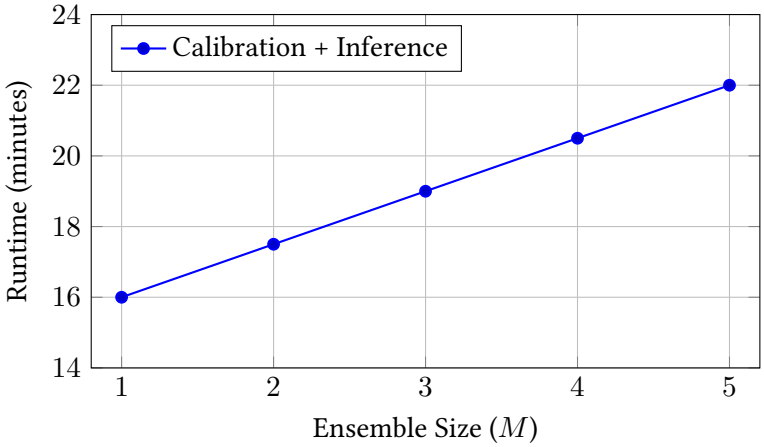


Fig. 5.2: Runtime scaling of conformal predictors on COCO as a function of ensemble size (M). Measured values at $M = 1$ and $M = 5$; intermediate values interpolated linearly.

5.2.2 Ablation Studies

To disentangle the contributions of different design choices within the proposed Ensemble Conformal Prediction (ECP) framework, a series of ablation studies were performed using the COCO dataset as the primary benchmark. The objective of these experiments is not only to validate the effectiveness of the final ECP design, but also to better understand the relative importance of its components. Four factors were systematically varied: (i) the ensemble size M, (ii) the aggregation strategy (majority voting versus probability averaging), (iii) the degree of model diversity (homogeneous versus heterogeneous ensembles), and (iv)

the target miscoverage rate α, which controls the level of coverage guarantees. Each of these ablations provides complementary insights into the mechanisms through which ECP achieves its improvements in predictive accuracy, calibration reliability, and efficiency of prediction sets.

Impact of Ensemble Size The first ablation investigates the role of ensemble size. Ensemble methods are well known to reduce variance and improve robustness, but increasing the number of base learners inevitably incurs computational cost. To explore this trade-off, homogeneous ensembles of logistic regression classifiers were constructed with ensemble sizes ranging from $M = 1$ (single-model CP baseline) up to $M = 10$.

Figure 5.3 provides a visual overview, while Table 5.4 reports detailed numerical results. The observed trends reveal three key findings. First, increasing the ensemble size from one to three members yields a marked boost in macro-F1 (0.5458 $\rightarrow$ 0.5630) while keeping prediction sets compact (7.25 labels on average). Second, the largest performance gain occurs between $M = 1$ and $M = 5$, where macro-F1 improves by approximately 2.2 points, accompanied by slightly more compact sets (7.39 $\rightarrow$ 7.23). Third, beyond $M = 5$, additional models contribute only marginal improvements, with performance essentially saturating around $M = 7$–10. Empirical coverage remains consistently within the range 0.89–0.91, indicating that calibration is stable even as ensemble size varies.

These findings suggest that ensembles of moderate size ($M \in [3, 5]$) represent the most efficient design point, striking a balance between predictive gains and computational overhead. Larger ensembles do not substantially improve accuracy, but linearly increase runtime (see Section 5.2.1), making them less attractive in practice.

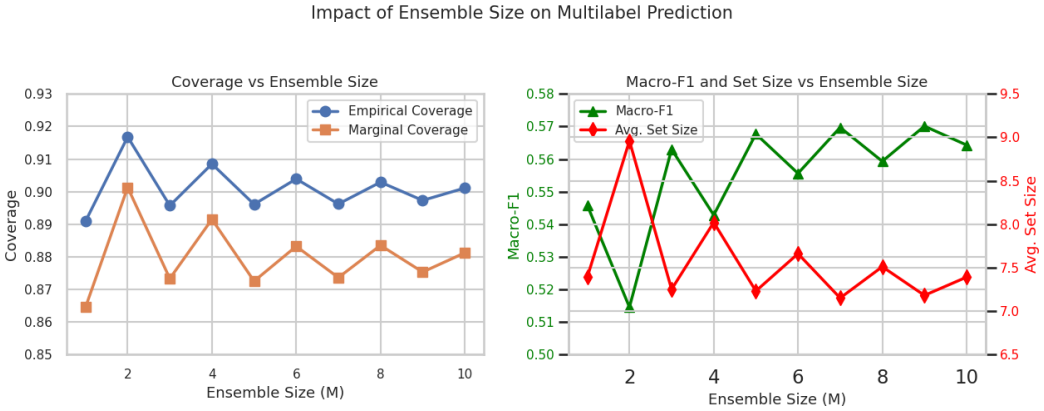


Fig. 5.3: Effect of ensemble size on prediction performance (COCO dataset). Left: empirical and marginal coverage as a function of M. Right: macro-F1 and average prediction set size. Gains saturate beyond $M = 5$, suggesting that ensembles of moderate size are sufficient.

Tab. 5.4: Performance across different ensemble sizes (M) on the COCO dataset.

M	EC	Avg. Size	Macro-F1	MC
1	0.8909	7.39	0.5458	0.8645
2	0.9168	8.96	0.5145	0.9013
3	0.8957	7.25	0.5630	0.8733
4	0.9085	8.02	0.5428	0.8915
5	0.8961	7.23	0.5677	0.8725
6	0.9039	7.66	0.5556	0.8832
7	0.8962	7.15	0.5697	0.8736
8	0.9029	7.51	0.5593	0.8835
9	0.8973	7.18	0.5701	0.8753
10	0.9011	7.39	0.5643	0.8812

Sensitivity to Miscoverage Rate The second ablation focuses on the effect of the miscoverage rate α, which directly specifies the desired reliability of conformal prediction. Lower values of α enforce stricter coverage guarantees, while higher values relax coverage to allow sharper and more selective predictions.

Table 5.5 demonstrates the trade-off clearly. At the strictest setting ($\alpha = 0.01$, 99% target coverage), empirical coverage is indeed nearly perfect (0.9877), but macro-F1 collapses to 0.2825 because the prediction sets become excessively large and conservative. In contrast, relaxing to $\alpha = 0.20$ (80% target coverage) yields compact sets and the highest macro-F1 (0.6700), but with lower reliability. Intermediate settings such as $\alpha = 0.10$ achieve a reasonable balance, with empirical coverage (0.8878) closely matching the nominal level and macro-F1 of 0.5875.

This analysis highlights α as a critical hyperparameter, governing the accuracy–coverage trade-off. Its selection should therefore be application-specific: safety-critical tasks (e.g., medical imaging) may require low α, while exploratory tasks (e.g., tagging large-scale multimedia data) may tolerate higher miscoverage in exchange for sharper predictions.

Tab. 5.5: Effect of target miscoverage rate (α) on empirical coverage and macro-F1 (COCO dataset).

α	Target Coverage	Empirical Coverage	Macro-F1
0.01	0.99	0.9877	0.2825
0.05	0.95	0.9399	0.4874
0.10	0.90	0.8878	0.5875
0.20	0.80	0.7887	0.6700

Homogeneous vs. Heterogeneous Ensembles The third ablation examines the effect of model diversity. Homogeneous ensembles are constructed from repeated instances of the same architecture (e.g., multiple MLPs), whereas heterogeneous ensembles combine distinct learners (LR, SGD, MLP).

Results across datasets (see Table 5.1) indicate that heterogeneous ensembles consistently deliver superior macro-F1 scores while maintaining comparable coverage. For example, on COCO, the heterogeneous weighted-voting ensemble achieves the best overall performance (F1 = 0.5745) with a compact prediction set size (7.22). Similarly, on Emotions, the *Stacked Heterogeneous Ensemble* achieves the highest F1 (0.6596), demonstrating that diversity across model families mitigates correlated errors and improves overall generalization.

Homogeneous ensembles nonetheless remain competitive baselines, particularly when paired with probabilistic aggregation methods (e.g., MLP–WA). However, they tend to produce slightly larger prediction sets (e.g., size = 3.17 on Emotions) and exhibit less robustness across datasets. These findings reinforce the value of incorporating diversity into ensemble design, especially in uncertainty-aware applications.

Summary. The ablation studies collectively highlight three principles for effective ECP: (i) ensembles of moderate size ($M = 3$–5) strike the best balance between predictive accuracy, coverage, and runtime; (ii) careful tuning of α is crucial, as it directly governs the accuracy–coverage trade-off; and (iii) heterogeneous ensembles consistently outperform homogeneous ones, underscoring the importance of model diversity in robust conformal prediction. Together, these results provide a systematic understanding of how ECP's components interact to yield strong overall performance.

5.3 Concept Detection: Experiments and Results

Roadmap. This section presents the experimental study of the **ImageCLEFmedical 2025 Concept Detection** task. The objective is to automatically identify relevant biomedical concepts (CUIs) from radiology images, a problem that is both high-dimensional and characterized by severe class imbalance. The section is organized as follows. First, the system architectures evaluated in this work are introduced, ranging from single CNN–FFNN pipelines with different convolutional backbones to ensemble configurations and threshold-tuning variants. Second, the evaluation methodology is described, including both the official F_1 score (averaged across all concepts per image) and the secondary F_1 score restricted to manually curated concept categories such as anatomy, topography, and imaging modality. Finally, experimental results are reported for both the internal development split and the official hidden test set, allowing direct comparison with competing systems in the challenge. The analysis emphasizes three axes of variation: (i) the influence of CNN backbone choice, (ii) the effect of ensemble aggregation strategies, and (iii) the role of per-label threshold optimization. Together, these experiments provide a comprehensive assessment of how architectural and methodological choices impact performance on large-scale medical concept detection.

5.3.1 Experimental Setup

For the Concept Detection task of ImageCLEFmedical 2025, a family of systems was designed around the **CNN–FFNN pipeline** introduced in Section 3.3.1. The overall strategy was to leverage strong convolutional feature extractors, followed by a lightweight feed-forward network responsible for multilabel classification over the 2,479 biomedical concepts. To examine the role of backbone architecture, three state-of-the-art CNN models were selected as encoders: *EfficientNet-B0* [TL19], known for its parameter efficiency and depth–width scaling; *DenseNet-121* [Hua+17b], which exploits dense skip connections to encourage feature reuse; and *ConvNeXt-Tiny* [Liu+22], a recent design inspired by vision transformers that modernizes convolutional networks with improved training stability and accuracy. Each backbone was initialized with ImageNet pre-trained weights and fine-tuned on the ImageCLEFmed dataset, ensuring a fair comparison across architectures.

To improve robustness against the severe label imbalance and long-tail distribution inherent to the dataset, several **ensemble strategies** were implemented (Section 3.3.2). These included: (i) *Union-based ensembling*, which maximizes recall by merging predicted labels across models; (ii) *Intersection-based ensembling*, which enforces high precision by restricting predictions to labels agreed upon by multiple models; (iii) *Dual-threshold ensembling*, a hybrid approach that applies a more conservative threshold to frequent concepts and a lower threshold to rare ones, thus balancing recall and precision; and (iv) *Partial-intersection*, which allows flexible consensus rules among ensembled models. These ensemble designs were motivated by the need to explore different points along the precision–recall trade-off, particularly in the presence of rare medical concepts.

Beyond ensembling, a **per-label threshold optimization** approach was also evaluated, inspired by the AUEB system in previous ImageCLEF editions [Cha+25]. In this method, the decision threshold for each concept was not fixed globally but instead tuned individually using a coordinate-ascent procedure on the development set. This adaptation is crucial in multi-label biomedical settings, where certain classes (e.g., *X-ray*, *MRI*) may benefit from conservative thresholds due to their frequency, while rare classes (e.g., *MRI venography*) require more permissive thresholds to avoid under-prediction.

In total, **sixteen systems** were shortlisted for submission, selected on the basis of development set performance (Section 4). These systems cover a representative spectrum of CNN backbones, ensemble aggregation strategies, and thresholding approaches, enabling a systematic comparison of architectural and methodological choices under the official evaluation metrics of the challenge.

5.3.2 Evaluation Metrics

The evaluation of systems in the Concept Detection task followed the official protocol of the ImageCLEFmedical 2025 campaign, which adopts the F_1 **score** as the primary performance metric. The F_1 score is particularly well-suited to multi-label biomedical tasks because it balances precision and recall, thereby penalizing both false positives (over-prediction of irrelevant concepts) and false negatives (failure to detect relevant ones). This trade-off is essential in clinical contexts: excessive false positives may clutter automated reports with irrelevant information, while false negatives risk omitting critical findings.

Formally, system outputs are represented as binary multi-hot vectors $y_{\text{pred}} \in \{0,1\}^L$, where L denotes the label vocabulary size, and compared against gold-standard vectors y_{true}. For each image $t \in T$ (where T is the test set), an individual F_1 score $\hat{f}_1(y_{\text{pred}}, y_{\text{true}})$ is computed by comparing the predicted and ground-truth concept sets. The overall metric is then obtained by averaging across all test samples:

$$F_1 = \frac{1}{|T|} \sum_{t \in T} \hat{f}_1(y_{\text{pred}}, y_{\text{true}}), \tag{5.1}$$

where implementation followed the standard `scikit-learn` procedure.[1]

In addition to the primary score, the organizers also introduced a **secondary F_1 metric**, restricted to a subset of manually selected concept categories, including imaging modalities (*angiogram*, *X-ray computed tomography*, *magnetic resonance imaging*, *positron-emission tomography*, *ultrasonography*, *plain X-ray*, *optical coherence tomography*), and anatomical structures (*upper extremity*, *lower extremity*, *vertebral column*, *pelvis*, *bone structure of cranium*, *chest*, *abdomen*, *breast*).

The rationale for this secondary evaluation is that not all UMLS concepts carry the same clinical importance: correctly identifying whether an image corresponds to an *MRI* of the *chest* is often more clinically useful than recognizing a highly specific but rare procedural label. To compute this secondary metric, both predictions and references are filtered to retain only concepts belonging to the selected categories, and images with no remaining ground-truth concepts are excluded. This yields a complementary perspective: while the primary F_1 score measures global tagging accuracy across the full label space, the secondary score emphasizes clinically actionable information.

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html`

Taken together, the use of these two F_1 measures ensures a balanced assessment: systems are rewarded for broad concept coverage while also being evaluated on their ability to capture the categories that are most relevant in clinical and diagnostic contexts.

5.3.3 Results and Submissions

Table 5.6 summarizes the results of all submitted systems to the ImageCLEFmedical 2025 Concept Detection task, reporting performance on both our internal development split and the official test set, together with competition rankings. The submissions span individual CNN–FFNN models, per-label thresholding, and a variety of ensemble configurations, including union, intersection, dual-threshold, and partial-intersection strategies.

Reading Table 5.6. A clear pattern emerges across submissions: ensembles (Runs 1980–1986) dominate the leaderboard, with Run 1980 achieving the highest overall performance ($F_1 = 0.5887$ on the official test set) and ranking **first place** among the nine participating teams. Secondary F_1 scores are particularly strong, often exceeding 0.95 and peaking at 0.9589 (Run 1986), which highlights the effectiveness of our methods on semantically critical categories such as imaging modality and anatomy. These consistently high secondary scores suggest that, in addition to achieving strong global tagging accuracy, our systems capture clinically meaningful concepts with notable reliability.

Comparison of Ensembles vs. Individual Models. Individual CNN–FFNN variants (e.g., Run 1971 with EfficientNet-B0, $F_1 = 0.5840$) were competitive but consistently outperformed by ensembles. The best-performing ensemble (Run 1980) combined Monte-Carlo EfficientNet-B0, DenseNet-121, ConvNeXt-Tiny, and an additional EfficientNet-B0 under a dual-threshold scheme, demonstrating the importance of architectural diversity as well as aggregation. Even smaller ensembles (e.g., Run 1979, Dual-2 with EfficientNet-B0 and DenseNet-121) achieved top-three rankings, showing that gains from ensembling are not limited to large model combinations but arise even with two or three complementary backbones.

Effect of Threshold Optimization. Threshold-per-label optimization (Run 1985, $F_1 = 0.5773$) provided modest improvements compared to naive global thresholds, but it remained below the performance of ensembles. This indicates that while fine-grained threshold tuning can sharpen decision boundaries, model diversity and aggregation exert a stronger influence on predictive performance.

Discussion. Taken together, these findings demonstrate three important conclusions. First, ensemble methods provide consistent and often substantial gains over individual CNN–FFNN models, confirming the benefit of combining multiple architectures and decision rules. Second, the secondary F_1 results indicate that improvements are not merely driven by frequent labels, but extend to clinically salient categories that are central to biomedical image retrieval and decision support. Finally, while threshold optimization offers incremental gains, it is markedly less impactful than ensemble diversity, suggesting that the most effective path toward robust performance lies in leveraging complementary architectural strengths. Overall, our submissions were highly competitive, with the top system ranking **first overall** in the 2025 competition and demonstrating state-of-the-art performance in both general and clinically focused metrics.

Tab. 5.6: Summary of submissions to the ImageCLEFmedical 2025 Concept Detection task. Results are reported on the internal development split and the official test set. Secondary F_1 scores correspond to manually selected clinical categories. Abbreviations: **MC**: Monte-Carlo, **EB0**: EfficientNet-B0, **D121**: DenseNet-121, **CN**: ConvNeXt-Tiny, **Dual-L**: dual-threshold with L base models.

Run ID	Method	F_1		Secondary F_1	Rank
		Dev	Test		
1980	Dual-3 (MC(EB0), D121, CN, EB0)	**0.5973**	**0.5887**	0.9484	1
1981	Dual-3 (MC(EB0), D121, EB0)	–	0.5880	0.9506	2
1979	Dual-2 (EB0, D121)	–	0.5873	0.9522	3
1977	Dual-2 (MC(EB0), EB0)	–	0.5867	0.9449	4
1982	Dual-3 (MC(EB0), EB0)	–	0.5866	0.9507	5
1978	Dual-2 (MC(EB0))	0.5945	0.5866	0.9465	6
1976	Dual-2 (MC(EB0), D121, EB0)	–	0.5864	0.9435	7
1975	Dual-2 (MC(EB0), D121, CN, B0)	0.5947	0.5858	0.9388	8
1983	Dual-3 (MC(B0))	0.5942	0.5855	0.9515	9
1986	Partial-Inter (MC(EB0), CN, D121)	0.5931	0.5853	**0.9589**	10
1971	CNN-FFNN (EB0)	0.5915	0.5840	0.9488	11
1970	Union(Inter(MC(EB0)), Inter(EB0, D121, CN))	0.5923	0.5819	0.9520	12
1973	CNN-FFNN (D121)	0.5909	0.5817	0.9462	13
1974	CNN-FFNN (CN)	0.5925	0.5808	0.9334	14
1985	Threshold-per-Label	0.5875	0.5773	0.9456	16
1984	Dual-3 (MC(EB0), D121)	0.5954	0.5755	0.9446	20

5.4 Discussion

The experiments conducted in this chapter provide a comprehensive evaluation of ensemble-based approaches across two distinct yet complementary domains—conformal prediction for multilabel classification and biomedical concept detection—and several unifying conclusions emerge. In the conformal prediction setting, ensemble conformal predictors (ECP) consistently improved upon single-model baselines by generating sharper and more compact prediction sets while preserving valid coverage guarantees, thereby addressing the dual challenge of predictive accuracy and principled uncertainty quantification. In contrast, the ImageCLEFmedical concept detection task highlighted the unique difficulties of extremely large, imbalanced label spaces, where ensembles of CNN–FFNN architectures proved most effective, delivering competitive F_1 scores and securing leading positions in

the 2025 evaluation campaign. Despite these differences in focus—uncertainty calibration versus large-scale biomedical tagging—both strands of experimentation converge on the central role of ensembling as a mechanism for enhancing robustness, with aggregation strategies such as weighted voting, dual-thresholding, and stacking emerging as decisive in balancing recall, precision, and overall system reliability. At a broader level, these findings reinforce the principle that model diversity is a powerful and general strategy for tackling the inherent complexity of multilabel problems across domains. At the same time, several limitations must be acknowledged: computational constraints limited the extent of hyperparameter exploration and restricted the ensemble sizes tested; reliance on pre-extracted CLIP embeddings in the conformal prediction experiments, while efficient, curtailed the investigation of end-to-end representation learning; and dataset-specific biases—such as strong co-occurrence patterns in COCO or publication-driven imbalances in ImageCLEFmedical—may affect the generalizability of results. Moreover, evaluation protocols in both contexts emphasized F_1 scores and coverage-based metrics, which, while informative, do not capture other important dimensions such as interpretability, cost sensitivity, or clinical usability. Nonetheless, the collective evidence demonstrates that ensemble methods, whether through conformal predictors in general-purpose classification or CNN-based systems in medical imaging, not only yield measurable improvements in predictive performance but also strengthen the reliability and trustworthiness of multilabel classification systems, underscoring their value as a broadly applicable methodological paradigm.

6 Conclusions and Future Work

This thesis investigated ensemble-based methods for multilabel classification across two complementary settings: **conformal prediction for uncertainty-aware classification** and **concept detection in biomedical imaging**. Despite their differences in domain and evaluation protocols, both case studies converge on a central insight: *model diversity and principled aggregation are key drivers of improved predictive performance and reliability in multilabel learning*.

In the conformal prediction part, a novel **Ensemble Conformal Prediction (ECP)** framework was introduced and evaluated on three benchmark datasets spanning vision (MS-COCO [Lin+14]), biology (Yeast [EW01]), and music information retrieval (Emotions [Tro+08]). The empirical results demonstrate that ECP consistently outperforms single-model conformal predictors and post-hoc conformalized ensembles. By integrating ensembling directly into the calibration pipeline, ECP achieves a more favorable trade-off between empirical coverage, compact prediction sets, and predictive accuracy. Detailed ablation studies further revealed that ensembles of moderate size ($M = 3$–5) provide the strongest balance between computational cost and predictive robustness, that heterogeneous ensembles outperform homogeneous counterparts due to their greater diversity, and that careful tuning of the miscoverage rate α enables practitioners to flexibly navigate the trade-off between reliability and informativeness. Together, these findings establish ECP as an effective and generalizable approach to multilabel classification under uncertainty [VGS05; AB21].

In the biomedical domain, the thesis addressed the **ImageCLEFmedical 2025 Concept Detection** task, which involves predicting large sets of medical concepts (UMLS CUIs [Bod04]) from radiology images. The experiments highlight the extreme challenges posed by this setting: a vocabulary of 2,479 labels, severe long-tail distributions, and high label co-occurrence. Systems based on CNN–FFNN architectures were developed with multiple convolutional backbones (EfficientNet [TL19], DenseNet [Hua+17b], ConvNeXt [Liu+22]), and their outputs were combined using ensemble strategies such as dual-thresholding, union, and partial-intersection aggregation. The results demonstrate that ensembles consistently outperformed individual models, both in terms of the primary F_1 score and a secondary F_1 score focused on clinically critical concepts (e.g., anatomy and modality). Importantly, the best ensemble system ranked **first overall** in the 2025 competition [**OverviewImageCLEF2025**], underlining the practical competitiveness of the proposed methodology. These findings underscore that ensembling not only improves

accuracy but also enhances robustness in the face of extreme class imbalance and complex label dependencies.

Although the two domains differ substantially—ECP focusing on principled uncertainty quantification and ImageCLEFmed concept detection addressing large-scale biomedical tagging—the experimental results lead to a unifying conclusion: **ensembling is a versatile and powerful paradigm for multilabel classification**. In both cases, diversity among base learners and carefully designed aggregation rules proved decisive, yielding more reliable, accurate, and trustworthy models. At the same time, limitations must be acknowledged: experiments were constrained by computational resources, restricting large-scale hyperparameter optimization and the exploration of deeper or larger ensembles; the reliance on pre-extracted embeddings in the conformal prediction setting (CLIP [Rad+21]) limited the potential of end-to-end feature learning; and dataset-specific biases, such as co-occurrence in COCO or publication-driven imbalances in ImageCLEFmedical, may have influenced results.

Future Work

Several promising directions arise from this work. For conformal prediction, extending ECP to end-to-end deep learning architectures (e.g., vision transformers [Dos+20] or multimodal encoders [AD+22]) would allow tighter integration of representation learning and calibration. Investigating adaptive or dynamic ensemble sizes—where the number of models varies with input difficulty—may further optimize the trade-off between computational cost and predictive performance. Moreover, exploring richer nonconformity measures, particularly those informed by uncertainty estimates from Bayesian deep learning [GG16] or ensemble-based uncertainty [LPB17], could strengthen the theoretical and practical guarantees of ECP.

For biomedical concept detection, future research should focus on more principled strategies to handle the extreme long-tail distribution of labels, such as incorporating label embeddings [RK18], graph-based regularization [KW17], or self-supervised pretraining on biomedical image–text pairs [Zha+20]. Another avenue lies in combining conformal prediction with concept detection, producing *uncertainty-aware biomedical tagging systems* that not only predict relevant concepts but also provide calibrated confidence sets to support clinical decision-making. Finally, broader evaluation metrics beyond F_1—such as interpretability, clinical utility, and cost-sensitive accuracy—would allow a more holistic assessment of system performance in real-world healthcare settings.

In conclusion, this thesis demonstrates that ensemble learning, when carefully integrated with conformal prediction or CNN architectures, provides a principled and effective strategy

for tackling the dual challenges of predictive accuracy and uncertainty quantification in multilabel classification. The findings not only advance the methodological understanding of ensemble-based approaches but also provide practical contributions to domains as diverse as computer vision, bioinformatics, and medical imaging, where reliable and trustworthy predictions are of paramount importance.

Bibliography

[AB21] Anastasios N Angelopoulos and Stephen Bates. "A gentle introduction to conformal prediction and distribution-free uncertainty quantification". In: *arXiv preprint arXiv:2107.07511* (2021).

[AD+22] Jean-Baptiste Alayrac, Dustin Donato, et al. "Flamingo: a visual language model for few-shot learning". In: *arXiv preprint arXiv:2204.14198* (2022).

[BBK19] Edmon Begoli, Tanmoy Bhattacharya, and A Gilad Kusne. "The need for uncertainty quantification in machine-assisted medical decision making". In: *Nature Machine Intelligence* 1.1 (2019), pp. 20–23.

[Bod04] Olivier Bodenreider. "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology". In: *Nucleic Acids Research* 32.suppl_1 (2004), pp. D267–D270.

[Bre96] Leo Breiman. "Bagging predictors". In: *Machine learning* 24.2 (1996), pp. 123–140.

[CEN14] Lars Carlsson, Martin Eklund, and Ulf Norinder. "Aggregated Conformal Prediction". In: *Artificial Intelligence Applications and Innovations (AIAI)*. Vol. 437. IFIP AICT. Springer, 2014, pp. 231–240.

[CGD21a] Maxime Cauchois, Chirag Gupta, and John C. Duchi. "Knowing What You Know: Valid and Validated Confidence Sets in Multiclass and Multilabel Prediction". In: *Journal of Machine Learning Research* 22.228 (2021), pp. 1–42.

[CGD21b] Maxime Cauchois, Suyash Gupta, and John C. Duchi. "Knowing What You Know: Valid and Validated Confidence Sets in Multiclass and Multilabel Prediction". In: *Journal of Machine Learning Research* 22.81 (2021), pp. 1–42.

[CGZ06] Nicolò Cesa-Bianchi, Claudio Gentile, and Luca Zaniboni. "Hierarchical classification: Combining Bayes with SVM". In: *Proceedings of the 23rd International Conference on Machine Learning*. ACM. 2006, pp. 177–184.

[Cha+15] Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. "Addressing imbalance in multilabel classification: Measures and random resampling algorithms". In: *Neurocomputing* 163 (2015), pp. 3–16.

[Cha+21] Foivos Charalampakos, Vasilis Karatzas, Vasiliki Kougia, John Pavlopoulos, and Ion 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, Bucharest, Romania, September 21-24*. Vol. 2936. CEUR Workshop Proceedings. 2021, pp. 1184–1200.

[Cha+22] Foivos Charalampakos, George Zachariadis, John Pavlopoulos, et al. "AUEB NLP Group at ImageCLEFmedical Caption 2022". In: *CLEF2022 Working Notes*. CEUR Workshop Proceedings. Bologna, Italy: CEUR-WS.or, 2022, pp. 1355–1373.

[Cha+25] Anna Chatzipapadopoulou, Ippokratis Pantelidis, Foivos Charalampakos, et al. "AUEB NLP Group at ImageCLEFmedical Caption 2025: Notebook for the AUEB NLP Group and Archimedes Unit at ImageCLEFmedical Caption 2025". In: *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*. CEUR-WS Workshop Proceedings, Vol. 4038, ISSN 1613-0073. Madrid, Spain, 2025.

[Cha25] Anna Chatzipapadopoulou. "Enhanced Biomedical Image Tagging". Bachelor's thesis. Athens University of Economics and Business, Department of Informatics, 2025.

[Che+19] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, et al. "Multi-label image recognition with graph convolutional networks". In: *CVPR*. 2019, pp. 5177–5186.

[Che19] Giovanni Cherubin. "Majority vote ensembles of conformal predictors". In: *Machine Learning* 108.3 (2019), pp. 501–527.

[DD09] Armen Der Kiureghian and Ove Ditlevsen. "Aleatory or epistemic? Does it matter?" In: *Structural Safety* 31.2 (2009), pp. 105–112.

[Den+09] Jia Deng, Wei Dong, Richard Socher, et al. "ImageNet: A large-scale hierarchical image database". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255.

[Dev+18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[Die00] Thomas G. Dietterich. "Ensemble methods in machine learning". In: *International workshop on multiple classifier systems* (2000), pp. 1–15.

[DN21] Prafulla Dhariwal and Alex Nichol. "Diffusion Models Beat GANs on Image Synthesis". In: *arXiv preprint arXiv:2105.05233* (2021).

[Dos+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: 2020.

[EW01] André Elisseeff and Jason Weston. "A kernel method for multi-labelled classification". In: *Advances in neural information processing systems*. Vol. 14. 2001, pp. 681–687.

[FRD21] Di Feng, Lars Rosenbaum, and Klaus Dietmayer. "A review of uncertainty estimation in deep learning for autonomous driving". In: *IEEE Transactions on Intelligent Vehicles* 6.2 (2021), pp. 195–209.

[FS97] Yoav Freund and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences*. Vol. 55. 1. Elsevier, 1997, pp. 119–139.

[GG16] Yarin Gal and Zoubin Ghahramani. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. 2016. arXiv: `1506.02142 [stat.ML]`.

[GR24] Matteo Gasparin and Aaditya Ramdas. *Conformal Online Model Aggregation*. arXiv:2403.15527. 2024.

[Guo+17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. "On calibration of modern neural networks". In: *ICML*. 2017, pp. 1321–1330.

[He+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *CVPR*. 2016, pp. 770–778.

[HG09] Haibo He and Edward A Garcia. "Learning from imbalanced data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284.

[Ho98] Tin Kam Ho. "The random subspace method for constructing decision forests". In: *IEEE transactions on pattern analysis and machine intelligence* 20.8 (1998), pp. 832–844.

[HS90] Lars Kai Hansen and Peter Salamon. "Neural network ensembles". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.10 (1990), pp. 993–1001.

[Hsu+09] Daniel Hsu, Sham Kakade, John Langford, and Tong Zhang. "Multi-label prediction via compressed sensing". In: *Advances in Neural Information Processing Systems*. Vol. 22. 2009.

[Hua+17a] Gao Huang, Yixuan Li, Geoff Pleiss, et al. "Snapshot ensembles: Train 1, get M for free". In: *International Conference on Learning Representations (ICLR)*. 2017.

[Hua+17b] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA, 2017, pp. 2261–2269.

[Hua+20] Jianqiang Huang, Lina Song, Richang Hong, Meng Wu, and Meng Wang. "Graph neural networks for multi-label classification". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.4 (2020), pp. 4453–4460.

[Jia+12] Xiaoqian Jiang, Megan Osl, Jin-Young Kim, and Lucila Ohno-Machado. "Calibrating predictive uncertainty in medical decision support systems". In: *PLOS ONE*. Vol. 7. 7. 2012, e41350.

[Kal+23a] Panagiotis Kaliosis, George Moschovis, Foivos Charalampakos, John Pavlopoulos, and Ion Androutsopoulos. "AUEB NLP Group at ImageCLEFmedical Caption 2023". In: *CLEF2023 Working Notes*. CEUR Workshop Proceedings. Thessaloniki, Greece: CEUR-WS.org, 2023.

[Kal+23b] Panagiotis Kaliosis, Georgios Moschovis, Foivos Charalampakos, John Pavlopoulos, and Ion Androutsopoulos. "AUEB NLP Group at ImageCLEFmedical Caption 2023". In: *Conference and Labs of the Evaluation Forum (CLEF)*. Thessaloniki, Greece, 2023.

[KB17] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[Kha+22] Firas Khader, Gustav Mueller-Franzes, Soroosh Tayebi Arasteh, et al. "Medical Diffusion: Denoising Diffusion Probabilistic Models for 3D Medical Image Generation". In: *arXiv preprint arXiv:2211.03364* (2022).

[KP24] Nikolaos Katsios and Harris Papadopoulos. "Multi-Label Conformal Prediction with Mahalanobis Nonconformity". In: *Proceedings of the 41st International Conference on Machine Learning*. 2024.

[KPA19] Vasiliki Kougia, John Pavlopoulos, and Ion Androutsopoulos. "AUEB NLP Group at ImageCLEFmed Caption 2019". In: *CLEF2019 Working Notes*. CEUR Workshop Proceedings. Lugano, Switzerland: CEUR-WS.org, 2019.

[KPA20] Vasiliki Kougia, John Pavlopoulos, and Ion Androutsopoulos. "Medical Image Tagging by Deep Learning and Retrieval". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020)*. Springer International Publishing, 2020, pp. 154–166.

[Kun14] Ludmila I Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons, 2014.

[KW03] Ludmila I. Kuncheva and Christopher J. Whitaker. "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy". In: *Machine learning* 51.2 (2003), pp. 181–207.

[KW17] Thomas N Kipf and Max Welling. "Semi-supervised classification with graph convolutional networks". In: *International Conference on Learning Representations*. 2017.

[Lei+18] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. "Distribution-Free Prediction Sets". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1094–1111.

[Lin+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. "Microsoft COCO: Common objects in context". In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[Lin+17] Henrik Linusson, Ulf Norinder, Henrik Boström, Ulf Johansson, and Tuve Löfström. "On the Calibration of Aggregated Conformal Predictors". In: *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications (COPA)*. Vol. 60. PMLR. 2017, pp. 135–151.

[Liu+17] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. "Deep learning for extreme multi-label text classification". In: *SIGIR* (2017), pp. 115–124.

[Liu+21] Jingzhou Liu, Shiyu Chang, Jie Fu, Boqing Wang, and Ronan Collobert. "Deep learning for extreme multi-label text classification". In: *SIGKDD Explorations* (2021).

[Liu+22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, et al. "A ConvNet for the 2020s". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 11966–11976.

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).

[LZ25] Rui Luo and Zhixin Zhou. *Weighted Aggregation of Conformity Scores for Classification*. arXiv:2407.10230. v2. 2025.

[MNA16] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *arXiv preprint arXiv:1606.04797* (2016).

[Naj22] Reabal Najjar. "Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging". In: *Journal of Radiology* (2022). Ed. by Michał Strzelecki, Adam Piorkowski, and Rafał Obuchowicz.

[Nea12] Radford M Neal. *Bayesian learning for neural networks*. Springer, 2012.

[OPT25] Eduardo Ochoa Rivera, Yash Patel, and Ambuj Tewari. *Conformal Prediction for Ensembles: Improving Efficiency via Score-Based Aggregation*. arXiv:2405.16246. v3. 2025.

[Ova+19] Yaniv Ovadia, Emily Fertig, Jie Ren, et al. "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift". In: *NeurIPS*. 2019, pp. 13991–14002.

[Pap14] Harris Papadopoulos. "A Cross-Conformal Predictor for Multi-label Classification". In: *Artificial Intelligence Applications and Innovations (AIAI) Workshops*. Vol. 437. IFIP AICT. Springer, 2014, pp. 241–250.

[Pap22] Harris Papadopoulos. *A Cross-Conformal Predictor for Multi-label Classification*. arXiv:2211.16238. 2022.

[PBG+21] Otto Pelka, Asma Ben Abacha, Alba García Seco de Herrera, et al. "Overview of the ImageCLEFmed 2021 Concept & Caption Prediction Task". In: *CLEF2021 Working Notes*. Bucharest, Romania: CEUR Workshop Proceedings, 2021.

[Pel+19] Otto Pelka, Christoph M. Friedrich, Alba García Seco de Herrera, and Henning Müller. "Overview of the ImageCLEFmed 2019 Concept Prediction Task". In: *CLEF2019 Working Notes*. Vol. 2380. Lugano, Switzerland: CEUR Workshop Proceedings, 2019.

[Pel+20] Otto Pelka, Christoph M. Friedrich, Alba García Seco de Herrera, and Henning Müller. "Overview of the ImageCLEFmed 2020 Concept Prediction Task: Medical Image Understanding". In: *CLEF2020 Working Notes*. Vol. 1166. Thessaloniki, Greece: CEUR Workshop Proceedings, 2020.

[Qia+24] Xiaoyu Qian, Jinru Wu, Ligong Wei, and Youwu Lin. *Random Projection Ensemble Conformal Prediction for High-Dimensional Classification*. SSRN preprint 4794962. 2024.

[Rad+21] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 2021, pp. 8748–8763.

[Raj+17] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning". In: *arXiv preprint arXiv:1711.05225* (2017).

[RBG+22] Jonas Rückert, Asma Ben Abacha, Alba García Seco de Herrera, et al. "Overview of ImageCLEFmedical 2022 - Caption Prediction and Concept Detection". In: *CLEF2022 Working Notes*. Bologna, Italy: CEUR Workshop Proceedings, 2022.

[Rea+11] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. "Classifier chains for multi-label classification". In: *Machine Learning*. 2011, pp. 333–359.

[RK18] Anthony Rios and Ramakanth Kavuluru. "Few-shot and zero-shot multi-label learning for structured label spaces". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (2018), pp. 3132–3142.

[RTC19] Filip Radenović, Giorgos Tolias, and Ondřej Chum. "Fine-Tuning CNN Image Retrieval with No Human Annotation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.7 (2019), pp. 1655–1668.

[Rüc+24] Johannes Rückert, Louise Bloch, Raphael Brüngel, et al. "ROCOv2: Radiology Objects in COntext Version 2, an Updated Multimodal Image Dataset". In: *Scientific Data* (2024).

[Sam+24] Marina Samprovalaki, Anna Chatzipapadopoulou, Georgios Moschovis, et al. "AUEB NLP Group at ImageCLEFmedical Caption 2024". In: *Notebook for the AUEB NLP Group at ImageCLEFmedical Caption 2024, Conference and Labs of the Evaluation Forum (CLEF)*. Athens, Greece, 2024.

[Sel+20] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *International Journal of Computer Vision (IJCV)* 128 (2020), pp. 336–359. arXiv: arXiv:1610.02391 [cs.CV].

[Shi+18] Hoo-Chang Shin, Neil A. Tenenholtz, Jameson K. Rogers, et al. "Medical Image Synthesis for Data Augmentation and Anonymization using Generative Adversarial Networks". In: *arXiv preprint arXiv:1807.10225* (2018).

[SS00] Robert E Schapire and Yoram Singer. "Boostexter: A boosting-based system for text categorization". In: *Machine Learning*. Vol. 39. 2-3. Springer, 2000, pp. 135–168.

[SS99] Robert E. Schapire and Yoram Singer. "Improved boosting algorithms using confidence-rated predictions". In: *Machine Learning*. Vol. 37. 3. 1999, pp. 297–336.

[SV08] Glenn Shafer and Vladimir Vovk. "A tutorial on conformal prediction". In: *Journal of Machine Learning Research* 9 (2008), pp. 371–421.

[TG23] Chhavi Tyagi and Wenge Guo. "Multi-label Classification under Uncertainty: A Tree-based Conformal Prediction Approach". In: *Conformal and Probabilistic Prediction with Applications (COPA)*. Vol. 204. PMLR. 2023, pp. 1–25.

[TG24] Aditya Tyagi and Chao Guo. "Hierarchical Conformal Prediction". In: *arXiv preprint arXiv:2404.19472* (2024).

[TKV10] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. "Mining multi-label data". In: *Data mining and knowledge discovery handbook* (2010), pp. 667–685.

[TL12] Farbound Tai and Hsuan-Tien Lin. "Multilabel classification with principal label space transformation". In: *Proceedings of the 2nd International Workshop on Statistical Relational AI*. 2012.

[TL19] Mingxing Tan and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Vol. 97. Proceedings of Machine Learning Research. 2019, pp. 6105–6114.

[Tro+08] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis Vlahavas. "Multi-label classification of music by emotion". In: *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*. 2008, pp. 325–330.

[TV07] Grigorios Tsoumakas and Ioannis Vlahavas. "Random k-labelsets for multi-label classification". In: *Machine Learning: ECML 2007*. Springer. 2007, pp. 406–417.

[TV11] Grigorios Tsoumakas and Ioannis Vlahavas. "Random k-labelsets for multilabel classification". In: *IEEE transactions on knowledge and data engineering*. 2011.

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. "Attention is all you need". In: *NeurIPS*. 2017, pp. 5998–6008.

[VGS05] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, Jan. 2005.

[Vov15] Vladimir Vovk. "Cross-conformal predictors". In: *Annals of Mathematics and Artificial Intelligence* 74.1 (2015), pp. 9–28.

[Wan+17] Xiaosong Wang, Yifan Peng, Le Lu, et al. “ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases”. In: *CVPR*. 2017.

[Wol92] David H. Wolpert. “Stacked generalization”. In: *Neural networks* 5.2 (1992), pp. 241–259.

[Wu+20] Zonghan Wu, Shirui Pan, Fengwen Chen, et al. “A comprehensive survey on graph neural networks”. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (2020), pp. 4–24.

[YK25] Yachong Yang and Arun Kumar Kuchibhotla. “Selection and Aggregation of Conformal Prediction Sets”. In: *Journal of the American Statistical Association* (2025). to appear; see arXiv:2104.13871.

[You+19] Ronghui You, Zihan Zhang, Shuo Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. “AttentionXML: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification”. In: *NeurIPS*. 2019, pp. 5820–5830.

[Zha+20] Sheng Zhang et al. “Contrastive learning of medical visual representations from paired images and text”. In: *arXiv preprint arXiv:2010.00747* (2020).

[ZZ07] Min-Ling Zhang and Zhi-Hua Zhou. “ML-KNN: A lazy learning approach to multi-label learning”. In: *Pattern recognition* 40.7 (2007), pp. 2038–2048.

[ZZ14] Min-Ling Zhang and Zhi-Hua Zhou. “A review on multi-label learning algorithms”. In: *IEEE Transactions on Knowledge and Data Engineering* 26.8 (2014), pp. 1819–1837.

List of Acronyms

AI	Artificial Intelligence
CNN	Convolutional Neural Network
CP	Conformal Prediction
FFNN	Feed-Forward Neural Network
F1	F1-score
MLC	Multi-Label Classification
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
CT	Computed Tomography
MS-COCO	Microsoft Common Objects in Context
ImageCLEFmedical	ImageCLEFmedical Caption Challenge
BR	Binary Relevance
CC	Classifier Chains
LP	Label Powerset
PLP	Pruned Label Powerset
RAkEL	Random k-Labelsets

ML-kNN	Multi-Label k-Nearest Neighbors
SVM	Support Vector Machine
RNN	Recurrent Neural Network
GNN	Graph Neural Network
BNN	Bayesian Neural Network
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
ECE	Expected Calibration Error
MCE	Maximum Calibration Error
OU	Observed Unconfidence
OF	Observed Fuzziness
FWER	Family-Wise Error Rate
MLC	Multilabel Classification
LR	Logistic Regression
SGD	Stochastic Gradient Descent
MLP	Multilayer Perceptron
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
FFNN	Feed-Forward Neural Network
CLIP	Contrastive Language–Image Pretraining
MV	Majority Voting

PA	Probability Averaging
CP	Conformal Prediction
ECP	Ensemble Conformal Prediction
BCE	Binary Cross-Entropy
CNN	Convolutional Neural Network
GeM	Generalized Mean
Adam	Adaptive Moment Estimation
StackECP	Stacked Ensemble Conformal Prediction
F1	F1 Score
ConvNeXt-Tiny	Convolutional Network Next - Tiny
ImageCLEF	Image Retrieval Evaluation Campaign

List of Figures

List of Tables

List of Algorithms