# Efficient Multilabel Uncertainty Quantification with Conformal Ensembles

No Author Given

No Institute Given

**Abstract.** In high-stakes domains such as medical diagnosis or fault detection, high accuracy alone is insufficient, missing a critical case (e.g., a false negative) can have severe consequences, making reliable uncertainty quantification essential. Conformal prediction (CP), which offers distribution-free coverage guarantees, help towards that direction, but it often produces unstable or overly large prediction sets, contradicting its practical utility, for instance in human-computer interaction where a clinician must review each predicted label. Multilabel classification (MLC) increases further the challenge, because the model predicts multiple labels per instance in a typically large and imbalanced label space. The increased uncertainty, compared to binary or multiclass tasks, motivates the following research question: *how can we obtain smaller, more informative prediction sets from trained MLC models while preserving marginal coverage and maintaining the theoretical guarantees of CP?* To address this question, we investigate ensembling, which can improve stability and efficiency yet its potential in MLC has not been fully explored. We conduct a systematic empirical study across standard MLC benchmarks (COCO, Yeast, Emotions, Scene) building ensembles under (i) majority voting, (ii) calibrated aggregation of nonconformity scores, and (iii) performance-weighted aggregation. We find that our ensembles for all three categories consistently improve over single-model CP yielding more efficient prediction sets (smaller and more informative), while maintaining target coverage and achieving higher macro-F1 scores. Majority-voting ensembles, however, also satisfy theoretical lower bounds.

**Keywords:** Conformal Prediction · Machine Learning · Ensembles · Uncertainty.

## 1 Introduction

Multilabel classification (MLC) arises in a wide range of applications such as image tagging, genomics, music emotion recognition, and clinical decision support [4, 31, 38, 9, 30, 24]. Compared to single-label prediction, MLC requires issuing decisions over a potentially large label set, typically with strong label imbalance and some labels appearing only very rarely [38, 31]. These characteristics amplify predictive uncertainty and make reliable confidence assessment especially important, particularly in high-stakes domains such as medicine and

autonomous systems, where models must not only be accurate but also convey when predictions may be unreliable [20, 1].

**Conformal prediction (CP)** provides a rigorous framework for uncertainty quantification with finite-sample marginal coverage guarantees [34, 29]. Rather than outputting a single label decision, CP produces a *prediction set* of candidate labels guaranteed to contain the true label with probability at least $1 - \alpha$, where $\alpha \in (0, 1)$ is a user-specified miscoverage level. Although CP can be extended to multi-label classification (MLC) via label-wise calibration [27, 28], the resulting prediction sets tend to grow with the number of labels, often becoming uninformatively large in practice [3, 6].

**Ensemble learning** can address this by combining multiple models to reduce variance, improve calibration, and yield stronger uncertainty estimates [8, 5, 35, 14, 20]. Building on this, prior work has explored ensemble CP in single and multi-class classification and regression settings, examining strategies such as majority voting, weighted aggregation, and score-based fusion of conformal scores [7, 10, 26].

Existing multilabel ensemble methods, such as CP-RAkEL [36], construct ensembles over random label subsets, but they neither combine multiple CP systems nor provide MLC coverage guarantees.

**In this work**, we present **the first empirical study** to systematically combine multiple CP systems via ensemble aggregation in the MLC setting, introducing label-wise calibration, MLC-aware aggregation rules, and formal label-wise coverage guarantees. Our contribution is threefold:

1. **Theory.** We extend known majority-vote coverage results [7] and recent score/weighted aggregation analyses [10, 26] to the MLC context, giving label-wise coverage bounds under independence and dependence assumptions.
2. **Framework.** We implement homogeneous and heterogeneous MLC ensembles using classical models (LR, SGD, MLP) and modern architectures (RNN, Transformer, MLP-Mixer), with conformal calibration applied at the label level before aggregation.
3. **Experiments.** On four benchmark MLC datasets, MS-COCO [18], Yeast [9], Scene [4] and Emotions [30], we show that MLC ensembles consistently improve F1, maintain or improve coverage, and reduce average prediction set size compared to single-model CP and post-hoc conformalized ensembles.

## 2 Background and Related Work

### 2.1 Multilabel Classification

MLC extends traditional classification by allowing each instance to be associated with multiple labels simultaneously [31, 38]. This formulation arises in domains as varied as image annotation [4], text categorization [17], genomics [9], and audio tagging [13].

Early methods include *binary relevance* (BR), which trains an independent binary classifier per label, and transformation-based approaches such as classifier chains and label powerset, which explicitly model label correlations [25, 32]. More recently, deep learning architectures, including CNNs, RNNs, and Transformers have achieved state-of-the-art performance on high-cardinality MLC tasks [19, 37, 15]. Despite these advances, reliable calibration and principled uncertainty quantification remain open challenges, particularly in the presence of label imbalance or rare labels.

### 2.2 Conformal Prediction

CP is a model-agnostic framework for constructing prediction sets with finite-sample marginal coverage guarantees [34, 29]. Given a target miscoverage rate $\alpha \in (0, 1)$, CP guarantees that the true label is included in the prediction set with probability at least $1-\alpha$, assuming only data exchangeability. This property makes CP particularly attractive in high-stakes domains where reliable uncertainty quantification is as important as predictive accuracy.

CP has been applied across a range of settings, including classification [21], regression [16], and structured prediction [3]. In the multilabel setting, label-wise calibration applies CP independently to each label by treating each as a binary task, leveraging the Mondrian CP framework [34]. While simple and computationally efficient, this approach can be conservative, as it ignores inter-label dependencies and often inflates prediction set size.

Several extensions have been proposed to improve the efficiency of conformal prediction sets: adaptive conformity scores that target both marginal and approximate conditional coverage [27, 2], and set-valued classification also studies ambiguity-efficient confidence sets with bounded error levels [28]. In the multiclass and multilabel setting, Cauchois et al [6] develop a coverage-efficiency framework for constructing confidence sets with finite-sample guarantees for a single underlying predictor. However, much of this literature focuses on calibrating one base predictor at a time rather than on ensemble design in which multiple conformalized predictors are combined. These observations motivate ensemble-aware conformal approaches that combine multiple models while preserving statistical guarantees.

### 2.3 Ensemble Methods and Conformal Ensembles

Ensemble learning is a well-established approach for improving predictive performance and robustness by aggregating multiple predictors [8, 5, 35]. Deep ensembles, in particular, have shown strong performance for uncertainty quantification under distribution shift [14, 20].

Integrating ensembles with CP has received increasing attention. Cherubin [7] analyzed majority-vote ensembles of conformal predictors, establishing coverage guarantees under independence and dependence assumptions in the single-label setting. Gasparin and Ramdas [10] studied weighted set-merging strategies, while

Rivera et al. [26] proposed score-based aggregation methods to reduce conservatism. Related work also considers fusing conformity scores across models [11] and provides practical guidance on conformal prediction in modern deep-learning pipelines [2]. Despite this progress, most existing conformal ensembling methods are developed and evaluated primarily for single-label prediction problems.

*Our focus.* In contrast to the broad progress, the use of ensembles in *multilabel* conformal prediction remains largely unexplored. Existing MLC-specific CP methods rely on single models and do not leverage ensemble diversity. Our work bridges this gap: we adapt majority voting, averaged nonconformity scores, and performance-weighted aggregation to the multilabel setting, provide theoretical label-wise coverage bounds, and validate their performance across benchmark MLC datasets.

## 3 Methodology

### 3.1 Problem Setup

We address the task of MLC, where each input $x \in \mathbb{R}^d$ may be associated with multiple relevant labels simultaneously. Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the input space and $\mathcal{Y} = \{0,1\}^L$ the label space, where $L$ is the total number of labels and $y \in \mathcal{Y}$ is a binary vector indicating the presence or absence of each label.

Given a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, our objective is to learn a function $f : \mathcal{X} \to [0,1]^L$ that outputs per-label confidence scores. These scores can be thresholded to obtain a predicted label set $\hat{y} \in \{0,1\}^L$ for each input $x$.

Beyond standard multilabel metrics such as macro-F1, we assess prediction quality using conformal metrics including empirical coverage, marginal coverage, and average prediction set size, each formally defined in §5.2. Our methodology integrates conformal prediction techniques with multilabel classifiers to produce calibrated prediction sets that balance predictive accuracy and uncertainty quantification.

### 3.2 Ensemble Learning

To improve predictive robustness and uncertainty quantification, we adopt ensemble learning strategies both as stand-alone baselines and as integral components of our conformal prediction framework. We consider three ensemble types: homogeneous, heterogeneous, and stacked.

**Homogeneous** A homogeneous ensemble trains $M$ instances of the *same* base classifier (e.g., Logistic Regression) on bootstrap-resampled versions of the training data. The diversity here comes from data resampling rather than model variation. For each label, predictions from the $M$ models are aggregated via majority voting or probability averaging. In *Majority Voting* (MV), binary predictions from all $M$ models are combined by simple majority; i.e., a label is included if

more than half the models predict it as positive. In *Probability Averaging* (PA), the probabilistic outputs of all $M$ models are averaged, and the result is thresholded at 0.5 to form the final prediction. This strategy is particularly effective when base models are sensitive to initialization or data sampling, as averaging reduces the impact of any single model's errors.

**Heterogeneous** A heterogeneous ensemble combines *diverse* model architectures, including both linear models (LR, SGD) and nonlinear learners (MLP, RNN, Transformer, MLP-Mixer). The intuition is that architecturally diverse models make different types of errors, so their combination is more robust than any single model. Each model is trained per label under the binary relevance assumption, and predictions are aggregated via MV or PA, as above, or with F1-weighted voting. In MV, binary label predictions from all models are combined via unweighted majority vote. In PA, probabilistic outputs are averaged across all models and thresholded at 0.5. In *F1-Weighted Voting* (WV), each model's contribution is weighted by its per-label F1 score on a held-out validation set. Final predictions are obtained via a weighted sum thresholded at 0.5. This introduces label-specific adaptivity, assigning greater influence to models that demonstrate stronger validation performance for a given label.

**Stacked** Stacking is a meta-learning strategy where the predictions of multiple base classifiers are used as input features to train a higher-level meta-classifier. Concretely, the outputs of base classifiers (LR, SGD, MLP) on a held-out calibration set serve as features to train a Logistic Regression meta-classifier for each label. At inference time, the meta-classifier produces calibrated probabilities, which are then post-processed using conformal prediction thresholds to form the final multilabel prediction set.

### 3.3 Conformal Prediction

We provide uncertainty-aware predictions by applying **label-wise conformal prediction** under the binary relevance assumption, treating each label as an independent binary classification task. Rather than predicting a single label set, our method produces a *prediction set* per instance. For each label $\ell$ independently, the prediction set is guaranteed to contain $\ell$ whenever it is a true label with probability at least $1 - \alpha$:

$$\mathbb{P}\big(\ell \in \hat{Y}(x) \mid \ell \in Y\big) \;\geq\; 1 - \alpha, \tag{1}$$

where $\alpha \in (0, 1)$ is a user-specified miscoverage level (e.g., $\alpha = 0.1$ for 90% per-label target coverage). We note that this is a *marginal per-label* guarantee — it holds independently for each label but does not directly imply simultaneous coverage of all true labels jointly.

**Nonconformity Scores and Calibration** The key quantity in conformal prediction is the *nonconformity score*, which measures how unusual a given prediction is relative to the calibration data. Intuitively, a high nonconformity score for a positive label means the model is uncertain or wrong about that label.

For each label $j$, we train a binary probabilistic classifier $f_j$ (e.g., LR, MLP, or SGD) on the training set. We then compute nonconformity scores on a separate calibration set, restricted to instances where label $j$ is present ($y_i^{(j)} = 1$):

$$s_i^{(j)} = 1 - f_j(x_i), \quad \text{for calibration samples } i \text{ with } y_i^{(j)} = 1. \tag{2}$$

Here, $f_j(x_i) \in [0, 1]$ is the predicted probability of label $j$ for instance $x_i$. We restrict calibration to positive instances because our goal is to ensure that *true* labels are included in the prediction set, we therefore measure nonconformity only where the ground truth label is present. A low predicted probability for a positive label corresponds to a high nonconformity score, reflecting model uncertainty.

Given these scores, we compute a per-label threshold $q_j$ as the $(1 - \alpha)$-quantile:

$$q_j = \text{Quantile}_{1-\alpha}\left(\{s_i^{(j)}\}_{i:\, y_i^{(j)}=1}\right). \tag{3}$$

This threshold captures the level of nonconformity that the model must not exceed in order to be considered conforming with the calibration data.

**Prediction Sets** At test time, for a new instance $x$, we include label $j$ in the predicted set $\hat{Y}(x)$ if its nonconformity score falls at or below the calibrated threshold:

$$\hat{Y}(x) = \{j \,:\, 1 - f_j(x) \leq q_j\}. \tag{4}$$

Equivalently, label $j$ is included if the model's predicted probability $f_j(x) \geq 1 - q_j$. This formulation ensures that the prediction set contains all labels the model is sufficiently confident about, with the confidence level controlled by $\alpha$.

### 3.4 Ensemble Conformal Prediction: Theoretical Guarantees

We now analyze the theoretical coverage properties of Ensemble Conformal Prediction (ECP) when combining $M$ base conformal models with *majority voting*. Our results extend known analyses of majority vote conformal ensembles [7] to the multilabel setting.

**Setup and Assumptions** Let $\mathcal{D} = \{(x_i, Y_i)\}_{i=1}^n$ be a dataset where $Y_i \subseteq \mathcal{L} = \{1, \ldots, L\}$. We assume data points are i.i.d. from an unknown distribution $\mathcal{P}$, and that $M$ base conformal models are trained independently (e.g., via bootstrapping). For each label $\ell$, model $m$ provides a calibrated threshold $\tau_\ell^{(m)}$ such that individual marginal coverage holds:

$$\mathbb{P}_{(x,Y)\sim\mathcal{P}}\left(\ell \in \hat{Y}^{(m)}(x) \;\middle|\; \ell \in Y\right) \;\geq\; 1 - \alpha. \tag{5}$$

Our goal is to characterize the coverage of the final ensemble prediction set $\hat{Y}^{\text{ens}}(x)$.

**Lemma 1 (Majority Vote Coverage Bounds, cf. [7]).** *Let $X_1^{(\ell)}, \ldots, X_M^{(\ell)} \in \{0,1\}$ be indicators, where $X_m^{(\ell)} = 1$ if model $m$ includes label $\ell$ in its prediction set, and $X_m^{(\ell)} = 0$ otherwise. Assume that, conditional on $\{\ell \in Y\}$, the indicators are independent and satisfy:*

$$\mathbb{P}\big(X_m^{(\ell)} = 1 \mid \ell \in Y\big) \;\geq\; 1 - \alpha.$$

*Then, for any voting threshold $k \in \{1, \ldots, M\}$:*

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

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

$$\mathbb{P}\left(\sum_{m=1}^{M} X_m^{(\ell)} \geq M \,\middle|\, \ell \in Y\right) \;\geq\; (1-\alpha)^M.$$

*Understanding the voting threshold $k$.* The threshold $k$ controls how many models must agree before a label is included in the ensemble prediction set. At one extreme, $k = 1$ means any single model suffices, this is the most permissive rule and yields the highest coverage. At the other extreme, $k = M$ requires unanimity, all $M$ models must include the label, giving the most conservative rule and the lowest coverage guarantee.

To make this concrete: with $\alpha = 0.1$ and $M = 5$ models, each including a true label with probability at least 0.9, the $k = 1$ bound is close to 1, while the unanimity bound gives $(0.9)^5 \approx 0.59$. This trade-off between conservatism and inclusiveness motivates choosing $k$ based on the desired balance between coverage and prediction set size.

**Theorem 1 (Ensemble Conformal Coverage Bounds).** *Under independence across models and a unanimity voting threshold ($k = M$), the ensemble prediction set satisfies:*

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

*For intermediate thresholds $k < M$, the binomial lower bound from Lemma 1 applies.*

*Proof Sketch.* The lower bound follows directly from Lemma 1 under independence. The upper bound is trivially 1. Full proofs are provided in the Appendix.

*Practical implication.* The independence assumption is an idealization. In practice, base models trained on overlapping data exhibit correlated errors, and empirical coverage lies between the theoretical lower bound and one. This motivates empirical validation of coverage alongside these theoretical guarantees, which we carry out in §5.

**Empirical Validation** We verify these guarantees by showing (§5) that ensemble methods preserve or improve empirical and marginal coverage while reducing average prediction set size relative to single-model conformal baselines.

## 4 Methods

We evaluate two broad categories of methods. **Baseline methods** include individual classifiers and post-hoc conformal strategies, where CP is applied after model training and aggregation. **Our proposed methods** fully integrate conformal calibration into the ensemble process, so that each model is individually calibrated *before* aggregation takes place.

### 4.1 Base Classifiers

We employ a diverse suite of base classifiers spanning linear, shallow, and deep architectures. All models are trained per label under the binary relevance assumption, treating each label as an independent binary classification task. Linear and shallow models (LR, SGD, MLP) operate directly on input features, while deep models (RNN, Transformer, MLP-Mixer) use pretrained CLIP embeddings [23] as input representations, providing rich semantic features suitable for high-cardinality label spaces.
**Logistic Regression (LR)** is a linear model optimized using the logistic loss. It produces well-calibrated probabilistic outputs and serves as a strong and interpretable baseline. **Stochastic Gradient Descent (SGD)** is a linear classifier trained with online stochastic updates. Raw outputs are post-hoc calibrated using Platt scaling [22] (`CalibratedClassifierCV`) to produce reliable per-label probability estimates. **Multilayer Perceptron (MLP)** is a shallow feedforward network with a single hidden layer of 256 units, ReLU activations, and binary cross-entropy loss. **Recurrent Neural Network (RNN)** is a unidirectional LSTM [12] that processes CLIP embeddings sequentially, followed by a sigmoid-activated output layer producing per-label probabilities. **Transformer Encoder** is a two-layer self-attention encoder [33] with multi-head attention, processing CLIP embeddings as input tokens and producing label-wise confidence scores via a shared linear classifier. **MLP-Mixer** is a compact architecture applying layer normalization and alternating token- and channel-mixing via feedforward layers to CLIP embeddings, with a sigmoid output layer for per-label probabilistic prediction. Each model produces per-label confidence scores in $[0, 1]$, which are either thresholded directly for standard classification or used to compute nonconformity scores within our conformal prediction procedures (as described in §4).

### 4.2 Baselines

**Standard Classification** refers to individual models trained and evaluated per label without any conformal calibration. Predictions are obtained by threshold-

ing per-label probabilities at 0.5. This baseline isolates raw predictive performance with no uncertainty quantification.

**Single-Model CP** applies label-wise conformal prediction to a single classifier using a held-out calibration set, with no ensembling. This isolates the benefit of conformal calibration alone, without any ensemble contribution, and serves as the primary reference point for evaluating the added value of ensembling.

**Post-Hoc ECP** first aggregates the outputs of multiple ensemble members (via probability averaging or majority voting) to form a single combined predictor, and then applies conformal calibration to this aggregated output. This treats the ensemble as a black-box single model for conformal purposes. It serves as a direct comparison point to our proposed approach, where calibration is applied *before* aggregation rather than after.

### 4.3 Proposed ECP Framework

Our proposed framework applies conformal calibration to each base model *independently and prior to* any aggregation. This design has two key advantages over post-hoc conformalization: it preserves individual per-label coverage guarantees for each ensemble member, and it allows the aggregation step to combine already-calibrated uncertainty estimates rather than calibrating a single potentially miscalibrated combined output.

**Conformal Ensembles** evaluate both homogeneous and heterogeneous configurations. In homogeneous ensembles, $M$ instances of the same base model are trained on bootstrap-resampled data, and their individually calibrated outputs are combined via majority voting (MV) or probability averaging (PA). In heterogeneous ensembles, architecturally diverse models (LR, SGD, MLP, RNN, Transformer, MLP-Mixer) are each conformally calibrated and their outputs are combined via MV, PA, or WV, as described in §3.2.

**Conformal Stacking (StackECP)** extends the ensemble framework with a meta-learning layer. Base model outputs on a dedicated meta-training split are used as features to train a logistic regression meta-classifier per label. The meta-training split and the conformal calibration set are kept strictly non-overlapping via a three-way data split (i.e., training, meta-training, conformal calibration), ensuring that conformal thresholds are computed on data never seen by the meta-classifier.

## 5 Empirical Analysis

We evaluate all methods across four diverse MLC benchmarks, covering a range of label cardinalities, label densities, and domain characteristics.

### 5.1 Datasets

The selected benchmarks span vision, biology, and audio domains and vary in label cardinality, density, and class imbalance, allowing us to test both predictive accuracy and uncertainty calibration across different domains.

**COCO** [18] is a popular benchmark in computer vision. We use the multi-label version with pre-extracted CLIP features [23]. COCO containts over $120,000$ images and each image is associated with multiple object labels (e.g., `person`, `car`, `dog`). The dataset contains 80 labels and exhibits severe label imbalance, making it a challenging benchmark for multilabel uncertainty quantification.
**Scene** [4] is a standard multi-label benchmark for image classification, evaluated here in its tabular form. It contains 2,407 instances, each represented by 294 visual features. The task is to associate each instance with a subset of 6 possible semantic labels: `Beach`, `Sunset`, `FallFoliage`, `Field`, `Mountain`, and `Urban`.
**Yeast** is a classic bioinformatics dataset introduced by [9] for protein function prediction. It consists of 2,417 samples and 14 labels, with moderate label density. Each instance represents a gene with various expression-based features.
**Emotions** is a music-related dataset that maps songs to a set of emotion labels such as `happy`, `sad`, and `relaxing`. It contains 593 instances and 6 labels, it is relatively balanced, and it is commonly used in multi-label benchmarking [30].

### 5.2 Evaluation Metrics

We evaluate the performance using the following five metrics:
**Empirical Coverage (EC)** is the proportion of true labels captured by the prediction set, averaged across all validation instances:

$$\text{Empirical Coverage} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{Y}(x_i)\cap Y_i|}{\max(1,|Y_i|)}. \tag{6}$$

**Marginal Coverage (MC)** is the average probability that a true label is included in the predicted set, computed separately for each label and averaged:

$$\text{Marginal Coverage} = \frac{1}{L}\sum_{j=1}^{L}\mathbb{P}\left(j\in\hat{Y}(x)\,\middle|\,j\in Y\right). \tag{7}$$

**Average Set Size (ASS):** The average number of labels predicted per instance:

$$\text{Average Set Size} = \frac{1}{N}\sum_{i=1}^{N}|\hat{Y}(x_i)|. \tag{8}$$

**Macro-F1 Score** computes the F1 score independently for each label (by comparing the predicted binary label assignments to the ground-truth assignments) and then averages these scores equally across all labels.
**Utility Score (US):** To summarize the trade-off between prediction quality, empirical coverage, marginal coverage, and set efficiency with a single scalar, we define the following utility:

$$US \;=\; \lambda_1\,\text{MacroF1} \;+\; \lambda_2\,\text{EC} \;+\; \lambda_3\,\text{MC} \;+\; \lambda_4\left(1-\frac{\text{ASS}}{L}\right), \tag{9}$$

where $L$ is the number of labels, and $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$ are weights satisfying:

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1,$$

so that $US$ is a convex combination of the four components bounded in $[0, 1]$.

This formulation ensures that $US \in [0, 1]$, since each component is normalized to lie in $[0, 1]$. In particular, MacroF1 measures classification quality, EC captures empirical coverage, MC captures marginal coverage, and $1 - \text{ASS}/L$ measures prediction-set compactness. Here, $\text{ASS}/L$ is the average prediction-set size normalized by the total number of labels, so it can be interpreted as the mean proportion of the label space included in the prediction set. Thus, $\text{ASS}/L \in [0, 1]$: values close to 0 indicate small and specific prediction sets, whereas values close to 1 indicate large and less informative sets. By using $1 - \text{ASS}/L$, we convert this quantity into a compactness measure, so that higher values correspond to smaller, more informative prediction sets.

Overall, the proposed utility increases with better classification performance, better coverage, and greater set compactness. Therefore, higher $US$ values indicate methods that achieve a more favorable balance between predictive accuracy, validity, and efficiency. Each of these metrics is applied to all individual classifiers and to aggregated predictions of ensembles using conformal prediction (§3.2).

### 5.3 Training Setup

All models are trained per label under the binary relevance assumption. For tabular datasets (Yeast, Scene, Emotions), input features are standardized; for COCO, we use precomputed CLIP embeddings (ViT-B/32). Data is partitioned into proper training, calibration, and held-out test sets, with 30% reserved for calibration. Linear models (LR, SGD) are implemented in `scikit-learn`; neural models (MLP, LSTM, Transformer, MLP-Mixer) in PyTorch, trained with Adam ($lr = 10^{-3}$), binary cross-entropy loss, batch size 128, for 10–15 epochs. Ensemble base models are trained independently via bootstrap sampling. Nonconformity scores are derived from predicted probabilities ($1 - \hat{p}$), and per-label conformal thresholds are estimated on the calibration set at miscoverage rate $\alpha$. Predictions are aggregated via one of three strategies: probability averaging before conformalization, majority voting over individual prediction sets, or F1-weighted voting. All experiments run on a single NVIDIA GPU.

### 5.4 Experimental Results

Table 1 reports results across four benchmark datasets. We organize our analysis around three axes: coverage validity, prediction-set efficiency, and classification quality, summarized by US (Eq. 9).

***ECP achieves the best overall utility.*** Our proposed ECP framework attains the highest US on every dataset: *Hom. CP (MLP-MV)* on Emotions (US = 0.738) and Scene (US = 0.823), *Hom. CP (MLP-WA)* on Yeast (US = 0.633), and *Hom. CP*

*(RNN-MV)* on COCO (US = 0.818). The second-best entry is also an ECP variant in three out of four cases, confirming that the advantage is systematic rather than dataset-specific.

**Coverage is maintained while set sizes shrink.** All ECP variants respect the target marginal coverage of $1-\alpha = 0.90$, with empirical and marginal coverage values consistently near or above the target threshold. By contrast, some post-hoc baselines deviate noticeably: *Hom. MLP MV-CP* on Yeast reaches EC = 0.992 and MC = 0.997, and *CP (SGD)* on Emotions collapses to the trivial predictor (EC = MC = 1.000, ASS = 6.00, i.e., all labels are always predicted). These cases illustrate that applying conformal calibration *after* aggregation, or to a poorly-calibrated single model, can inflate prediction sets to the point of being uninformative. ECP avoids this pathology by calibrating each model *before* aggregation: on Emotions the best ECP variant produces ASS = 3.25 (vs. 6.00 for CP (SGD) and 4.11 for the post-hoc homogeneous ensemble), and on Scene ASS = 1.48 (vs. 2.70 for the post-hoc MV variant), both while maintaining valid coverage. On the largest and most challenging dataset, COCO (80 labels), ECP *Hom. CP (RNN-MV)* reduces ASS to 6.93, the lowest among all conformal methods, again without sacrificing coverage.

**ECP improves classification quality over single-model CP.** Beyond coverage and set size, ECP consistently raises macro-F1 over its single-model counterparts. On **Emotions**, *Hom. CP (MLP-MV)* achieves F1=0.668, compared with 0.542 for *CP (MLP)*. On **Yeast**, the best ECP variants, *Hom. CP (MLP-WA)* and *Het. CP (MV)*, reach F1=0.471, slightly improving over the strongest single-model baseline, *CP (MLP)* at 0.468. On **Scene**, *Hom. CP (MLP-MV)* attains F1=0.759 versus 0.743 for *CP (MLP)*. Similarly, on **COCO**, *Hom. CP (RNN-MV)* obtains F1=0.577, outperforming the best single-model conformal baseline, *CP (RNN)* at 0.541. These gains reflect the well-known variance-reduction effect of ensembling: combining multiple independently calibrated predictors smooths out label-wise errors and produces sharper, more informative prediction sets.

**ECP surpasses non-conformal methods in predictive accuracy on three out of four datasets**. Although the non-conformal baselines (BR, Ensemble (LR), Label Bagging) cannot be compared on conformal metrics (EC, MC, ASS, and US are undefined for them, as they carry no coverage guarantee), their macro-F1 scores serve as a useful upper-bound reference for raw classification quality. Notably, ECP exceeds the best non-conformal F1 on three datasets: on Emotions (ECP 0.668 vs. Label Powerset 0.640), on Yeast (ECP 0.471 vs. Label Powerset 0.467), and most clearly on Scene (ECP 0.759 vs. the non-conformal ensemble at 0.701). This result is somewhat surprising, because conformal thresholding is designed to control coverage rather than to optimize F1. To further investigate the behavior of our proposed ECP framework, we conduct ablation experiments on the COCO dataset examining the impact of ensemble size, sensitivity to the miscoverage rate $\alpha$, and computational runtime. Full results, proofs, and implementation code are provided in the supplementary material to support complete reproducibility.

Table 1: Multi-label conformal prediction results on four benchmark datasets. **Metrics:** EC = empirical coverage; MC = marginal coverage; ASS = avg. set size (↓); F1 = macro-F1 (↑); US = utility score (↑). **Bold** = best per column; underline = 2nd-best. — = not applicable (EC, MC, ASS, US are undefined for non-conformal methods). = proposed ECP (ours). †Evaluated on COCO only.

**(a) Emotions**

| Method | EC↑ | MC↑ | ASS. | F1↑ | US↑ |
|---|---|---|---|---|---|
| *Single CP* | | | | | |
| CP (LR) | 0.890 | 0.876 | 3.45 | 0.646 | 0.709 |
| CP (MLP) | 0.903 | 0.893 | 4.42 | 0.542 | 0.650 |
| CP (SGD) | 1.000 | 1.000 | 6.00 | 0.469 | 0.617 |
| *Post-hoc CP Ensemble* | | | | | |
| Het. Ens.–CP | 0.966 | 0.894 | 3.49 | 0.628 | 0.726 |
| Hom. MLP MV-CP | 0.873 | 0.909 | 4.11 | 0.602 | 0.674 |
| Stacked Het.–CP | 0.886 | 0.877 | 3.26 | 0.644 | 0.715 |
| **ECP (ours) — proposed method** | | | | | |
| Hom. CP (MLP-WA) | 0.871 | 0.862 | 3.17 | 0.646 | 0.712 |
| Hom. CP (LR-MV) | 0.899 | 0.886 | 3.36 | 0.632 | 0.714 |
| Hom. CP (MLP-MV) | 0.911 | 0.915 | 3.25 | 0.668 | **0.738** |
| Het. CP (MV) | 0.904 | 0.893 | 3.61 | 0.602 | 0.699 |
| Het. CP (WV) | 0.893 | 0.875 | 3.52 | 0.642 | 0.705 |
| Stacked Het. CP | 0.899 | 0.883 | 3.40 | 0.659 | 0.718 |
| *Non-Conformal Baselines* | | | | | |
| BR (LR) | — | — | — | 0.610 | — |
| Ensemble (LR) | — | — | — | 0.620 | — |
| Label Bag. (3) | — | — | — | 0.640 | — |

**(b) Yeast**

| Method | EC↑ | MC↑ | ASS. | F1↑ | US↑ |
|---|---|---|---|---|---|
| *Single CP* | | | | | |
| CP (LR) | 0.891 | 0.882 | 10.61 | 0.453 | 0.617 |
| CP (MLP) | 0.902 | 0.892 | 10.54 | 0.468 | 0.627 |
| CP (SGD) | 0.897 | 0.870 | 10.62 | 0.455 | 0.615 |
| *Post-hoc CP Ensemble* | | | | | |
| Het. Ens.–CP | 0.903 | 0.892 | 10.71 | 0.462 | 0.623 |
| Hom. MLP MV-CP | 0.992 | 0.997 | 13.93 | 0.425 | 0.604 |
| Stacked Het.–CP | 0.874 | 0.870 | 10.17 | 0.466 | 0.620 |
| **ECP (ours) — proposed method** | | | | | |
| Hom. CP (MLP-WA) | 0.913 | 0.893 | 10.38 | 0.471 | **0.633** |
| Hom. CP (LR-MV) | 0.908 | 0.872 | 10.35 | 0.463 | 0.625 |
| Hom. CP (MLP-MV) | 0.912 | 0.890 | 10.54 | 0.461 | 0.627 |
| Het. CP (MV) | 0.910 | 0.870 | 10.10 | 0.471 | 0.632 |
| Het. CP (WV) | 0.907 | 0.900 | 11.08 | 0.449 | 0.616 |
| Stacked Het. CP | 0.900 | 0.889 | 10.62 | 0.457 | 0.621 |
| *Non-Conformal Baselines* | | | | | |
| BR (LR) | — | — | — | 0.349 | — |
| Ensemble (LR) | — | — | — | 0.349 | — |
| Label Bag. (5) | — | — | — | 0.467 | — |

**(c) Scene**

| Method | EC↑ | MC↑ | ASS. | F1↑ | US↑ |
|---|---|---|---|---|---|
| *Single CP* | | | | | |
| CP (LR) | 0.910 | 0.907 | 2.33 | 0.583 | 0.752 |
| CP (MLP) | 0.896 | 0.892 | 1.58 | 0.743 | 0.816 |
| CP (SGD) | 0.904 | 0.904 | 3.86 | 0.481 | 0.661 |
| *Post-hoc CP Ensemble* | | | | | |
| Het. Ens.–CP | 0.900 | 0.902 | 1.82 | 0.686 | 0.796 |
| Hom. MLP WA-CP | 0.901 | 0.895 | 1.56 | 0.749 | 0.821 |
| Hom. MLP MV-CP | 0.892 | 0.893 | 2.70 | 0.663 | 0.749 |
| Stacked Het.–CP | 0.900 | 0.907 | 1.84 | 0.710 | 0.802 |
| **ECP (ours) — proposed method** | | | | | |
| Hom. CP (MLP-WA) | 0.894 | 0.890 | 1.52 | 0.753 | 0.820 |
| Hom. CP (MLP-MV) | 0.895 | 0.888 | 1.48 | 0.759 | **0.823** |
| Hom. CP (LR-MV) | 0.913 | 0.905 | 2.28 | 0.601 | 0.759 |
| Het. CP (MV) | 0.906 | 0.901 | 1.98 | 0.660 | 0.784 |
| Het. CP (WV) | 0.912 | 0.907 | 2.16 | 0.644 | 0.775 |
| Stacked Het. CP | 0.908 | 0.905 | 1.82 | 0.711 | 0.805 |
| *Non-Conformal Baselines* | | | | | |
| BR (LR) | — | — | — | 0.670 | — |
| Ensemble (LR) | — | — | — | 0.701 | — |
| Label Bag. (3) | — | — | — | 0.680 | — |
| Label Bag. (4) | — | — | — | 0.683 | — |

**(d) COCO**

| Method | EC↑ | MC↑ | ASS. | F1↑ | US↑ |
|---|---|---|---|---|---|
| *Single CP* | | | | | |
| CP (LR) | 0.910 | 0.899 | 8.56 | 0.525 | 0.806 |
| CP (MLP) | 0.909 | 0.899 | 9.34 | 0.510 | 0.800 |
| CP (SGD) | 0.911 | 0.900 | 8.67 | 0.527 | 0.807 |
| CP (RNN)† | 0.911 | 0.900 | 8.08 | 0.541 | 0.812 |
| *Post-hoc CP Ensemble* | | | | | |
| Het. Ens.–CP | 0.904 | 0.893 | 7.83 | 0.548 | 0.811 |
| Hom. MLP WA–CP | 0.903 | 0.893 | 8.43 | 0.529 | 0.804 |
| **ECP (ours) — proposed method** | | | | | |
| Hom. CP (MLP-WA) | 0.904 | 0.889 | 7.96 | 0.543 | 0.809 |
| Hom. CP (LR-MV) | 0.896 | 0.897 | 7.16 | 0.567 | 0.817 |
| Het. CP (MV) | 0.906 | 0.887 | 7.68 | 0.555 | 0.813 |
| Het. CP (WV) | 0.887 | 0.858 | 7.22 | 0.575 | 0.807 |
| Hom. CP (MLP-MV) | 0.890 | 0.871 | 7.31 | 0.568 | 0.809 |
| Hom. CP (RNN-MV) | 0.900 | 0.882 | 6.93 | 0.577 | **0.818** |
| *Non-Conformal Baselines* | | | | | |
| BR (LR) | — | — | — | 0.698 | — |
| Ensemble (LR) | — | — | — | 0.696 | — |
| CLIP-RNN† | — | — | — | 0.700 | — |
| Label Bag. (40) | — | — | — | 0.694 | — |

— not applicable  ECP (ours)  **Bold** = best  underline = 2nd-best  ASS ↓  F1, US ↑

## 6 Conclusion

We proposed Ensemble Conformal Prediction (ECP), a framework that applies conformal calibration to each base model independently, before ensemble aggregation. This design ensures every ensemble member enters aggregation with valid per-label coverage guarantees, and admits formal label-wise coverage bounds under majority voting, which we derived by extending Cherubin [7] to the multilabel setting. Empirically, ECP achieves the highest Utility Score on all four benchmarks, reduces average prediction set size relative to both single-model CP and post-hoc conformalized ensembles, and maintains valid marginal coverage throughout. On Emotions, the best ECP variant reduces ASS from 4.42 to 3.25 while improving macro-F1 from 0.542 to 0.668; on COCO, Hom. CP (RNN-MV) achieves the lowest ASS among all conformal methods (6.93 vs. 8.08) without sacrificing coverage. Notably, ECP also matches or exceeds the macro-F1 of non-conformal baselines on three out of four datasets, despite being designed to control coverage rather than optimize classification accuracy.

*Limitations and future work.* Label-wise calibration under binary relevance ignores inter-label dependencies, which are pronounced in high-cardinality settings such as COCO. Extending ECP to structured label spaces via classifier chains, Mondrian calibration over label clusters, or graph-based nonconformity scores is a natural next step. Additionally, the equal-weighted Utility Score is a neutral summary statistic; future work could explore task-specific weighting or direct coverage-compactness trade-off curves to better reflect application priorities.

**Use of Generative AI** The authors used a Generative AI tool exclusively to improve readability through minor language editing (grammar and proofreading). No scientific content, results, claims were generated by the tool. The authors carefully reviewed all edits and take full responsibility for the final manuscript.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. arXiv preprint arXiv:1606.06565 (2016)
2. Angelopoulos, A., Bates, S., Malik, J., Jordan, M.I.: Uncertainty sets for image classifiers using conformal prediction. arXiv preprint arXiv:2009.14193 (2020)
3. Angelopoulos, A.N., Bates, S.: Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning **16**(4), 494–591 (2023)
4. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern recognition **37**(9), 1757–1771 (2004)
5. Breiman, L.: Bagging predictors. Machine learning **24**(2), 123–140 (1996)
6. Cauchois, M., Gupta, S., Duchi, J.C.: Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. Journal of machine learning research **22**(81), 1–42 (2021)

7. Cherubin, G.: Majority vote ensembles of conformal predictors. Machine Learning **108**(3), 475–488 (2019)
8. Dietterich, T.G.: Ensemble methods in machine learning. In: International workshop on multiple classifier systems. pp. 1–15. Springer (2000)
9. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. Advances in neural information processing systems **14** (2001)
10. Gasparin, M., Ramdas, A.: Merging uncertainty sets via majority vote. arXiv preprint arXiv:2401.09379 (2024)
11. Gauraha, N., Spjuth, O.: Synergy conformal prediction. In: Conformal and Probabilistic Prediction and Applications. pp. 91–110. PMLR (2021)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
13. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D.: Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 2880–2894 (2020)
14. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
15. Lanchantin, J., Wang, T., Ordonez, V., Qi, Y.: General multi-label image classification with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16478–16488 (2021)
16. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-free predictive inference for regression. Journal of the American Statistical Association **113**(523), 1094–1111 (2018)
17. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. Journal of machine learning research **5**(Apr), 361–397 (2004)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
19. Nam, J., Kim, J., Loza Mencía, E., Gurevych, I., Fürnkranz, J.: Large-scale multi-label text classification—revisiting neural networks. In: Joint european conference on machine learning and knowledge discovery in databases. pp. 437–452. Springer (2014)
20. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems **32** (2019)
21. Papadopoulos, H.: Inductive conformal prediction: Theory and application to neural networks. INTECH Open Access Publisher Rijeka (2008)
22. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
24. Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., et al.: Scalable and accurate deep learning with electronic health records. NPJ digital medicine **1**(1), 18 (2018)

25. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine learning **85**(3), 333–359 (2011)
26. Rivera, E.O., Patel, Y., Tewari, A.: Conformal prediction for ensembles: Improving efficiency via score-based aggregation (2025), https://arxiv.org/abs/2405.16246
27. Romano, Y., Sesia, M., Candes, E.: Classification with valid and adaptive coverage. Advances in neural information processing systems **33**, 3581–3591 (2020)
28. Sadinle, M., Lei, J., Wasserman, L.: Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association **114**(525), 223–234 (2019)
29. Shafer, G., Vovk, V.: A tutorial on conformal prediction. Journal of machine learning research **9**(3) (2008)
30. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multi-label classification of music by emotion. EURASIP Journal on Audio, Speech, and Music Processing **2011**(1), 4 (2011)
31. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining **3**, 1–13 (09 2009). https://doi.org/10.4018/jdwm.2007070101
32. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. Data mining and knowledge discovery handbook pp. 667–685 (2010)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
34. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world. Springer (2005)
35. Wolpert, D.H.: Stacked generalization. Neural networks **5**(2), 241–259 (1992)
36. Yang, F., Gan, X., Wang, H., Feng, L., Lai, Y.: Cp-ra $k$ el: Improving random $k$-labelsets with conformal prediction for multi-label classification. In: Conformal and Probabilistic Prediction and Applications. pp. 266–279. PMLR (2017)
37. Yang, P., Sun, X., Li, W., Ma, S., Wu, W., Wang, H.: Sgm: sequence generation model for multi-label classification. In: Proceedings of the 27th international conference on computational linguistics. pp. 3915–3926 (2018)
38. Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering **26**(8), 1819–1837 (2013)