On Arbitrary Predictions from Equally Valid Models

Sarah Lockfisch[1], Kristian Schwethelm[2], Martin Menten[2,3,4], Rickmer Braren[1,5], Daniel Rückert[2,3,4], Alexander Ziller[2*], Georgios Kaissis[6*†]

[1]School of Medicine and Health, Technical University of Munich, Germany
[2]Chair for AI in Healthcare and Medicine, Technical University of Munich, Germany
[3]BioMedIA, Imperial College London, United Kingdom
[4]Munich Center for Machine Learning, Germany
[5]Universitätsklinikum Hamburg-Eppendorf, Germany
[6]Hasso Plattner Institute for Digital Engineering, University of Potsdam, Germany
sarah.lockfisch@tum.de, georg.kaissis@hpi.de

Abstract

Model multiplicity describes the existence of multiple models that fit the data equally well but can produce different predictions for 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 statistically indistinguishable performance show variability in patient-level predictions, resulting in arbitrary and potentially harmful outcomes under any single model. However, predictive multiplicity does not affect samples equally, and the converse can be used to reduce predictive multiplicity. We find that (3) high model capacity decreases predictive multiplicity by improving accuracy. Lastly, (4) ensembles with an abstention strategy enhance expected per-sample accuracy and stability.

Together, these findings highlight that predictive multiplicity is not merely 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.

Code — https://github.com/tofooschnitzel/mm4mi

1 Introduction and Prior Work

Model multiplicity (Black, Raghavan, and Barocas 2022) describes the existence of many *plausible* models for the same dataset without a principled way to determine a single optimal model. In practice, multiple machine learning models can fit the same data equally well according to a given 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 a 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 in practice.

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 and Collins 2023). Prior work has discussed its opportunities (Rudin et al. 2024) and challenges, notably *predictive multiplicity* (Marx, Calmon, and Ustun 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 and 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 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 and Collins 2023; Riley et al. 2023) that are largely infeasible for modern ML; see Appendix C for details.

In this study, we address this gap through a comprehensive empirical analysis of predictive multiplicity across multiple medical tasks (abdominal CTs, blood cell images, breast ultrasounds and OCT scans) and architectures (ResNet50, GC ViT, EfficientNet, and ConvNeXt). Results for the ResNet50 appear in the main text; additional architectures are in the Appendix A. Recognizing and addressing model multiplicity not only exposes limitations of the "single model" paradigm but also offers a path to improve

*These authors contributed equally.
†This work was partially conducted at Google DeepMind.

Workshop on Navigating Model Uncertainty and the Rashomon Effect: From Theory and Tools to Applications and Impact (AAAI 2026)

predictive stability and accuracy by levering inter-model (dis)agreement. Our large-scale study, based on the training of 1,400 models, leads to the following conclusions that form the core contributions of our work:

1. Validation performance is an unreliable indicator of generalization and (thus) fails to identify an optimal model.
2. Relying on a single model exposes some patients to arbitrary predictions. Yet, the underlying structure can be exploited to improve stability and accuracy.
3. Higher-capacity models, when improving accuracy, reduce predictive multiplicity.
4. Ensembles with abstention eliminate measurable predictive multiplicity and improve accuracy.

2 Methodology

Deployment typically relies on a single, "optimal" model, without considering other, equally valid alternatives. We instead examine the set of models that fit the data equally well and thus represent equally plausible solutions, the so-called *Rashomon set* (Breiman 2001). Following prior work (e.g., D'Amour et al. 2022; Black, Leino, and Fredrikson 2022), we explore this set empirically by randomizing the initialization of 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 model instances, which differ only in the weight initialization of the last layer while keeping all other components fixed. This results in a total of 1,400 models—1,000 for the main experiments and 400 to analyze model capacity. In brief, our experiments cover aforementioned medical imaging datasets and model architectures and perform competitively (and often surpass) prior results; see Appendix B for details on datasets, architectures, training procedure, and performance).

To assess whether models are of equal quality, we avoid an *ad hoc* threshold for performance differences and instead apply the hypothesis testing framework of Paes et al. (2023). We use the Clopper-Pearson (CP) interval (Clopper and Pearson 1934), an exact method for constructing confidence intervals for binomial error rates, and apply it to model accuracies (the inverse of error). The model with the highest accuracy – i.e., the lowest empirical error—serves as the reference ϵ_0. We then compare each model's confidence interval to the reference interval: if two intervals overlap at the 95% significance level, we consider the models statistically indistinguishable. The Rashomon parameter ϵ corresponds to the smallest decrease in empirical accuracy at which the condition no longer holds, representing the minimal deviation required to reject the null hypothesis of equal true accuracy. We compute ϵ numerically using bisection (see Appendix D for implementation details).

To quantify prediction stability at the per-sample level, we define Adjusted Pairwise Prediction Agreement (APPA). APPA measures the probability that two models, drawn uniformly at random from the empirical Rashomon set, assign the same prediction to a sample x. The measure is normalized by the number of models M and classes K (see Appendix E for a detailed derivation). By definition, $\mathrm{APPA}(\mathrm{x}) \in [0, 1]$, where $APPA = 1$ indicates maximal expected agreement between two models (high stability), $APPA = 0$ maximal disagreement (low stability), and intermediate values indicate partial expected agreement.

Lastly, we evaluate the effectiveness of ensembles to reduce predictive multiplicity by comparing prediction stability and coverage rates between two single models or two ensembles (of size two or five). Prediction stability is measured as the expected pairwise agreement between two distinct models (or ensembles of equal size) on the test set and averaged over 100 repetitions. To avoid zero-inflation, we draw model or ensemble pairs without replacement from the empirical Rashomon set. For ensembles, we apply a conservative decision rule: a prediction is made only if all constituent models agree; otherwise, the ensemble abstains.

3 Results

The illusion of an optimal model

The true Rashomon set is the set of statistically indistinguishable "good" models. In practice, we face two constraints: model quality can only be assessed on finite samples, and for complex model classes, an exhaustive characterization of the Rashomon set is infeasible—e.g., in neural networks the size of the Rashomon set is tied to the number of local minima, which grows exponentially with the number of parameters (Auer, Herbster, and Warmuth 1995). Thus, we explore the Rashomon empirically by varying the random seed used to initialize the weights of the last layer. This approach results in a set of models with substantial variation in their performance on the finite validation and test set: accuracy differs by up to 16% (Breast Ultrasound in Figure 1). While prior work often designates a fixed 1% tolerance in loss from a reference model as the criterion for indistinguishability (e.g., Coston, Rambachan, and Chouldechova 2021), we instead apply the hypothesis testing framework of Paes et al. (2023). Using the most accurate model among the 50 in the empirical Rashomon set as a reference, we determine the smallest decrease in accuracy that leads to the rejection of the null hypothesis of equal true error rates, based on validation and test performance. Figure 1 visualizes the *indistinguishability region* between ϵ_0 (reference performance) and ϵ (rejection threshold). The width of this region depends on the Type-I error level (α=0.05), the baseline loss ϵ_0 of the best empirical model, and the number of validation and test samples n; see Appendix D for details.

As shown in Figure 1, all obtained models are statistically indistinguishable at the 95% significance level. In other words, while their validation and test accuracies vary, statistical testing reveals no evidence of true performance differences on the underlying distribution. Moreover, 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 in Figure 1) for which we performed the hyperparameter search is in no way "special" regarding validation or test set performance relative to the other models. Better and worse models exist despite this being the model that has been explicitly optimized for.

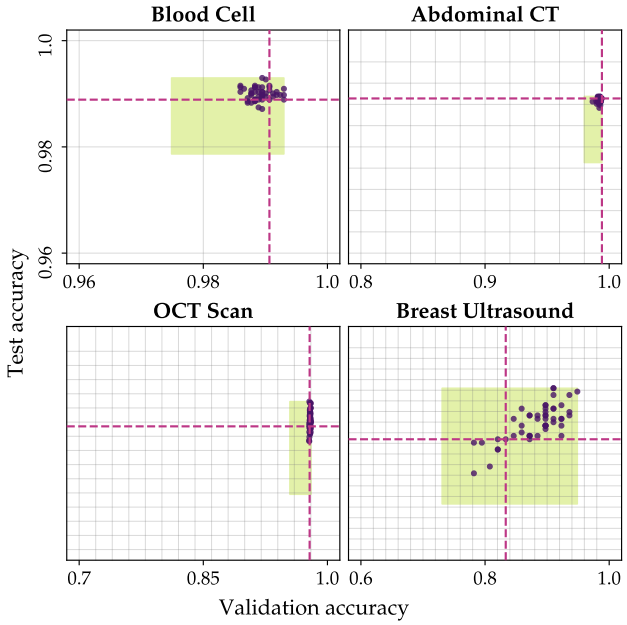

Figure 1: Variation in accuracy within the empirical Rashomon set. Each plot shows a dataset, with points representing models trained from different random initialization. Models in the green region are indistinguishable at the 95% significance level. Dashed lines mark the initial model for which we performed a hyperparameter search.

In summary, these findings imply that validation and test performance cannot uniquely identify an "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.

On arbitrary predictions under any single model

So far, we have focused on the average performance metrics within the empirical Rashomon set. We now turn to model predictions to examine the *arbitrariness of diagnostic outcomes when relying on a single model*. We use APPA (as defined in Section 2) to quantify prediction stability at the per-sample level, i.e., *vis-à-vis* a patient. The pink points in Figure 2 reveal that arbitrary predictions occur in all datasets but to varying degrees—from 2.5% of samples in Blood Cell to 48.7% in Breast Ultrasound. Importantly, (dis)agreement is not uniformly distributed across the data but concentrated on *specific* samples: those that are frequently predicted correctly (high accuracy) also exhibit high inter-model agreement (high APPA). This positive relationship is expected as both metrics depend on how models distribute their predictions across classes; intuitively, when multiple models predict the correct class, they necessarily agree more often.

Although per-sample accuracy and agreement are correlated by definition, they capture distinct aspects of model behavior. *Predictive multiplicity is not about whether a prediction is correct, but whether it could have been different under an equally valid model.* Accuracy measures correctness relative to the ground truth, while predictive multiplicity reflects the consistency of predictions across models—regardless of correctness. Consequently, high APPA is often associated with correct predictions but can also occur for incorrect ones, indicating systematic bias shared across models (see top-left regions in Figure 2). In contrast, low APPA implies that a prediction depends strongly on the specific model instance and is associated with higher error.

The relationship between accuracy and predictive multiplicity extends to the model level. Intuitively, error provides the "space" for disagreement: a higher error rate creates more opportunities for conflicting predictions. Indeed, the Generalization Disagreement Equality (GDE) (Jiang et al. 2021) formalizes this link, stating that the expected disagreement rate between independently trained models (e.g. with different random seeds) approximately equals their error rate. More precisely, this equality holds for well-calibrated ensembles, a property that SGD-trained networks, such as ours, naturally exhibit (Lakshminarayanan, Pritzel, and Blundell 2017).

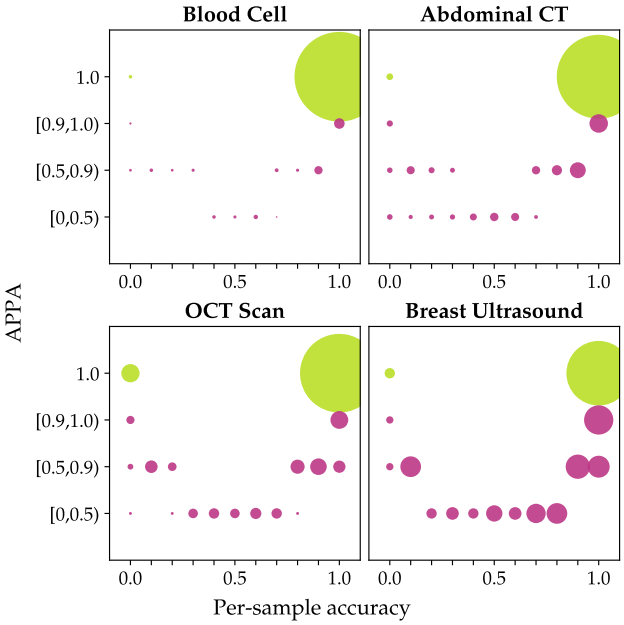

Figure 2: Prediction stability as a function of accuracy. For clarity, samples are binned by accuracy and APPA; point size reflects their relative frequency (normalized by dataset size). Color encodes stability (green: APPA = 1.0, i.e., stable samples; pink: APPA < 1.0). Across datasets, correctly predicted samples tend to exhibit high APPA. Samples in the top-left corner are, however, consistently misclassified revealing systematic failure modes.

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 reveals where and why models converge or diverge, offering both diagnostic insight (e.g., bias detection) and practical means (e.g., stability-based filtering) for building more reliable systems. Importantly, *predictive multiplicity is not inherently negative—it can help identify bias and expose misclassifications that would otherwise go unnoticed*.

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 model parameters exceeds the number of training samples—have been shown to generalize effectively (e.g., Allen-Zhu, Li, and Liang 2019). According to the GDE (Jiang et al. 2021), we expect networks with lower test error to exhibit less predictive multiplicity. However, Black et al. (2021) theoretically demonstrate 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. This relationship is further supported by their empirical findings (Black and Fredrikson 2021; Black et al. 2021), which compares low-complexity linear models to (highly) expressive deep neural networks. We build on those findings by examining predictive multiplicity in neural networks of different capacities. To operationalize capacity, we use EfficientNet variants: EfficientNetB0 (5.3M parameters) and B4 (19.5M). Both EfficientNet variants operate in the overparameterized regime for datasets with 546 to 97,477 training samples.

Table 3 summarizes the overall effect of increasing model capacity. For Blood Cell and Abdominal CT, using a higher-capacity model (B4 instead of B0) results in little change in accuracy and APPA ($\leq$ 0.6%). In contrast, for OCT Scan and Breast Ultrasound we see clear improvements—4.4% improvement in accuracy and up to 12.8% more stable samples. This pattern is also reflected in the proportion of affected samples—those whose stability or correctness changes between models (see last row in Table 3).

To understand how the composition changes by increasing model capacity, we examine how the affected samples are distributed across categories in Figure 3. Two categories are of specific interest: Previously stable samples that become unstable (pink in Figure 3). These represent the *cost of model switching*—cases that were previously handled reliably but now produce inconsistent predictions which depend on the specific, selected model. However, when we use more than one model, we can detect these cases as they are unstable across multiple models and refer them to manual review (see subsection 3). Further, the fraction of previously *undetectable errors*—samples that were both stable and incorrect under the lower-capacity models (green)—decreases after switching to the higher-capacity model for the datasets

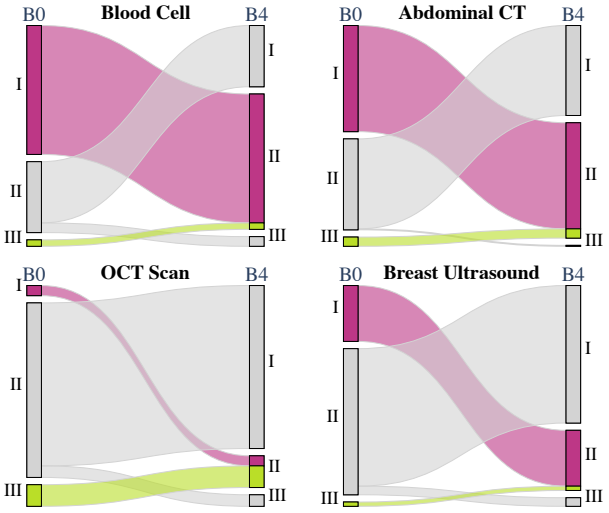

Figure 3: Changes by increasing model capacity (only affected samples). Samples are grouped as (I) correct-stable, (II) unstable, and (III) incorrect-stable. Bundles connect category transitions from EfficientNetB0 (left) to EfficientNetB4 (right), illustrating how samples shift between groups. Pink 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 Abdominal CT and OCT Scan, overall utility improves; for others the effect is minimal (for Blood Cell only 2.6% of samples are affected).

which achieve higher accuracy (Abdominal CT and OCT Scan). Reducing the size of this category is desirable, as these samples are consistently misclassified across models and cannot be identified under a single model.

In summary, our results suggest that in already overparameterized models capacity alone does not determine predictive multiplicity. When increasing capacity does not improve accuracy, we observe only marginal changes in predictive multiplicity (affecting only a small fraction of samples); however, when higher capacity leads to improved accuracy, predictive multiplicity consistently decreases. This indicates that the relationship between capacity and multiplicity is mediated by accuracy: greater expressiveness can either introduce minor additional variability when performance remains stagnant or, when it enhances generalization, substantially reduce predictive multiplicity. Yet, awareness and systematic monitoring of such shifts in the composition of affected samples—and of which individuals/groups are impacted—remain essential, particular in high-stakes domains like healthcare, where collective performance gains must be carefully balanced against individual rights.

Prediction reliability requires more than one model

Using more than one model allows the detection of predictions that would be arbitrary under a single model. Fur-

	Blood Cell	Abdominal CT	OCT Scan	Breast Ultrasound
Accuracy B0	99.0 ± 0.1	95.2 ± 0.3	86.7 ± 1.3	84.7 ± 3.7
Accuracy B4	99.1 ± 0.1	95.9 ± 0.3	91.1 ± 0.8	89.1 ± 1.4
$\uparrow$ Acc	0.1	0.7	4.4	4.4
Δ APPA (raw / binarized)	-0.1 / -0.6	-0.1 / -0.6	2.7 / 9.9	12.8 / 12.8
Affected (raw / binarized)	3.9 / 1.9	16.0 / 5.4	25.9 / 14.3	66.0 / 30.8

Table 1: Performance and stability across model capacities. The first two rows report *mean test accuracy* ($\pm$ std) in % for EfficientNetB0 (relatively low capacity) and B4 (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 change in mean accuracy ($\uparrow$ Acc), the change in mean APPA, and the corresponding proportion of samples whose APPA values differ between models (Affected). Raw denotes the continous APPA values; binarized APPA is set to 1 when raw $\text{APPA} = 1.0$, 0 otherwise. When higher-capacity models achieve higher accuracy, stability increases (higher APPA); otherwise, the aggregated effects are minimal.

ther, we can abstain from predicting when there is insufficient consensus (Black, Leino, and Fredrikson 2022), and flag ambiguous and potentially harmful predictions. The capacity to abstain does not come without costs: while it improves reliability and robustness, it requires multiple models and may reduce coverage, as not all samples receive predictions. To evaluate the effectiveness, we compare *coverage rates*, *correctness* (across covered samples), and *predictive stability* across single models and ensembles consisting of two, five, and ten models. To assess predictive stability, we compute the *expected pairwise agreement* between two distinct models or ensembles of equal size (see Appendix F for more details). Intuitively, this metric captures how often two equally plausible models/ ensembles agree, complementing existing measures of disagreement and discrepancy (Black, Raghavan, and Barocas 2022; Marx, Calmon, and Ustun 2020; D'Amour et al. 2022). For the ensembles, we apply a conservative decision rule: a prediction is made *only if all constituent models agree*; otherwise, the ensemble abstains. Note that we can assess agreement between ensembles only on samples where both ensembles made a prediction; we additionally display the fraction of samples that cannot be evaluated because they are not predicted by the alternative ensemble in (purple, low opacity in Figure 3).

As shown in Figure 4, using ensembles reduces predictive multiplicity across all datasets and model architectures. The main exception is the Breast Ultrasound dataset where stability decreases (see Figure F and Figure 8 in the Appendix for more details). The reported stability values for ensembles (in comparison to single models) should be interpreted as conservative estimates, since samples for which another ensemble of the same size would abstain from making a prediction are excluded; for these cases, stability cannot be assessed. With respect to accuracy, and in line with prior work, ensembles improve predictive performance (see, for example, Dietterich (2000) for an overview). This improvement is reflected in an increased proportion of correct predictions among the covered samples, as indicated by the expansion of the darker regions within the green bars in Figure 4 (see also Figure 8 in the Appendix for a sample-wise analysis). This demonstrates that ensembles not only reduce predictive multiplicity (compared to an alternative, equally plausible predictor), but also concentrate their predictions on cases

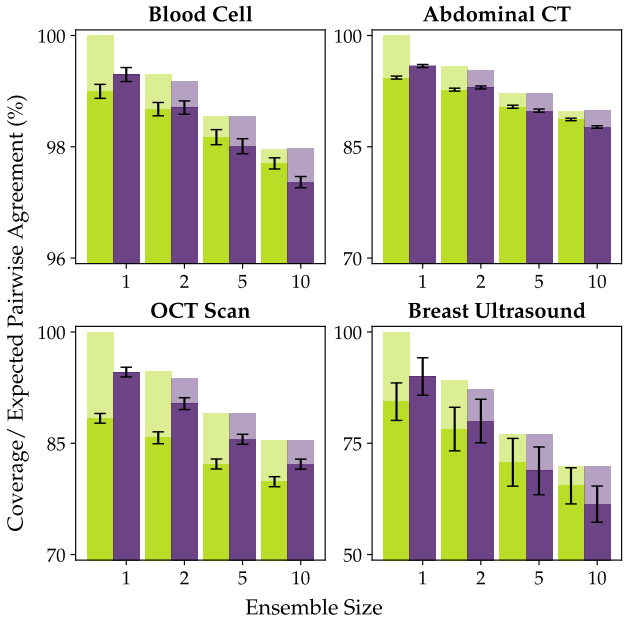

Figure 4: Ensembles substantially reduce predictive multiplicity. Coverage rate (green) and expected pairwise agreement (purple) of the test set across ensembles of sizes 1, 2, 5, and 10. For *coverage*, the darker segment indicates the correctly covered samples. For *expected pairwise agreement*, the darker segment represents stable samples, while the lighter segment corresponds to samples for which no judgment can be made due to missing coverage from the other ensemble. Error bars show standard deviation.

where they are more likely to be correct.

To sum up, using ensembles instead of single models improves both the stability of predictions (with respect to alternative models or ensembles) and predictive accuracy. This improvement comes at the cost of reduced coverage. Notably, ensembles improve correctness only for samples that are ambiguous across ensemble members (i.e., those that suffer from predictive multiplicity); they do not aid in detecting samples that are misclassified consistently.

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 Rashomon effect consider only models with small performance differences as functionally equivalent. Using the hypothesis testing 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 aligns with previous work (Jordan 2024), which demonstrates that apparent performance variance across finite test sets often reflects finite-sample noise rather than true generalization differences—a problem amplified in data-limited domains like medicine. Consequently, sampling variability may distort the empirical Rashomon set by overfitting idiosyncrasies of the finite test set. If performance differences between plausible models grow too large, this may indicate that the problem itself is ill-suited for reliable classification. These findings highlight important directions for future research and call for rethinking statistical and methodological practices in model evaluation (see also Appendix C).

The existence of multiple models with equal quality renders the selection of a single, supposedly "best"-performing model effectively arbitrary, undermining the justification for the model and its predictions. Relying on one such model exposes some patients to effectively arbitrary and potentially harmful outcomes—those whose predictions would differ under another, equally valid model. Creel and 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[1] and to consistent, evidence-based, and non-arbitrary treatment (Olejarczyk and Young 2024; Varkey 2021). While we identified technical sources of arbitrariness, questions such as *to what extent should 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—such as the EU Artificial Intelligence Act and related data governance provisions—already capture these concerns warrants closer examination. If left unaddressed, model multiplicity, and particularly predictive multiplicity, may complicate the ethical and legal justification of AI-assisted decision-making in healthcare.

Contrary to Black, Raghavan, and Barocas (2022), we find that higher-capacity models can enhance both accuracy and predictive stability. We agree, however, that "accuracy is not an antidote to multiplicity, and model selection cannot simply be reduced to accuracy-maximization" (Black, Raghavan, and Barocas 2022) – even more, accuracy-maximization is an insufficient criterion for model selection. However, 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 beyond the single-model paradigm. Using ensembles and predicting only on consistent samples improves both accuracy and stability. An ensemble with selective abstention—deferring unstable cases to human review—eliminates measurable multiplicity and aligns predictive confidence with clinical accountability. Our consensus-based approach, which requires unanimous model agreement, is intentionally simple. More sophisticated methods, such as the statistical consistency test by Black, Leino, and Fredrikson (2022), may be better suited for deployment. The optimal agreement criterion and ensemble size should depend on the application domain, balancing computational cost and desired confidence.

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 that, 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 landscape—rather than a nuisance—calls for a

[1]While the form and legal status of informed consent vary across cultural and regulatory contexts, its foundation in patient self-determination is broadly acknowledged (Angell 1988). Arbitrariness in model outcomes risks violating this principle, since patients cannot meaningfully consent to decisions based on shifting rationales among equally valid models.

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. By acknowledging and characterizing predictive multiplicity, we seek to catalyze the development of diagnostic machine learning systems that are not only accurate, but also robust, equitable, and trustworthy.

A Generalization across Architectures

Figures 5, 6, and 7 present results across four architectures (ResNet50, GC ViT, EfficientNetB2, and ConvNeXtBase). Overall, the findings described in the main text generalize across architectures. The only notable difference is that two models from the Abdominal CT/ConvNeXtBase combination would have been excluded from the empirical Rashomon set due to their performance (see Figure 5). Interestingly, based on accuracy alone, these models would likely appear functionally equivalent—underscoring the value of a formal approach for verifying model-quality equivalence.

B General Methodology

The following section describes the datasets, model architectures, and training procedures in greater detail. We used four 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), and OCT Scan (Kermany et al. 2018). See Table B for an overview of dataset properties. All experiments rely on the official training, validation, and test splits to ensure comparability with prior work (see Yang et al. 2023).

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

Table 2: Number of training, validation, test samples and classes for each dataset.

We evaluated four model architectures ResNet50 (He et al. 2016), GC ViT (Hatamizadeh et al. 2023), EfficientNet(Tan and Le 2019) (variants B0 and B4 in subsection 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 and 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 induces 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 and 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 and 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. We used sparse categorical cross-entropy for all other datasets. Classification accuracy served as the primary performance metric. Table B 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.

C 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 multiple equally performing models is particularly relevant with respect to their effect and consequences in the real world. Several works highlight the opportunities that model multiplicity offers (Rudin et al. 2024), like the selection of fairer (Dutta et al. 2020; Wick, Tristan 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, Broelemann, and Kasneci 2020), the risk of fair-washing explanations (Anders et al. 2020) or fairness metrics (Black, Gillis, and Hall 2024), and predictive multiplicity (Marx, Calmon, and Ustun 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 (i.e., a cause for model and predictive multiplicity)

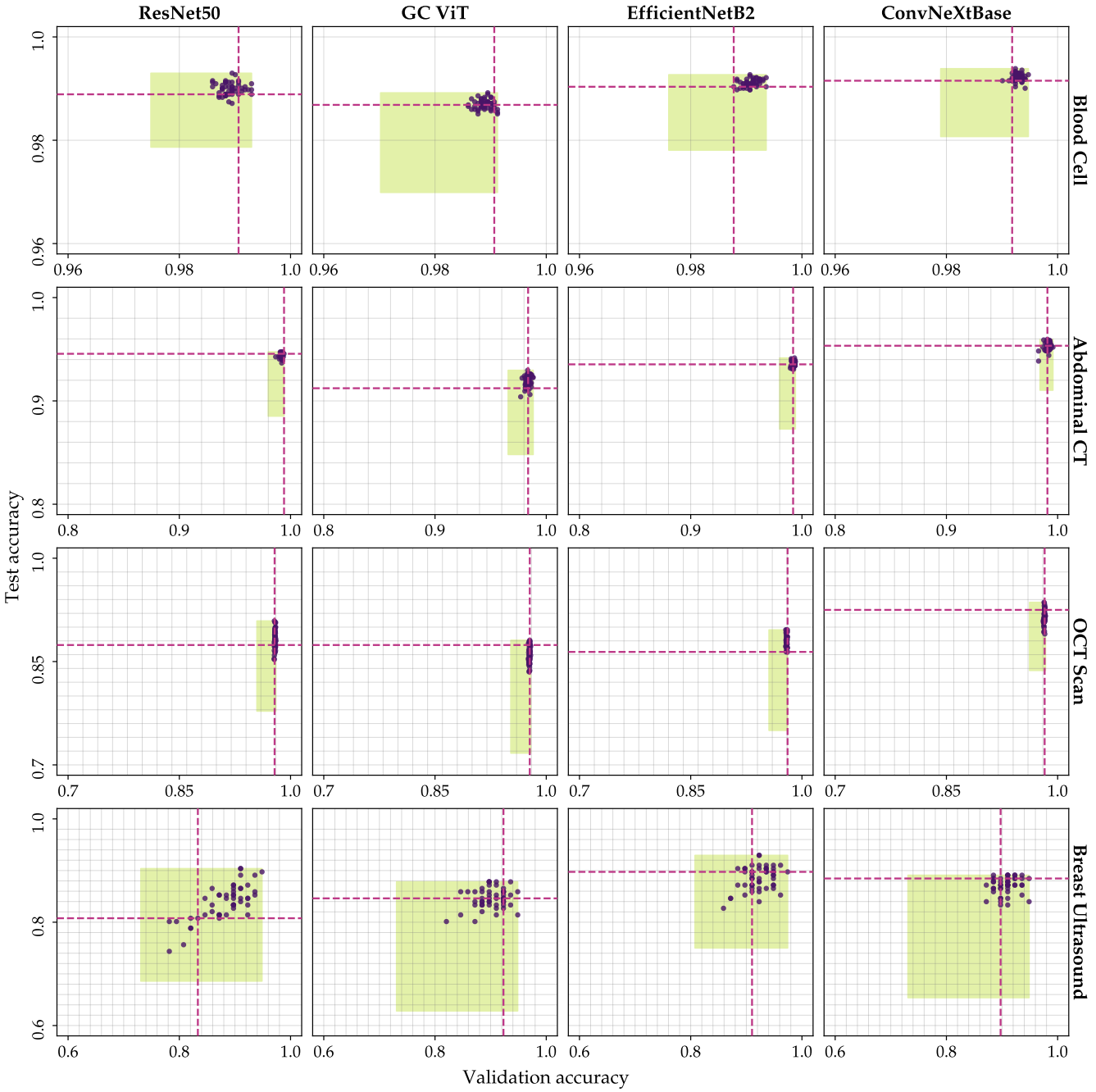

Figure 5: Variation in accuracy within the empirical Rashomon set. Each plot corresponds to a dataset/architecture combination, with points representing models trained from different random initialization. Models within the green region cannot be distinguished at the 95% significance level. Notably, only two models from the Abdominal CT/ConvNeXtBase combination show a statistically significant performance difference—an effect that would likely remin 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).

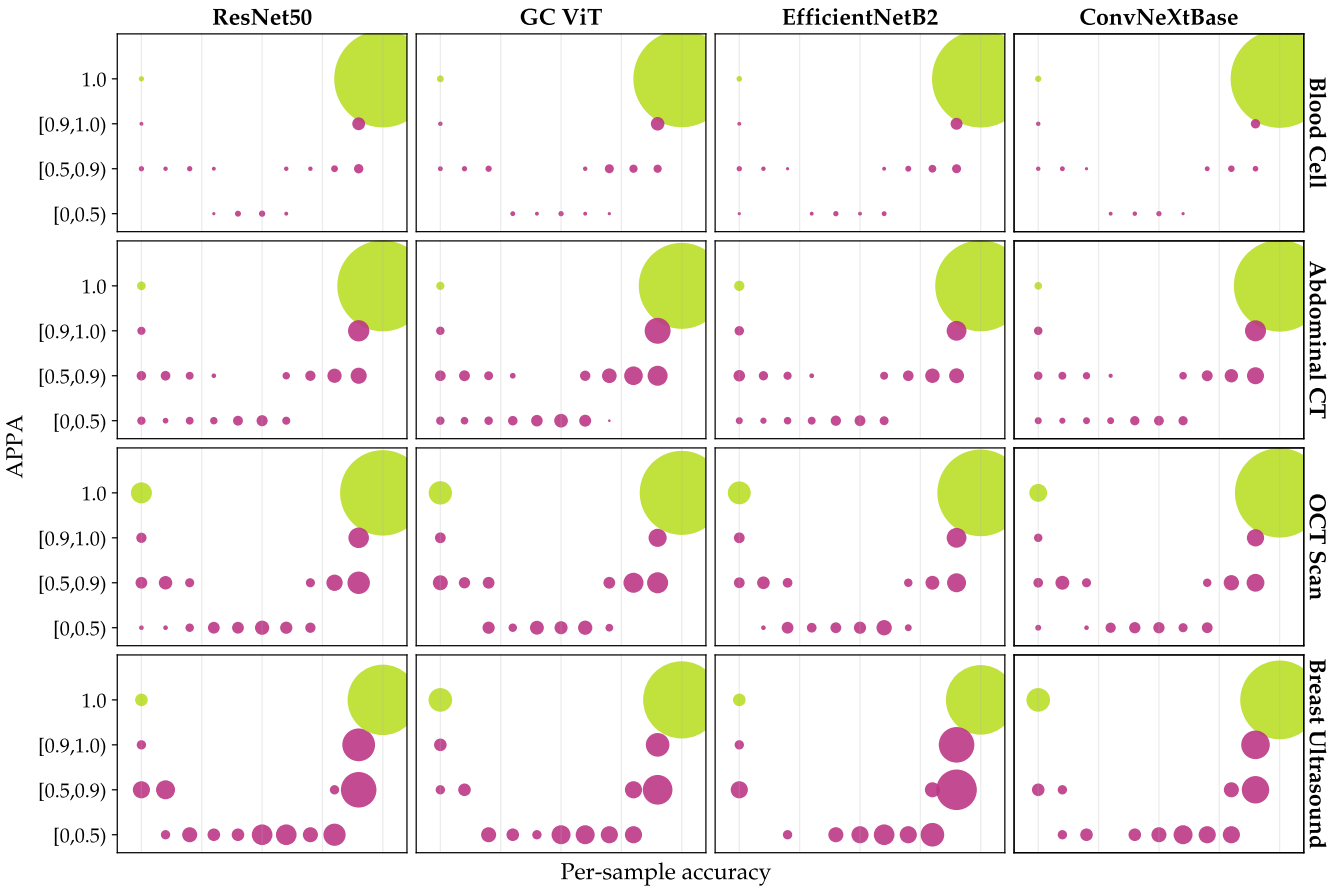

Figure 6: Prediction stability as a function of accuracy. Each plot corresponds to a dataset/architecture combination, for visual clarity, samples are binned by accuracy and APPA. Point size reflects their relative frequency, normalized by dataset size for comparability. Color encodes stability (green: $APPA - 1.0$, i.e., stable samples, pink: $APP < 1.0$). Across all datasets and architectures, samples frequently predicted correctly show high APPA. Samples in the in the top-left corner are consistently misclassified, indicating systematic bias.

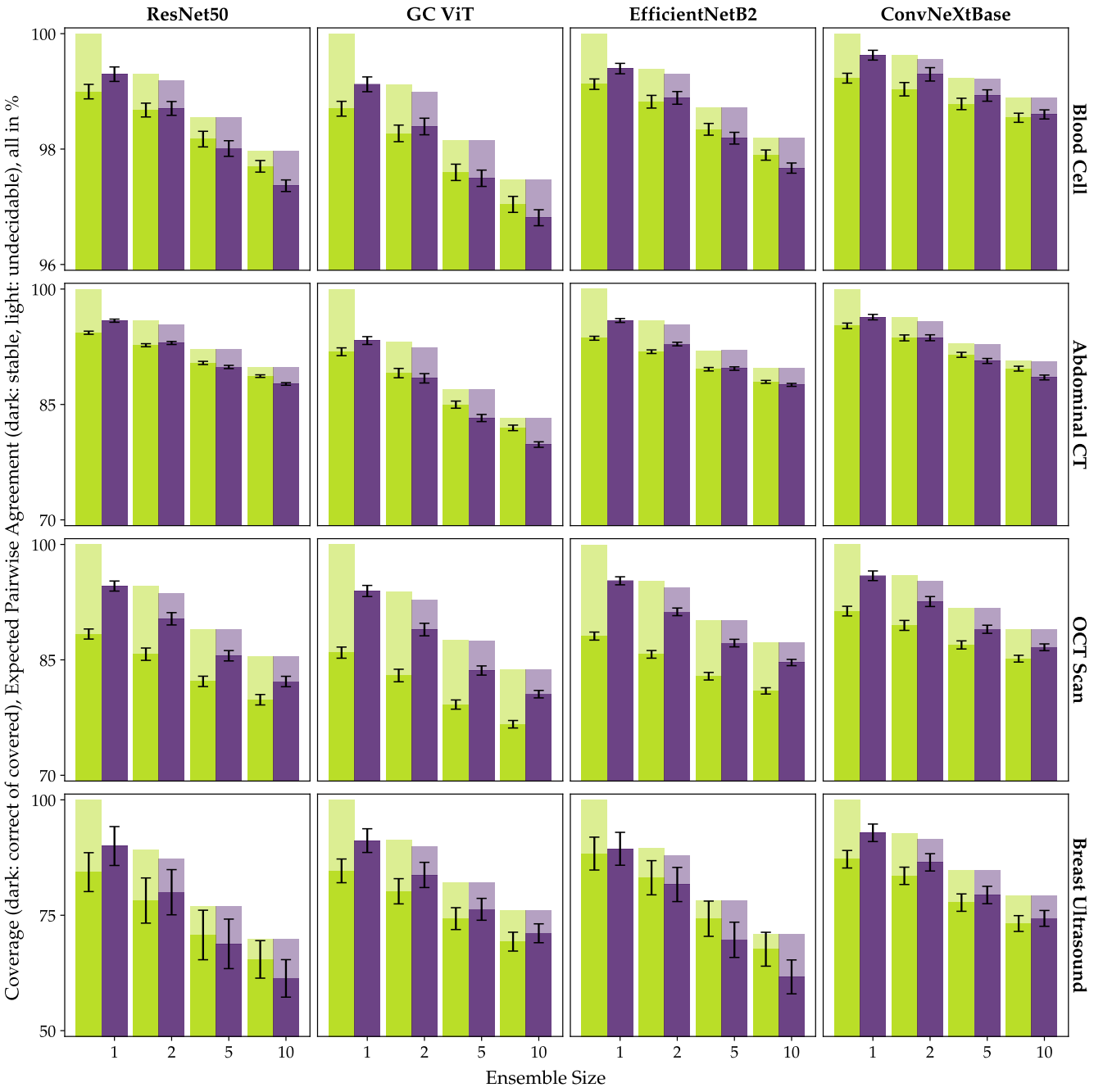

Figure 7: Ensembles substantially reduce predictive multiplicity. Coverage rate (green) and expected pairwise agreement (purple) of the test set across single models and ensembles of sizes 2, 5, and 10. For *coverage*, the darker segment indicates the correctly covered samples. For *expected pairwise agreement*, the darker segment represents stable samples, while the lighter segment corresponds to samples for which no judgment can be made due to missing coverage from the other ensemble. Error bars show standard deviation.

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

Table 3: For each dataset (columns) and model architecture (rows), we report the initial model's test performance and, in parenthesis, the mean test accuracy across 50 models from the empirical Rashomon set. Best-performing initial models per dataset are highlighted in bold. Benchmark results are from Yang et al. (2023) (highest accuracy across seven architectures).

(D'Amour et al. 2022) without addressing how such instability affects individual predictions, or proposed bootstrapping as a remedy to predictive multiplicity (Riley et al. 2023; Riley and Collins 2023).

Bootstrapping trains ensemble members on different resampled versions of the training data to induce diversity. While effective for deterministic models lacking intrinsic sources of randomness, modern machine learning models inherently incorporate stochasticity, e.g., through random initialization, data shuffling, and optimization dynamics. Further, bootstrapping is computationally intensive and often not economically sensible; Riley and Collins (2023) recommend the training of at least 200 models. Moreover, it is impractical in medical applications, where datasets are typically small due to privacy constraints, regulatory limits, and the high cost of expert annotation (Kaissis et al. 2020). When models have multiple local optima, bootstrapping can even degrade performance, as each ensemble member only observes a fraction of the (already limited) training data. Empirically, Lakshminarayanan, Pritzel, and Blundell (2017) found that training each model on the entire dataset—with random initialization and data shuffling—achieved better performance than bootstrapping. In summary, while bootstrapping is conceptually simple, ensembling full-data models offers a more practical and effective approach for deep learning models with inherent stochasticity and in data-limited medical settings. In our experiments (see Section 3), we found that ensembles of as few as five models with an abstention capability substantially reduce predictive multiplicity while improving accuracy.

D Numerical Rashomon Parameter Estimation

To identify the range of statistically indistinguishable models within the empirical Rashomon set, we determine the smallest error rate ϵ for which the CP confidence intervals of two models' empirical errors no longer overlap. This point marks the boundary at which we can reject the null hypothesis that both models have equal true error rates at significance level α (we choose $\alpha = 0.05$).

We compute ϵ numerically using a bisection search. Given the number of samples in the respective (finite) dataset n and ϵ_0, which we define as the lowest obtained error rate among the 50 models in the empirical Rashomon set, we first compute its upper CP bound $UB(\alpha, n\epsilon_0, n)$ (see Table B for the number of samples per dataset and split; see Table D for ϵ_0 and ϵ values). We then search for the smallest

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

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

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. Note that overlapping confidence intervals are no formal statistical test for evaluate the equality of means (e.g., Schenker and 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).

The Rashomon parameter ϵ is determined by the baseline error ϵ_0, the dataset size n and the confidence level $(1 - \alpha)$. Intuitively, a smaller α (i.e., higher confidence) leads to wider CP intervals, which in turn increase the indistinguishability region. A larger ϵ_0 (worse baseline performance) also enlarges the region, since performance differences must be larger before they become statistically meaningful. Finally, increasing n narrows the region, because more data reduces statistical uncertainty and makes small

differences detectable.

Model selection is often guided by *marginal* error rates, i.e., a model's overall accuracy, which is well-captured by the CP confidence intervals presented in the main body. McNemar's test (McNemar 1947) complements this perspective by assessing whether two models differ on the same instances: rather than comparing average error levels, it evaluates whether two models disagree on the *same* instances, thereby revealing differences that marginal errors may obscure.

We first apply McNemar's test to the *first* model—the one selected via hyperparameter search—and assess how many of the 49 alternative models in the empirical Rashomon set cannot be distinguished from it at the 95% significance level. Table 5 reports the proportion of alternative models that are statistically indistinguishable on the validation and test sets, that is, the proportion for which the null hypothesis cannot be rejected ($p > 0.05$). These results corroborate our conclusions in the subsection 3: the first model is not exceptional. Across datasets, 85-100% of models are statistically indistinguishable from it on the validation set, and the majority remain so on the test set (besides the Abdominal CT/ GCViT and OCT Scan/ EfficientNetB2). For the datasets Abdominal CT and OCT Scan, we see more variability, especially w.r.t. the test set, suggesting split instability. Overall, according to McNemar's test, the first model appears far from unique, reinforcing that model selection is highly underdetermined.

We next apply the same procedure using the *best* model among the 50 candidates as the reference model. Relative to the *first* model, the proportion of statistically indistinguishable models becomes smaller, slightly for Blood Cell, Breast Ultrasound and CheXpert, but more substantially for Abdominal CT and OCT Scan. For example, on Abdominal CT with ConvNeXtBase, only 16.3% of models are indistinguishable from the best model on the validation split, and 12.2% on the test split.

While varying model initialization results in models that perform better according to McNemar's test, even the "best" model is far from unique: in every combination, other models exist that cannot be distinguished from it, indicating that it does not represent a single "optimal" solution.

E APPA

We assess per-sample prediction stability using Adjusted Pairwise Prediction Agreement, which quantifies the probability that two models, sampled uniformly at random from the empirical Rashomon set produce the same predictions.

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

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

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 $\mathrm{PPA}_{\min}(M, K)$. 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 $\mathrm{APPA(x)}$ 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.

F Ensembling predictions

We evaluate the effectiveness of ensembles to reduce predictive multiplicity by comparing prediction stability and coverage rates pairwise between single models and ensembles of size two, five, and ten. Prediction stability is measured as the expected pairwise agreement between two distinct models (or ensembles of equal size) on the test set and averaged over 100 repetitions. To avoid zero-inflation, we draw model or ensemble pairs without replacement from the empirical Rashomon set. For ensembles, we apply a conservative decision rule: a prediction is made only if all constituent models agree; otherwise, the ensemble abstains.

Figure 8 shows per-sample distributions of coverage, stability (measured by APPA), and correctness for increasing ensemble sizes across all datasets (for ResNet50). For each dataset and ensemble size $k \in \{1, 2, 5, 10\}$, predictions are obtained from multiple random disjoint ensembles, and per-sample metrics are computed by aggregating over 100 ensemble draws. To enable distributional comparison, samples are sorted for each metric and ensemble size, and plotted as a function of the fraction of samples sorted by the respective metric. *Coverage* quantifies the probability that a sample receives a prediction from the ensemble. As expected, increasing ensemble size leads to lower coverage, indicating that larger ensembles are more conservative in issuing predictions. This is reflected by *stability*, measured via APPA, which captures the consistency of predictions across ensembles. As ensemble size increases, stability improves for a growing fraction of samples in all datasets. The increase in stability with more ensemble members is little for the Blood Cell dataset which exhibits uniformly high stability even for a single model, indicating that predictions are already consistent and that additional ensemble members provide only marginal gains. *Correctness*, defined as the probability of being correct conditional on making a prediction, reveals a more nuanced behavior. While increasing ensemble size generally improves correctness, the effect is heterogeneous across samples. In particular, the sorted correctness curves exhibit a change in curvature. For lower-quantile

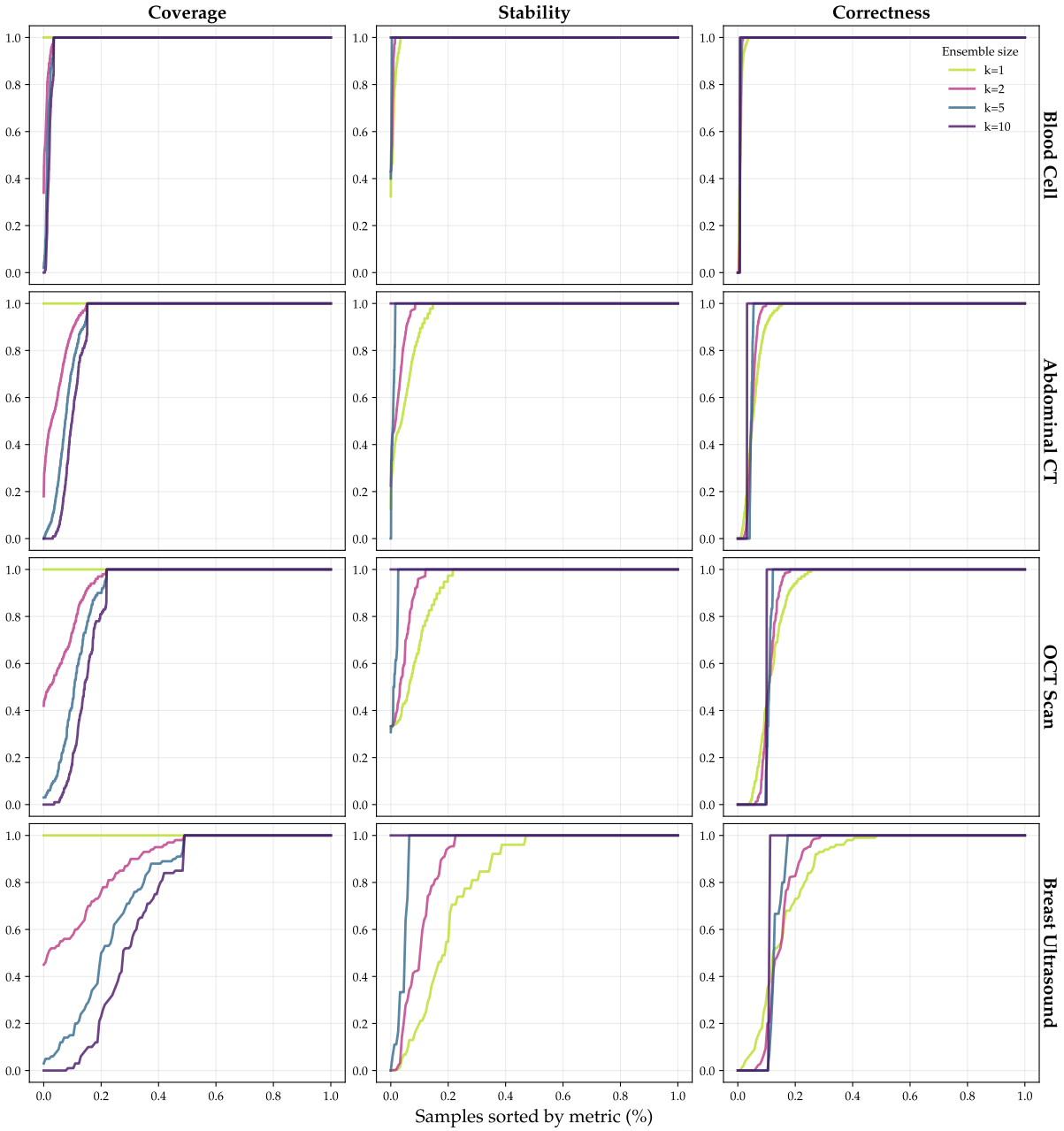

Figure 8: Per-sample coverage, stability, and correctness as a function of ensemble size across datasets, exemplary for ResNet50. For each dataset and ensemble size ($k \in \{1, 2, 5, 10\}$), we compute per-sample metrics over random disjoint single models or ensembles (of equal size) and visualize their empirical distributions by sorting samples according to the respective metric value. The x-axis denotes the fraction of samples, and the y-axis shows the corresponding metric value. *Coverage* measures the probability that a sample is predicted by the ensemble, *stability* (APPA) captures the agreement of predictions across sampled ensembles, and *correctness* measures the probability that a prediction is correct given that a prediction is made. Increasing ensemble size consistently reduces coverage while improving stability across datasets, with dataset-specific saturation behavior. Correctness exhibits a heterogeneous response to ensembling, revealing distinct regimes of samples with limited versus substantial gains from larger ensembles.

First model	ResNet50	GCViT	EfficientNetB2	ConvNeXtBase
Blood Cell	100.0 / 95.9	98.0 / 100.0	91.8 / 100.0	100.0 / 95.9
Abdominal CT	61.2 / 85.7	95.9 / 34.7	100.0 / 89.8	85.7 / 67.3
OCT Scan	100.0 / 63.3	100.0 / 55.1	83.7 / 40.8	95.9 / 51.0
Breast Ultrasound	83.7 / 75.5	100.0 / 100.0	100.0 / 98.0	100.0 / 100.0
Best model	ResNet50	GCViT	EfficientNetB2	ConvNeXtBase
Blood Cell	79.6 / 38.8	93.9 / 91.8	87.8 / 98.0	95.9 / 81.6
Abdominal CT	61.2 / 46.9	51.0 / 16.3	83.7 / 26.5	16.3 / 12.2
OCT Scan	85.7 / 18.4	71.4 / 34.7	83.7 / 36.7	63.3 / 16.3
Breast Ultrasound	63.3 / 40.8	87.8 / 85.7	83.7 / 59.2	100.0 / 91.8

Table 5: Percentage of models in the Rashomon set under McNemar's test for each architecture/dataset pair. For each entry, the left value corresponds to the validation split, and the right value corresponds to the test split. The upper block uses the first model as a reference, while the lower block uses the best model.

	ResNet50	GCViT	EfficientNetB2	ConvNeXtBase
Blood Cell	99.30 / 99.39	99.12 / 99.32	99.39 / 99.46	99.63 / 99.70
Abdominal CT	95.89 / 97.54	93.31 / 95.88	95.91 / 97.60	96.35 / 97.65
OCT Scan	94.59 / 96.24	93.95 / 96.19	95.28 / 97.04	95.93 / 97.42
Breast Ultrasound	89.99 / 87.85	91.17 / 93.48	89.40 / 86.97	92.86 / 93.73

Table 6: Percentage of stable predictions are higher for ensembles than for single models, except for ResNet50 and EfficientNetB2 on the Breast Ultrasound dataset. We report the mean percentage of samples with stable predictions for a single model and an ensemble of size ten models over 100 repetitions. For a single model, stability is computed over all test samples (full coverage) by comparing predictions to those of another single model. For an ensemble, stability is computed only on samples for which all ensemble members agree and is evaluated by comparison with an ensemble of the same size. Note, reported ensemble percentages are conservative estimates, as samples on which another ensemble of the same size would abstain from prediction are excluded; for these samples, no stability assessment can be made.

samples, correctness improves slowly with ensemble size, indicating errors that are largely irreducible and likely dominated by systematic bias or intrinsic ambiguity. In contrast, higher-quantile samples benefit substantially from ensembling, with correctness increasing rapidly as ensemble size grows.

Taken together, these results show that increasing ensemble size trades coverage for stability and selectively improves correctness. While ensembles enhance predictive consistency and accuracy for a substantial subset of samples, they also expose a class of samples for which errors remain largely irreducible, highlighting the importance of per-sample analysis when assessing ensemble behavior.

G Acknowledgments

SL received support from the Research and Development Program Information and Communication Technology Bavaria, DIK0444/03. KS received support from the German Ministry of Education and Research and the Medical Informatics Initiative as part of the PrivateAIM Project, from the Bavarian Collaborative Research Project PRIPREKI of the Free State of Bavaria Funding Programme "Artificial Intelligence – Data Science". This project was funded by the German Ministry of Education and Research under the PrivateAIM Project (reference 01ZZ2316C).

References

Acevedo, A.; Merino, A.; Alférez, S.; Molina, Á.; Boldú, L.; and Rodellar, J. 2020. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in brief*, 30: 105474.

Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; and Fahmy, A. 2020. Dataset of breast ultrasound images. *Data in brief*, 28: 104863.

Allen-Zhu, Z.; Li, Y.; and Liang, Y. 2019. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32.

Anders, C.; Pasliev, P.; Dombrowski, A.-K.; Müller, K.-R.; and Kessel, P. 2020. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*, 314–323. PMLR.

Angell, M. 1988. Ethical imperialism? *New England Journal of Medicine*, 319(16): 1081–1083.

Auer, P.; Herbster, M.; and Warmuth, M. K. 1995. Exponentially many local minima for single neurons. *Advances in neural information processing systems*, 8.

Bilic, P.; Christ, P.; Li, H. B.; Vorontsov, E.; Ben-Cohen, A.; Kaissis, G.; Szeskin, A.; Jacobs, C.; Mamani, G. E. H.; Chartrand, G.; et al. 2023. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84: 102680.

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

Black, E.; Gillis, T.; and Hall, Z. Y. 2024. D-hacking. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 602–615.

Black, E.; Leino, K.; and Fredrikson, M. 2022. Selective Ensembles for Consistent Predictions. In *10th International Conference on Learning Representations, ICLR 2022*.

Black, E.; Raghavan, M.; and Barocas, S. 2022. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 850–863.

Black, E.; Wang, Z.; Fredrikson, M.; and Datta, A. 2021. Consistent counterfactuals for deep models. *arXiv preprint arXiv:2110.03109*.

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

Chen, C.; Lin, K.; Rudin, C.; Shaposhnik, Y.; Wang, S.; and Wang, T. 2018. An interpretable model with globally consistent explanations for credit risk. *arXiv preprint arXiv:1811.12615*.

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

Coston, A.; Rambachan, A.; and Chouldechova, A. 2021. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, 2144–2155. PMLR.

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

D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M. D.; et al. 2022. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226): 1–61.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dietterich, T. G. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, 1–15. Springer.

Dutta, S.; Wei, D.; Yueksel, H.; Chen, P.-Y.; Liu, S.; and Varshney, K. 2020. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*, 2803–2813. PMLR.

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

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

Hatamizadeh, A.; Yin, H.; Heinrich, G.; Kautz, J.; and Molchanov, P. 2023. Global context vision transformers. In *International Conference on Machine Learning*, 12633–12646. PMLR.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jiang, Y.; Nagarajan, V.; Baek, C.; and Kolter, J. Z. 2021. Assessing generalization of SGD via disagreement. *arXiv preprint arXiv:2106.13799*.

Jordan, K. 2024. On the Variance of Neural Network Training with respect to Test Sets and Distributions. *ICLR*.

Kaissis, G. A.; Makowski, M. R.; Rückert, D.; and Braren, R. F. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6): 305–311.

Kermany, D. S.; Goldbaum, M.; Cai, W.; Valentim, C. C.; Liang, H.; Baxter, S. L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5): 1122–1131.

Kobylińska, K.; Krzyziński, M.; Machowicz, R.; Adamek, M.; and Biecek, P. 2024. Exploration of the Rashomon Set Assists Trustworthy Explanations for Medical Data. *IEEE Journal of Biomedical and Health Informatics*, 28(11): 6454–6465.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Marx, C.; Calmon, F.; and Ustun, B. 2020. Predictive multiplicity in classification. In *International Conference on Machine Learning*, 6765–6774. PMLR.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2): 153–157.

Olejarczyk, J. P.; and Young, M. 2024. Patient rights and ethics. *StatPearls [Internet]*.

Paes, L. M.; Cruz, R.; Calmon, F. P.; and Diaz, M. 2023. On the inevitability of the Rashomon effect. In *2023 IEEE International Symposium on Information Theory (ISIT)*, 549–554. IEEE.

Pawelczyk, M.; Broelemann, K.; and Kasneci, G. 2020. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*, 809–818. PMLR.

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

Riley, R. D.; Pate, A.; Dhiman, P.; Archer, L.; Martin, G. P.; and Collins, G. S. 2023. Clinical prediction models and the multiverse of madness. *BMC medicine*, 21(1): 502.

Rudin, C.; Zhong, C.; Semenova, L.; Seltzer, M.; Parr, R.; Liu, J.; Katta, S.; Donnelly, J.; Chen, H.; and Boner, Z. 2024. Position: amazing things come from having many good models. In *Proceedings of the 41st International Conference on Machine Learning*, 42783–42795.

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

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

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

Wick, M.; Tristan, J.-B.; et al. 2019. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32.

Yang, J.; Shi, R.; Wei, D.; Liu, Z.; Zhao, L.; Ke, B.; Pfister, H.; and Ni, B. 2023. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1): 41.