Beyond Accuracy: Controlling Broad Error Types in Selective Classification

Emilien Jemelen INRIA & Epiconcept, Paris
Sandrine Katsahian AP-HP, Paris
Francisco Orchard Epiconcept, Paris
Agathe Guilloux INRIA, Paris

Abstract

Selective prediction allows classifiers to abstain on uncertain inputs, improving reliability in high-stakes applications. Existing frameworks, such as Selection with Guaranteed Risk (SGR), provide tight statistical guarantees on overall misclassification risk, but this is inadequate for many applications—particularly in medicine—where guarantees on specific error types, such as false positives, false negatives, or positive predictive value (PPV), are required. We propose a general framework for selective binary classification with metric-specific guarantees. Our theory extends risk control from 0/1 loss to arbitrary binary losses and derives new high-probability bounds for the corresponding metrics under abstention. We instantiate the framework with neural networks and validate it on public imaging datasets, showing that our metric-aware selective classifiers better capture domain-specific trade-offs than accuracy-based approaches. Code and reproducibility artifacts are released at `https://github.com/EmilienJemelen/selective-classification`.

1 Introduction

Predictive systems in high-stakes settings must recognize when their outputs may be unreliable. *Selective prediction* enables models to abstain on uncertain inputs and defer to safer processes such as human review [Chow, 1957, El-Yaniv and Wiener, 2010]. This "reject option" reduces harmful errors and improves trustworthiness in domains such as autonomous systems, finance, and medicine [Kompa et al., 2021, Leibig et al., 2022, Dvijotham et al., 2023].

A key development is the Selection with Guaranteed Risk (SGR) framework [Geifman and El-Yaniv, 2017], which provides distribution-free guarantees (that hold for any underlying probability distribution) on overall misclassification risk at a given coverage. Yet accuracy alone is insufficient when error costs are asymmetric—for example, false negatives vs. false positives in medicine, fraud detection, or autonomous driving.

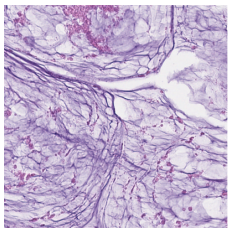
Figure 1 shows that SGR can guarantee low overall risk while still allowing clinically critical false positives. This highlights the need for guarantees on *conditional metrics*—including False Positive Rate (FPR), False Negative Rate (FNR), Positive Predictive Value (PPV), sensitivity (SE), and specificity (SP)—rather than accuracy alone.

Related Work. Early abstention methods include *reject-option* formulations [Chow, 1957, El-Yaniv and Wiener, 2010], uncertainty-based heuristics [Gal and Ghahramani, 2016, Liu et al., 2019], and selective networks [Geifman and El-Yaniv, 2019, Corbière et al., 2019]. While effective, most lack formal guarantees under abstention. Other works provide guarantees for overall misclassification risk or surrogate losses [Cortes et al., 2016, Geifman and El-Yaniv, 2017, Cao et al., 2022]. Narasimhan et al. [2024] address long-tail and class imbalance settings, and Wu et al. [2024] propose contrastive learning for selective classification. Conformal selection methods such as FDR control [Jin and Candes, 2023] offer rigorous guarantees but do not explicitly model abstention in classification. To our knowledge, no prior work guarantees type-I and type-II conditional metrics under abstention in binary classification—this is our focus.

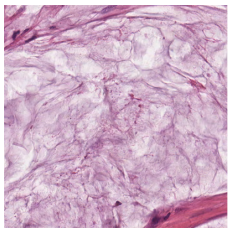
Our Contributions. We propose a general framework for metric-aware selective binary classification:

1. **Generalized guarantees.** We extend 0/1-loss guarantees to arbitrary binary losses (e.g.,

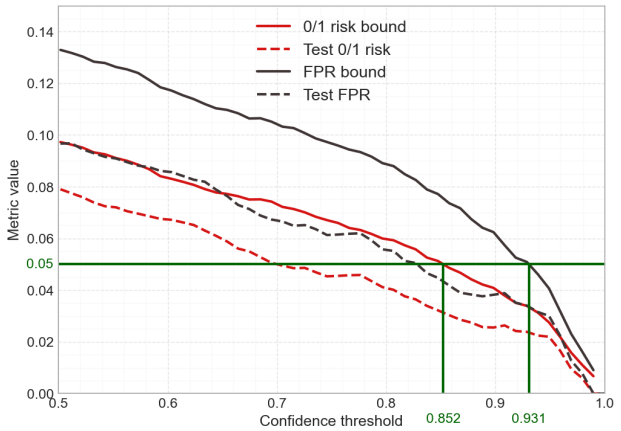
Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco.



(a) **False positive cancer prediction** with model confidence = 0.911.



(b) **True negative cancer prediction** with model confidence = 0.934.



(c) Error rates guarantees with respect to model confidence

Figure 1: **Limitations of accuracy-based selective prediction.** (a) False positive: model predicts "cancer" with confidence 0.911. (b) True negative with confidence 0.934. (c) Risk curves vs. confidence threshold: SGR guarantees overall misclassification risk (red) $< 5\%$ when abstaining below 0.852, but the False Positive Rate (FPR; black) remains uncontrolled, allowing errors such as (a). Guaranteeing FPR $< 5\%$ requires a stricter threshold of 0.931, which excludes (a) while retaining 95% of true negatives as in (b). Dataset: H&E images [Kather, 2019].

false positives and false negatives), deriving high-probability, distribution-free (that hold for any underlying probability distribution) monotonic bounds.

2. **Conditional metrics.** We derive new bounds for FPR, FNR, PPV, SE, and SP.

3. **Algorithms.** We introduce (i) a dichotomic search leveraging monotonic bounds for binary losses, and (ii) a greedy search for non-monotonic conditional metrics.

4. **Empirical results.** Experiments on CIFAR-10 [Krizhevsky and Hinton, 2009] and colorectal cancer imaging data show improved reliability and practical trade-offs.

2 Problem setting: binary classification, Selective Prediction and associated performance metrics

We consider the classical binary classification framework. Let $\mathbb{P}$ be the distribution of random pairs $(X, Y) \in \mathcal{X} \times \{0, 1\}$, where X are features and Y the label. A classifier is any function $f : \mathcal{X} \to \{0, 1\}$.

Risks. Given a loss function $L : \{0, 1\}^2 \to \mathbb{R}^+$, the L-risk of f is

$$R_{\mathbb{P},L}(f) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}[L(f(X), Y)].$$

Important risks include:

- **Misclassification risk $\mathbf{R}_{\mathbb{P},\mathbf{L}_{0/1}}$:** defined as the $\mathrm{L}_{0/1}$-risk with $\mathrm{L}_{0/1} : (y, y') \in \{0, 1\}^2 \mapsto \mathbb{1}(y \neq y')$.
- **False positive risk $\mathbf{R}_{\mathbb{P},\mathbf{L}_{\mathbf{FP}}}$:** defined as the L_{FP}-risk with $\mathrm{L}_{\mathrm{FP}} : (y, y') \in \{0, 1\}^2 \mapsto \mathbb{1}(y = 1, y' = 0)$.
- **False negative risk $\mathbf{R}_{\mathbb{P},\mathbf{L}_{\mathbf{FN}}}$:** defined as the L_{FN}-risk with $\mathrm{L}_{\mathrm{FN}} : (y, y') \in \{0, 1\}^2 \mapsto \mathbb{1}(y = 0, y' = 1)$.

Conditional metrics. In medicine, one is often interested in conditional performance metrics that distinguish error types relative to the true class or to the prediction of f, e.g.:

- False Positive Rate (FPR): $\mathbb{P}(f(X) = 1 \mid Y = 0)$,
- False Negative Rate (FNR): $\mathbb{P}(f(X) = 0 \mid Y = 1)$,
- Positive Predictive Value (PPV): $\mathbb{P}(Y = 1 \mid f(X) = 1)$,
- Sensitivity (SE) $= 1-$FNR,
- Specificity (SP) $= 1-$FPR.

These conditional metrics are crucial in high-stakes domains: for example, controlling FNR avoids missed diagnoses, while controlling FPR prevents overdiagnosis.

Selective classification. Selective prediction augments a classifier f with a *confidence function* $\kappa_f : \mathcal{X} \to \mathrm{Im}(\kappa_f) \subset \mathbb{R}$, intended to correlate monotonically with prediction reliability. Here, $\mathrm{Im}(\kappa_f)$ denotes the image of κ_f. Common examples include the maximum softmax probability and Monte Carlo dropout–based scores [Gal and Ghahramani, 2016]. This assumption

is formalized in Hypothesis 1 and is standard in the selective prediction literature [Geifman and El-Yaniv, 2017, Wu et al., 2024]. In our setting, it is empirically supported by calibration experiments (Figure 9, Supplementary D.3).

Hypothesis 1 (Loss–confidence monotonicity)**.** *Given a loss function L and a confidence function κ_f, for any $(x_1, y_1), (x_2, y_2) \overset{iid}{\sim} \mathbb{P}$,*

$$\kappa_f(x_1) \leq \kappa_f(x_2) \quad \Longleftrightarrow \quad L(f(x_1), y_1) \geq L(f(x_2), y_2)\,.$$

A *selection function* is defined by thresholding κ_f:

$$g_\theta(x) = \mathbb{1}(\kappa_f(x) \geq \theta).$$

The corresponding *selective classifier* is

$$(f, g_\theta)(x) = \begin{cases} f(x) & \text{if } g_\theta(x) = 1, \\ \text{abstain} & \text{otherwise.} \end{cases}$$

The *coverage* of a selection function is $\mathbb{E}_{\mathbb{P}}[g_\theta(X)]$, i.e. the probability of making a prediction. The *selective risk* is the risk conditional on predicting:

$$R_{\mathbb{P},L}(f, g_\theta) = \mathbb{E}_{\mathbb{P}}[L(f(X), Y) \mid g_\theta(X) = 1].$$

Its empirical counterpart follows analogously.

We express all previously introduced metrics in their selective form. For example, the selective False Positive Rate, denoted by $\mathrm{FPR}_{\mathbb{P}}(f, g_\theta)$, is defined by restricting attention to instances for which the model does not abstain:

$$\mathrm{FPR}_{\mathbb{P}}(f, g_\theta) = \mathbb{P}(f(X) = 1 \mid Y = 0, g_\theta(X) = 1).$$

The remaining selective metrics are defined in an analogous manner; see Supplementary Table 1 for a complete list.

The SGR framework. Geifman and El-Yaniv [2017] proposed the Selection with Guaranteed Risk (SGR) algorithm, which provides high-probability bounds on the *selective misclassification risk* $\mathrm{R}_{\mathbb{P},\mathrm{L}_{0/1}}(f, g_\theta)$. Their result, based on the distribution-free bound of Gascuel and Caraux [1992], ensures that with probability (w.p.) at least $1 - \delta$:

$$\mathrm{R}_{\mathbb{P},\mathrm{L}_{0/1}}(f, g_{\theta^*}) \leq B(\theta^*).$$

The threshold θ^* is chosen via a dichotomic search, exploiting the intuition that the bound $B(\theta)$ behaves monotonically with θ.

However, their framework applies only to the 0/1 loss, i.e. overall misclassification risk.

3 Control over the selective metrics

In this section, we propose bounds for a broader range of risks, and derive bounds for all the selective metrics introduced in Section 2. Specifically, we prove that an approach similar to Gascuel and Caraux [1992] can be applied to arbitrary binary losses (e.g. $\mathrm{L_{FP}}$, $\mathrm{L_{FN}}$), yielding guaranteed upper bounds B^*_δ (explicitly defined in Supplementary Proposition A1) granting control over $\mathrm{R}_{\mathbb{P},\mathrm{L_{FP}}}$ and $\mathrm{R}_{\mathbb{P},\mathrm{L_{FN}}}$.

Control of the selective risk with general binary losses. We introduce the L-empirical risk on a dataset $S_n = (x_j, y_j)_{j=1}^n$ drawn i.i.d. from $\mathbb{P}$:

$$\hat{R}_{\mathrm{L}}(f \mid S_n) = \tfrac{1}{n} \sum_{j=1}^{n} \mathrm{L}(f(x_j), y_j).$$

Proposition 1. *Using the notations introduced in Section 2, we have, for any pair $(\theta, \delta) \in \mathrm{Im}(\kappa_f) \times (0, 1)$, w.p. at least $1 - \delta$,*

$$R_{\mathbb{P},L}(f, g_\theta) \leq B^*_\delta(g_\theta(S_n), L, f).$$

Monotonicity of the bound in θ. For any binary loss L and fixed $\delta \in (0, 1)$, the selective risk bound in Proposition 1 is decreasing in the confidence threshold $\theta \in \mathrm{Im}(\kappa_f)$ (Lemma A3, Supplementary A). Under the reasonable assumption that $\mathrm{Im}(\kappa_f)$ is an interval (Supplementary C.3), this justifies a dichotomy search over $\mathrm{Im}(\kappa_f)$: Algorithm 1 (Supplementary C) finds a threshold $\theta^* = \theta^*(S_n)$ such that $B^*_\delta(g_{\theta^*}(S_n), L, f)$ matches a user-defined target $r^* \in (0, 1)$. For example, in cancer screening, one may require the false positive risk $\mathrm{R}_{\mathbb{P},\mathrm{L_{FP}}}(f, g_{\theta^*})$ to be at most $r^* = 1\%$.

In Proposition 2 we derive upper bounds for conditional selective metrics such as FPR (bounds for FNR, PPV, SE and SP are presented in Supplementary A).

Proposition 2. *Considering the notations introduced in Section 3, for any pair $(\theta, \delta) \in \mathrm{Im}(\kappa_f) \times (0, 1)$, and with B^* defined as in Proposition A1, w.p. at least $1 - \delta$,*

$FPR_{\mathbb{P}}(f, g_\theta)$ is less than

$$\min\left\{\, 1, \frac{B^*_{\delta/2}(g_\theta(S_n), L_{FP}, f)}{\left(1 - \frac{\sum_{j=1}^n y_j g_\theta(x_j)}{|g_\theta(S_n)|} - \frac{\sqrt{nlog(2/\delta)/2}}{|g_\theta(S_n)|}\right)_+} \,\right\}.$$

Since these conditional metric bounds are non-monotonic with respect to the threshold θ (Supplementary C.1), the dichotomy procedure described in Algorithm 1 is no longer applicable. Instead, we propose a greedy search approach, detailed in Algorithm 2 (Supplementary C.1).

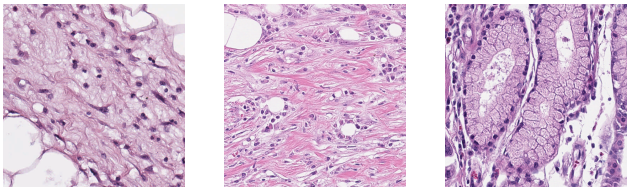
4 Empirical experiments

We evaluate our selective metrics bounds on two image datasets: CIFAR-10 [Krizhevsky and Hinton, 2009] and H&E-stained colorectal cancer images [Kather, 2019]. All the airplane detection experiments on CIFAR-10 data are fully available in Supplementary E.1. Unless stated otherwise, we fix $\delta = 0.005$, ensuring bounds hold with high probability. In all our experiments we found $K_2 = 50$ to be enough to guarantee tight bounds with Algorithm 2. To ensure similar bounds tightness between Algorithms 1-2 we chose $K_1 = 6$ (such that $\delta/K_2 \approx \delta/(2^{K_1} - 1)$).

Common setup. Each dataset is partitioned into three disjoint subsets: 1/3 for training the classifier, 1/2 for fitting the selective metric bounds, and 1/6 for evaluating them. This design (i) produces moderately accurate classifiers, leaving room for abstention to improve performance, and (ii) allocates sufficiently large bound-fitting sets to obtain tighter guarantees (see Supplementary E.3 for an analysis of how parameters affect bound tightness). To ensure reproducibility and prevent favorable splits, we repeat the bound-fitting/testing partition for each metric using fixed random seeds.

Colorectal cancer detection: task & model. We consider a colorectal cancer histopathology dataset comprising 11,977 Hematoxylin and Eosin (H&E) image patches [Kather, 2019]. The patches are annotated as ADIMUC/STRMUS (non-tumor, 0) and TUMSTU (tumor, 1), with tumor samples representing 33% of the dataset. Figure 2 displays randomly selected 512×512 patches. The same lightweight CNN is trained but stopped after one epoch (91% accuracy on balanced validation set) to avoid saturation, since highly accurate models leave little room for abstention gains. SR and MCD are again considered; MCD yields looser bounds than SR, so its results are reported in Supplementary E.2.

Results. Figure 3 shows that FP risk dominates misclassification risk across coverages (golden and red curves nearly overlap). We also consider sensitivity (SE) to improve cancer detection capability. By Proposition A2, all SR-based SE–FP risk trade-offs hold w.p. at least 0.99: for instance, $(f, g_{0.91})$ achieves SE $\geq 98.5\%$ and FP risk $\leq 3.5\%$ at $\sim 75\%$ coverage, compared to SE $= 93\%$ and FP risk $= 8.5\%$ at full coverage. In practice, such a low guaranteed FP risk under abstention would allow clinicians to accept more easily a positive result when the model outputs one, freeing them to focus on more complex cases without increasing economic burden or patient stress associated with false positives.



(a) ADIMUC (b) STRMUS (c) TUMSTU

Figure 2: Random 512×512 H&E image patches.

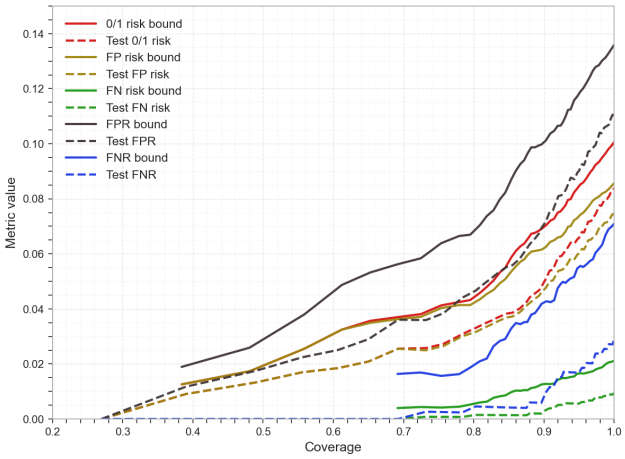


Figure 3: Upper bounds and test-set estimates of main selective metrics as a function of coverage. Confidence function: SR. Dataset: H&E patch images.

5 Concluding remarks

We presented theoretical bounds and practical algorithms for abstention in binary neural network classifiers, enabling control of diverse performance metrics. Empirically, abstention improves selective metrics—individually or jointly—with customizable targets and probability levels.

The effectiveness of post-hoc abstention depends on the confidence function κ_f, baseline accuracy, sample size, class imbalance, and the chosen confidence level δ. While our analysis focused on neural networks, the guarantees extend to other models (e.g., logistic regression, SVMs, decision trees) that may achieve competitive performance at lower cost. Further directions include refining the search for conditional metric guarantees by exploiting structural properties of bound evolution with θ, and integrating asymmetric risks into training so that networks learn to use abstention more effectively.

The framework is broadly applicable: any binary classification task in which controlling error types is critical can benefit from abstention. Medical applications, in particular, stand to gain from models that abstain on uncertain diagnoses. Doing so helps prevent harm-

ful errors for patients (false negatives), reduces unnecessary interventions, costs, and patient stress (false positives), and supports clinical workflows by improving the trustworthiness of AI-based diagnoses with abstention—ultimately freeing clinicians to focus on more complex cases.

References

Akshay Balsubramani, Sanjoy Dasgupta, yoav Freund, and Shay Moran. An adaptive nearest neighbor rule for classification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Nicolas Bourbaki. *Topologie générale*. Springer, Berlin, Heidelberg, 1 edition, 1971. ISBN 978-3-540-33982-3. doi: 10.1007/978-3-540-33982-3.

Lawrence D Brown, T Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical science*, 16(2):101–133, 2001.

Yuzhou Cao, Tianchi Cai, Lei Feng, Lihong Gu, Jinjie GU, Bo An, Gang Niu, and Masashi Sugiyama. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 521–534. Curran Associates, Inc., 2022.

C. K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254, 1957. doi: 10.1109/TEC.1957.5222035.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In Ronald Ortner, Hans Ulrich Simon, and Sandra Zilles, editors, *Algorithmic Learning Theory*, pages 67–82, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46379-7.

Krishnamurthy Dvijotham, Jim Winkens, Melih Barsbey, Sumedh Ghaisas, Robert Stanforth, Nick Pawlowski, Patricia Strachan, Zahra Ahmed, Shekoofeh Azizi, Yoram Bachrach, et al. Enhancing the reliability and accuracy of ai-enabled diagnosis via complementarity-driven deferral to clinicians. *Nature Medicine*, 29(7):1814–1820, 2023.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

Olivier Gascuel and Gilles Caraux. Distribution-free performance bounds with the resubstitution error estimate. *Pattern Recognition Letters*, 13(11):757–764, 1992. ISSN 0167-8655. doi: https://doi.org/10.1016/0167-8655(92)90125-J.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4885–4894, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option, 2019. URL `https://arxiv.org/abs/1901.09192`.

Shani Goren, Ido Galil, and Ran El-Yaniv. Hierarchical selective classification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016b. doi: 10.1109/CVPR.2016.90.

Ying Jin and Emmanuel J. Candes. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023. URL `http://jmlr.org/papers/v24/22-1176.html`.

Jakob Nikolas Kather. Histological images for tumor detection in gastrointestinal cancer. Zenodo, 2019. URL `https://doi.org/10.5281/zenodo.2530789`.

Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue (v0.1), 2018. URL `https://doi.org/10.5281/zenodo.1214456`. Data set.

Achim Klenke and Lutz Mattner. Stochastic ordering of classical discrete distributions. *Advances in Applied probability*, 42(2):392–410, 2010.

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4(1):4, 2021. ISSN 2398-6352. doi: 10.1038/s41746-020-00367-3.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

Christian Leibig, Moritz Brehmer, Stefan Bunk, Danalyn Byng, Katja Pinker, and Lale Umutlu. Combining the strengths of radiologists and ai for breast cancer screening: a retrospective analysis. *The Lancet Digital Health*, 4(7):e507–e519, 2022.

Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Harikrishna Narasimhan, Aditya Krishna Menon, Wittawat Jitkrittum, Neha Gupta, and Sanjiv Kumar. Learning to reject meets long-tail learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Representation Learning*, volume 2024, pages 45256–45283, 2024.

Andrea Pugnana, Lorenzo Perini, Jesse Davis, and Salvatore Ruggieri. Deep neural network benchmarks for selective classification. *Journal of Data-centric Machine Learning Research*, 2024.

Tyss Santosh, Irtiza Chowdhury, Shanshan Xu, and Matthias Grabmair. The craft of selective prediction: Towards reliable case outcome classification - an empirical study on European court of human rights cases. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3656–3674, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.208.

Galen R Shorack and Jon A Wellner. *Empirical processes with applications to statistics*. SIAM, 2009.

Yu-Chang Wu, Shen-Huan Lyu, Haopu Shang, Xiangyu Wang, and Chao Qian. Confidence-aware contrastive learning for selective classification. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

A Theoretical results with proofs

In this section, we present the proofs of our theoretical results. We rely on Theorem 1.a. from [Klenke and Mattner, 2010], which we recall here for completeness:

Lemma A1 (Stochastic ordering of binomial distributions)**.** *Let $\mathbb{P}_i = \mathcal{B}(n_i, p_i)$ with $p_i \in (0,1)$, $n_i \in \mathbb{N}$ and $i = 1, 2$. Then $\mathbb{P}_1 \leq_{st} \mathbb{P}_2$ if and only if the left and right tail conditions hold:*

$$\begin{cases} \textit{Left tail condition: } (1-p_1)^{n_1} \geq (1-p_2)^{n_2} \\ \textit{Right tail condition: } n_1 \leq n_2 \end{cases}$$

Proposition A1. *Let $S_n \overset{iid}{\sim} \mathbb{P}$ and consider a loss function $L : \mathcal{Y}^2 \to \{0,1\}$. For a fixed $\delta \in (0,1)$, define the random variable $B^*_\delta(S_n, L, f)$ as follows:*

- *if $n \geq 1$ and $\hat{R}_L(f \mid S_n) < 1$, then $B^*_\delta(S_n, L, f)$ is the unique $b \in (0,1)$ solving*

$$\sum_{k=0}^{n\hat{R}_L(f|S_n)} \binom{n}{k} b^k (1-b)^{n-k} = \delta.$$

- *if $n = 0$ (empty dataset) or $\hat{R}_L(f \mid S_n) = 1$, then $B^*_\delta(S_n, L, f) = 1$.*

Then

$$\mathbb{P}\big(R_{\mathbb{P},L}(f) \leq B^*_\delta(S_n, L, f)\big) \geq 1 - \delta.$$

Note that in Proposition A1, B^*_δ is a quantile of the Beta distribution Brown et al. [2001].

Proof of Proposition A1. First we study the pathological cases. For a fixed S_n, when $n = 0$ (S_n is empty) or if $\hat{R}_L(f \mid S_n) = 1$ (f classifies all empirical samples wrong) then the equation equates to $1 = \delta$ and admits no solution, so no risk bound can be guaranteed and we set $B^*_\delta(S_n, L, f) = 1$, such that:

$$\mathbb{P}(\underbrace{R_{\mathbb{P},L}(f)}_{\in[0,1]} \leq \underbrace{B^*_\delta(S_n, L, f)}_{=1} \mid \underbrace{\{S_n = \emptyset \cup \hat{R}_L(f \mid S_n) = 1\}}_{\textbf{noted event } A}) = 1.$$

We now consider the case where $n \geq 1$ and $\hat{R}_L(f \mid S_n) < 1$ (the complement of event A, noted $\bar{A}$). Let $K < n$ and introduce, for $x \in (0,1)$, the value at K of the cumulative distribution function (CDF) of the binomial distribution $\mathcal{B}(n, x)$:

$$F(K; n, x) = \sum_{j=0}^{K} \binom{n}{j} x^j (1-x)^{n-j}.$$

Seeing this as a continuous polynomial function of x, Lemma A1 applied with $n_1 = n_2 = n$ shows that it is strictly decreasing from $(0,1)$ to $(0,1)$.

Hence, given event $\bar{A}$ we can uniquely define the solution on $(0,1)$ of

$$F(n\hat{R}_L(f \mid S_n); n, x) = \delta.$$

which is noted $B^*_\delta(S_n, L, f)$. This proves the first point in Proposition A1.

As we proceed, one may notice that the decreasing behaviour of F with respect to x implies that if $\delta_1 \leq \delta_2$, then

$$B^*_{\delta_1}(S_n, L, f) \geq B^*_{\delta_2}(S_n, L, f).$$

We now apply the decreasing bijection $x \in (0,1) \mapsto F(n\hat{R}_L(f \mid S_n); n, x)$ to the following inequality:

$$R_{\mathbb{P},L}(f) \leq B^*_\delta(S_n, L, f) \iff F(n\hat{R}_L(f \mid S_n); n, R_{\mathbb{P},L}(f)) \geq \underbrace{F(n\hat{R}_L(f \mid S_n); n, B^*_\delta(S_n, L, f))}_{=\delta \text{ by definition.}} \quad (*)$$

By combining both cases A and $\bar{A}$, we have

$$\begin{aligned}\mathbb{P}(\underbrace{R_{\mathbb{P},L}(f) \leq B^*_\delta(S_n, L, f)}_{\text{noted } E}) &= \mathbb{P}(E \cap \{A \cup \bar{A}\}) \\ &= \underbrace{\mathbb{P}(E \mid A)}_{=1}\mathbb{P}(A) + \underbrace{\mathbb{P}(E \mid \bar{A})}_{=\mathbb{P}(*)}\mathbb{P}(\bar{A}) \\ &= \mathbb{P}(A) + \mathbb{P}(*)\mathbb{P}(\bar{A})\end{aligned}$$

Since F is a CDF, we have [Shorack and Wellner, 2009, Chapter 1, Proposition 2]:

$$\mathbb{P}\big(F(n\hat{R}_L(f \mid S_n); n, R_{\mathbb{P},L}(f)) \leq \delta\big) \leq \delta.$$

So

$$\begin{aligned}\mathbb{P}(*) &= 1 - \mathbb{P}\big(F(n\hat{R}_L(f \mid S_n); n, R_{\mathbb{P},L}(f)) < \delta\big) \\ &\geq 1 - \mathbb{P}\big(F(n\hat{R}_L(f \mid S_n); n, R_{\mathbb{P},L}(f)) \leq \delta\big) \\ &\geq 1 - \delta.\end{aligned}$$

And then

$$\begin{aligned}\mathbb{P}(R_{\mathbb{P},L}(f) \leq B^*_\delta(S_n, L, f)) &\geq \mathbb{P}(A) + (1-\delta)\mathbb{P}(\bar{A}) \\ &\geq 1 - \mathbb{P}(\bar{A})\,\Big(\,1 - 1 + \delta\,\Big) \\ &\geq 1 - \delta.\end{aligned}$$

which concludes this proof.

□

Proposition 1. *Using the notations introduced in Section 2, we have, for any pair $(\theta, \delta) \in \mathrm{Im}(\kappa_f) \times (0,1)$, w.p. at least $1-\delta$,*

$$R_{\mathbb{P},L}(f, g_\theta) \leq B^*_\delta(g_\theta(S_n), L, f).$$

Proof of Proposition 1. Write $g = g_\theta$ and define the (random) index set $I := \{i \in \{1, \dots, n\} : g_\theta(X_i) = 1\}$ If $N = |g(S_n) = |I| = 0$, then $g(S_n) = \emptyset$ and by definition $B^*_\delta(g(S_n), L, f) = 1$, hence $R_{\mathbb{P},L}(f, g) \in [0,1] \leq 1$ almost surely, so the desired inequality holds trivially. Assume now that $N \geq 1$. Define the conditional (selected) distribution

$$\mathbb{P}_g(\cdot) := \mathbb{P}(\,\cdot \mid g(X) = 1\,).$$

Note that the selective risk is exactly the L-risk under $\mathbb{P}_g$:

$$R_{\mathbb{P},L}(f, g) = \mathbb{E}[L(f(X), Y) \mid g(X) = 1] = R_{\mathbb{P}_g,L}(f).$$

Let $g_i := g(X_i) \in \{0,1\}$ and let $G := (g_1, \dots, g_n)$. Conditionally on G, the collection of selected pairs $\{(X_i, Y_i) : i \in I\}$ is independent, and for each $i \in I$ we have

$$(X_i, Y_i) \mid (g_i = 1) \;\sim\; \mathbb{P}_g.$$

Moreover, by independence of (X_i, Y_i) across i, and since g_i depends only on X_i, the conditional law of (X_i, Y_i) given g_i factorizes across i, so that

$$\big((X_{i_1}, Y_{i_1}), \dots, (X_{i_N}, Y_{i_N})\big) \;\big|\; G \quad \overset{d}{=} \quad (\tilde{X}_1, \tilde{Y}_1), \dots, (\tilde{X}_N, \tilde{Y}_N) \;\text{ with }\; (\tilde{X}_j, \tilde{Y}_j) \overset{iid}{\sim} \mathbb{P}_g.$$

Now apply Proposition A1 to the i.i.d. sample of size N from $\mathbb{P}_g$ (conditionally on G). Using that $g(S_n)$ is exactly the selected sample, we obtain, for every realization of G with $N \geq 1$,

$$\mathbb{P}\big(R_{\mathbb{P}_g,L}(f) \leq B^*_\delta(g(S_n), L, f) \;\big|\; G\big) \;\geq\; 1 - \delta.$$

Since the event is sure when $N = 0$, the same inequality holds for all G. The conclusion comes from the total probability law.

□

Lemma A2. *Let $n \in \mathbb{N}^*$ such that $n \geq 2$, $(K_1, K_2) \in \mathbb{N}^{*2}$ such that $K_2 < K_1$. Then for every $b \in (0,1)$,*

$$\sum_{i=1}^{K_1} \binom{n}{i} b^i (1-b)^{n-i} \geq \sum_{i=1}^{K_2} \binom{n-1}{i} b^i (1-b)^{n-1-i}.$$

Proof (coupling). Let $X \sim \mathcal{B}(n-1, b)$ and $Y \sim \mathcal{B}(b)$ independent, and set $X' := X + Y$. Then $X' \sim \mathcal{B}(n, b)$ and we have:

$$\{1 \leq X \leq K_2\} \subseteq \{1 \leq X' \leq \underbrace{K_2 + 1}_{\leq K_1}\}.$$

So

$$\mathbb{P}(1 \leq X \leq K_2) \leq \mathbb{P}(1 \leq X' \leq K_1).$$

Which concludes the proof.

□

Lemma A3. *Considering the notations introduced in Section 3 and a fixed $\delta \in (0,1)$, we have:*

$$\forall (\beta_1, \beta_2) \in \mathrm{Im}(\kappa_f)^2, \quad \beta_1 \leq \beta_2 \Rightarrow B^*_\delta(g_{\beta_1}(S_n), L, f) \geq B^*_\delta(g_{\beta_2}(S_n), L, f).$$

Proof of Lemma A3. First, we prove by induction that, letting $(\theta_i)_{i \in 1,..,n}$ be the sequence $(\kappa_f(x_i))_{i \in 1,..,n}$ sorted in ascending order, we have for any $i \in (1,..,n-1)$:

$$B^*_\delta(g_{\theta_i}(S_n), L, f) \geq B^*_\delta(g_{\theta_{i+1}}(S_n), L, f).$$

The case $\theta_i = \theta_{i+1}$ is trivial, so we now assume $\theta_i < \theta_{i+1}$.

By definition of $(\theta_i)_{i \in 1,..,n}$, we have

$$|g_{\theta_{i+1}}(S_n)| = |g_{\theta_i}(S_n)| - 1.$$

We observe that $g_{\theta_{i+1}}(S_n)$ is obtained from $g_{\theta_i}(S_n)$ by removing its lowest-confidence sample. Under Hypothesis 1, this corresponds to removing the sample with the highest loss. Therefore,

$$\begin{aligned}
\hat{R}_L(f \mid g_{\theta_{i+1}}(S_n)) &= \frac{1}{|g_{\theta_{i+1}}(S_n)|} \sum_{(x,y) \in g_{\theta_{i+1}}(S_n)} L(f(x), y) \\
&= \frac{1}{|g_{\theta_i}(S_n)| - 1} \left(\underbrace{\sum_{(x,y) \in g_{\theta_i}(S_n)} L(f(x), y)}_{= |g_{\theta_i}(S_n)| \hat{R}_L(f | g_{\theta_i}(S_n))} - \max_{(x,y) \in g_{\theta_i}(S_n)} L(f(x), y) \right) \\
&= \hat{R}_L(f \mid g_{\theta_i}(S_n)) + \frac{1}{|g_{\theta_i}(S_n)| - 1} \underbrace{\left(\hat{R}_L(f \mid g_{\theta_i}(S_n)) - \max_{(x,y) \in g_{\theta_i}(S_n)} L(f(x), y) \right)}_{=:h}.
\end{aligned}$$

We now distinguish two cases:

$$\begin{cases}
\max_{(x,y) \in g_{\theta_i}(S_n)} L(f(x), y) = 0 & \Rightarrow \hat{R}_L(f \mid g_{\theta_i}(S_n)) = 0 \Rightarrow h = 0, \\
\max_{(x,y) \in g_{\theta_i}(S_n)} L(f(x), y) = 1 & \Rightarrow h \leq 0.
\end{cases}$$

In either case, it follows that

$$\hat{R}_L(f \mid g_{\theta_{i+1}}(S_n)) \leq \hat{R}_L(f \mid g_{\theta_i}(S_n)).$$

By definition of B^*_δ (see Proposition A1), the case $\hat{R}_L(f \mid g_{\theta_i}(S_n)) = 0$ immediately yields

$$B^*_\delta(g_{\theta_i}(S_n), L, f) = 1 - \delta^{1/|g_{\theta_i}(S_n)|} \geq 1 - \delta^{1/(|g_{\theta_i}(S_n)|-1)} = B^*_\delta(g_{\theta_{i+1}}(S_n), L, f).$$

We therefore assume $\hat{R}_L(f \mid g_{\theta_i}(S_n)) > 0$, in which case

$$\underbrace{|g_{\theta_{i+1}}(S_n)|}_{=|g_{\theta_i}(S_n)|-1} \hat{R}_L(f \mid g_{\theta_{i+1}}(S_n)) < |g_{\theta_i}(S_n)| \, \hat{R}_L(f \mid g_{\theta_i}(S_n)).$$

To simplify the notation, let

$$F_i : b \in (0,1) \mapsto F\big(|g_{\theta_i}(S_n)|\hat{R}_L(f \mid g_{\theta_i}(S_n)); \, |g_{\theta_i}(S_n)|, \, b\big).$$

The function F_{i+1} is defined analogously, replacing θ_i with θ_{i+1}. As shown earlier in the proof of **Proposition A1**, both F_i and F_{i+1} are decreasing bijections from $(0,1)$ to $(0,1)$.

With these elements, Lemma A2, applied with $n = |g_{\theta_i}(S_n)|$ and

$$K_2 = |g_{\theta_{i+1}}(S_n)|\hat{R}_L(f \mid g_{\theta_{i+1}}(S_n)) < |g_{\theta_i}(S_n)|\hat{R}_L(f \mid g_{\theta_i}(S_n)) = K_1,$$

yields, for every $b \in (0,1)$,

$$F_{i+1}(b) \leq F_i(b).$$

In particular, taking $b = B^*_\delta(g_{\theta_i}(S_n), L, f)$ (provided that $B^*_\delta(g_{\theta_i}(S_n), L, f) < 1$; see the end of the proof for the case where it equals 1), we obtain

$$F_{i+1}(B^*_\delta(g_{\theta_i}(S_n), L, f)) \leq \underbrace{F_i(B^*_\delta(g_{\theta_i}(S_n), L, f))}_{=\delta \text{ by definition of } B^*_\delta(g_{\theta_i}(S_n), L, f)}.$$

Moreover, we also have $\delta = F_{i+1}(B^*_\delta(g_{\theta_{i+1}}(S_n), L, f))$, so that

$$\begin{aligned}
&F_{i+1}(B^*_\delta(g_{\theta_i}(S_n), L, f)) \leq F_{i+1}(B^*_\delta(g_{\theta_{i+1}}(S_n), L, f)) \\
&\Rightarrow F^{-1}_{i+1} \circ F_{i+1}(B^*_\delta(g_{\theta_i}(S_n), L, f)) \geq F^{-1}_{i+1} \circ F_{i+1}(B^*_\delta(g_{\theta_{i+1}}(S_n), L, f)) \\
&\Rightarrow B^*_\delta(g_{\theta_i}(S_n), L, f) \geq B^*_\delta(g_{\theta_{i+1}}(S_n), L, f).
\end{aligned}$$

By the definition of the function B^*_δ (see Proposition A1), the case $B^*_\delta(g_{\theta_i}(S_n), L, f) = 1$ corresponds either to

- $g_{\theta_i}(S_n) = \emptyset$; in this case, since $g_{\theta_{i+1}}(S_n) \subseteq g_{\theta_i}(S_n)$, we also have $g_{\theta_{i+1}}(S_n) = \emptyset$, and thus $B^*_\delta(g_{\theta_{i+1}}(S_n), L, f) = 1 \leq B^*_\delta(g_{\theta_i}(S_n), L, f)$;
- or $\hat{R}_L(f \mid g_{\theta_i}(S_n)) = 1$, meaning that f makes an error on all samples in $g_{\theta_i}(S_n)$. Consequently, f also makes an error on all samples in $g_{\theta_{i+1}}(S_n) \subseteq g_{\theta_i}(S_n)$, and by definition $B^*_\delta(g_{\theta_{i+1}}(S_n), L, f) = 1 \leq B^*_\delta(g_{\theta_i}(S_n), L, f)$.

We have thus shown that the bounds are non-increasing along the sequence of empirical confidences $(\theta_i)_{i \in 1,..,n}$. To conclude the proof, observe that for any pair $(\beta_1, \beta_2) \in \mathrm{Im}(\kappa_f)^2$ such that $\beta_1 \leq \beta_2$, there exist $i, j \in (1,..,n)$ with $i \leq j$ such that $g_{\beta_1}(S_n) = g_{\theta_i}(S_n)$ and $g_{\beta_2}(S_n) = g_{\theta_j}(S_n)$. Consequently,

$$B^*_\delta(g_{\beta_1}(S_n), L, f) = B^*_\delta(g_{\theta_i}(S_n), L, f) \geq B^*_\delta(g_{\theta_j}(S_n), L, f) = B^*_\delta(g_{\beta_2}(S_n), L, f).$$

Which concludes the proof.

□

Proposition 2. *Considering the notations introduced in Section 3, for any pair $(\theta, \delta) \in \mathrm{Im}(\kappa_f) \times (0,1)$, and with B^* defined as in Proposition A1, w.p. at least $1-\delta$,*

$$\begin{aligned}
&FPR_{\mathbb{P}}(f, g_\theta) \text{ is less than} \\
&\min\left\{ 1, \frac{B^*_{\delta/2}(g_\theta(S_n), L_{FP}, f)}{\big(1 - \frac{\sum_{j=1}^n y_j g_\theta(x_j)}{|g_\theta(S_n)|} - \frac{\sqrt{nlog(2/\delta)/2}}{|g_\theta(S_n)|}\big)_+} \right\}.
\end{aligned}$$

$$\mathrm{FNR}_{\mathbb{P}}(f, g_\theta) \leq \min\left\{ 1, \frac{B^*_{\delta/2}(g_\theta(S_n), \mathrm{L_{FN}}, f)}{\left(\frac{\sum_{j=1}^n y_j g_\theta(x_j)}{|g_\theta(S_n)|} - \frac{\sqrt{n\log(2/\delta)/2}}{|g_\theta(S_n)|}\right)_+} \right\}.$$

$$\mathrm{PPV}_{\mathbb{P}}(f, g_\theta) \geq \max\left\{ 0, 1 - \frac{B^*_{\delta/2}(g_\theta(S_n), \mathrm{L_{FP}}, f)}{\left(\frac{\sum_{j=1}^n f(x_j) g_\theta(x_j)}{|g_\theta(S_n)|} - \frac{\sqrt{n\log(2/\delta)/2}}{|g_\theta(S_n)|}\right)_+} \right\}.$$

$$\mathrm{SE}_{\mathbb{P}}(f, g_\theta) \geq \max\left\{ 0, 1 - \frac{B^*_{\delta/2}(g_\theta(S_n), \mathrm{L_{FN}}, f)}{\left(\frac{\sum_{j=1}^n y_j g_\theta(x_j)}{|g_\theta(S_n)|} - \frac{\sqrt{n\log(2/\delta)/2}}{|g_\theta(S_n)|}\right)_+} \right\}.$$

$$\mathrm{SP}_{\mathbb{P}}(f, g_\theta) \geq \max\left\{ 0, 1 - \frac{B^*_{\delta/2}(g_\theta(S_n), \mathrm{L_{FP}}, f)}{\left(1 - \frac{\sum_{j=1}^n y_j g_\theta(x_j)}{|g_\theta(S_n)|} - \frac{\sqrt{n\log(2/\delta)/2}}{|g_\theta(S_n)|}\right)_+} \right\}.$$

Proof of Proposition 2. By proposition 1 applied to loss $\mathrm{L_{FP}}$: $\forall(\theta, \delta) \in \mathrm{Im}(\kappa_f) \times (0, 1)$

$$\mathbb{P}(\underbrace{R_{\mathbb{P},\mathrm{L_{FP}}}(f, g_\theta) \leq B^*_\delta(g_\theta(S_n), \mathrm{L_{FP}}, f)}_{=\text{ event } \mathrm{E}_1}) \geq 1 - \delta.$$

To simplify notations, g_θ and $B^*_\delta(g_\theta(S_n), \mathrm{L_{FP}}, f)$ will be noted g and B respectively for the rest of the proof. First we simply rewrite event E_1 to make an upper bound on FPR appear:

$$\begin{aligned}
\mathrm{E}_1 &\iff \mathbb{E}_{\mathbb{P}}[\mathrm{L_{FP}}(f(X), Y) \mid g(X) = 1] \leq B \\
&\iff \mathbb{P}(\mathrm{L_{FP}}(f(X), Y) = 1 \mid g(X) = 1) \leq B \\
&\iff \mathbb{P}(f(X) = 1 \cap Y = 0 \;\Big|\; g(X) = 1) \leq B \\
&\iff \underbrace{\mathbb{P}(f(X) = 1 \;\Big|\; Y = 0, g(X) = 1)}_{=\mathrm{FPR}_{\mathbb{P}}(f,g)} P(Y = 0 \;\Big|\; g(X) = 1) \;\leq B \\
&\iff \mathrm{FPR}_{\mathbb{P}}(f, g) \leq \frac{B}{1 - p}.
\end{aligned}$$

With

$$p = \mathbb{P}(Y = 1 \;\Big|\; g(X) = 1) = \frac{\mathbb{E}[Y g(X)]}{\mathbb{E}[g(X)]}$$

In total,

$$\mathbb{P}\left(\mathrm{FPR}_{\mathbb{P}}(f, g) \leq \frac{B}{1 - p} \right) = \mathbb{P}(\mathrm{E}_1) \geq 1 - \delta.$$

We now have an upper bound on FPR, but p is an unknown quantity and as such we cannot build a search algorithm for a bound of the form $\frac{B}{1-p}$ based on the knowledge of S_n. As a consequence we need to build an S_n-measurable upper bound of $\frac{B}{1-p}$, which is what we do below:

Let $(z_j)_{j\in 1,..,n}$ be the collection of random variables defined as:

$$z_j = g(x_j)\left(y_j - p \right)$$

Variables $(z_j)_{j\in 1,..,n}$ are iid since $(x_j, y_j)_{j\in 1,..,n}$ are sampled iid from $\mathbb{P}(X, Y)$. Variables $(z_j)_{j\in 1,..,n}$ are centered: $\forall j \in 1, .., n$,

$$\begin{aligned}
\mathbb{E}[z_j] &= \mathbb{E}[Y g(X)] - p\mathbb{E}[g(X)] \\
&= \mathbb{E}[Y g(X)] - \left(\frac{\mathbb{E}[Y g(X)]}{\mathbb{E}[g(X)]} \right) \mathbb{E}[g(X)] \\
&= 0.
\end{aligned}$$

Finally, $\forall j \in 1,..,n, z_j \overset{\text{a.s.}}{\in} \{-p, 1-p\}$.

We can therefore apply Hoeffding inequality to the collection $(z_j)_{j\in 1,..,n}$:

$\forall \varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{j=1}^{n} z_j \leq -\varepsilon\Big) \leq e^{-\frac{2\varepsilon^2}{n}} \\
&\iff \mathbb{P}\Big(\sum_{j=1}^{n} y_j g(x_j) - p\sum_{j=1}^{n} g(x_j) \leq -\varepsilon\Big) \leq e^{-\frac{2\varepsilon^2}{n}} \\
&\iff \mathbb{P}\Big(\underbrace{\sum_{j=1}^{n} y_j g(x_j) - p|g(S_n)| \leq -\varepsilon}_{\text{event A}}\Big) \leq e^{-\frac{2\varepsilon^2}{n}} \\
&\implies \mathbb{P}\Big(\underbrace{p \leq \frac{\sum_{j=1}^{n} y_j g(x_j)}{|g(S_n)|} + \frac{\varepsilon}{|g(S_n)|}}_{\text{event } \mathrm{E}_2}\Big) \geq 1 - e^{-\frac{2\varepsilon^2}{n}}.
\end{aligned}$$

Indeed:

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}\Big(\underbrace{A \cap |g(S_n)| = 0}_{\subseteq [0\leq -\varepsilon] = \emptyset}\Big) + \mathbb{P}\Big(A \cap |g(S_n)| > 0\Big) \\
&= 0 + \mathbb{P}\Big(p \geq \frac{\sum_{j=1}^{n} y_j g(x_j)}{|g(S_n)|} + \frac{\varepsilon}{|g(S_n)|} \cap |g(S_n)| > 0\Big)
\end{aligned}$$

And on the other hand:

$$\begin{aligned}
\mathbb{P}\Big(E_2\Big) &\geq \mathbb{P}\Big(p < \frac{\sum_{j=1}^{n} y_j g(x_j)}{|g(S_n)|} + \frac{\varepsilon}{|g(S_n)|}\Big) \\
&= 1 - \mathbb{P}\Big(p \geq \frac{\sum_{j=1}^{n} y_j g(x_j)}{|g(S_n)|} + \frac{\varepsilon}{|g(S_n)|}\Big) \\
&= 1 - \underbrace{\mathbb{P}\Big(p = +\infty \cap |g(S_n)| = 0\Big)}_{=0} - \underbrace{\mathbb{P}\Big(p \geq \frac{\sum_{j=1}^{n} y_j g(x_j)}{|g(S_n)|} + \frac{\varepsilon}{|g(S_n)|} \cap |g(S_n)| > 0\Big)}_{=\mathbb{P}(A) \text{ as shown above}} \\
&= 1 - \mathbb{P}(A) \geq 1 - e^{-\frac{2\varepsilon^2}{n}}
\end{aligned}$$

Now define $(u)_+ := \max(u, 0)$ and note that on E_2, we have

$$1 - p \geq \big(1 - \frac{\sum_{j=1}^{n} y_j g(x_j)}{|g(S_n)|} + \frac{\varepsilon}{|g(S_n)|}\big)_+$$

By applying Bonferroni inequality at order 2 to events E_1 and E_2, ie $\mathbb{P}(\mathrm{E}_1 \cap \mathrm{E}_2) \geq \underbrace{\mathbb{P}(\mathrm{E}_1)}_{\geq 1-\delta} + \underbrace{\mathbb{P}(\mathrm{E}_2)}_{\geq 1-e^{-\frac{2\varepsilon^2}{n}}} - 1$, and since

$$\mathrm{E}_1 \cap \mathrm{E}_2 \implies \mathrm{FPR}(f, g) \leq \min\Big(1, \frac{B}{\big(1 - \frac{\sum_{j=1}^{n} y_j g(x_j)}{|g(S_n)|} - \frac{\varepsilon}{|g(S_n)|}\big)_+}\Big)$$

we get that, $\forall \varepsilon > 0$:

$$\mathbb{P}\left(\text{FPR}(f,g) \leq \min\left(1, \frac{B}{\left(1 - \frac{\sum_{j=1}^{n} y_j g(x_j)}{|g(S_n)|} - \frac{\varepsilon}{|g(S_n)|}\right)_+} \right)\right) \geq \mathbb{P}(\text{E}_1 \cap \text{E}_2) \geq 1 - \delta - e^{-2\varepsilon^2/n}.$$

By taking $\varepsilon = \sqrt{\frac{n\log(1/\delta)}{2}}$, we have:

$$\mathbb{P}\left(\text{FPR}(f,g) \leq \min\left(1, \frac{B}{\left(1 - \frac{\sum_{j=1}^{n} y_j g(x_j)}{|g(S_n)|} - \frac{\varepsilon}{|g(S_n)|}\right)_+} \right)\right) \geq 1 - 2\delta.$$

So taking $\delta/2$ instead of δ finally yields:

$$\mathbb{P}\left(\text{FPR}(f,g) \leq \min\left(1, \frac{B^*_{\delta/2}(g_\theta(S_n), \text{L}_{\text{FP}}, f)}{\left(1 - \frac{\sum_{j=1}^{n} y_j g(x_j)}{|g(S_n)|} - \frac{\sqrt{n\log(2/\delta)/2}}{|g(S_n)|}\right)_+} \right)\right) \geq 1 - \delta.$$

The proof for $\text{FNR}_{\mathbb{P}}(f, g_\theta)$ and $\text{PPV}_{\mathbb{P}}(f, g_\theta)$ relies on the same steps. For PPV it involves exactly the same steps but take $p = \mathbb{P}(f(X) = 1 \mid g_\theta(X) = 1)$ and adapt the definition of z_j: $z_j = g(x_j)(f(x_j) - p)$.

Control over $\text{SE}_{\mathbb{P}}(f, g_\theta)$ and $\text{SP}_{\mathbb{P}}(f, g_\theta)$ is obtained directly through control over $\text{FNR}_{\mathbb{P}}(f, g_\theta)$ and $\text{FPR}_{\mathbb{P}}(f, g_\theta)$ respectively (see metrics definitions).

□

Proposition A2. *Let $M = \{m_1, .., m_v\} \subseteq \{R_{\mathbb{P},L_{0/1}}, R_{\mathbb{P},L_{FP}}, R_{\mathbb{P},L_{FN}}, FPR, FNR, PPV, SE, SP\}$ be a set of selective metrics ($1 \leq v \leq 8$). Let $\{\mathcal{O}_1, .., \mathcal{O}_v\}$ be the outputs of Algorithm 2 corresponding to the metrics in M, and $\{\Theta_1, .., \Theta_v\}$ the associated projections on the first coordinate: for instance $\Theta_1 = \{\theta : (\theta, \nu) \in \mathcal{O}_1\}$. Then,*

If $\bigcap_{t\in 1,..,v} \Theta_t \neq \emptyset$, any $\theta \in \bigcap_{t\in 1,..,v} \Theta_t$ satisfies g_θ controlling all the metrics in M with probability at least $1 - v\delta$. Additionally, the coverage-maximizing selection function g^* to control all metrics in M is g_{θ^*} with $\theta^* = \min(\bigcap_{t\in 1,..,v} \Theta_t)$.

Proof of Proposition A2. For $t \in 1, .., v$, we consider $\mathcal{O}_t$ the output of Algorithm 2 for metric $m_t \in M$. For all $(\theta, \nu) \in \mathcal{O}_t$, let event $E_t(\theta, \nu)$ be defined as

$$E_t(\theta,\nu) = \begin{cases} [m_t(f, g_\theta) \leq \nu] \text{ if } m_t \in \{\text{R}_{\mathbb{P},\text{L}_{0/1}}, \text{R}_{\mathbb{P},\text{L}_{\text{FP}}}, \text{R}_{\mathbb{P},\text{L}_{\text{FN}}}, \text{FPR}, \text{FNR}\} \\ [m_t(f, g_\theta) \geq \nu] \text{ if } m_t \in \{\text{PPV}, \text{SE}, \text{SP}\} \end{cases}$$

By Propositions 1-2 applied to Algorithms 1-2, $\forall t \in 1, .., v$:

$$\mathbb{P}\Big(\bigcap_{(\theta,\nu)\in\mathcal{O}_t} E_t(\theta,\nu)\Big) \geq 1 - \delta \qquad (*)$$

For all $t \in 1, .., v$, let Θ_t be defined as the projection set of $\mathcal{O}_t$ on its first coordinate: $\Theta_t = \{\theta : (\theta, \nu) \in \mathcal{O}_t\}$. Then let θ^* be defined as the minimum of the intersection of the projections, if not empty:

$$\theta^* = \min\Big(\bigcap_{t\in 1,..,v} \Theta_t\Big)$$

By definition of θ^*, $\forall t \in 1, .., v, \exists!\, \nu_t^* \in [0, 1] : (\theta^*, \nu_t^*) \in \mathcal{O}_t$, so

$$\bigcap_{t\in 1,..,v}\bigcap_{(\theta,\nu)\in\mathcal{O}_t} E_t(\theta,\nu) \subseteq \bigcap_{t\in 1,..,v} E_t(\theta^*,\nu_t^*)$$

$$\implies \mathbb{P}\Big(\bigcap_{t\in 1,..,v} E_t(\theta^*,\nu_t^*)\Big) \geq \mathbb{P}\Big(\bigcap_{t\in 1,..,v}\bigcap_{(\theta,\nu)\in\mathcal{O}_t} E_t(\theta,\nu)\Big)$$

$$\implies \mathbb{P}\Big(\bigcap_{t\in 1,..,v} E_t(\theta^*,\nu_t^*)\Big) \geq 1 - v + \sum_{t=1}^{v} \underbrace{\mathbb{P}\Big(\bigcap_{(\theta,\nu)\in\mathcal{O}_t} E_t(\theta,\nu)\Big)}_{\geq 1-\delta \text{ by } (*)}$$

$$\implies \mathbb{P}\Big(\bigcap_{t\in 1,..,v} E_t(\theta^*,\nu_t^*)\Big) \geq 1 - v\delta$$

Which concludes the proof. □

B Supplementary definitions of selective metrics

In this section, we present explicit definitions of all selective metrics, formulated in terms of both probabilities and expectations.

Table 1: Selective metrics definitions.

Selective metric	**Definition as probability**	**Definition as expectation**
$\mathrm{R}_{\mathbb{P},\mathrm{L}_{0/1}}(f,g_\theta)$	$\mathbb{P}(f(X)\neq Y \mid g_\theta(X)=1)$	$\mathbb{E}_{\mathbb{P}}[\mathrm{L}_{0/1}(f(X),Y) \mid g_\theta(X)=1]$
$\mathrm{R}_{\mathbb{P},\mathrm{L}_{\mathrm{FP}}}(f,g_\theta)$	$\mathbb{P}(f(X)=1 \cap Y=0 \mid g_\theta(X)=1)$	$\mathbb{E}_{\mathbb{P}}[\mathrm{L}_{\mathrm{FP}}(f(X),Y) \mid g_\theta(X)=1]$
$\mathrm{R}_{\mathbb{P},\mathrm{L}_{\mathrm{FN}}}(f,g_\theta)$	$\mathbb{P}(f(X)=0 \cap Y=1 \mid g_\theta(X)=1)$	$\mathbb{E}_{\mathbb{P}}[\mathrm{L}_{\mathrm{FN}}(f(X),Y) \mid g_\theta(X)=1]$
$\mathrm{FPR}(f,g_\theta)$	$\mathbb{P}(f(X)=1 \mid Y=0, g_\theta(X)=1)$	$\mathbb{E}_{\mathbb{P}}[\mathrm{L}_{\mathrm{FP}}(f(X),Y) \mid Y=0, g_\theta(X)=1]$
$\mathrm{FNR}(f,g_\theta)$	$\mathbb{P}(f(X)=0 \mid Y=1, g_\theta(X)=1)$	$\mathbb{E}_{\mathbb{P}}[\mathrm{L}_{\mathrm{FN}}(f(X),Y) \mid Y=1, g_\theta(X)=1]$
$\mathrm{PPV}(f,g_\theta)$	$\mathbb{P}(Y=1 \mid f(X)=1, g_\theta(X)=1)$	$1-\mathbb{E}_{\mathbb{P}}[\mathrm{L}_{\mathrm{FP}}(f(X),Y) \mid f(X)=1, g_\theta(X)=1]$
$\mathrm{SE}(f,g_\theta)$	$\mathbb{P}(f(X)=1 \mid Y=1, g_\theta(X)=1)$	$1-\mathbb{E}_{\mathbb{P}}[\mathrm{L}_{\mathrm{FN}}(f(X),Y) \mid Y=1, g_\theta(X)=1]$
$\mathrm{SP}(f,g_\theta)$	$\mathbb{P}(f(X)=0 \mid Y=0, g_\theta(X)=1)$	$1-\mathbb{E}_{\mathbb{P}}[\mathrm{L}_{\mathrm{FP}}(f(X),Y) \mid Y=0, g_\theta(X)=1]$

C Algorithms Implementation

In addition to the algorithms themselves, this section presents the terminal conditions and computational considerations (handling large binomial coefficients, discretization in the greedy search, execution times, and the code environment and hardware) used to run Algorithms 1–2.

C.1 Algorithms to compute guaranteed selective metrics bounds

Application of Algorithm 1 to control FP and FN risks. Running Algorithm 1 with inputs S_n, f, L_{FP}, κ_f, δ, r^*, $(\theta_{\min},\theta_{\max})$, and K_1 returns (f, g_{θ^*}) together with

$$B^* = B^*_{\delta/(2^{K_1}-1)}(g_{\theta^*}(S_n), \mathrm{L}_{\mathrm{FP}}, f)\,.$$

Applying Proposition 1 to the grid $\mathcal{G}_1$ of all thresholds explored during the K_1 dichotomy iterations, with $|\mathcal{G}_1| = 2^{K_1}-1$, a union bound gives

$$\begin{aligned}\mathbb{P}(R_{\mathbb{P},\mathrm{L}_{\mathrm{FP}}}(f,g_{\theta^*}) > B^*) &\leq \sum_{\theta\in\mathcal{G}_1} \mathbb{P}(R_{\mathbb{P},\mathrm{L}_{\mathrm{FP}}}(f,g_\theta) > B_\theta)\\ &\leq |\mathcal{G}_1|(\delta/(2^{K_1}-1)) \leq \delta.\end{aligned}$$

Thus, $R_{\mathbb{P},\mathrm{L}_{\mathrm{FP}}}(f,g_{\theta^*}) \leq B^*$ with probability at least $1-\delta$. Figure 4 reports bounds for 50 targets $r^* \in [0, 0.2]$ (linear grid) as a function of θ^*; the same procedure with loss L_{FN} yields control of $R_{\mathbb{P},\mathrm{L}_{\mathrm{FN}}}$.

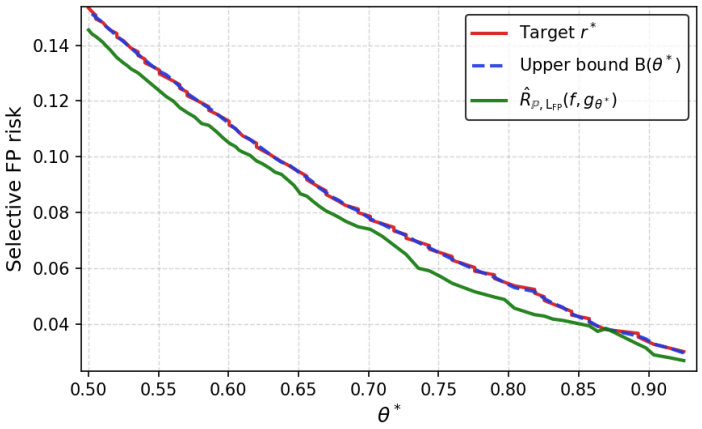


Figure 4: Guaranteed FP risk bounds from Algorithm 1 as a function of the associated confidence threshold θ^*. Dataset: CIFAR-10.

Non monotonicity of the conditional metrics bounds. From Proposition 2, we obtain bounds for FPR, FNR, PPV, SE, and SP that hold with probability at least $1-\delta$. Conversely to the bounds for $\mathrm{R}_{\mathbb{P},\mathrm{L}_{0/1}}$, $\mathrm{R}_{\mathbb{P},\mathrm{L_{FP}}}$, and $\mathrm{R}_{\mathbb{P},\mathrm{L_{FN}}}$, these conditional bounds are not monotonic in $\theta \in \mathrm{Im}(\kappa_f)$ (see Figure 5a). Their numerator decreases with θ (Lemma A3), but the denominator depends on $\sum_{j=1}^{n} y_j g_\theta(x_j)/|g_\theta(S_n)|$ (proportion of positives), $\sum_{j=1}^{n} f(x_j)g_\theta(x_j)/|g_\theta(S_n)|$ (proportion of positive predictions), and the term $-\sqrt{n\log(2/\delta)/2}/|g_\theta(S_n)|$, all of which may decrease as $g_\theta(S_n)$ shrinks with higher θ. Thus monotonicity cannot be ensured without stronger assumptions on f and S_n.

Figure 5b shows cases where numerator and denominator decrease together. Studying the denominator's evolution with θ may allow excluding regions of steep decline, mitigating the instability of Figure 5a.

Greedy search to control any selective metric In the absence of monotonicity, we resort to a greedy search strategy (Algorithm 2) to identify regions of θ within $\mathrm{Im}(\kappa_f)$ for which the metric bounds meet a user-specified target $r^* \in (0,1)$. As an illustration, Figure 5a highlights the regions I_{FNR}, corresponding to FNR upper bounds not exceeding 0.13, and I_{PPV}, corresponding to PPV lower bounds of at least 0.55.

Let $\mathcal{O}$ denote the set of (random S_n-dependent) outputs of Algorithm 2 for a chosen `metric` among those in Section 2. As in Algorithm 1 we uniformly guarantee all values in the fixed grid $\mathcal{G}_2$ by computing bounds at δ/K_2, such that any pair (θ, ν) from $\mathcal{O}$ is guaranteed with probability at least $1-\delta$.

Thus, for any `metric`, if $\mathcal{O}$ is nonempty, choosing

$$\theta^* = \min\{\theta : (\theta, \nu) \in \mathcal{O}\}$$

yields the coverage-maximizing selective classifier (f, g_{θ^*}) guaranteeing control of the chosen `metric` with probability at least $1-\delta$.

Remarks on Algorithms 1–2. For any target collection $(r_i^*)_{i=1}^q \in (0,1)^q$ with $q \in \mathbb{N}^*$, the proposed algorithms provide optimal thresholds $(\theta_i^*)_{i=1}^q$, their bounds, and the corresponding empirical coverages. Since the confidence threshold θ maps to dataset coverage, it is possible to draw metrics evolution with coverage. Figure 10 shows the resulting bounds–coverage curves on a specific dataset.

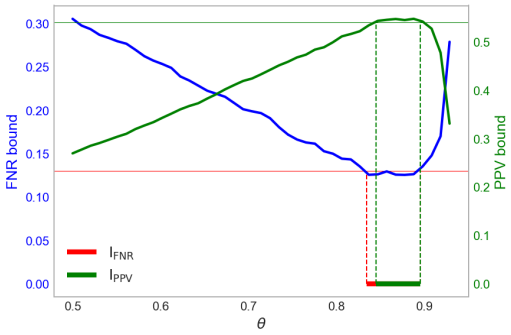
For risks control, both Algorithms 1–2 apply. Algorithm 1 is simpler (execution time comparison in Supplementary C.3), yielding a single threshold θ^* and bound B^* close to the target r^*. Algorithm 2, by contrast, returns a set $\mathcal{O}$ of thresholds and bounds. Thus, Algorithm 1 is preferable for individual risks, while Algorithm 2 is useful when multiple thresholds are required, e.g., to identify common ones for joint control across metrics (Proposition A2, Supplementary A).

One should note that both algorithms guarantee that the returned bounds satisfy the user-specified target r^*; however, no explicit guarantee is provided on the deviation $|r^* - B^*|$. In our experiments, this deviation is consistently small.

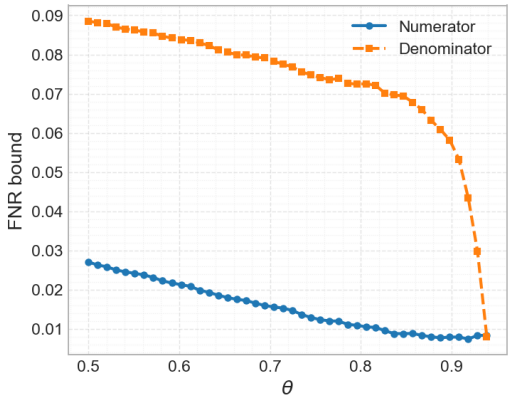
Algorithm 1 Dichotomy search returning $(f, g_{\theta^*}), B^*$ such that $R_{\mathbb{P},\mathrm{L}}(f, g_{\theta^*}) \leq B^*$ with probability at least $1-\delta$, where B^* is close to the user-specified target r^*. **Inputs**: dataset S_n; classifier f; loss $\mathrm{L} : \mathcal{Y}^2 \to \{0,1\}$; confidence function κ_f; probability level $\delta \in (0,1)$; target r^*; a fixed S_n-independent interval $(\theta_{\min}, \theta_{\max})$ representing $\mathrm{Im}(\kappa_f)$; and K_1, the number of iterations. **Terminal conditions** (line 6) handle pathological cases; their justification and implementation details are provided in Supplementary C.

Require: $S_n, f, \mathrm{L}, \kappa_f, \delta \in (0,1), r^* \in (0,1), \theta_{\min} < \theta_{\max} \in \mathrm{Im}(\kappa_f), K_1 \in \mathbb{N}^*$

1: **for** $k \in 1, .., K_1$ **do**
2: $\quad \theta^* \leftarrow \frac{\theta_{\min}+\theta_{\max}}{2}$
3: $\quad g_{\theta^*}(S_n) \leftarrow \{(x,y) \in S_n : g_{\theta^*}(x) = 1\}$
4: $\quad \hat{r} \leftarrow \hat{R}_L(f \mid g_{\theta^*}(S_n))$
5: $\quad B^* \leftarrow B^+_{\delta/(2^{K_1}-1)}(g_{\theta^*}(S_n), L, f)$
6: $\quad$ **if** $B^* = 1 \vee (\hat{r} = 0 \wedge B^* \geq r^*)$ **then**
7: $\quad\quad$ **return** $\emptyset$
8: $\quad$ **else**
9: $\quad\quad$ **if** $B^* < r^*$ **then**
10: $\quad\quad\quad \theta_{\max} \leftarrow \theta^*$
11: $\quad\quad$ **else**
12: $\quad\quad\quad \theta_{\min} \leftarrow \theta^*$
13: $\quad\quad$ **end if**
14: $\quad$ **end if**
15: **end for**
16: **return** $(f, g_{\theta^*}), B^*$



(a) Evolution of FNR and PPV bounds with respect to θ. Regions $\mathrm{I}_{\mathrm{FNR}}$ and $\mathrm{I}_{\mathrm{PPV}}$ correspond to FNR bound $\in (0, 0.13]$ and PPV bound $\in [0.55, 1)$. $0.84 \in \mathrm{I}_{\mathrm{FNR}} \cap \mathrm{I}_{\mathrm{PPV}}$ so the selective classifier $(f, g_{0.84})$ guarantees FNR ≤ 0.13 and PPV ≥ 0.55 with probability at least $1 - 2\delta = 0.99$.



(b) Detailed view of the selective FNR bound ratio as a function of the confidence threshold.

Figure 5: CIFAR-10 airplane detection with a small CNN. Panel (a) shows the evolution of FNR and PPV bounds with respect to the confidence threshold θ, and panel (b) provides a detailed view of the FNR bound components.

C.2 Terminal Conditions of Algorithms 1–2

The algorithms include explicit terminal conditions to handle degenerate cases where the search cannot progress toward the target r^*. Below we detail the rationale behind each condition.

First terminal condition: $B^* = 1$. This condition prevents the search from continuing when the bound has become maximal and can no longer decrease. It is used in both the dichotomy algorithm (Algorithm 1, line 6) and the greedy algorithm (Algorithm 2, line 14). By definition, $B^* = 1$ occurs if either $|g_{\theta^*}(S_n)| = 0$ or $\hat{r} = 1$.

- If $B^* = 1$ because $|g_{\theta^*}(S_n)| = 0$, then θ increases at the next iteration ($B^* = 1 \Rightarrow \theta_{\min} \leftarrow \theta$ in Algorithm

Algorithm 2 Greedy search granting control over a `metric`. $\text{bound}_{\texttt{metric}}(\delta, g_\theta, \mathrm{L}, S_n)$ in line 18 is the corresponding `metric` bound from Proposition 2. Computation details are provided in Supplementary C.3.

Require: S_n, f, κ_f, $\delta \in (0,1)$, $r^* \in (0,1)$, $\texttt{metric} \in \{\mathrm{R}_{\mathbb{P},\mathrm{L}_{0/1}}, \mathrm{R}_{\mathbb{P},\mathrm{L}_{\mathrm{FP}}}, \mathrm{R}_{\mathbb{P},\mathrm{L}_{\mathrm{FN}}}, \mathrm{FPR}, \mathrm{FNR}, \mathrm{PPV}, \mathrm{SE}, \mathrm{SP}\}$, $K_2 \in \mathbb{N}^*$, $\mathcal{G}_2 = \{\theta_1 \leq \theta_2 \leq .. \leq \theta_{K_2}\} \in \mathrm{Im}(\kappa_f)^{K_2}$

1: $\mathcal{O} \leftarrow \emptyset$
2: **if** $\texttt{metric} \in \{\mathrm{R}_{\mathbb{P},\mathrm{L}_{\mathrm{FP}}}, \mathrm{FPR}, \mathrm{PPV}, \mathrm{SP}\}$ **then**
3: $\quad \mathrm{L} \leftarrow \mathrm{L}_{\mathrm{FP}}$
4: **else if** $\texttt{metric} \in \{\mathrm{R}_{\mathbb{P},\mathrm{L}_{\mathrm{FN}}}, \mathrm{FNR}, \mathrm{SE}\}$ **then**
5: $\quad \mathrm{L} \leftarrow \mathrm{L}_{\mathrm{FN}}$
6: **else**
7: $\quad \mathrm{L} \leftarrow \mathrm{L}_{0/1}$
8: **end if**
9: **for** $\theta \in \mathcal{G}_2$ **do**
10: $\quad g_\theta(S_n) \leftarrow \{(x,y) \in S_n : g_\theta(x) = 1\}$
11: $\quad \hat{r} \leftarrow \hat{R}_{\mathrm{L}}(f \mid g_\theta(S_n))$
12: $\quad$ **if** $\texttt{metric} \in \{\mathrm{R}_{\mathbb{P},\mathrm{L}_{0/1}}, \mathrm{R}_{\mathbb{P},\mathrm{L}_{\mathrm{FP}}}, \mathrm{R}_{\mathbb{P},\mathrm{L}_{\mathrm{FN}}}\}$ **then**
13: $\quad\quad \nu \leftarrow B^*_{\delta/K_2}(g_\theta(S_n), \mathrm{L}, f)$
14: $\quad\quad$ **if** $\nu = 1$ **then**
15: $\quad\quad\quad$ **return**
16: $\quad\quad$ **end if**
17: $\quad$ **else**
18: $\quad\quad \nu \leftarrow \text{bound}_{\texttt{metric}}(\delta/K_2, g_\theta, \mathrm{L}, S_n)$
19: $\quad$ **end if**
20: $\quad$ **if** $\texttt{metric} \in \{\mathrm{PPV}, \mathrm{SE}, \mathrm{SP}\}$ {lower-bounded metrics} **then**
21: $\quad\quad$ **if** $\nu \geq r^*$ **then**
22: $\quad\quad\quad \mathcal{O} \leftarrow \text{append}(\mathcal{O}, (\theta, \nu))$
23: $\quad\quad$ **end if**
24: $\quad$ **else** {upper-bounded metrics}
25: $\quad\quad$ **if** $\nu \leq r^*$ **then**
26: $\quad\quad\quad \mathcal{O} \leftarrow \text{append}(\mathcal{O}, (\theta, \nu))$
27: $\quad\quad$ **end if**
28: $\quad$ **end if**
29: **end for**
30: **return** $\mathcal{O}$

1, and by construction in Algorithm 2), so the selected set is empty again at the next iteration and we still have $B^* = 1$. By recurrence, the upper bound remains fixed at 1 and never reaches the target r^*.

- If $B^* = 1$ because $\hat{r} = 1$, then θ increases at the next iteration (again, $B^* = 1 \Rightarrow \theta_{\min} \leftarrow \theta$ in Algorithm 1, and by construction in Algorithm 2), reducing the selected set at the next iteration. Since f was mistaken on all samples of $g_{\theta^*}(S_n)$ (as $\hat{r} = 1$), it will also be mistaken on all samples of the reduced selected set at the next iteration, so $\hat{r} = 1$ and again $B^* = 1$. Again, by recurrence, the upper bound never reaches the target r^*.

Second terminal condition: ($\hat{r} = 0$ **and** $B^* \geq r^*$)**.** It appears only in Algorithm 1 (line 6). The logic is as follows: if $\hat{r} = 0$ then, by Proposition A1, $B^* = 1 - \delta^{\frac{1}{n}}$. Moreover, as $B^* \geq r^*$, the next dichotomy step increases the threshold θ, thus yielding a non-superior empirical risk at the next iteration, therefore equal to 0. Hence again $B^* = 1 - \delta^{\frac{1}{n}}$. The two consecutive iterations are identical, and by recurrence the upper bound never reaches r^*.

C.3 Computation details of Algorithms 1–2

Binomial coefficients computation details. In Algorithms 1–2, the computation of B^* requires the evaluation of binomial coefficients $\binom{n}{k}$ (see Proposition A1), where n denotes the size of the selected subset. Depending

on the dataset, n can be on the order of 10^2, 10^5, or even larger. In such cases, computing $\binom{n}{k}$ directly from its factorial definition

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

is numerically infeasible. A classical remedy is to rely on the Gamma function, which generalizes the factorial through the identity $\Gamma(n+1) = n!$. When n is large, it is more stable to work in logarithmic form by using the log-gamma function, denoted $\log\Gamma(\cdot)$. In this representation, the logarithm of the binomial coefficient is expressed as

$$\log\binom{n}{k} \;=\; \log\Gamma(n+1) \;-\; \log\Gamma(k+1) \;-\; \log\Gamma(n-k+1).$$

This formulation avoids overflow and significantly improves numerical accuracy in practical computations. In particular, it is the approach we adopt to evaluate the binomial terms involved in the sum appearing in Proposition A1.

Discretization of $\mathrm{Im}(\kappa_f)$. In Algorithms 1–2, the confidence space $\mathrm{Im}(\kappa_f)$ is explored using discrete grids $\mathcal{G}_1$ and $\mathcal{G}_2$ that are independent of S_n. This approach relies on the underlying assumption that $\mathrm{Im}(\kappa_f)$ is an interval of $\mathbb{R}$.

More precisely, under the following assumptions:

1. $\mathcal{X}$ is a connected set;

2. $\kappa_f : \mathcal{X} \to \mathbb{R}$ is continuous;

classical results in topology state that the image of a connected set under a continuous function is connected, and that the connected subsets of $\mathbb{R}$ are precisely intervals [Bourbaki, 1971]. Consequently, under Assumptions (1)–(2), $\mathrm{Im}(\kappa_f)$ is an interval.

Assumption (2) holds when κ_f is the Softmax response of a convolutional network f, since it is a composition of continuous functions and is therefore continuous. Continuity also holds when κ_f corresponds to the Monte Carlo dropout negative variance; in this case, it follows from an application of Lebesgue's Dominated Convergence Theorem. Assumption (1), however, depends on the data-generating distribution $\mathbb{P}$, since $\mathcal{X} = X(\Omega)$. As $\mathbb{P}$ is unknown, this assumption cannot be verified in practice.

In practice, we discretize $\mathrm{Im}(\kappa_f)$ using $K_2 = 50$ points, which provides a satisfactory trade-off between numerical precision and computational cost.

Execution times of Algorithms 1–2. We compare the execution times of Algorithm 1 and Algorithm 2 using $K_2 = 30$, 50, 70, and 90. Synthetic datasets are generated with sample sizes ranging from $n = 10{,}000$ to $n = 100{,}000$ (step size 10,000). Each synthetic dataset consists of "true" binary labels noted `y_true`, "predicted" labels noted `y_pred`, and associated confidence scores noted `kappa`, simulated as follows.

1. **True labels (`y_true`).** Ground-truth labels $(y_i^{\text{true}})_{i=1}^n$ are sampled independently and uniformly from $\{0, 1\}$, so the dataset is balanced in expectation. Indeed, as our analysis focuses solely on relative execution times across different sample sizes, there was no methodological justification for simulating imbalanced datasets.

2. **Confidence scores (`kappa`).** For each sample i, a confidence score $\kappa_i \in (0, 1)$ is drawn from a mixture of two Beta distributions:

$$\kappa_i \sim \begin{cases} \mathrm{Beta}(9, 1), & \text{with probability } \pi \in (0, 1), \\ \mathrm{Beta}(3, 2), & \text{with probability } 1 - \pi. \end{cases}$$

   The $\mathrm{Beta}(9, 1)$ component is concentrated near 0.9, representing high-confidence predictions, while $\mathrm{Beta}(3, 2)$ has mean 0.6 and larger variance, representing lower-confidence predictions. π is the proportion of high-confidence predictions.

3. **Predicted labels (`y_pred`).** Given κ_i, the predicted label y_i^{pred} is generated by introducing errors with probability proportional to $(1 - \kappa_i)$:

$$y_i^{\text{pred}} = \begin{cases} y_i^{\text{true}}, & \text{with probability } \kappa_i, \\ 1 - y_i^{\text{true}}, & \text{with probability } 1 - \kappa_i. \end{cases}$$

Thus, higher confidence scores correspond to a lower probability of prediction error.

Overall, the simulation yields a dataset

$$\{(y_i^{\text{true}}, y_i^{\text{pred}}, \kappa_i)\}_{i=1}^{n}$$

with expected accuracy

$$\pi.\mathbb{E}[\kappa \mid \kappa \sim \text{Beta}(9, 1)] + (1 - \pi).\mathbb{E}[\kappa \mid \kappa \sim \text{Beta}(3, 2)].$$

In practice we used $\pi = 0.7$ to obtain an expected accuracy of 0.8, consistent with the accuracies of our models reported in Section 4). To go into details, choosing $\pi = 0.7$ yields an expected accuracy of:

$$0.7 \times 0.9 + 0.3 \times 0.6 = 0.81.$$

The final dataset is returned as a table with three columns: `y_true`, `y_pred`, and `kappa`. Figure 6 shows that the empirical misclassification error rate (= 1−accuracy) evolves monotonically with respect to the threshold on `kappa`.

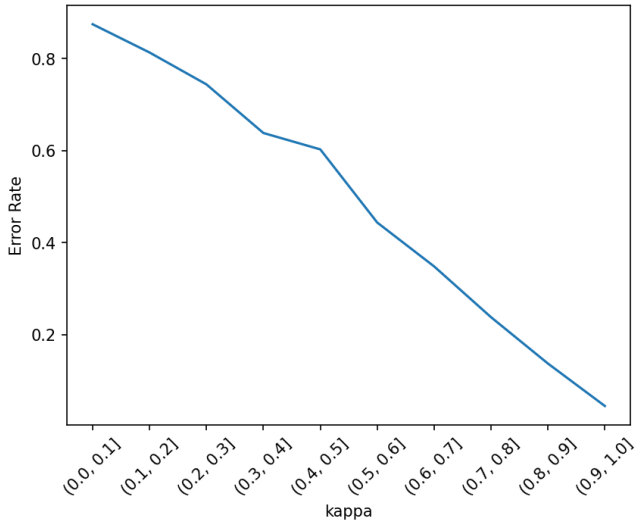


Figure 6: Evolution of the empirical misclassification risk rate in a β-mixture simulated dataset as a function of the threshold on `kappa`. In this setting we chose a sample size $n = 10,000$.

We then compute the execution times of Algorithm 1 and Algorithm 2 using $K_2 = 30$, 50, 70 or 90, with a fixed target $r^* = 5\%$ and the misclassification risk chosen as the `metric` to control. For each n value, execution times are computed over 20 simulated datasets with fixed random seeds to ensure full reproducibility. Figure 7 reports the mean computation times across seeds, with confidence intervals obtained from the 25th and 75th percentiles.

As expected, the dichotomy search is the fastest. For the greedy search, execution time grows approximately linearly both with the number of discretization points and with n. Importantly, even the most demanding case ($K_2 = 90$ and $n = 100{,}000$) requires under two minutes, showing that Algorithm 2 remains practical for large datasets.

Code and hardware specifications.

- The code, along with detailed instructions for reproducing all results, is publicly available at `https://github.com/EmilienJemelen/selective-classification`.

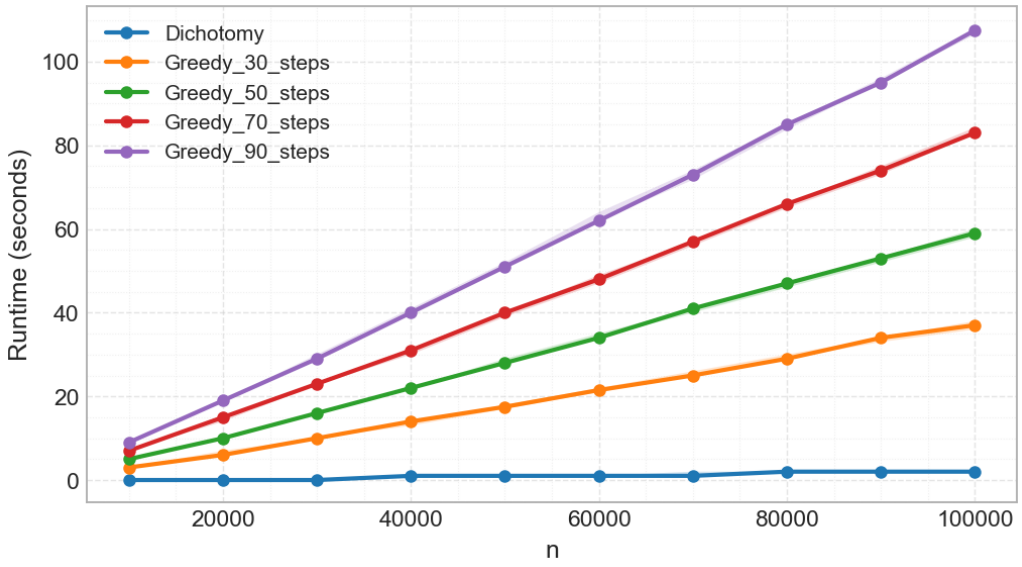


Figure 7: Execution times of Algorithm 1 and Algorithm 2 with respect to the dataset sample size n. Results are averaged over 20 simulated datasets per n value, with empirical confidence intervals.

- All experiments were conducted in Python 3.13.2 using the VS Code IDE.
- Experiments were run on a local machine equipped with a small GPU (RTX 4060). This setup was sufficient for both model training (see Section D.2) and repeated runs of Algorithms 1–2.

D Modeling details

This section presents the architectures employed in our empirical experiments, their training protocols, and the approach used to derive the κ_f confidence functions.

D.1 Architectures

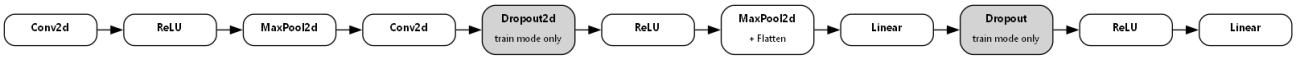


Figure 8: CNN used for the binary classification tasks. CIFAR-10 input is a 32×32 RGB image. Layer-wise tensor shapes (for CIFAR-10, the same logic applies for H&E images, which are 512×512 RGB images): $(3, 32, 32) \xrightarrow[\text{ReLU}]{\text{Conv2d: } 3\to20,\ k=4,\ p=1} (20, 31, 31) \xrightarrow{\text{MaxPool } 2\times2} (20, 15, 15) \xrightarrow[\text{ReLU}]{\text{Conv2d } 20\to32,\ k=4,\ p=1} (32, 14, 14) \xrightarrow{\text{MaxPool } 2\times2} (32, 7, 7) \xrightarrow{\text{AdaptiveAvgPool } (4,4)} (32, 4, 4) \xrightarrow{\text{Flatten}} 512 \xrightarrow[\text{train only}]{\text{Dropout}} \xrightarrow{\text{Linear } 512\to64} \xrightarrow{\text{ReLU}} \xrightarrow{\text{Linear } 64\to2} (2)$. k denotes the convolution kernel size and p the padding parameter. The final 2-D output corresponds to the logits of the two classes. The adaptive pooling layer enforces a fixed feature size, allowing the model to accommodate varying input resolutions and thus be used for both CIFAR-10 and H&E images classification tasks.

Simple convolutional network (CNN). To perform binary classification on CIFAR-10 images and H&E images, we use the CNN architecture shown in Figure 8. The model includes two dropout layers: Dropout2d after the second convolution and Dropout before the final fully connected layer. Both use a dropout probability of 0.2, meaning that each neuron is independently masked with probability 0.2 during training. Dropout is disabled at test time, ensuring fully reproducible results.

ResNet-18. To illustrate the behavior of selective prediction on a strong baseline classifier, we trained a ResNet-18 model [He et al., 2016a] for binary classification (airplane vs. non-airplane) on the CIFAR-10 dataset. This experiment is intended purely as an illustration, so we restricted it to CIFAR-10 without extending training to the H&E dataset. The corresponding training setup and test results are reported in Section D.2.

D.2 Models training recipes

CNN on CIFAR-10. The CIFAR-10 dataset was loaded using the `torchvision.datasets` module. Class labels were binarized into *airplane* vs. *non-airplane*. Data splits followed the procedure described in the Common Setup section of the Empirical Experiments (Section 4). Model training and per-epoch validation set ($1,000$ images out of the training set) were balanced by oversampling the minority class (airplanes). The learning rate was initialized at 10^{-4} and reduced by a factor of 0.1 upon a validation-accuracy plateau (no patience). Optimization used Adam with standard binary cross-entropy (BCE) loss and no regularization. Training was stopped after 10 epochs, reaching 81.78% training accuracy and 81.88% validation accuracy. Final weights are stored in `models_weights/cnn_cifar_binary_MCD_epoch9.pth` in the code repository of this paper. For the remaining $40,000$ unseen samples, we computed Softmax Response and Monte Carlo Dropout negative variance confidence functions (see Section D.3).

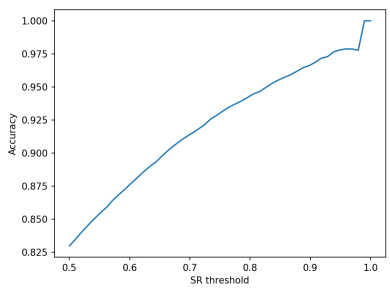
CNN on H&E images for colorectal cancer detection. H&E-stained colorectal tissue images were obtained from the Zenodo archive [Kather et al., 2018] (`https://zenodo.org/records/2530789`). Data splits followed the procedure described at the beginning of Section 4. The training set contained $5,274$ images, with a per-epoch validation set of 806 images balanced by oversampling the minority (cancer) class (37%). The learning rate was initialized at 10^{-4} and reduced by a factor of 0.1 upon validation-accuracy plateau (no patience). The Adam optimizer and BCE loss were used. Training stopped after one epoch, achieving 85.53% training accuracy and 91.69% validation accuracy. Final weights are stored in `models_weights/cnn_wsi_binary_epoch0.pth`. For the remaining $6,788$ unseen samples, we computed Softmax Response and Monte Carlo Dropout negative variance confidence functions (see Section D.3).

ResNet-18 on CIFAR-10. ResNet-18 was trained for 20 epochs using the same training and validation sets as the CNN. The learning rate started at 10^{-4} and was reduced by a factor of 0.1 upon a validation-accuracy plateau (no patience). Optimization again used Adam with BCE loss. The last epoch model achieved 98.46% validation accuracy. Since this ResNet-18 experiment was meant to illustrate that high-performance models gain limited benefit from abstention, we reported results only for CIFAR-10 data using the Softmax Response confidence function (Monte Carlo Dropout was omitted as it would have required an additional $40,000 \times T$ forward passes, with $T \geq 10$ Gal and Ghahramani [2016], see D.3). Final weights are available at `models_weights/resnet18_cifar_binary_epoch19.pth`.

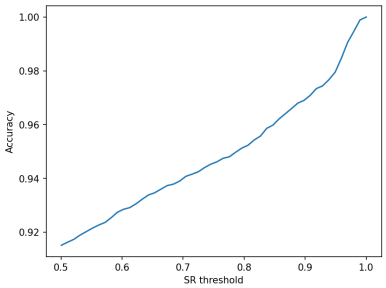
D.3 Softmax Response (SR) and Monte Carlo Dropout negative variance (MCD) computation details

Softmax Response (SR). For the CNN trained on CIFAR-10 and H&E datasets, we extracted the *softmax response* (SR) defined as the maximum of the two softmax outputs. In binary classification, SR values lie in $(0.5, 1)$. For all samples unseen at training in both datasets, we empirically verified the monotonic relationship between classification accuracy and the confidence threshold, which indicates a well-calibrated confidence function [Gal and Ghahramani, 2016]. Figures 9a-9b-9e show that SR yields well-calibrated confidence estimates for our models and datasets.

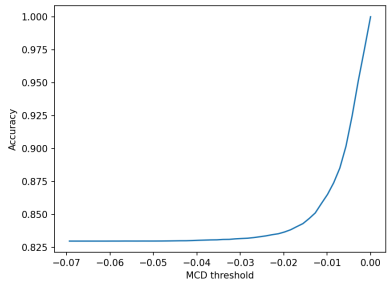
Monte Carlo Dropout (MCD) negative variance. Following Gal and Ghahramani [2016], the CNN was trained with dropout layers (see Figure 8). Applying dropout after the second convolutional and the first linear layer provided stable monotonic calibration, as shown in Figures 9c–9d. After training, we computed the MCD confidence function for all unseen samples from the CIFAR-10 and H&E datasets as follows: for each sample, we recorded the maximum softmax probability corresponding to the model's predicted class (obtained without dropout), then performed $T = 30$ forward passes with dropout enabled and computed the variance of the original prediction. Intuitively, the variance reflects the classifier's uncertainty in its original prediction, while its negative serves as a confidence function used in our selective prediction framework.



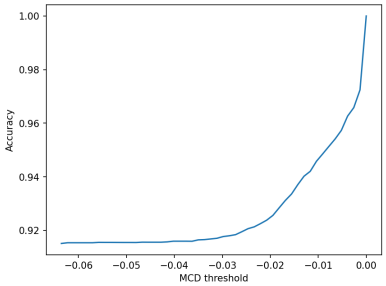
(a) CIFAR-10 CNN accuracy vs. SR confidence.



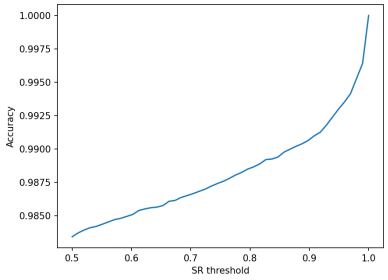
(b) H&E CNN accuracy vs. SR confidence.



(c) CIFAR-10 CNN accuracy vs. MCD confidence.



(d) H&E CNN accuracy vs. MCD confidence.



(e) CIFAR-10 ResNet-18 accuracy vs. SR confidence.

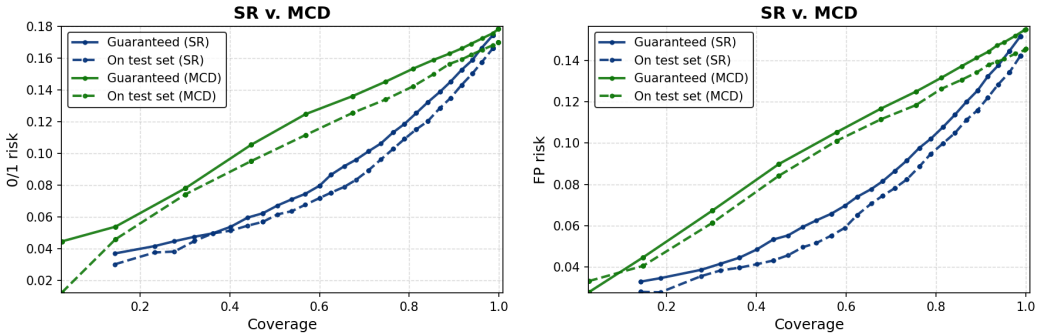
Figure 9: Calibration of the Softmax Response (SR) and Monte Carlo Dropout negative variance (MCD) confidence functions across datasets. (a)–(d): empirical calibration results for CNN models on CIFAR-10 and H&E datasets. (e): example of a highly accurate model (ResNet-18 on CIFAR-10) where thresholding yields only marginal gains.

E Supplementary experimental results

E.1 CIFAR-10 experiments

Task & models. We study binary airplane detection, where airplanes represent 10% of CIFAR-10's 60k images. A lightweight CNN with two convolutional layers (44k parameters; training details in Supplementary D) is trained, and confidence scores are derived from Softmax Response (SR) and Monte Carlo dropout variance (MCD) [Gal and Ghahramani, 2016, Geifman and El-Yaniv, 2017, Pugnana et al., 2024, Goren et al., 2024, Santosh et al., 2024] (see Supplementary D.3). For comparison, we also train a ResNet-18 [He et al., 2016b] (11M parameters; results in Supplementary E.1), showing that highly accurate models gain less from abstention.

Risks guarantees. As shown in Section C.1, selective risks bounds from Algorithm 1 hold with probability at least $1 - \delta = 0.995$. Figure 10 shows that false positive risk dominates misclassification risk, as panels 10a and 10b are nearly identical; SR achieves consistently better risk–coverage trade-offs and tighter bounds than MCD (see Supplementary E.1 for additional results).



(a) Selective risk-coverage trade-offs.

(b) FP risk-coverage trade-offs.

Figure 10: CIFAR-10 airplane detection. Bounds closely match estimates, and SR outperforms MCD across all coverages. Results use 50 target values $r^* \in [0, 0.2]$ (linear grid) as inputs to Algorithm 1 (see remarks at the end of Section C.1).

Conditional metrics guarantees. Algorithm 2 provides conditional guarantees valid with probability at least $1 - \delta = 0.995$ (Section C.1). In Figure 11, abstaining on $\sim 60\%$ of samples (coverage $= 40\%$) reduces the guaranteed FNR from 30% (at full coverage) to 14%.

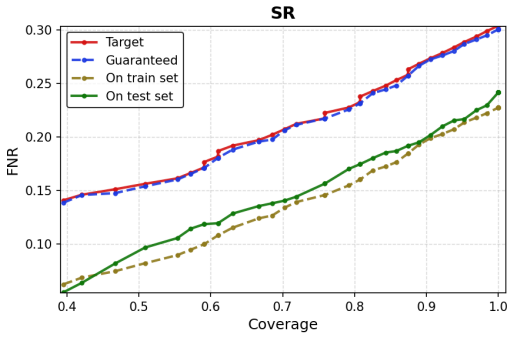


Figure 11: CIFAR-10 selective FNR bounds with SR. Computed on 50 targets $r^* \in [0, 0.3]$ via Algorithm 2. Each point uses the best pair (θ^*, ν^*) from Section C.1.

Joint control. We aim to jointly guarantee low FP risk (dominant in this setting) and low FNR (to limit missed detections). At full coverage, the baseline classifier has FP risk $= 15\%$ and FNR $= 30\%$ (Figure 12). As showed in Proposition A2, all joint FNR/FP risk trade-offs in Figure 12 hold with probability at least $1 - 2\delta = 0.99$.

For instance, censoring predictions with SR ≤ 0.89 gives a selective classifier $(f, g_{0.89})$ with guaranteed FP risk $\leq 4\%$ and FNR $\leq 13\%$.

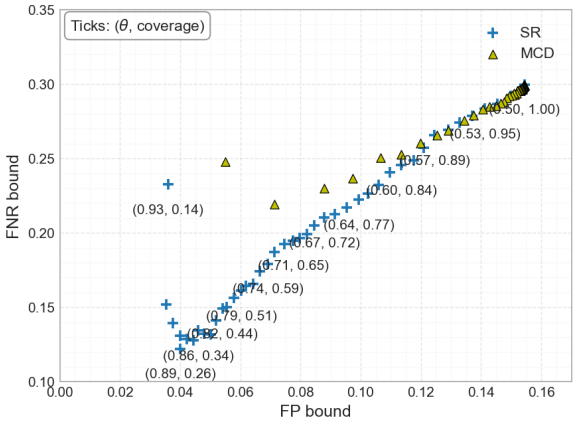


Figure 12: FNR–FP risk bounds trade-offs obtained over K_2 confidence thresholds θ (linear grid) for each confidence function (SR or MCD). Ticks indicate the corresponding $(\theta, \text{coverage})$ pairs. SR provides stronger guarantees. Dataset: CIFAR-10.

ResNet-18 results. Figure 13 shows the evolution of the selective metrics bounds and the corresponding test-set estimates as a function of coverage on the CIFAR-10 dataset with the trained ResNet-18 model (see Section D.2). The guaranteed performance gains obtained through abstention are generally very limited, and almost negligible for all metrics except FNR (and its counterpart SE) and PPV. For most other metrics, the bound–coverage curves remain relatively flat. Detailed results for each metric with ResNet-18 and SR confidence are provided in Figure 14. Overall, these findings indicate that although high-performing baseline models benefit less from abstention, our framework still provides valid and tight guarantees for all selective metrics.

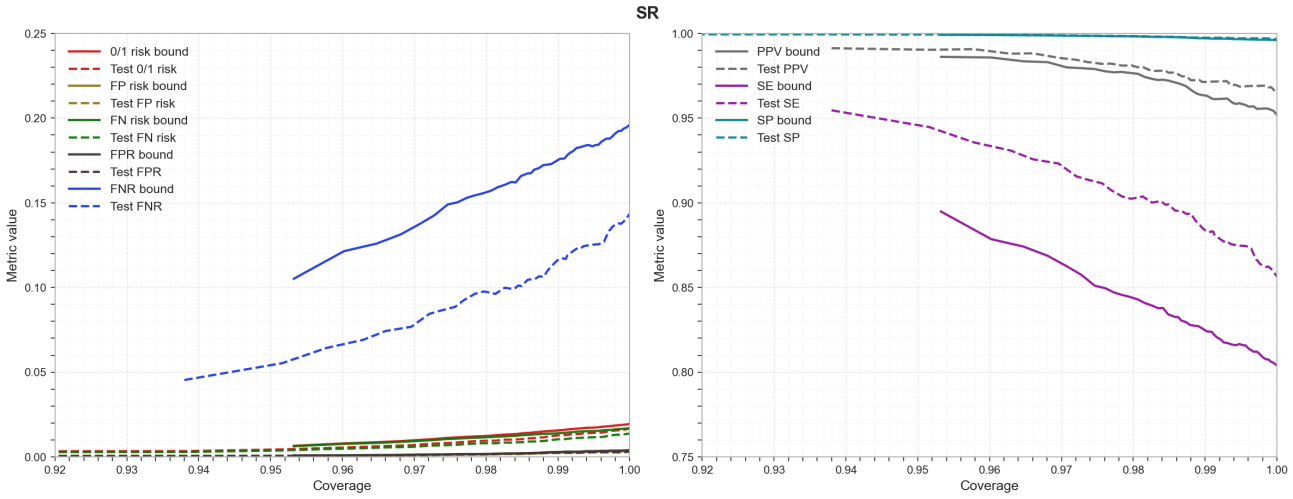
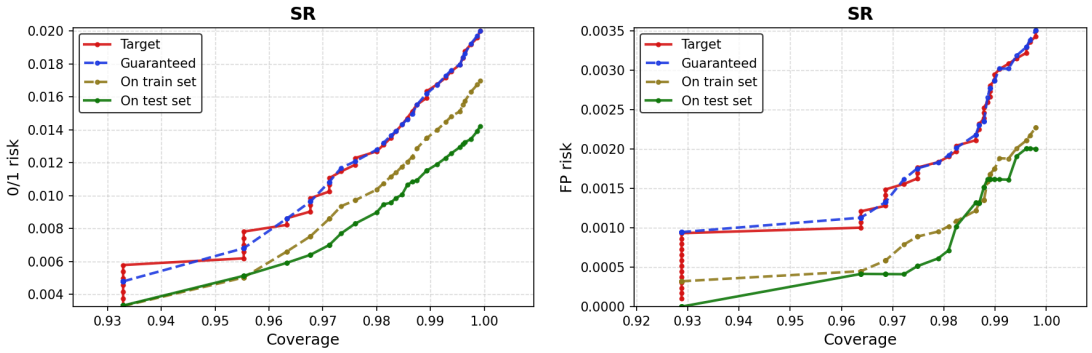


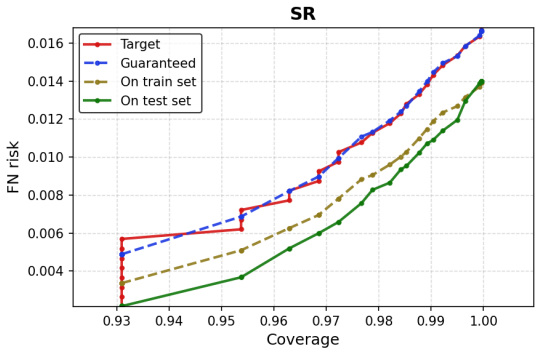
Figure 13: **CIFAR-10 evolution of selective metrics bounds and test-set estimates, with ResNet-18 model and SR confidence**. The left panel displays upper-bounded metrics on a common scale: misclassification risk, false-positive (FP) risk, false-negative (FN) risk, false positive rate (FPR), and false negative rate (FNR). The right panel shows lower-bounded metrics: positive predictive value (PPV), sensitivity (SE), and specificity (SP).

CNN additional results. Figures 15–16 show the evolution of selective metric bounds and the corresponding test-set estimates as a function of coverage on the CIFAR-10 dataset using the trained CNN model (see Section D.2). Each figure reports results for a different confidence function (SR or MCD; see Section D.3). In this

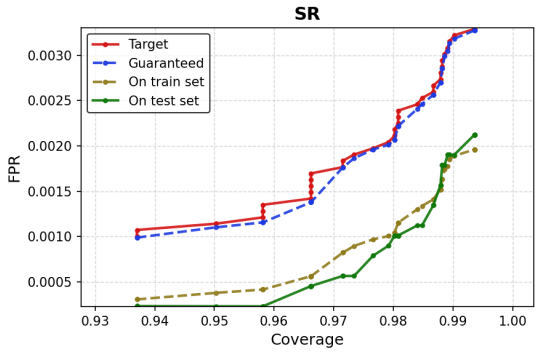


(a) Misclassification risk control

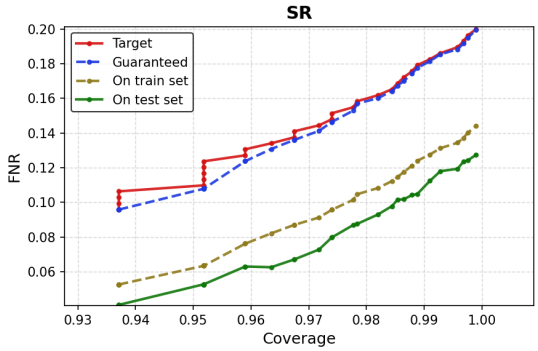
(b) FP risk control



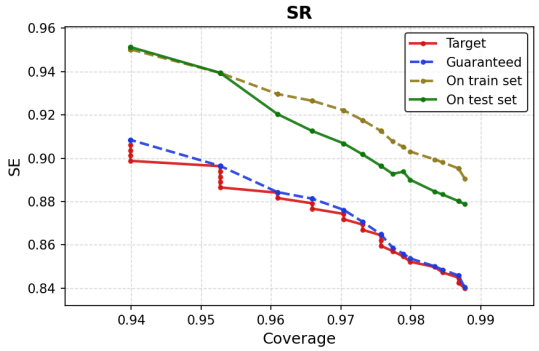
(c) FN risk control



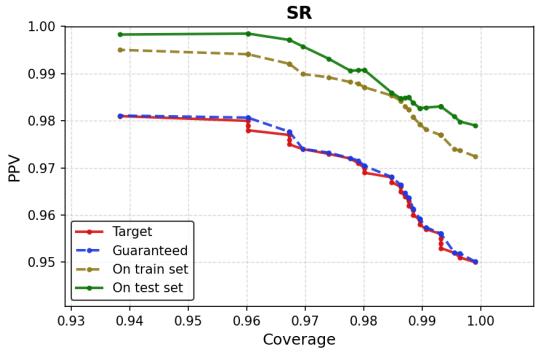
(d) FPR control



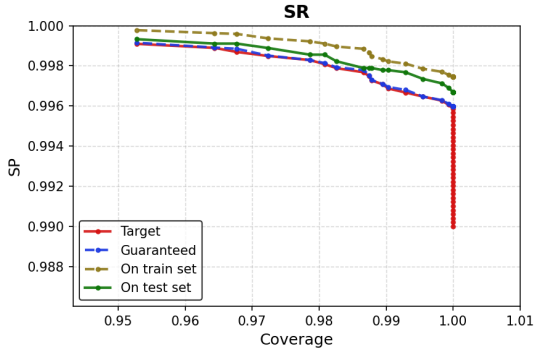
(e) FNR control



(f) SE control



(g) PPV control



(h) SP control

Figure 14: **CIFAR-10 evolution of selective metrics bounds** obtained from Algorithms 1–2 (computed over 50 target values r^*, sampled on a linear grid defined by the metric range), along with the corresponding train- and test-set metric estimates, as a function of coverage. **Model: ResNet-18. Confidence function: SR.**

setting, the guaranteed performance gains obtained through abstention are substantial. Detailed results for each metric with the CNN model are provided in Figure 17.

For example, Figure 17-(h) shows that abstention can yield a guaranteed specificity (SP) above 96% at approximately 20% coverage, whereas the baseline SP was around 83% at 100% coverage. Similar improvements are observed for other metrics. Notably, across the CIFAR-10 experiments, SR produces tighter guarantees than MCD.

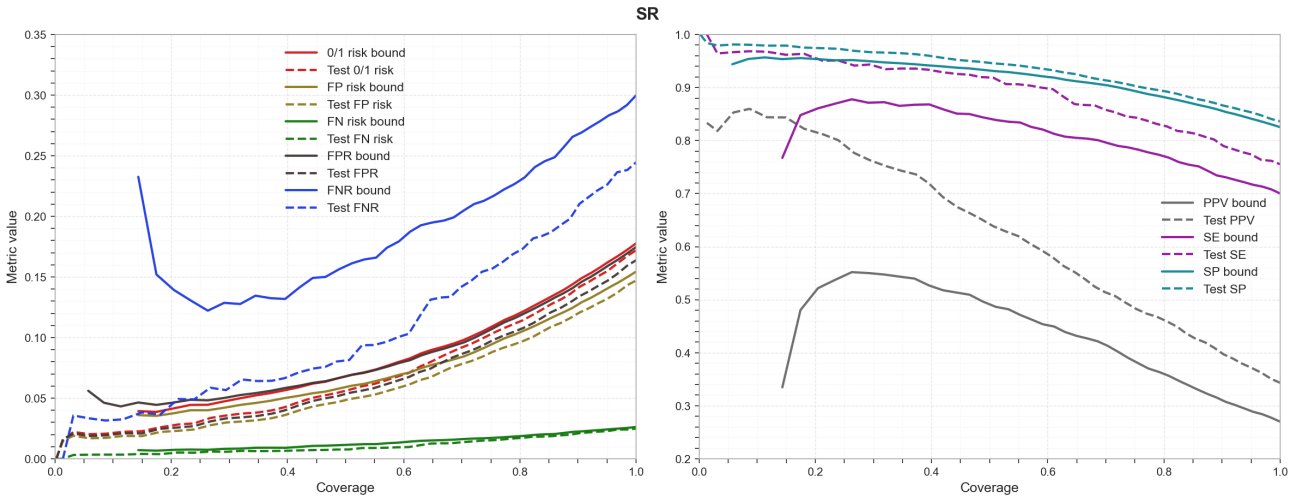


Figure 15: **CIFAR-10 evolution of selective metrics bounds and test-set estimates, with CNN model and SR confidence**. The left panel displays upper-bounded metrics on a common scale: misclassification risk, false-positive (FP) risk, false-negative (FN) risk, false positive rate (FPR), and false negative rate (FNR). The right panel shows lower-bounded metrics: positive predictive value (PPV), sensitivity (SE), and specificity (SP).

E.2 Colorectal cancer detection—additional results

Figures 18–19 show the evolution of selective metric bounds and the corresponding test-set estimates as a function of coverage on the H&E images dataset using the trained CNN model (see Section D.2). Each figure reports results for a different confidence function (SR or MCD; see Section D.3). In this setting, the guaranteed performance gains obtained through abstention are substantial. Detailed results for each metric with the CNN model are provided in Figure 20. For example, Figure 20-(b) shows that abstention can yield a guaranteed FP risk below 1% at approximately 35% coverage, whereas the baseline FP risk was around 8% at 100% coverage. Similar improvements are observed for other metrics. Notably, across the H&E experiments, SR generally produces tighter guarantees than MCD.

E.3 Impact of δ, sample size proportion N, and class imbalance on bound tightness

This section examines how the sample size proportion n, the probability parameter δ, and class imbalance influence the tightness of the proposed selective metric bounds. Tightness is quantified using the *Average Distance Between bounds and test Curves* (ADBC), defined as the mean gap between the guaranteed bounds and the corresponding test-set curves (i.e., the average distance between continuous and dashed lines in Figure 3). ADBC is computed over 30 uniformly spaced confidence thresholds θ spanning the observed range of confidence values in each dataset. For each configuration of n, δ, or class imbalance, we perform 50 independent random splits (with fixed seeds) into bounds-fitting and test sets, and report the mean ADBC along with 95% confidence intervals.

In Figures 21–23, ADBC is plotted as a function of δ (left column), N (middle column), and class imbalance, expressed as the proportion of positive samples (right column). Each row of plots corresponds to a distinct selective metric: the first row shows the standard misclassification risk, the second the false-positive (FP) risk, and so on. Orange curves represent bounds obtained using Softmax Response (SR) confidence, and blue curves

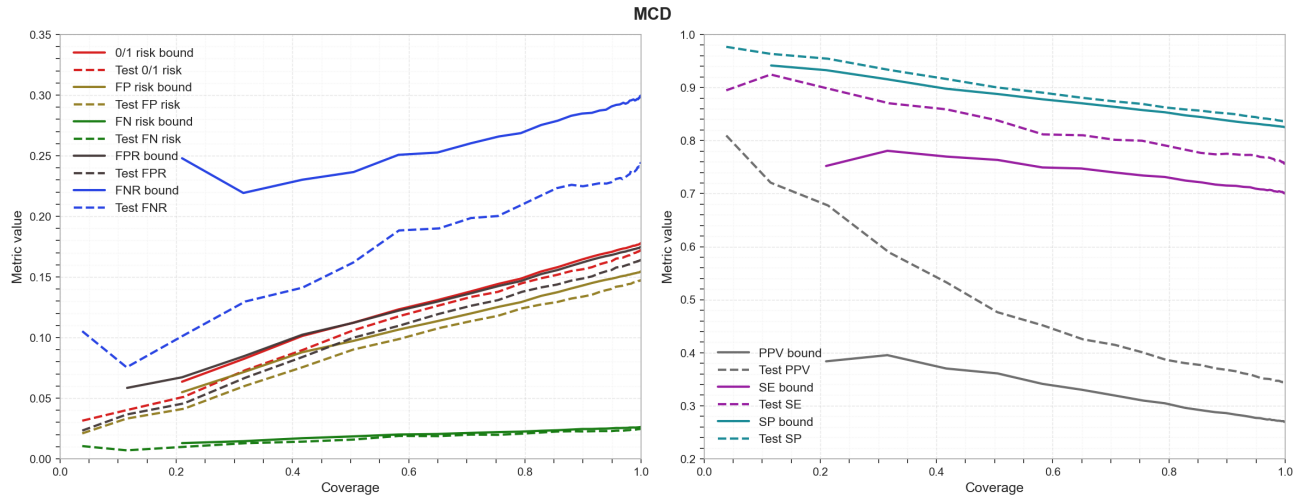


Figure 16: **CIFAR-10 evolution of selective metrics bounds and test-set estimates, with CNN model and MCD confidence**. The left panel displays upper-bounded metrics on a common scale: misclassification risk, false-positive (FP) risk, false-negative (FN) risk, false positive rate (FPR), and false negative rate (FNR). The right panel shows lower-bounded metrics: positive predictive value (PPV), sensitivity (SE), and specificity (SP).

those obtained using Monte Carlo Dropout (MCD). Confidence intervals were estimated over 50 random seeds, each corresponding to an independent bounds fitting/testing split.

Figure 21 reports results for **CIFAR-10** with the **CNN** model, Figure 22 for **CIFAR-10** with the **ResNet-18**, and Figure 23 for the **H&E histology images** dataset with the **CNN**.

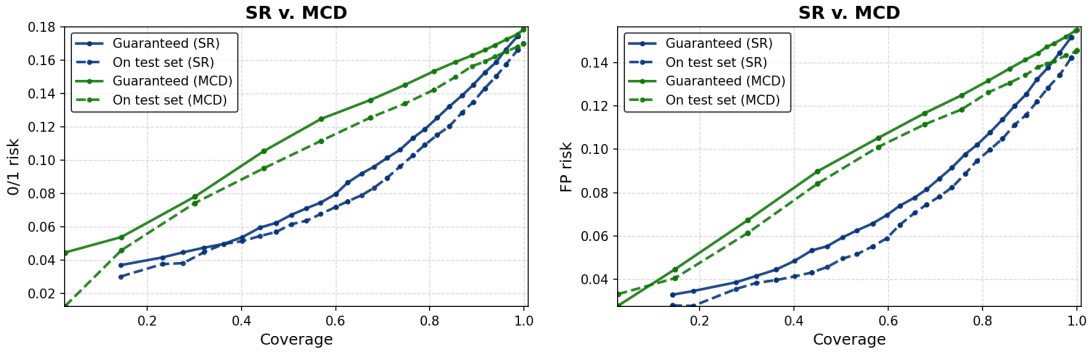
Interpretation of the results. For both CIFAR-10 and H&E datasets, and for both SR and MCD confidence functions, increasing either δ or the sample size N leads to tighter bounds. The effect of class imbalance on the tightness of misclassification-risk bounds is less straightforward, as this risk is inherently asymmetric. No significant effect can be inferred for FN and FP risks due to the large variability of their confidence intervals. As expected, greater imbalance (a higher proportion of positive samples) empirically tightens the FNR bounds—dependent on true positives, which become more numerous as imbalance increases—while loosening the FPR bounds, which rely on the decreasing negative class.

E.4 Comparison with Hoeffding-derived upper bounds

We finally compare the tightness of the bounds proposed in this work with general-purpose methods for conditional control of binomial proportions, such as Equation (11) in Balsubramani et al. [2019], which rely on Hoeffding-type concentration inequalities. More precisely, we compare our bounds for the selective FPR and FNR with the corresponding upper bounds derived from Hoeffding's inequality as given in Equation (11) of Balsubramani et al. [2019]. In particular, for the FPR, Equation (11) yields:

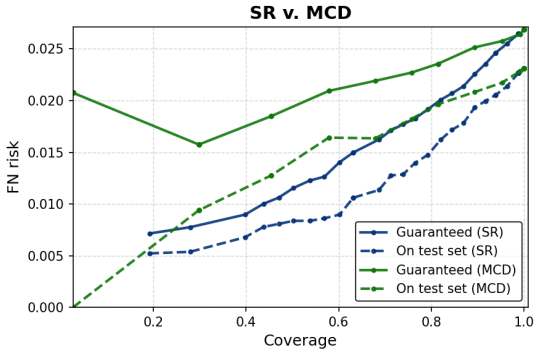
$$\mathbb{P}\left(\mathrm{FPR}(f, g_\theta) \leq \frac{\sum_{i=1}^{n} f(x_i)\, g_\theta(x_i)\, (1-y_i)}{\sum_{i=1}^{n} (1-y_i)\, g_\theta(x_i)} + \sqrt{\frac{2\log(1/\delta)}{\sum_{i=1}^{n} (1-y_i)\, g_\theta(x_i)}} \right) \geq 1-\delta.$$

We evaluated our FPR bounds—guaranteed with probability at least $1-\delta$—against these bounds, and ours consistently appear to be significantly tighter across all thresholds. For example, on the CIFAR/CNN/SR setting (see Figure 24), our lowest guaranteed FPR bound reaches 3.4% at $\theta = 0.93$, whereas the lowest guaranteed bound from Equation (11) is 6.5% at $\theta = 0.90$. In Figure 24 we observe similarly improved results for other selective metrics.

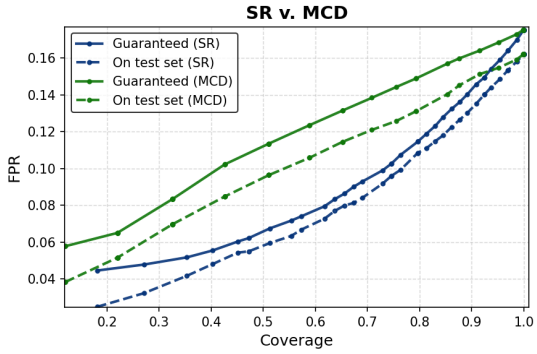


(a) Misclassification risk control

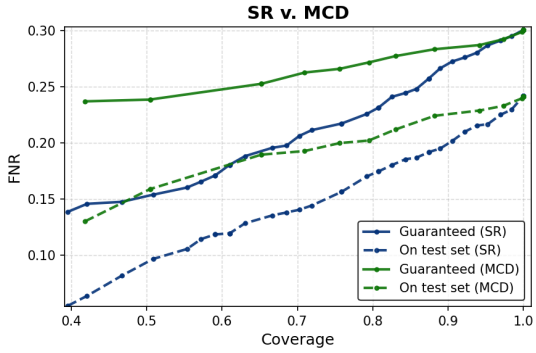
(b) FP risk control



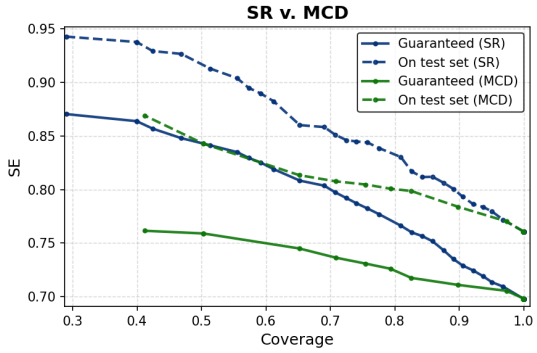
(c) FN risk control



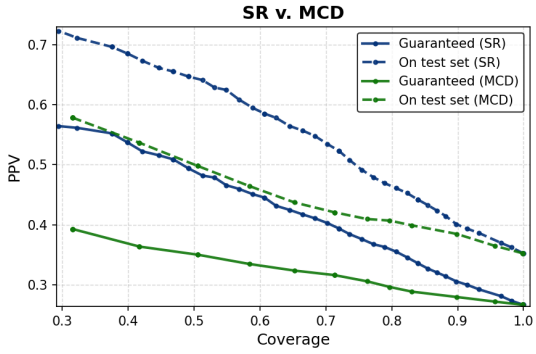
(d) FPR control



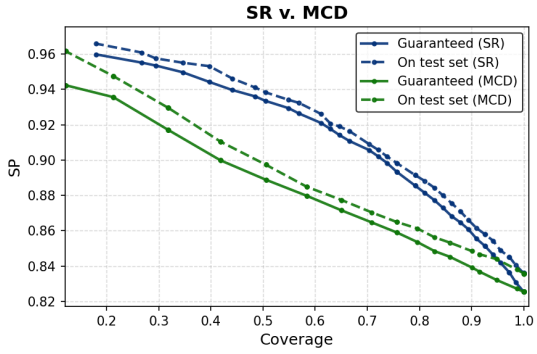
(e) FNR control



(f) SE control



(g) PPV control



(h) SP control

Figure 17: **CIFAR-10 evolution of selective metrics bounds** obtained from Algorithms 1–2 (computed over 50 target values r^*, sampled on a linear grid defined by the metric range), along with the corresponding and test-set metric estimates, as a function of coverage. **Model: CNN. Confidence function: SR and MCD.**

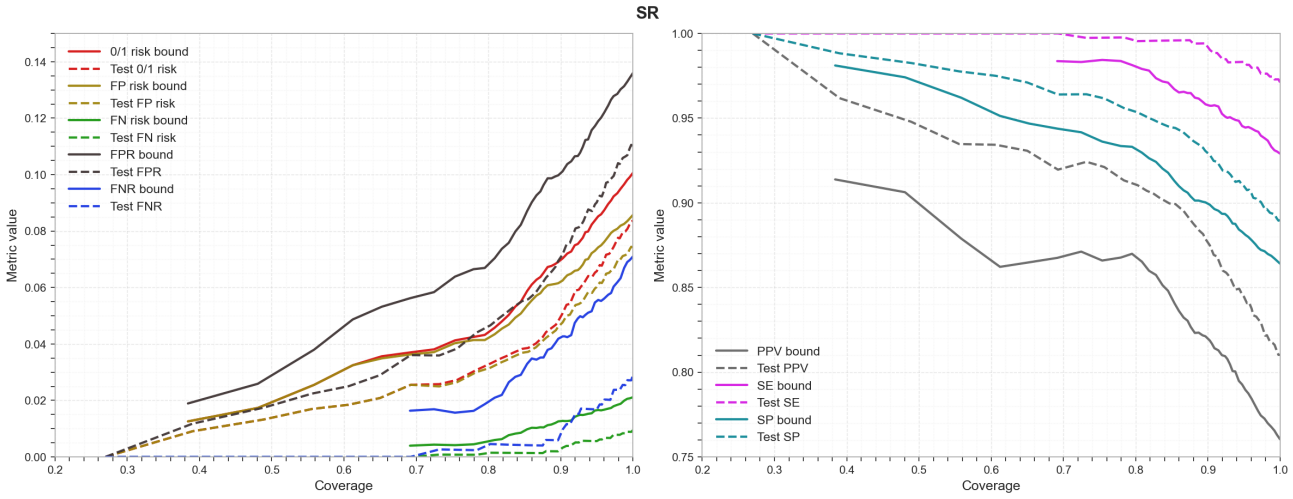


Figure 18: **H&E evolution of selective metrics bounds and test-set estimates, with CNN model and SR confidence**. The left panel displays upper-bounded metrics on a common scale: misclassification risk, false-positive (FP) risk, false-negative (FN) risk, false positive rate (FPR), and false negative rate (FNR). The right panel shows lower-bounded metrics: positive predictive value (PPV), sensitivity (SE), and specificity (SP).

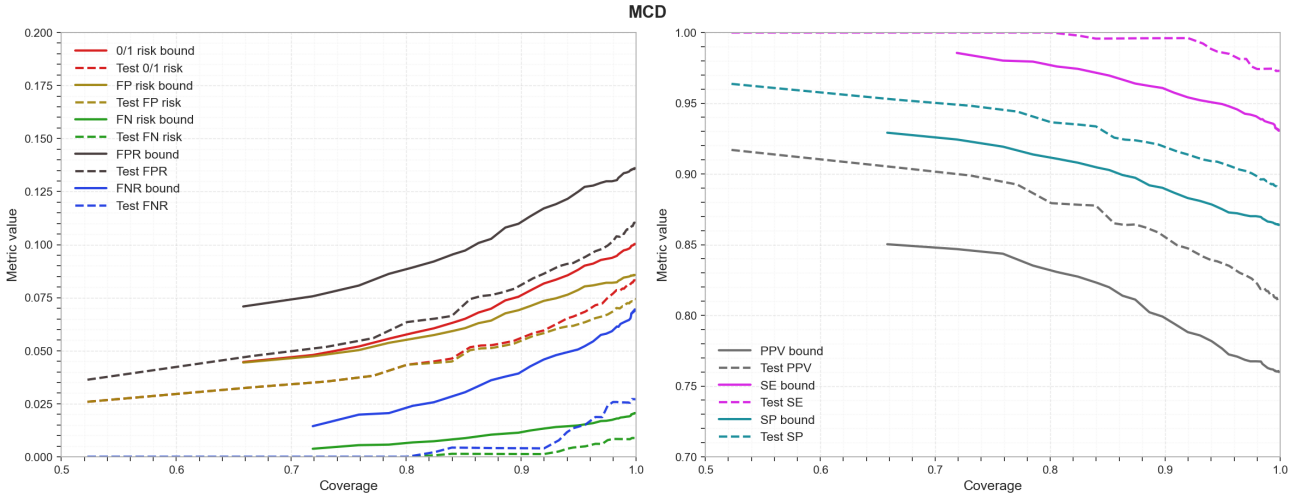
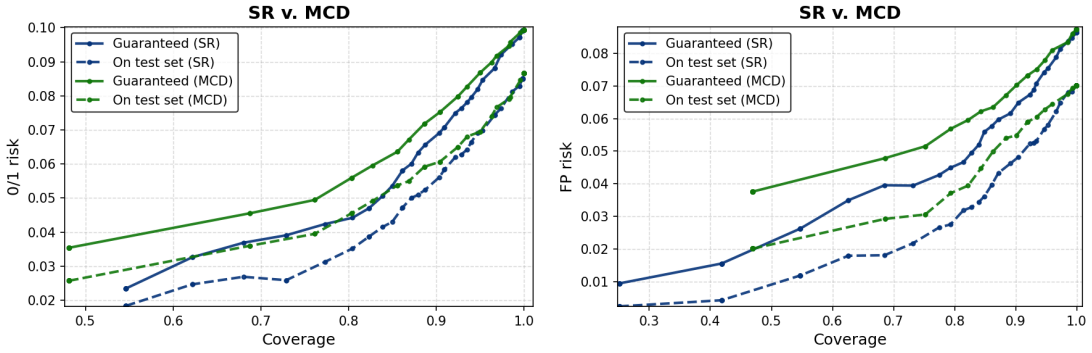
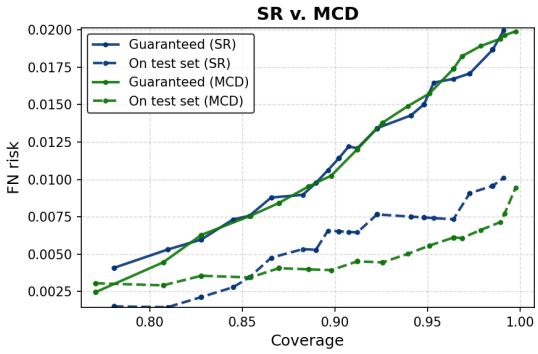


Figure 19: **H&E evolution of selective metrics bounds and test-set estimates, with CNN model and MCD confidence**. The left panel displays upper-bounded metrics on a common scale: misclassification risk, false-positive (FP) risk, false-negative (FN) risk, false positive rate (FPR), and false negative rate (FNR). The right panel shows lower-bounded metrics: positive predictive value (PPV), sensitivity (SE), and specificity (SP).

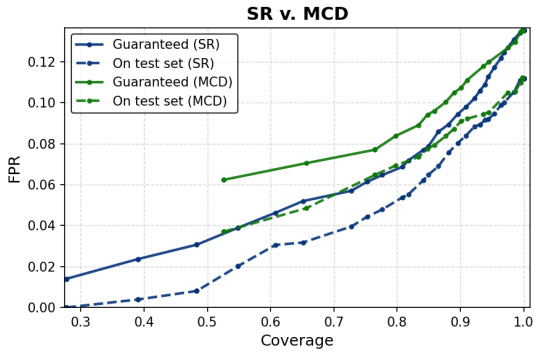


(a) Misclassification risk control

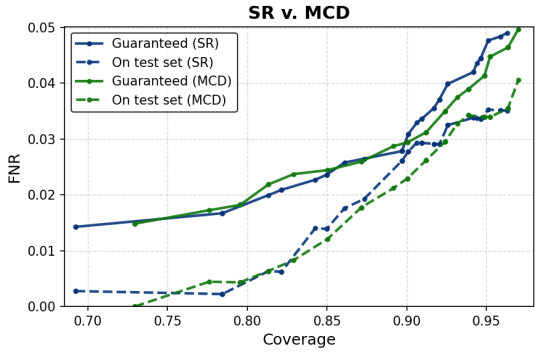
(b) FP risk control



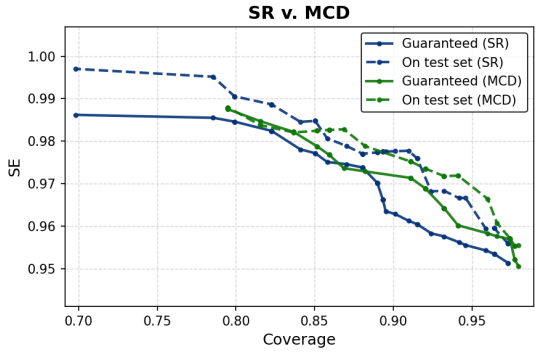
(c) FN risk control



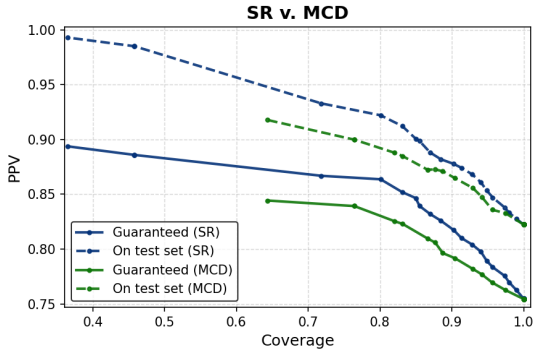
(d) FPR control



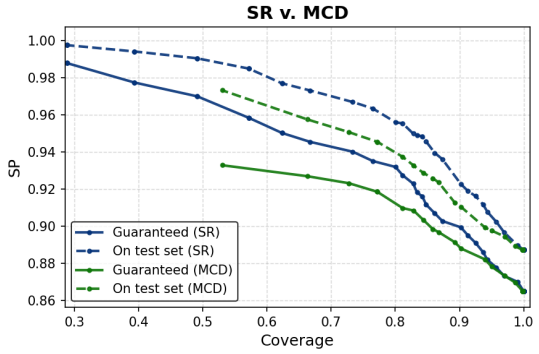
(e) FNR control



(f) SE control



(g) PPV control



(h) SP control

Figure 20: **H&E evolution of selective metrics bounds** obtained from Algorithms 1–2 (computed over 50 target values r^*, sampled on a linear grid defined by the metric range), along with the corresponding and test-set metric estimates, as a function of coverage. **Model: CNN. Confidence function: SR and MCD.**

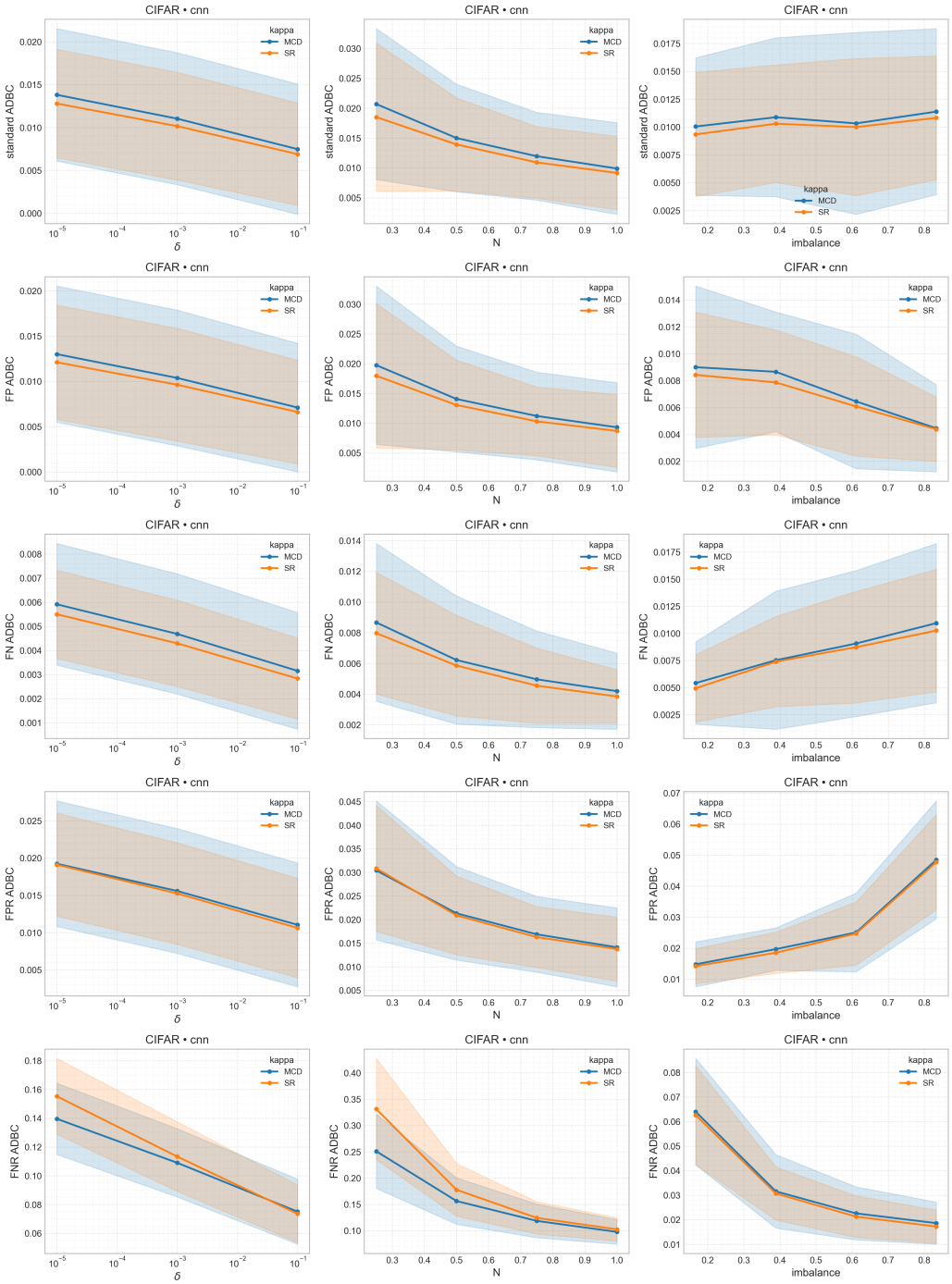


Figure 21: Effect of δ, N, and class imbalance on bound tightness (Average Distance Between Curves; ADBC). **Dataset: CIFAR-10. Model: CNN.**

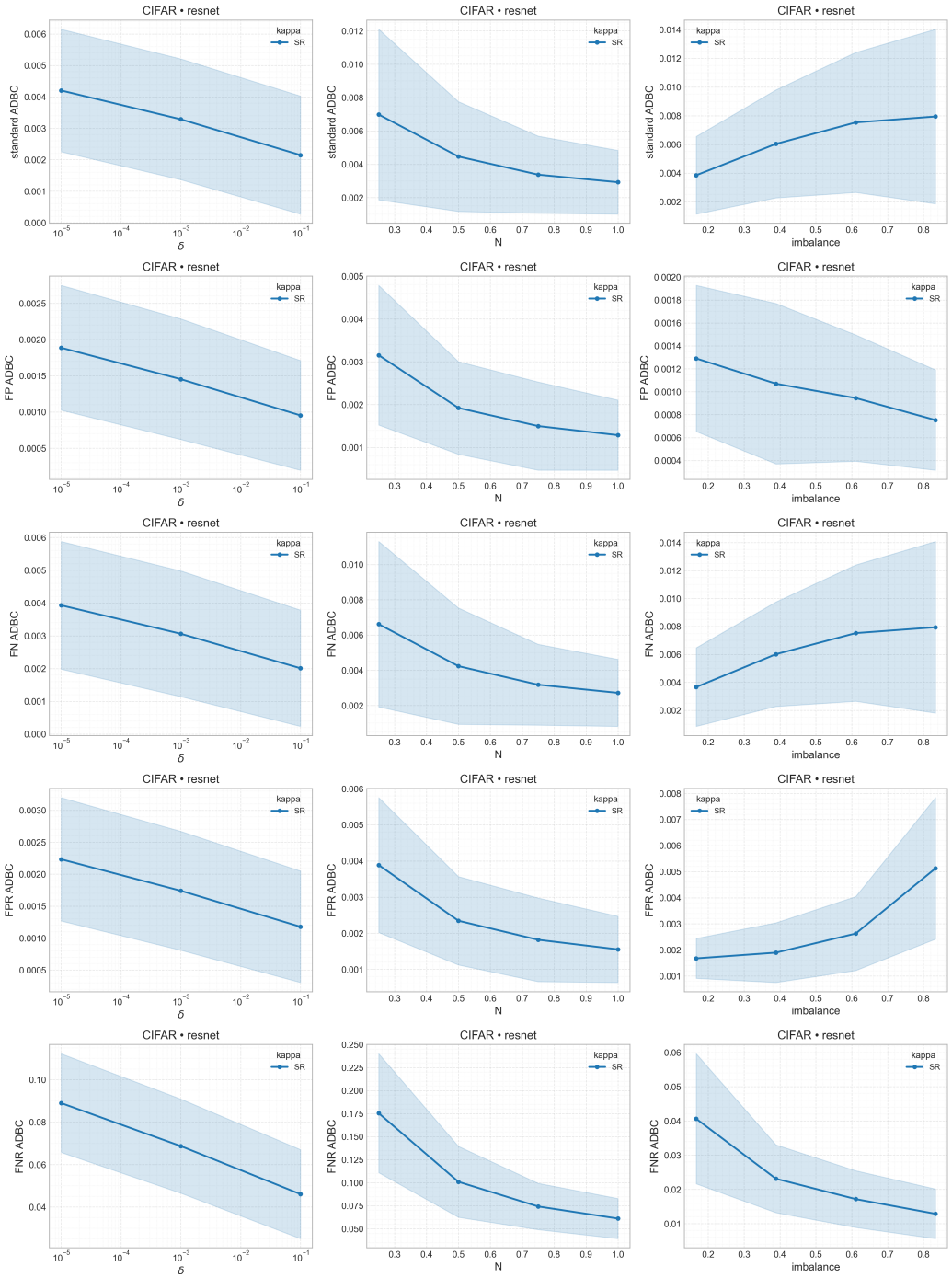


Figure 22: Effect of δ, N, and class imbalance on bound tightness (Average Distance Between Curves; ADBC). **Dataset: CIFAR-10. Model: ResNet-18.**

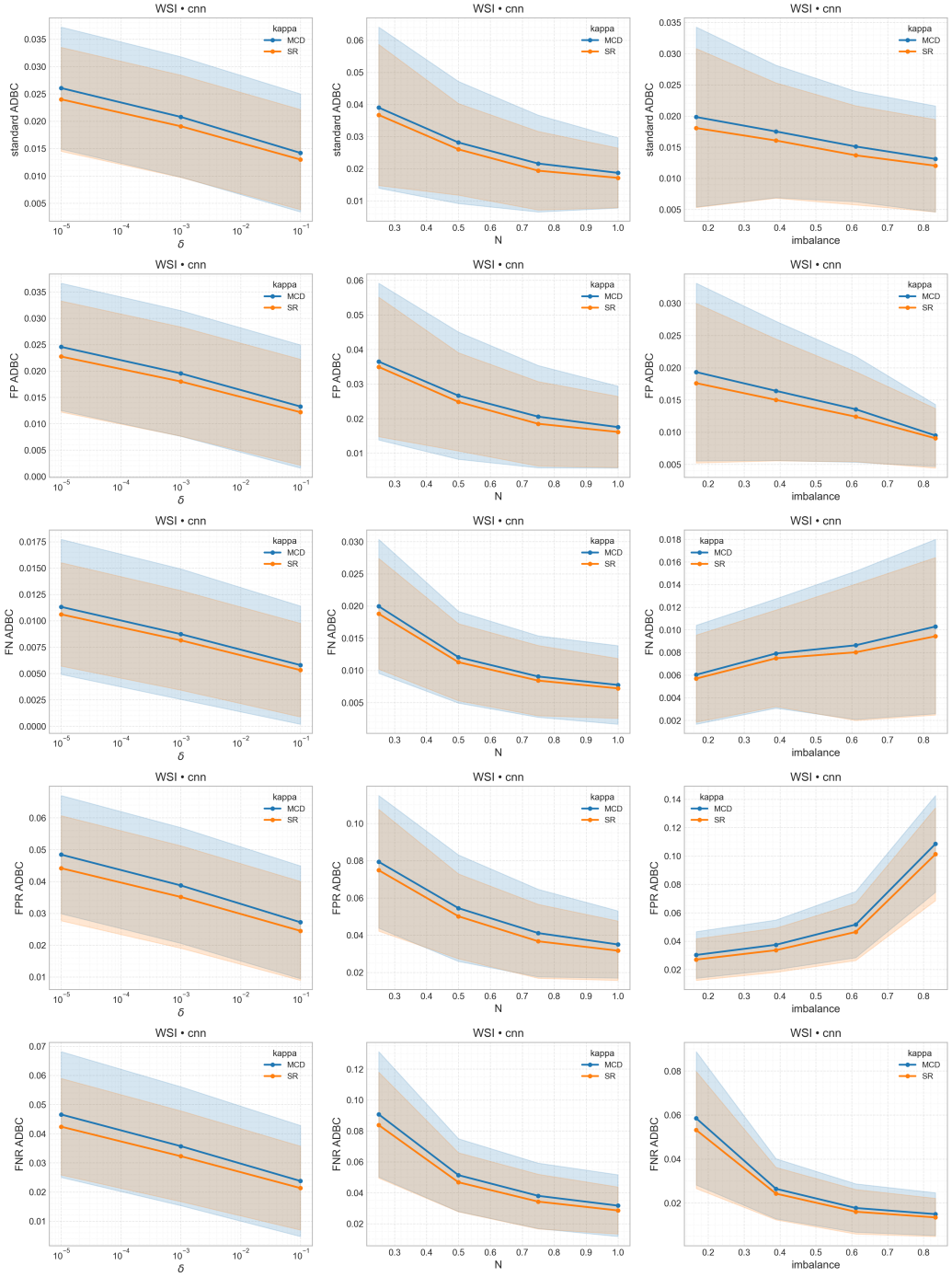


Figure 23: Effect of δ, N, and class imbalance on bound tightness (Average Distance Between Curves; ADBC). **Dataset: H&E images. Model: CNN.**

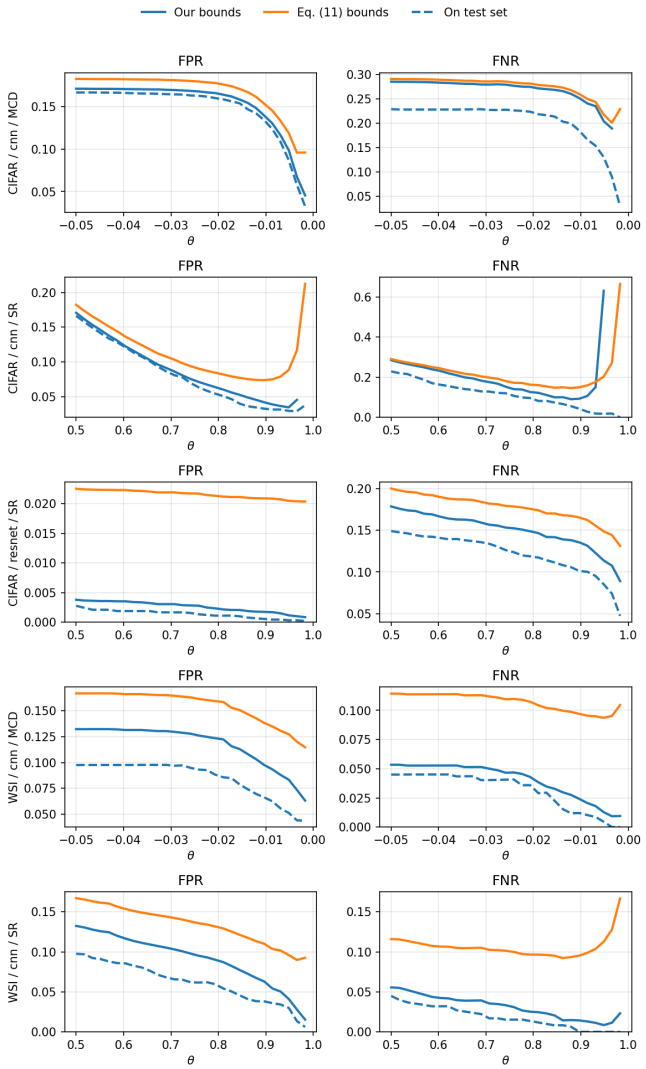


Figure 24: Comparison of our bounds and the bounds from Equation (11) in [Balsubramani et al., 2019]. Left column plots is FPR, right column is FNR. Each row of plots corresponds to a dataset/model/confidence combination.