On Arbitrary Predictions from Equally Models

Anonymous authors
Paper under double-blind review

Abstract

Model multiplicity describes the existence of multiple models that fit the data equally well but can produce different predictions on individual samples, so-called predictive multiplicity. In medicine, these models can admit conflicting predictions for the same patient—a risk that is poorly understood and insufficiently addressed. In this study, we empirically analyze predictive multiplicity across multiple medical tasks and model architectures and show practical strategies to mitigate it. Our analysis reveals that (1) standard validation metrics fail to identify a uniquely optimal model. (2) Models with equivalent population-level performance yield conflicting patient-level predictions, resulting in arbitrary and potentially harmful outcomes under any single model. (3) For already overparameterized models, higher capacity reduces predictive multiplicity when it improves accuracy. Disagreement between models is not at random and (4) can be leveraged as an indicator of genuine uncertainty, identifying patients for whom the current evidence is insufficient. Together, our findings highlight that predictive multiplicity is not a theoretical curiosity but a pervasive and practically significant issue in medical AI. We argue that accounting for multiplicity should be considered a core component of model evaluation and deployment in safety-critical domains.

1 Introduction and Prior Work

Model multiplicity (Black et al., 2022b) describes the existence of many *plausible* models for the same dataset without a principled way to determine a single optimal model, reflecting the underlying uncertainty in how the data should be interpreted.[1] In practice, multiple machine learning models can fit the same data equally well according to a performance metric (e.g., loss or accuracy), but may differ in their internal structure (e.g., the value of their parameters) and, more critically, in their individual predictions. Yet, for deployment, typically a *single* model is chosen—commonly, without consideration for other, equally valid options. Using such an arbitrary[2] model is particularly problematic in high-stakes scenarios when other, equally well-performing models exist that produce different predictions on the same data point(s). If such a model is deployed in a clinical setting, *a patient's diagnosis—and their treatment—may ultimately depend on the choice of this specific model rather than on relevant properties of the patient's data.* This raises critical concerns about the justification for deploying this model. We illustrate the problem in Figure 1.

The phenomenon of model multiplicity is not novel and has appeared under various names, *inter alia*, the Rashomon Effect (Breiman, 2001), underspecification (D'Amour et al., 2022) or instability (Riley & Collins, 2023). Prior work has discussed its opportunities (Rudin et al., 2024) and challenges, notably *predictive multiplicity* (Marx et al., 2020), where equally valid models produce conflicting predictions. While predictive multiplicity may be negligible in low-stakes settings or beneficial in some contexts (e.g., to avoid systemic exclusion; Creel & Hellman 2022), it poses serious risks in medicine, where model predictions inform counseling, resource allocation, and clinical care. Conflicting predictions from equally valid models can erode

[1] In theory, the Bayes optimal predictor represents the optimal solution, i.e., the model that minimizes the expected loss with full knowledge of the data-generating distribution. However, this ideal predictor is not attainable in practice, as the true underlying distribution is unknown, and only finite data is available.

[2] In contrast, selecting the/ a single best model would be non-arbitrary from the perspective of instrumental rationality: the model that performs as well as it can according to its own metric of success (Citron & Pasquale, 2014). However, model multiplicity describes the existence of multiple equally optimal models that produce in different predictions. Thus, there is no self-interested reason to choose one optimal model over another (Creel & Hellman, 2022).

trust, cause inconsistent treatment, and ultimately harm patients. Despite its relevance, systematic studies of predictive multiplicity in medicine remain scarce. Existing work identifies underspecificaton as a general source of instability (D'Amour et al., 2022) without addressing its impact on individual predictions, or proposes bootstrapping approaches (Riley & Collins, 2023; Riley et al., 2023) that are largely infeasible for modern machine learning algorithms; see Appendix A.1 for details.

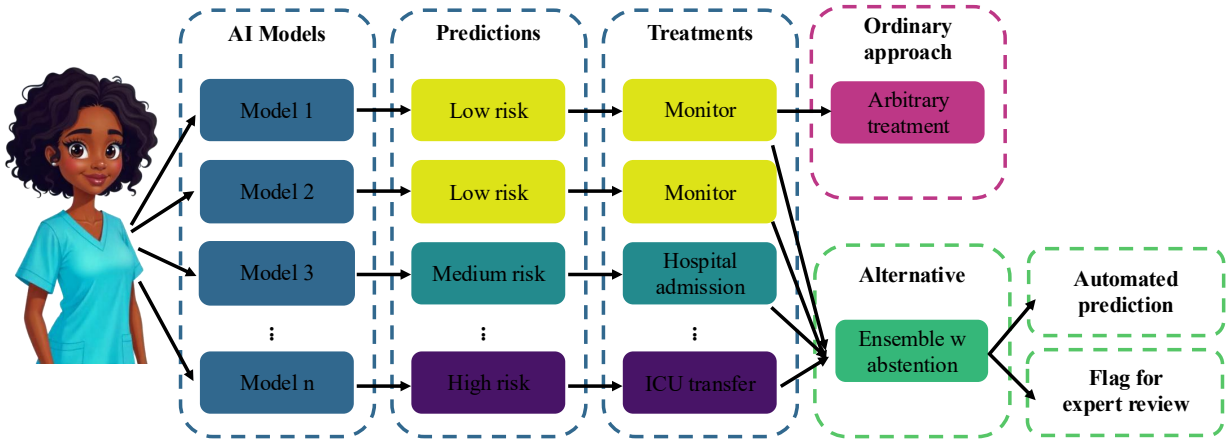

Figure 1: **Model multiplicity in clinical decision-making.** Many machine learning models achieve similar overall performance and are thus equally valid choices (model 1 to n) for deployment. Yet, for the same patient, these models may produce different risk predictions (low, medium, high), leading to divergent treatment recommendations (monitor, hospital admission, ICU transfer). A patient's care pathway may thus depend on which model happens to be deployed and result in an arbitrary treatment. Rather than relying on a *single* model, we propose an ensemble-based strategy: cases with high inter-model agreement may be suitable for automated prediction, whereas cases with low agreement should be routed to clinical experts.

In this study, we address this gap through a comprehensive empirical analysis of predictive multiplicity across five medical imaging modalities (abdominal CTs, blood cell images, breast ultrasounds, OCT scans, and X-rays) and four model architectures (ResNet50, GC ViT, EfficientNet, and ConvNeXt). Recognizing and addressing model multiplicity not only exposes limitations of the "single model" paradigm but also offers a path to improve predictive stability and accuracy by levering inter-model (dis)agreement. Our large-scale study, based on the training of 1,500 models, leads to the following conclusions that form the core contributions of our work:

1. **Validation performance is an unreliable basis for selecting a single model.** Choosing the seemingly best-performing model becomes an arbitrary choice.
2. **Single-model deployment exposes some patients to arbitrary predictions.** However, the structure of disagreement can be exploited.
3. **Increasing model capacity reduces predictive multiplicity when it improves accuracy.** Yet the burden of inconsistent predictions is redistributed, requiring monitoring of who is affected.
4. **Ensembles with abstention offer a principled approach to predictive multiplicity.** High consensus predictions may be automated, low-consensus ones prompt deferral to human experts.

2 Methodology

Deployment typically relies on a single, "optimal" model, without considering other, equally valid alternatives. The set of alternative models, that fit the data equally, are commonly referred to as the *Rashomon set* (Breiman, 2001). Predictive multiplicity arises from disagreement *within* this set, thus the characterization of the Rashomon set is central to our question. In practice, characterizing the *true* Rashomon set

remains an open challenge. First, model quality can only be assessed on finite validation and test samples. Second, for large models an exhaustive enumeration is infeasible—for neural networks, for example, the number of local minima grows exponentially with the number of parameters (Auer et al., 1995). Following prior work (D'Amour et al., 2022; Black et al., 2022a), we therefore explore the Rashomon set *empirically* by randomizing the initialization of the model weights. Specifically, we replace the classification head of an ImageNet-pretrained model with a randomly initialized one and train the entire model. For each dataset/architecture pair we train 50 such instances, resulting in a total of 1,500 models—1,000 for our main experiments and 500 to analyze model capacity. In brief, our experiments cover the aforementioned medical imaging datasets and model architectures. The trained models perform competitively with, and often surpass established benchmarks; see Appendix A.2 for details.

Prior work often designated an *ad hoc* tolerances, for example a fixed 1% deviation in loss or accuracy to determine membership in the empirical Rashomon set. In contrast, we determine membership using an interval-overlap criterion based on Clopper-Pearson (CP) binomial confidence intervals (Clopper & Pearson, 1934) following Paes et al. (2023). We used the best-performing model of the 50 models as the reference model and consider any model whose 95% confidence interval overlaps with that of the reference model as statistically indistinguishable in accuracy and thus a member of the empirical Rashomon set. The Rashomon parameter ϵ is then defined as the boundary of this indistinguishability region—the largest decrease in accuracy at which the model's confidence interval still overlaps with that of the reference model. The width of the indistinguishability region depends on the significance level (α=0.05), the accuracy of the best empirical model, and the number of validation and test samples; see Appendix A.3 for details.

To quantify per-sample prediction stability, we define Adjusted Pairwise Prediction Agreement (APPA): the chance-corrected probability that two models drawn uniformly at random (without replacement) from the empirical Rashomon set produce the same prediction for a sample.. The raw agreement is normalized by the number of models and classes to discount agreement that would occur by chance even between unrelated classifiers (see Appendix A.4 for the full derivation). Note that if class prevalence were available, normalizing by the expected agreement under the empirical class distribution may provide a more appropriate correction than uniform chance. By definition, values of APPA lie between 0 and 1: a value of 1 indicates that all model pairs agree (maximal stability), a value of 0 indicates agreement no better than chance, while intermediate values reflect partial agreement.

Predictive multiplicity arises from several sources, only some of which can be resolved through technical means or additional data; others persist regardless.[3] We therefore reframe predictive multiplicity as a governance problem. Rather than seeking to eliminate disagreement among equally valid models, we treat it as a signal of genuine uncertainty—an indication that the available data are insufficient to support a prediction in a high-stakes scenario. Concretely, for each sample we measure the agreement across a subset of models from the empirical Rashomon set and compare it against a predefined threshold. Samples above the threshold are candidates for automatic prediction; those below are deferred to expert review. The threshold can be adjusted: lowering it may reduce reliability (potentially higher predictive multiplicity), raising it may increase human effort (by referring more cases to clinicians).

3 Results

3.1 The illusion of an uniquely optimal solution

We find that models that differ solely in the initial values of their final-layer weights prior to training show considerable variation in performance on finite validation and test sets, with accuracy differences up to 16%. In Figure 2, we visualize an *indistinguishability region* (●) from the reference accuracy to the largest decrease

[3]Parikh (2026) groups the causes underlying the Rashomon Effect into three categories. *Statistical* sources arise from finite samples and noise in the data-generating process, and diminish as more data becomes available. *Structural* sources from non-convexity of optimization objectives and unobservable variables, and cannot be resolved. *Procedural* sources from limitations of optimization algorithms and deliberate restrictions to suboptimal model classes; these are resolvable in principle—for instance, by increasing model capacity, as our experiments demonstrate on certain datasets.

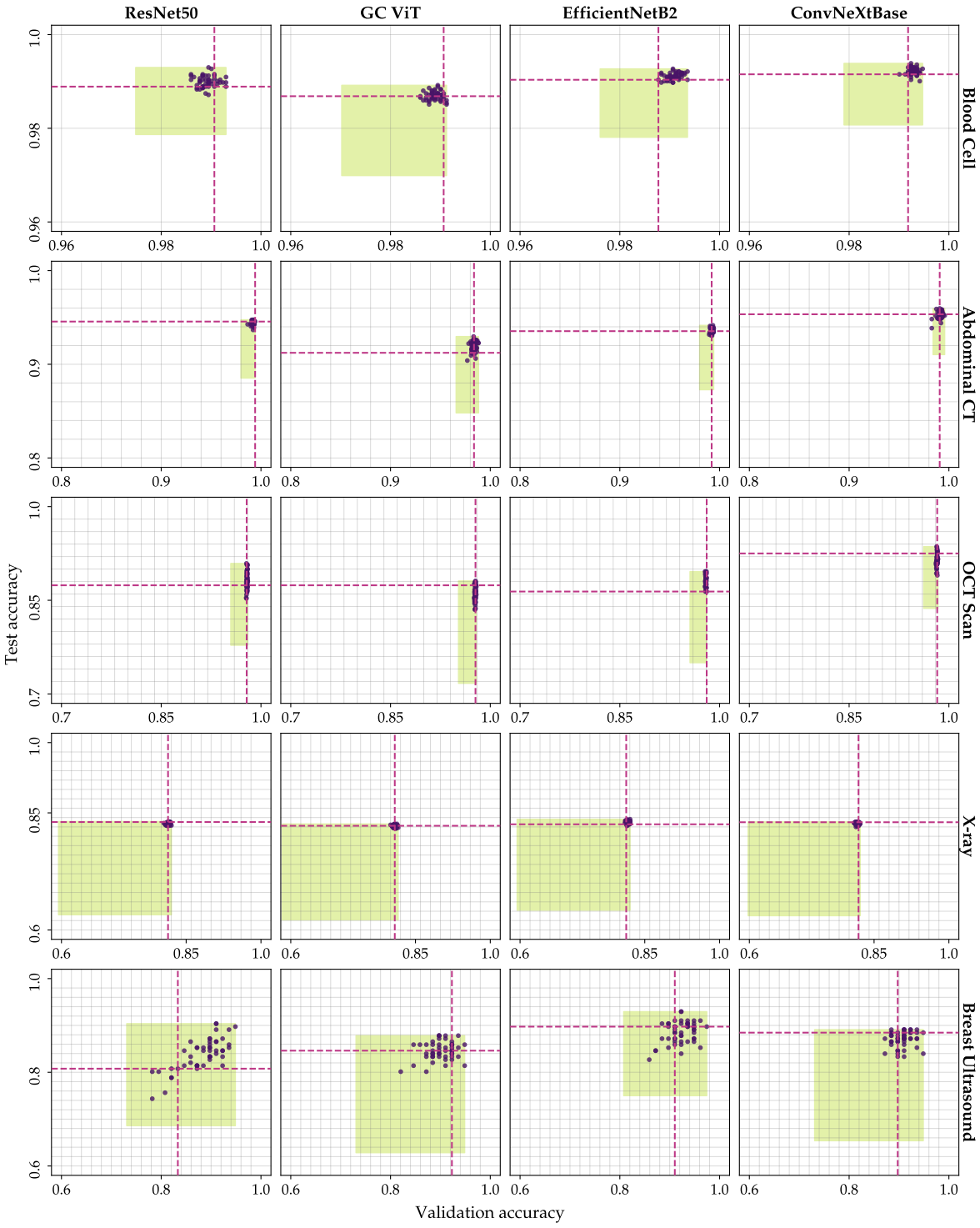

Figure 2: **Variation in accuracy within the empirical Rashomon set.** Each plot shows a dataset/architecture combination with points (•) representing models trained from different random initialization. Models within the green region (•) are indistinguishable at the 95% significance level. Only two models (Abdominal CT/ConvNeXt) exhibit a statistically significant difference, an effect that would likely remain undetected without formal statistical testing. The dashed line (•) marks the model for which we performed a hyperparameter search. Axes are scaled uniformly within each dataset (0.02 units per grid cell).

in accuracy (for α=0.05) Nearly all models fall within this region and are therefore statistically indistinguishable at the 95% significance level; the only exceptions are two models in the abdominal CT/ConvNeXt setting. In other words, while variation in validation and test accuracies may suggest varying quality, *formal statistical testing provides no evidence that their performance on the underlying true distribution is actually different.* Moreover, the two models that *are* of statistically inferior quality are not obvious outliers in the empirical Rashomon set and would likely have gone unnoticed without formal testing. Overall, within the empirical Rashomon set, validation performance poorly predicts test performance: models that perform well on the validation set often underperform on the test set, and *vice versa*. Notably, the initial model (dashed line, •) for which we performed the hyperparameter search is in no way "special" regarding validation or test set performance relative to the other models. According to validation and test accuracy, better and worse models exist despite this being the model that has been explicitly optimized for.

In summary, these findings imply that validation and test performance cannot identify a uniquely "optimal" model. Crucially, even the initial model for which we perform a hyperparameter search is as good a draw as any other model from the set. Selecting the model with the highest validation performance becomes an arbitrary decision among other equally valid alternatives. Consequently, the standard selection criteria—choosing the model with the highest validation performance—is not only inadequate but potentially harmful to the patients receiving inferior predictions.

3.2 Individual predictions are arbitrary under any single model

So far we have focused on aggregate performance metrics, which obscure outcomes for individual patients. We thus turn to per-sample predictions to quantify how much a patient's diagnosis depends on a specific model, using APPA as stability measure. As depicted by the pink points (•) in Figure 3, arbitrary predictions occur across all datasets and architectures, though to varying degrees: the fraction of affected samples ranges from 1.2% (Blood Cell/ConvNeXt) to 50% (Breast Ultrasound/EfficientNetB2). For these patients, the prediction is not driven by meaningful patterns in the data but ultimately determined by a random seed.

Yet, predictive multiplicity is not uniformly distributed across all samples but concentrated on *specific* ones: the samples that are frequently misclassified (low accuracy) also show low APPA. APPA and accuracy are correlated—when multiple models predict the correct class (high accuracy), they necessarily agree (high APPA). However, they metrics capture distinct properties: accuracy measures correctness against the ground truth, whereas APPA measures consistency across equally valid models. *Predictive multiplicity is not about whether a prediction is correct, but whether it could have been different under an equally valid model.* The two extremes of the combination of accuracy and APPA are informative. Low APPA and low accuracy marks samples whose diagnosis is genuinely arbitrary—the regime that motivates governance. High APPA with low accuracy marks samples that all models agree on (no predictive multiplicity) but are predicted incorrectly, indicating systematic bias shared across models in the Rashomon set rather than underdetermination (which is of distinct concern; see D'Amour et al. (2022, Section 2.3) for a detailed discussion).

In summary, predictive multiplicity poses a serious risk for high-stakes applications: Relying on any single model exposes some patients to arbitrary outcomes, *predictions not driven by meaningful patterns in the data but ultimately determined by a random seed.* Yet, analyzing predictive multiplicity alongside accuracy offers a lens to analyze where and why models converge/diverge, offering practical means (e.g., stability-based filtering, as we will show in subsection 3.4) for building more reliable systems. Importantly, *predictive multiplicity is not inherently negative—it can help expose misclassifications that would otherwise go unnoticed.*

3.3 Reducing Predictive Multiplicity Through Accuracy Maximization

Before moving on how to leverage predictive multiplicity in practice, we take a closer look at the relationship between model capacity, accuracy, and predictive multiplicity. Overparameterized networks—where the number of models exceeds the number of training samples—have been shown to generalize effectively. The Generalization Disagreement Equality (GDE) (Jiang et al., 2022) states that, for well-calibrated ensembles, the disagreement rate between independently trained models equals their error rate in expectation. SGD-trained networks, like ours, satisfy this calibration property in practice (Lakshminarayanan et al., 2017), which is consistent with the per-sample correlation we observe: error provides the space for disagreement

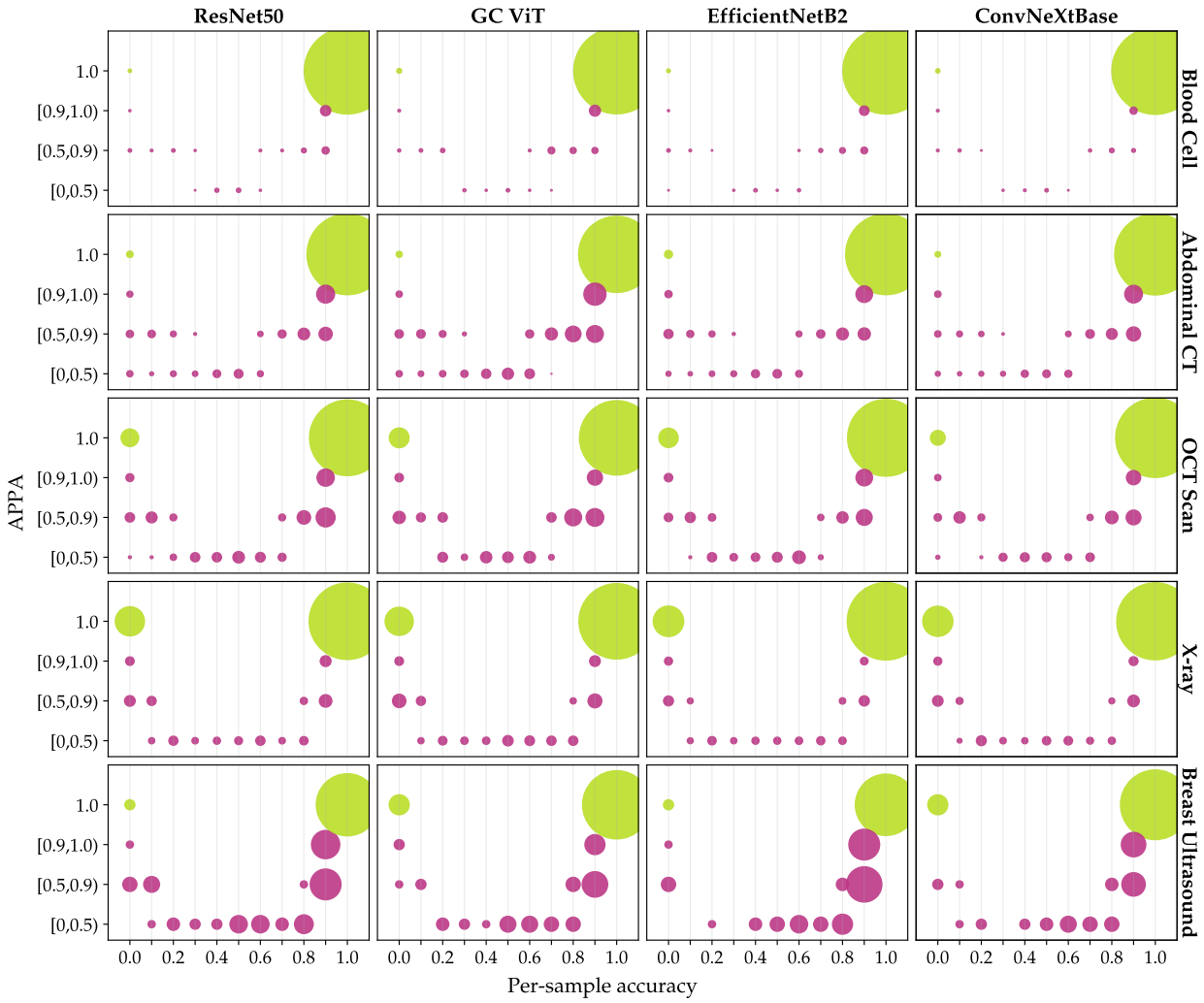

Figure 3: **Prediction stability as a function of accuracy.** For clarity, samples are binned by per-sample accuracy and APPA; point size reflects their relative frequency (normalized by dataset size). Color encodes stability: (●) marks stable samples (APPA = 1.0), whereas (●) correspond to unstable samples (APPA < 1.0). Across datasets and architectures, high accuracy across the models in the empirical Rashomon set consistently exhibits lower APPA (top-right corner). Samples in the top-left corner are consistently misclassified across models; a systematic failure mode that is distinct from predictive multiplicity.

to occur. We thus expect networks with lower test error to exhibit less predictive multiplicity. However, Black et al. (2022b) show that multiplicity is closely liked to variance. In particular, when higher accuracy is achieved by increasing model complexity (and thus variance), we should expect an increase in predictive multiplicity. Intuitively, models with greater capacity can learn more intricate decision boundaries, which, in theory, allow for higher predictive multiplicity. This relationship is further supported by their empirical findings (Black & Fredrikson, 2021; Black et al., 2021), which compares lower-complexity linear models to expressive deep neural networks. We build on those findings by examining predictive multiplicity in deep neural networks of different capacities. To operationalize capacity, we utilize EfficientNet variants, specifically EfficientNetB0 (5.3M parameters) and EfficientNetB4 (19.5M parameters). Both EfficientNet variants operate in the overparameterized regime for datasets consisting of 546 to 223,414 training samples.

We find that accuracy mediates the effect. Table 1 summarizes the population-level results of increasing model capacity. For Blood Cell, Abdominal CT, and X-ray using a higher-capacity model (B4 instead of B0)

Table 1: **Performance and stability across model capacities—when higher-capacity models achieve higher accuracy, stability increases (higher APPA); otherwise the aggregated effects are minimal.** The first two rows report *mean test accuracy* (in % ± std) for EfficientNetB0 (relatively low capacity) and B4 (relatively high capacity) across 50 models from the empirical Rashomon set. Subsequent rows show the *effects of increasing model capacity from B0 to B4* (in %): the increase in mean accuracy (↑ Acc), the change in mean APPA, and the corresponding proportion of samples whose APPA values differ between models (Affected). For Δ APPA and Affected we report actual/binarized APPA values; in the binarized setting, APPA=1 when APPA is actually 1, 0 otherwise.

	Blood Cell	Abdominal CT	OCT Scan	X-ray	Breast Ultrasound
Accuracy B0	99.0 ± 0.1	95.2 ± 0.3	86.7 ± 1.3	83.0 ± 0.3	84.7 ± 3.7
Accuracy B4	99.1 ± 0.1	95.9 ± 0.3	91.1 ± 0.8	82.8 ± 0.2	89.1 ± 1.4
↑ Acc	0.1	0.7	4.4	-0.2	4.4
Δ APPA	-0.1 / -0.6	-0.1 / -0.6	2.7 / 9.9	0.2 / 0.1	12.8 / 12.8
Affected	3.9 / 1.9	16.0 / 5.4	25.9 / 14.3	13.4 / 5.2	66.0 / 30.8

results in little change in both, accuracy and APPA ($\leq 0.6\%$). On OCT Scan and Breast Ultrasound, where the higher-capacity model improves accuracy (up to 4.4%), up to 12.8% of more samples become stable. To sum up, we find that higher capacity reduces predictive multiplicity only when it improves accuracy; otherwise, the effects at the population level are minimal.

Yet, population-level stability hides individual-level churn. Even when aggregate APPA barely changes, *which* samples are affected shifts—cases that were stable under the lower capacity model may become unstable under the higher capacity model, and *vice versa*. Figure 4 depicts the changes in two consequential categories. First, the *cost of model switching* (pink bundles, •), i.e., samples that were stable under B0 but become unstable under B4. These cases were previously handled reliably; under the higher-capacity model they are now predicted inconsistently. Such samples are detectable only when more than one model is considered. Second, *newly detectable errors* (green bundles, •), i.e., samples that were consistently misclassified under B0 but become inconsistently predicted under B4. This category shrinks for the datasets where B4 also raises accuracy (Abdominal CT and OCT Scan).

In summary, in the overparameterized regime higher capacity reduces predictive multiplicity when it increases overall accuracy as predicted by the GDE (Jiang et al., 2022). When higher capacity does not improve accuracy, the fraction of samples affected changes only minimally, yet changes which patients bear the arbitrariness. In high-stakes scenarios, where individual outcomes matter, monitoring *who* is affected is as important as monitoring how many patients are affected: collective performance gains can mask shifts in the population that bears the uncertainty. Further, within the range of capacities we test, multiplicity is not eliminated; predictive multiplicity arises from different sources, not only from limited model expressiveness, and thus cannot be solved purely by technical means. This motivates the shift from technical mitigation to risk governance in the next section.

3.4 Prediction reliability requires more than one model

In the previous sections we showed that some predictions are arbitrary under any single model and that predictive multiplicity concentrates on the samples which are likely to be classified incorrectly. This insight can be leveraged: when we consider multiple equally valid models, the structure of their disagreement carries information about which patient's predictions are underdetermined by the current data. In evidence-based medicine, genuine uncertainty about which of two treatments is more effective is termed clinical equipoise, prompting approaches such as shared decision-making or multidisciplinary review (e.g., tumor boards) (Freedman, 1987). We propose to treat predictive multiplicity as its computational analogue: an *in-silico* signal of equipoise that tells the system when not to act. Operationally, we replace single model deployment with a small ensemble (e.g., $N = 5$) drawn from the empirical Rashomon set. For each patient, the system returns an automated prediction only if the models reach a chosen consensus threshold; otherwise

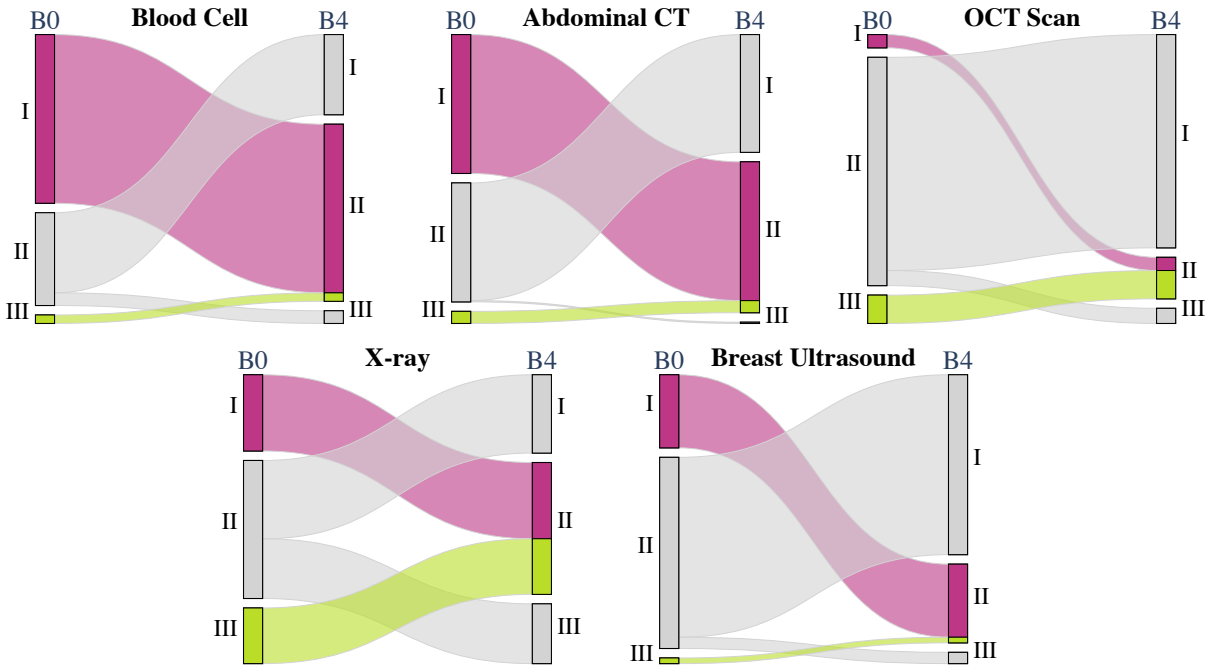

Figure 4: **Changes by increasing model capacity (only affected samples).** Samples are grouped as (I) correct-stable, (II) unstable, and (III) incorrect-stable. Bundles connect category transitions from EfficientNetB0 (left) to EfficientNetB4 (right), illustrating how samples shift between groups. Pink bundles mark the cost of model switching (previously stable-correct samples becoming unstable) while green bundles indicate newly detectable errors (previously consistently misclassified samples become unstable and thus identifiable). For Blood Cell the effect is minimal (only 2.6% of samples are affected), for Abdominal CT, OCT Scan, and Breast Ultrasound overall utility improves, for X-ray utility stays the same, but different samples are affected.

it abstains and routes the case to a clinician. The threshold is a *governance parameter*: stricter consensus raises diagnostic reliability at the cost of automated coverage, and *vice versa*. Unlike standard ensembling, which averages disagreement away and emits a prediction in every case, abstention is grounded: disagreement among models from the Rashomon set marks underdetermined evidence in the current setting rather than noise to be smoothed. Figure 5 traces this trade-off across thresholds. For Blood Cell/ResNet50, for example, requiring unanimous agreement across the five models raises reliability at the cost of only deferring 1.5% of cases. The threshold allows administrators to calibrate automation against clinical risk.

To summarize, the cases deferred to human review are not random. They are precisely those for which the available data fail to discriminate among equally valid models, i.e., where the current evidence is genuinely insufficient. This makes the system function as in-silico triage: it identifies, transparently and at scale, the patients for whom further workup (additional imaging, biopsy, multidisciplinary review) is warranted before any decision is made. More broadly, rather than forcing a prediction, the system makes a decision only when the data warrants it, and otherwise refers to human review.

4 Discussion

We presented a comprehensive evaluation of the empirical Rashomon set in the medical domain and show that the existence of multiple equally valid models challenges conventional practices of model selection and deployment. While grounded in healthcare, our findings likely extend to other high-stakes domains facing similar constraints of uncertainty, limited data, and ethical responsibility. Traditionally, studies of the

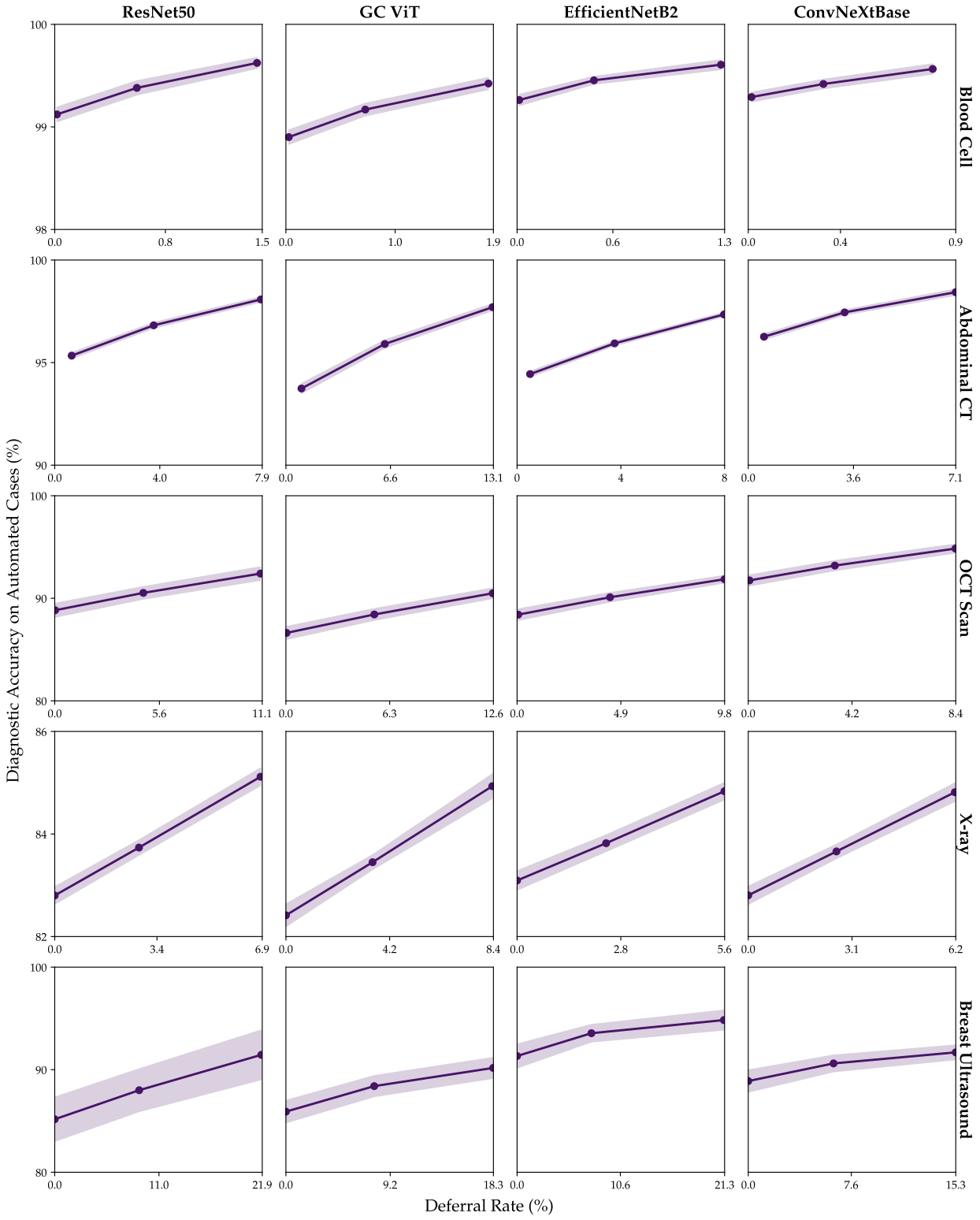

Figure 5: **Risk governance trade-off between diagnostic reliability and human effort.** Deferral rate (patients routed to human experts under low consensus) vs. accuracy on automated cases, for ensembles of five. Points mark consensus thresholds, from majority vote (abstaining only on ties) to unanimous agreement, each step requiring one more model to concur; the line is mean and the shaded region standard deviation.

Rashomon effect consider only models with small performance differences as statistically indistinguishable. Using the framework of Paes et al. (2023), we find that even models with substantial differences can be statistically indistinguishable, underscoring the limits of conventional performance-based comparisons. This fits with earlier work (Jordan, 2024) showing that when models seem to perform differently on a test set, the gap is often just random noise from having limited test data, not a real difference in how well the models generalize. The problem may be worse in medicine: datasets are getting larger overall, but still contain very few examples of rare conditions; sampling variability may distort the empirical Rashomon set by overfitting idiosyncrasies of the finite test set.

The existence of multiple models of equal quality renders the selection of a single, supposedly "best" performing model arbitrary, undermining the justification for the choice of this model and its predictions. Relying on one model exposes some patients to effectively arbitrary and potentially harmful outcomes—those whose predictions would differ under another, equally valid model. Creel & Hellman (2022) argue that isolated arbitrary decisions are not inherently morally problematic, except when other rights make non-arbitrariness normatively relevant. In medicine, such rights arguably exist: patients are entitled to informed consent and to consistent, evidence-based, and non-arbitrary treatment (Olejarczyk & Young, 2024; Varkey, 2021). While the form and legal status of informed consent vary across cultural and regulatory contexts, its underlying commitment to respecting patient self-determination is widely recognized (Angell, 1988). Arbitrariness in model selection may compromise this principle: a patient cannot meaningfully consent to a decision when its outcome depends not on any clinical rationale but on incidental factors such as the random seed used during training. We show the extend of predictive multiplicity in medical imaging datasets, yet questions such as *to what extent should/do patients have a right to non-arbitrary, consistent treatment?* and *how should informed consent be interpreted under model multiplicity?* require further ethical and legal analysis. In how far existing regulatory frameworks already capture these concerns warrants domain expert examination. Frohnapfel et al. (2026) argue that the arbitrariness due to predictive multiplicity conflicts with the EU AI Act conflicts with the EU AI Act (European Parliament and Council of the European Union, 2024), which requires providers of high-risk AI systems to report performance not only at the dataset level (Article 15 (1)) but also for specific individuals (Article 11(1) and Annex IV(3), 13(3)). If left unaddressed, model multiplicity and predictive multiplicity in particular, may complicate the ethical and legal justification of AI-assisted decision-making in high-stakes settings, like healthcare.

In the already overparameterized regime, we find that higher-capacity models can improve both accuracy and predictive stability. However, "accuracy is not an antidote to multiplicity, and model selection cannot simply be reduced to accuracy-maximization" (Black et al., 2022b)—accuracy-maximization is an insufficient criterion for model selection. In real-world applications, where overparameterized models are the norm, the number of trainable parameters may, in fact, have limited relevance for predictive multiplicity.

Samples with higher expected accuracy are predicted more consistently across models, suggesting that predictive multiplicity is not inherently detrimental but can be leveraged when we rely on more than one model. Using multiple models and restricting predictions to samples which are consistent across multiple models improves accuracy and prediction stability. A objection to our risk governance framework may be that clinicians also make inconsistent judgments, so deferral merely relocates the arbitrariness rather than solving it. The objection misreads the asymmetry. Clinical arbitrariness is governable: clinician decisions must be justified, can be challenged, and are subject to second opinions and review. Arbitrariness under predictive multiplicity is not: when many equally valid models exist, there is no principled basis for why one model's output should apply to a specific patient—and therefore no meaningful way to contest it. Deferral does not eliminate arbitrariness; it moves it to a process that has the institutional infrastructure to handle it.

Our work is not without limitations. We do not fully characterize the broader impact of predictive multiplicity on medical diagnosis tasks, and we focus exclusively on classification problems thereby omitting the full diversity of clinical scenarios and modeling paradigms. In line with this, we do not assess how predictive multiplicity influences downstream clinical decision-making or patient outcomes—an important area for future research. Our aim is to demonstrate the implications of predictive multiplicity in medical application domains and to motivate further investigation into mitigating the risks that model multiplicity poses to the adoption of machine learning in high-stakes domains.

5 Conclusion

In this study we show that predictive multiplicity is both pervasive and consequential in medical AI. Small, seemingly inconsequential training variations can lead to different predictions for individual patients despite statistically indistinguishably overall performance of the models. This finding challenges the widespread assumption that a single "best" model can reliably guide decisions in high-stakes (clinical) contexts. Our results reveal that multiplicity is not a rare anomaly but a fundamental property of modern predictive modeling. Recognizing it as a structural property of the learning process—rather than a nuisance—calls for a rethinking of how models are evaluated, selected, and deployed in high-stakes settings. Ultimately, predictive multiplicity matters most where it is least tolerable, at the point of care, where treatment decisions depend on individual predictions. Identifying cases of predictive disagreement lets models signal when the available evidence is insufficient for safe automation and when human expertise, additional data, or both are required.

References

Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in brief*, 30:105474, 2020.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*, pp. 314–323. PMLR, 2020.

Marcia Angell. Ethical imperialism?, 1988.

Peter Auer, Mark Herbster, and Manfred KK Warmuth. Exponentially many local minima for single neurons. *Advances in neural information processing systems*, 8, 1995.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680, 2023.

Emily Black and Matt Fredrikson. Leave-one-out unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 285–295, 2021.

Emily Black, Zifan Wang, Matt Fredrikson, and Anupam Datta. Consistent counterfactuals for deep models. In *Ninth International Conference on Learning Representations*, 2021.

Emily Black, Klas Leino, and Matt Fredrikson. Selective ensembles for consistent predictions. In *10th International Conference on Learning Representations, ICLR 2022*, 2022a.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 850–863, 2022b.

Emily Black, Talia Gillis, and Zara Yasmine Hall. D-hacking. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 602–615, 2024.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. An interpretable model with globally consistent explanations for credit risk. *arXiv preprint arXiv:1811.12615*, 2018.

Danielle Keats Citron and Frank Pasquale. The scored society: Due process for automated predictions. *Wash. L. Rev.*, 89:1, 2014.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Kathleen Creel and Deborah Hellman. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, 52(1):26–43, 2022.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*, pp. 2803–2813. PMLR, 2020.

European Parliament and Council of the European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending certain union legislative acts (artificial intelligence act). Official Journal of the European Union, L 2024/1689, 2024. URL `http://data.europa.eu/eli/reg/2024/1689/oj`.

Benjamin Freedman. Equipoise and the ethics of clinical research. *The New England Journal of Medicine*, 1987.

Karolin Frohnapfel, Mara Seyfert, Sebastian Bordt, Ulrike von Luxburg, and Kristof Meding. Using predictive multiplicity to measure individual performance within the ai act. *arXiv preprint arXiv:2602.11944*, 2026.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Leif Hancox-Li. Robustness in machine learning explanations: does it matter? In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 640–647, 2020.

Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. In *International Conference on Machine Learning*, pp. 12633–12646. PMLR, 2023.

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*, pp. 770–778, 2016.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of sgd via disagreement. In *International Conference on Learning Representations*, 2022.

Keller Jordan. On the variance of neural network training with respect to test sets and distributions. *ICLR*, 2024.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

Katarzyna Kobylińska, Mateusz Krzyziński, Rafał Machowicz, Mariusz Adamek, and Przemysław Biecek. Exploration of the rashomon set assists trustworthy explanations for medical data. *IEEE Journal of Biomedical and Health Informatics*, 28(11):6454–6465, 2024.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning*, pp. 6765–6774. PMLR, 2020.

Johannes Mehrer, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1):5725, 2020.

Jacob P Olejarczyk and Michael Young. Patient rights and ethics. *StatPearls [Internet]*, 2024.

Lucas Monteiro Paes, Rodrigo Cruz, Flavio P Calmon, and Mario Diaz. On the inevitability of the rashomon effect. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pp. 549–554. IEEE, 2023.

Harsh Parikh. Why are there many equally good models? an anatomy of the rashomon effect. *arXiv preprint arXiv:2601.06730*, 2026.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*, pp. 809–818. PMLR, 2020.

Richard D Riley and Gary S Collins. Stability of clinical prediction models developed using statistical or machine learning methods. *Biometrical Journal*, 65(8):2200302, 2023.

Richard D Riley, Alexander Pate, Paula Dhiman, Lucinda Archer, Glen P Martin, and Gary S Collins. Clinical prediction models and the multiverse of madness. *BMC medicine*, 21(1):502, 2023.

Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. Position: amazing things come from having many good models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 42783–42795, 2024.

Nathaniel Schenker and Jane F Gentleman. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55(3):182–186, 2001.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Basil Varkey. Principles of clinical ethics and their application to practice. *Medical principles and practice*, 30(1):17–28, 2021.

Michael Wick, Jean-Baptiste Tristan, et al. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32, 2019.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

A Appendix

A.1 Related Literature

In the following we provide a more detailed discussion of related literature that complements the brief overview presented in the main text. The phenomenon of model multiplicity was first described by Breiman (2001) as the *Rashomon Effect*. They observed that small perturbations in the training set for decision trees and different weight initializations for small neural networks can lead to different solutions while having approximately equal error rates. More recent work showed that model multiplicity is ubiquitous in modern machine learning and a key obstacle to reliable training models that behave as expected in deployment (D'Amour et al., 2022). The existence of many models that achieve indistinguishable overall performance is particularly relevant w.r.t. their effect and consequences in the real world. Several works highlight the opportunities that model multiplicity offers (Rudin et al., 2024); such as the selection of fairer (Dutta et al., 2020; Wick et al., 2019), more interpretable (Chen et al., 2018), or more robust models (D'Amour et al., 2022) without impairing predictive performance. Challenges arising from model multiplicity are among others the inconsistency of explanations (Hancox-Li, 2020; Pawelczyk et al., 2020), the risk of fair-washing explanations (Anders et al., 2020) or fairness metrics (Black et al., 2024), and predictive multiplicity (Marx et al., 2020)—the main focus of this paper.

Despite the relevance of predictive multiplicity in the medical domain, systematic investigations into its risks and mitigation remains limited. To the best of our knowledge previous work in the medical domain has leveraged model multiplicity for trustworthy explanations (Kobylińska et al., 2024), explored the role of underspecification in model robustness (D'Amour et al., 2022) and examined representational variability (Mehrer et al., 2020) (both works highlight the cause for predictive multiplicity without addressing how such instability/ variability affects predictions), and proposed bootstrapping as a remedy to predictive multiplicity (Riley et al., 2023; Riley & Collins, 2023).

Bootstrapping introduces diversity by training ensemble members on different resampled subsets (with replacement) of the original training data. While effective for deterministic models that lack intrinsic sources of randomness, modern machine learning models are inherently stochastic, with randomness arising from initialization, data shuffling, and optimization dynamics. Further, bootstrapping is computationally intensive and thus often not economically sensible; Riley & Collins (2023) recommend the training of at least 200 models. Empirically, Lakshminarayanan et al. (2017) found that training each model on the entire dataset—using random initialization and data shuffling—achieved better performance than bootstrapping. This may be even more pronounced in medical settings with rare and underrepresented conditions. In summary, while bootstrapping is conceptually simple, ensembling full datasets is a more practical and effective approach for deep learning models with inherent stochasticity, especially in medical settings where data for certain conditions is scarce. In our experiments (see Section 3.4), we found that ensembles of as few as five models with an abstention capability substantially reduce predictive multiplicity while improving accuracy.

A.2 Datasets, Model Architectures, and Model Training

In this section we describe the datasets, model architectures, and training procedures in more detail. We used five medical imaging datasets spanning diverse modalities and classification tasks: Abdominal CT (Bilic et al., 2023), Breast Ultrasound (Al-Dhabyani et al., 2020), Blood Cell (Acevedo et al., 2020), OCT Scan (Kermany et al., 2018), and X-Ray (Irvin et al., 2019). See Table 2 for an overview of dataset properties. For X-ray, we adopted the protocol of Irvin et al. (2019), training multi-label binary classifiers across all 14 classes and evaluating performance on the five pathologies with annotated validation/test labels: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. Each image was treated as an independent sample. Uncertain labels were converted to binary: mapped to positive for Atelectasis, Edema, and Pleural Effusion, and to negative for all other classes. All experiments rely on the official training, validation, and test splits to ensure comparability with prior work (see Yang et al. 2023; Irvin et al. 2019). X-ray images were resized to $1\times224\times224$. All other datasets followed the MedMNISTv2 preprocessing (Yang et al., 2023), resulting in $C\times224\times224$ inputs ($C=3$ for Blood Cell; $C=1$ for the others). Single-channel inputs were triplicated to match the input dimensionality expected by ImageNet-pretrained models.

Table 2: Number of training, validation, test samples and classes for each dataset. *The training set contains 14 labels, with five designated as target labels (details provided in the text.)

Dataset	# Train	# Val	# Test	Labels
Abdominal CT	34,561	2,392	8,825	11
Blood Cell	11,959	1,712	3,421	8
Breast Ultras.	546	78	156	2
OCT Scan	97,477	10,832	1,000	4
X-ray	223,414	234	500	14/5*

Table 3: For each dataset (columns) and model architecture (rows), we report the initial model's test performance, along with the mean test accuracy across 50 models from the empirical Rashomon set, in parentheses. The best-performing initial model for each dataset is highlighted in bold. Benchmark results are from Irvin et al. (2019) for X-ray (highest ROC-AUC across five uncertainty methods) and from Yang et al. (2023) for all other datasets (highest accuracy across seven architectures).

	Benchmark	ResNet50	GC ViT	EfficientNetB2	ConvNeXt
Abdominal CT	92.0	94.6 (94.4)	91.2 (91.6)	93.5 (93.6)	**95.3** (95.2)
Blood Cell	96.6	98.9 (99.0)	98.7 (98.7)	99.0 (99.1)	**99.2** (99.2)
Breast Ultrasound	86.8	80.8 (83.9)	84.6 (84.6)	**89.7** (88.5)	88.5 (87.1)
OCT Scan	77.6	87.4 (88.3)	87.4 (86.0)	86.4 (88.1)	**92.5** (91.4)
X-ray	89.5	88.9 (88.6)	87.2 (87.6)	**89.8** (89.9)	89.7 (89.7)

We evaluated four model architectures ResNet50 (He et al., 2016), GC ViT (Hatamizadeh et al., 2023), EfficientNet(Tan & Le, 2019) (variants B0 and B4 in subsection 3.3; B2 for all others), and ConvNeXtBase (Liu et al., 2022), all pretrained on ImageNet (Deng et al., 2009). To differ random weight initialization, we replaced the final classification layer with a randomly initialized dense layer matching the number of classes in the respective dataset, using a Glorot uniform initializer (Glorot & Bengio, 2010).

All training was conducted under deterministic conditions with fixed random seeds. Within each dataset/architecture combination, variation across models (i.e., the Rashomon set exploration) was induced solely by varying the random seed, which determined the initial weights of the final classification layer via the Glorot uniform initializer. All other factors were fixed to ensure reproducibility. All models were trained with a batch size of 64 using the AdamW optimizer (Loshchilov & Hutter, 2017), with exponential decay rates of 0.9 and 0.999 for the first and second-moment estimates, respectively. To select the initial learning rate, we performed a sweep over 0.01, 0.001, 0.0001 with a fixed random seed (seed = 0); the best-performing learning rate was used in all subsequent experiments without further tuning. We employed a cosine decay learning rate schedule (Loshchilov & Hutter, 2016), the decay steps matched the number of epochs. We trained for a fixed number of epochs without early stopping, with the number of epochs depending on the dataset: 15 epochs for Breast Ultrasound, five epochs for Blood Cell, OCT Scan, and Abdominal CT, one epoch for X-ray. We used sparse categorical cross-entropy for all other datasets. Classification accuracy served as the primary performance metric. Table 3 reports both the performance of the initial model and the mean performance across the 50 models from the empirical Rashomon set. Across datasets, our models perform competitively and often surpass reported benchmarks.

All models and training procedures were implemented using Keras 3.8. We used different GPUs for different dataset/architecture combinations; however, all model instances within one empirical Rashomon set were trained on the same GPU to ensure consistency.

A.3 Numerical Rashomon Parameter Estimation

Following Paes et al. (2023), we defined the Rashomon parameter ϵ as the largest error rate for which the CP confidence intervals (of a predefined reference model and all other empirically obtained models) no longer

overlap at the significance level $\alpha = 0.05$. We computed ϵ numerically using a bisection search. Given the number of samples in the respective (finite validation and test) datasets n and ϵ_0—the lowest obtained error rate among the 50 models in the empirical Rashomon set—we first compute the upper CP bound $UB(\alpha, n\epsilon_0, n)$ for ϵ_0. We then search for the smallest ϵ, such that $LB(\alpha, n\epsilon, n) = UB(\alpha, n\epsilon_0, n)$, where LB and UB are the lower and upper CP confidence limits respectively. The bisection procedure iterates until convergences (tolerance $< 10^{-8}$) or after 10,000 steps. We adopt an exact binomial model by rounding $n\epsilon$ to the nearest integer, ensuring that the CP bounds are computed from valid discrete sample counts; see Table 4 for ϵ_0 and ϵ values. Note that overlapping confidence intervals are no formal statistical test for evaluate the equality of means (e.g., Schenker & Gentleman 2001); the approach is conservative and may miss small but statistically significant differences. However its simplicity and graphical interpretability makes it appealing for practitioners (see Paes et al. 2023 for more details and alternatives).

Table 4: Numerical values of the Rashomon parameter ϵ and the reference model ϵ_0.

Dataset	Model	Validation		Test	
		ϵ_0	ϵ	ϵ_0	ϵ
Breast	ResNet	5.1	21.8	9.6	21.8
	GC ViT	5.1	21.8	12.2	25.0
	EB2	2.6	16.7	7.1	17.9
	Conv	5.1	21.8	10.9	23.7
Blood Cell	ResNet	0.7	1.8	0.7	1.4
	GC ViT	0.9	2.1	1.1	1.9
	EB2	0.6	1.8	0.7	1.5
	Conv	0.5	1.6	0.6	1.3
Abdominal CT	ResNet	0.6	1.4	5.2	6.2
	GCViT	1.2	2.3	7.0	8.2
	EB2	0.6	1.4	5.8	6.9
	Conv	0.4	1.2	4.0	4.9
OCT Scan	ResNet	2.0	2.6	9.1	13.1
	GCViT	2.1	2.7	11.9	16.4
	EB2	1.9	2.5	10.4	14.6
	Conv	1.6	2.2	6.4	9.9
X-ray	ResNet50	17.9	22.6	16.8	19.9
	GCViT	18.5	23.2	17.4	20.5
	EfficientNetB2	17.9	22.6	16.4	19.4
	ConvNeXt	17.8	22.5	16.9	20.0

A.4 Adjusted Pairwise Prediction Agreement (APPA)

We assess per-sample prediction stability using APPA, which quantifies the probability that two models, sampled uniformly at random from the empirical Rashomon set produce the same prediction. Formally, let $M > 1$ denote the number of models, each predicting a class $c \in \{1, ..., K\}$ for the same sample x. For a given input x, let $n_c(x)$ represent the number of models that predict class c. The unnormalized pairwise prediction agreement is defined as

$$\text{PPA}(x) = \frac{\sum_{c=1}^{K} n_c(x)(n_c(x) - 1)}{M(M-1)}. \tag{1}$$

As the minimal attainable pairwise agreement depends on both the number of models M and classes K (intuitively, with fewer classes than models, some models must necessarily coincide in their predictions), we normalize by the achievable minimum $\text{PPA}_{\min}(M, K)$. This minimum corresponds to model predictions being distributed as uniformly as possible across classes; with $q, r = M \bmod K$, r classes receive $q + 1$

predictions, and $(K - r)$ classes receive q predictions, resulting in

$$\mathrm{PPA}_{\min}(M, K) = \frac{(K - r)q(q - 1) + r(q + 1)q}{M(M - 1)}. \tag{2}$$

We then compute APPA(x) as

$$\mathrm{APPA}(x) = \frac{\mathrm{PPA}(x) - \mathrm{PPA}_{\min}(M, K)}{1 - \mathrm{PPA}_{\min}(M, K)}, \tag{3}$$

which captures the excess agreement beyond what is expected from maximally uniform predictions, enabling comparison across tasks with differing numbers of models or classes. $\mathrm{PPA}_{\min}$ is a reasonable default in the absence of prior knowledge about the distribution. When information about the true distribution is available (such as the prevalence), using this prior can be a more appropriate choice.