

On Arbitrary Predictions from Equally Valid Models

Sarah Lockfisch^{1*}, Kristian Schwethelm¹, Martin Menten¹,
Rickmer Braren², Daniel Rückert^{1,3,4}, Alexander Ziller^{1†},
Georgios Kaissis^{5*†}

^{1*}Chair for AI in Healthcare and Medicine, Technical University of Munich (TUM) and TUM University Hospital, Munich, Germany.

²Institute of Diagnostic and Interventional Radiology, Technical University of Munich, School of Medicine, Munich, Germany.

³Department of Computing, Imperial College London, UK.

⁴Munich Center for Machine Learning (MCML), Munich, Germany.

^{5*}Google DeepMind, London, UK.

*Corresponding author(s). E-mail(s): sarah.lockfisch@tum.de;
gkaissis@google.com;

†These authors contributed equally to this work.

Abstract

Model multiplicity refers to the existence of multiple machine learning models that describe the data equally well but may produce different predictions on individual samples. 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 the extent, drivers, and ramifications of predictive multiplicity across diverse medical tasks and model architectures, and show that even small ensembles can mitigate/eliminate predictive multiplicity in practice. Our analysis reveals that (1) standard validation metrics fail to identify a uniquely optimal model and (2) a substantial amount of predictions hinges on arbitrary choices made during model development. Using multiple models instead of a single model reveals instances where predictions differ across equally plausible models – highlighting patients that would receive arbitrary diagnoses if any single model were used. In contrast, (3) a small ensemble paired with an abstention strategy can effectively mitigate measurable predictive multiplicity in practice; predictions with high inter-model consensus may thus be amenable to automated classification. While accuracy is not a principled antidote to predictive multiplicity, we find that (4) higher accuracy achieved through increased model capacity reduces predictive multiplicity.

Our findings underscore the clinical importance of accounting for model multiplicity and advocate for ensemble-based strategies to improve diagnostic reliability. In cases where models fail to reach sufficient consensus, we recommend deferring decisions to expert review.

Keywords: Model Multiplicity, Predictive Multiplicity, Rashomon Effect, Rashomon Set

1 Introduction and Prior Work

Model multiplicity [9] describes the existence of many *plausible* models of the same dataset without a principled way to determine a single “true” 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 given performance metric (e.g., loss or accuracy), yet may differ substantially 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 any consideration for other, equally valid and plausible model options. Using such an arbitrary² 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 ultimately their treatment – may depend on the randomness in the training rather than on relevant properties of the patient’s data, raising critical concerns about the justification of the model and its predictions. We illustrate the problem in Figure 1.

The phenomenon of model multiplicity was first described by Breiman [11] in 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 [16]. 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 [42], like the selection of fairer [19, 48], more interpretable [13], or more robust models [16] without impairing predictive performance. Challenges arising from model multiplicity are, *inter alia*, the inconsistency of explanations [22, 40], the risk of fair-washing explanations [3] or fairness metrics [10], and predictive multiplicity [9, 39] – the phenomenon of equally optimal models producing conflicting predictions. While predictive multiplicity may be irrelevant (e.g., in low-stakes settings) or even desirable (for avoiding systemic exclusion) [15] in some cases, it poses a serious

¹In theory, the Bayes optimal predictor represents the true solution, i.e., the model that minimizes 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 are available.

²In 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 [14]. However, model multiplicity highlights the existence of multiple equally optimal models that yield different predictions for individual cases. In such scenarios, there is no *self-interested* reason to choose one optimal model over another [15].

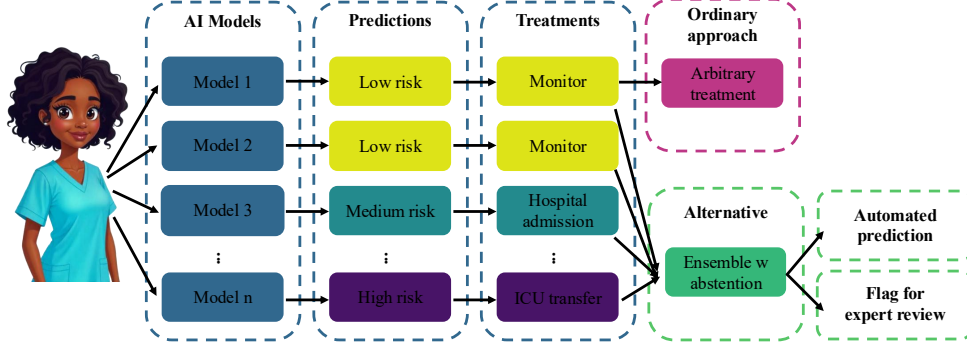


Fig. 1 Predictive multiplicity in clinical decision-making. To illustrate, we assume multiple models (Model 1 to n) to evaluate a patient’s data. The individual models achieve similar on-average accuracy and are thus equally valid model options. However, these models may produce different predictions (e.g., low, medium, high) for the same patient, which results in divergent treatment recommendations (e.g., monitor, hospital admission, ICU transfer). As a consequence, the patient’s care pathway can vary markedly – ultimately resulting in an arbitrary treatment – depending on which specific model is deployed. Rather than relying on a *single* model, we can adopt an *ensemble*-based approach, i.e., relying on the predictions of multiple models. If there is sufficient consensus among models, decisions may be automated; cases with insufficient agreement can be flagged for expert review.

challenge to leveraging machine learning in medical applications: When decisions differ across equally valid models, outcomes can depend on random factors in training rather than on meaningful data characteristics, rendering such decisions effectively arbitrary.

The arbitrariness of predictions can be analyzed using the *Rashomon set* [11], which comprises models that explain the data equally well. This set typically consists of models whose performance on a validation set is statistically indistinguishable, often measured by metrics such as average loss or accuracy. However, this approach relies on the assumption that validation performance reliably reflects generalization to the underlying data distribution. Jordan [27] demonstrated that for CIFAR-10 [32] and ImageNet [17], models exhibiting *significant variance* across two test *sets* drawn from the same distribution show minimal variance on the underlying test *distribution* itself. The variability on the test sets in performance is largely due to finite-sample noise arising from the limited size of the test sets – a problem that becomes even more pronounced with smaller datasets. Unlike CIFAR-10 and ImageNet, which contain (tens of) thousands of validation and test samples, medical datasets are often substantially smaller due to privacy concerns, regulatory constraints, and the high cost of expert annotation [29]. With such limited validation data, sampling variability can distort the composition of the Rashomon set. Models with genuinely similar generalization performance may appear different (according to a validation metric), while others with subtle but meaningful differences may seem indistinguishable. A narrowly defined Rashomon set therefore risks both false exclusion and false inclusion – potentially favoring models that overfit idiosyncrasies of the small dataset.

1.1 Overview of the Paper

Although the consequences of predictive multiplicity are particularly critical in the medical domain, systematic investigations into its risks remain limited.³ To the best of our knowledge, prior studies in the medical context have largely focused on leveraging the Rashomon set for trustworthy explanations [31] and exploring the role of underspecification in model robustness [16].

In this work, we address this gap through a comprehensive evaluation of the *empirical Rashomon set* across multiple medical modalities (abdominal CT, blood cell image, breast ultrasound, OCT scan, and X-ray) and diverse model architectures (ResNet50, GC ViT, EfficientNet, ConvNeXt). We follow the common practice of *empirically*⁴ exploring the Rashomon set for complex models by varying the random seed used to initialize model weights (e.g., [16, 20]). Unlike prior work, we do not constrain the Rashomon set to models within a narrow performance range. Instead, we include all models obtained through random seed variation, in line with the considerations outlined above. Concretely, we train 50 model instances for each configuration that differ only in the random seed used to initialize the final layer, resulting in a total of 1,400 models – 1,000 for our main experiments and 400 for investigating the impact of model complexity. These sets of models constitute the empirical Rashomon set for each configuration. Our contributions are twofold: (1) we demonstrate that the proposed approach yields a meaningful and useful characterization of the empirical Rashomon set to analyze the arbitrariness of individual predictions, and (2) we highlight the ubiquity of predictive multiplicity, its implications in the medical domain, and a strategy for addressing it. In more detail, this large-scale study yields the following conclusions that form the core contributions of our work:

1. **Validation performance is an unreliable indicator of generalization for models within the empirical Rashomon set.** Thus selecting a single seemingly best-performing model becomes an arbitrary choice among equally valid alternatives.
2. **Predictive outcomes for individual patients hinge on arbitrary decisions.** Deploying any single model leads to ambiguous and potentially harmful diagnoses.
3. **Ensembling combined with selective abstention can eliminate measurable arbitrary predictions in practice.** Predictions with high inter-model consensus may be amenable to automated prediction, while insufficient consensus prompts deferral to human experts instead.
4. **Higher model capacity that improves accuracy, reduces predictive multiplicity.** While accuracy alone is not sufficient for model selection, accuracy maximization can help to minimize arbitrary predictions.

³While medicine poses the additional challenge of small dataset sizes, the underlying issues are broadly relevant to other high-stakes decision-making scenarios and most findings and conclusions may be transferable.

⁴An exhaustive characterization of the Rashomon set is infeasible for complex models; in neural networks, for instance, its size is determined by the number of local minima, which grows exponentially with the number of parameters [4].

2 Results

We start at a typical point in the model development pipeline: the problem is clearly defined, and a hyperparameter sweep is performed to identify the best-performing model for deployment. Rather than stopping at this *single model*, we investigate the *set of alternative models* that we could have obtained through changing only the random seed – the resulting model instances thus represent equally probable outcomes of the training process.

Specifically, we fix all components of the training pipeline and vary only the random seed, which affects the initialization of the final layer in an ImageNet-pretrained model. Then we unfreeze and train the entire model. To systematically investigate the empirical Rashomon sets in medical imaging, we evaluate five diverse modalities – abdominal CT, blood cell images, breast ultrasound, OCT scans, and X-rays – spanning a range of classification difficulties (external baseline accuracies: 86.8–96.6%; see Table 1) and dataset sizes (two with small validation/test sets <1,000 samples, three with medium-sized sets <11,000; see subsection 5.1). Across the following three experiments, we use four widely adopted architectures: ConvNeXtBase [35], EfficientNetB2 [44], GC ViT [23], and ResNet50 [24]; for the last experiment we use three EfficientNet variants.

2.1 The Illusion of Consistency: Variability Within the Empirical Rashomon Set

As displayed in Figure 2, the validation performance of individual models does not reliably reflect their generalization behavior. In more detail, we observe three types of behavior: First, models trained on the OCT Scan dataset (row four in Figure 2) exhibit low variance on the validation set but high variance on the test set. While validation accuracy varies by only 0.2–0.3%, test accuracy fluctuates by 3–6% – nearly an order of magnitude more. Although higher test set variance is expected due to its smaller size (validation: 10,832 samples; test: 1,000 samples), differences in dataset size do not explain the observed variability. Taking the different dataset sizes into account, we simulate the expected variation in accuracy under the assumption of independent errors on individual samples (see Appendix Appendix A). This additional experiment shows that the observed validation variance is significantly lower than expected under the assumption of independent errors ($p = 0.001$), suggesting strong correlations in model errors on the validation set, indicating convergence to *seemingly* similar solutions. In contrast, test set variance aligns closely with expected variance under the assumption of independent errors. These findings align with those of D’Amour et al. [16], who show that model differences often become apparent only under “stress tests”. Here, the shift from a class-imbalanced validation set (which mirrors the training distribution) to a class-balanced test set acts as such a stress test, uncovering *substantial performance differences across models*. In summary, low validation variance may misleadingly suggest stability: it can obscure genuine variability in model behavior. This highlights the limitations of validation-based model selection and underscores the challenge of underspecification. Further, we find models where *validation performance partially reflects generalization behavior*, such as in Breast Ultrasound/ResNet50 (bottom-left of Figure 2). In this case, the random seed functions analogously to a

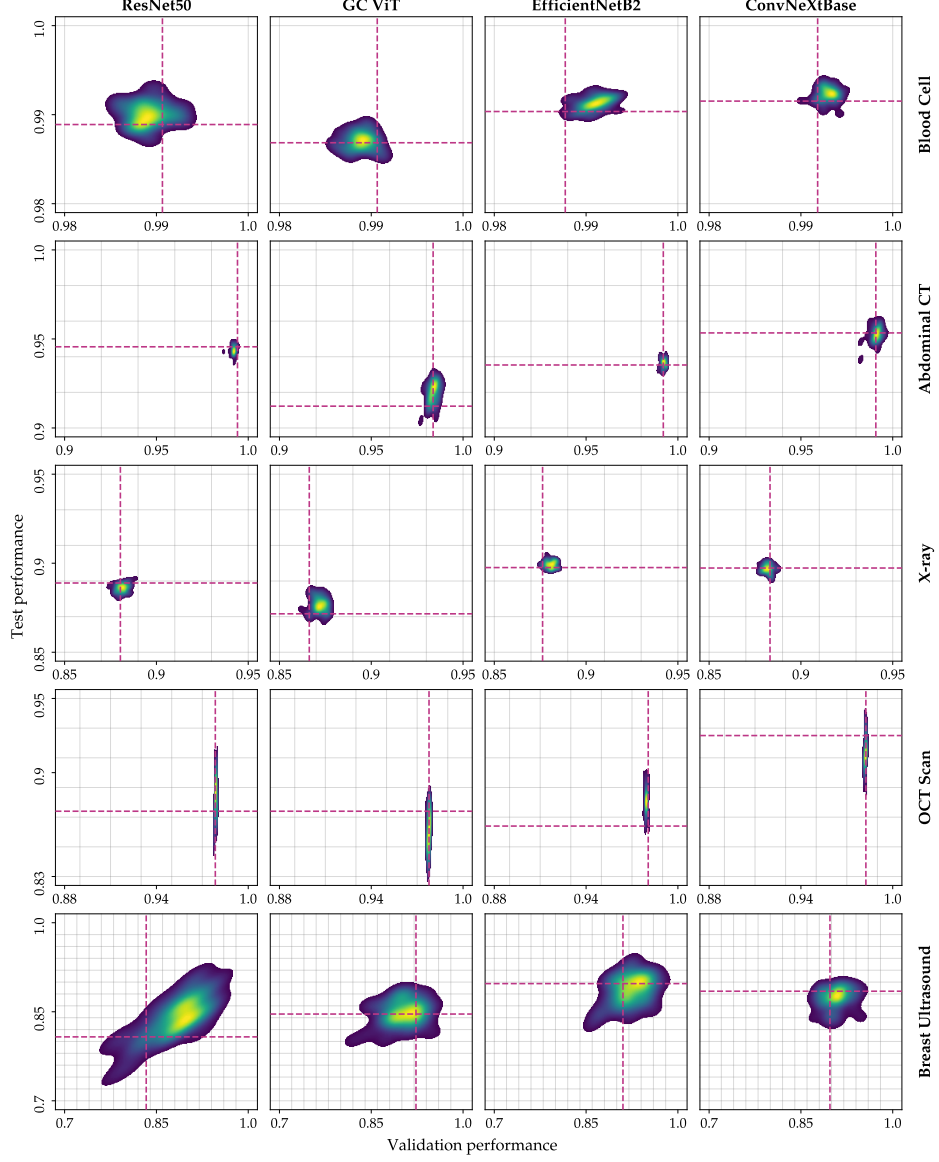


Fig. 2 Validation performance fails to reliably predict test performance. Each plot shows validation (x-axis) vs. test performance (y-axis) across datasets (rows) and architectures (columns), using ROC-AUC for X-ray and accuracy elsewhere. A kernel density estimate visualizes the distribution of model performance with color indicating density (● ● ● high to low). The dashed line (●) marks the model for which we performed a hyperparameter search. Axes are uniformly scaled per dataset (0.02 units per grid cell).

tunable hyperparameter, improving performance – a phenomenon which Jordan [27] terms “training with pathological instability”. However, even in this setting, validation

accuracy remains an unreliable predictor of generalization for most models. Lastly, in all other cases, *validation and test performance exhibit similar levels of variability*. More specifically, we observe cases where models perform statistically indistinguishably on the validation set and cases with measurable differences in validation performance, however, in both cases, variation in validation performance reflects little genuine variation in model quality as these differences are not reflected on the test set. In summary, for all dataset/model combinations, we find models that perform better/worse than average on the validation set but show the opposite behavior on the test set. Interestingly, the initial model (dashed line • in Figure 2) for which we performed the hyperparameter search is in no way “special” in terms of validation or test set performance relative to the other models. In other words, *better and worse models exist despite this being the model that has been optimized for*. Overall, validation performance proves to be an unreliable predictor of test performance across all analyzed settings.

These findings challenge the assumption that high validation performance is a reliable indicator of a model’s generalization ability. Since validation performance does not consistently predict test performance, selecting the model with the highest validation performance becomes an *arbitrary decision among several equally valid alternatives*. Crucially, even the initial model for which we perform a hyperparameter search is as good a draw as any other model from the set. Consequently, the standard model selection criteria – namely, choosing the model with the highest validation set performance – is not only inadequate but potentially harmful to the patients receiving inferior predictions.

Traditionally, the (empirical) Rashomon set comprises models with statistically indistinguishable performance (e.g., [25, 39, 43, 47]). However, we find that within this set, superior validation performance does not reliably translate to superior test performance. This aligns with findings from Jordan [27], who shows that models with significant variance on their test *sets* have little variance in performance on the underlying test *distribution*. Methodologically, we argue that the traditional definition of the Rashomon set may be too restrictive for identifying arbitrary decisions among equally valid models.⁵ In our case, including all models obtained through minor perturbations to the training pipeline provides a more comprehensive view of model multiplicity.

2.2 Individual predictions are arbitrary under a single model

So far, we have focused on the average performance metrics of the models within the empirical Rashomon set. We now turn to their predictions to examine the *arbitrariness of diagnostic outcomes when relying on any single model*. To quantify arbitrariness, we measure the proportion of test samples, i.e., *vis-a-vis* new patients, that receive conflicting predictions across multiple models. Following Marx et al. [39], we adopt the notion of *ambiguity*, which formally quantifies predictive multiplicity. Ambiguity is defined as the fraction of samples that receive different predictions from any competing

⁵Overall, the Rashomon set is a tool, and the appropriate definition thus depends on the objective. For instance, Kulynych et al. [34] avoid performance-based restrictions when analyzing model multiplicity under differential privacy constraints, where such filtering could compromise privacy. In contrast, Semenova et al. [43] focus on identifying simpler yet accurate models, for which broader inclusion would be misaligned with their goal.

classifiers.⁶ Unlike Marx et al. [39], who restrict competing classifiers (which comprise the Rashomon set) to those who perform similarly, we use all models obtained by small perturbations to the training pipeline irrespective of their validation performance following the reasoning outlined in the previous section.

As displayed in Figure 3, the number of data points affected by predictive multiplicity varies by dataset and model architecture, with some combinations being more prone to ambiguous predictions when considering multiple models from the empirical Rashomon set. Importantly, even in high-accuracy regimes (see, for example, the Blood Cell dataset in the top row in Figure 3), a non-trivial proportion of samples receive conflicting predictions, underscoring the risk of relying on a single model, especially if arbitrarily chosen. If any single model were to be deployed, a considerable number of patients would receive arbitrary predictions – predictions which are effectively determined by the “choice” of a random seed.

Before moving on to a potential solution for predictive multiplicity, we want to draw the attention to two important points. First, our choice to include all models results in a broader spectrum of performance levels. Accordingly, we examine how this variability influences predictive multiplicity. We use validation performance to incrementally add models to the set. Specifically, we compare three strategies: (1) *random* selection from the remaining models in the empirical Rashomon set, (2) adding in *ascending*, and (3) *descending* order of validation performance. For most datasets, *the order in which we add models has minimal impact*. The main exception is the Breast Ultrasound dataset (which has the greatest variation in performance). Surprisingly, adding the worst-performing models first in the Breast Ultrasound/ResNet50 setting results in fewer ambiguous predictions. More in-depth investigation reveals that these models predict the same samples incorrectly, resulting in high agreement but low accuracy. As more models are added (beyond eight), some errors are corrected, causing a step increase in predictive multiplicity. For GC ViT and EfficientNetB2, considering lower-performing models first leads to higher ambiguity; random and best-first strategies affect a similar proportion of samples. For ConvNeXtBase, model order has little effect and aligns with the results in the previous section: variability in accuracy across the empirical Rashomon set is smallest here compared to the other architectures. In general, model order has little to no impact on most tasks. However, sensitivity to model order – such as observed in the Breast Ultrasound dataset – indicates high variance within the Rashomon set. This could serve as a rough signal that the training setup fails to learn consistent patterns in the data, suggesting that such tasks or configurations may not yet be suitable for reliable automated classification.

Second, we highlight the connection between accuracy and predictive multiplicity. While related, they are of distinct concern and can vary independently. Models can be highly accurate yet still disagree on individual predictions; *vice versa*, low-accuracy models may exhibit consistent agreement (as pointed out above). *Predictive multiplicity is not about whether a prediction is wrong, but whether it could easily have been different under an equally valid model*. Still, accuracy and predictive multiplicity are related: Intuitively, error provides the “space” for disagreement between models: the higher the

⁶See Appendix Appendix C for a relaxed version where ambiguity requires a certain percentage of classifier to disagree.

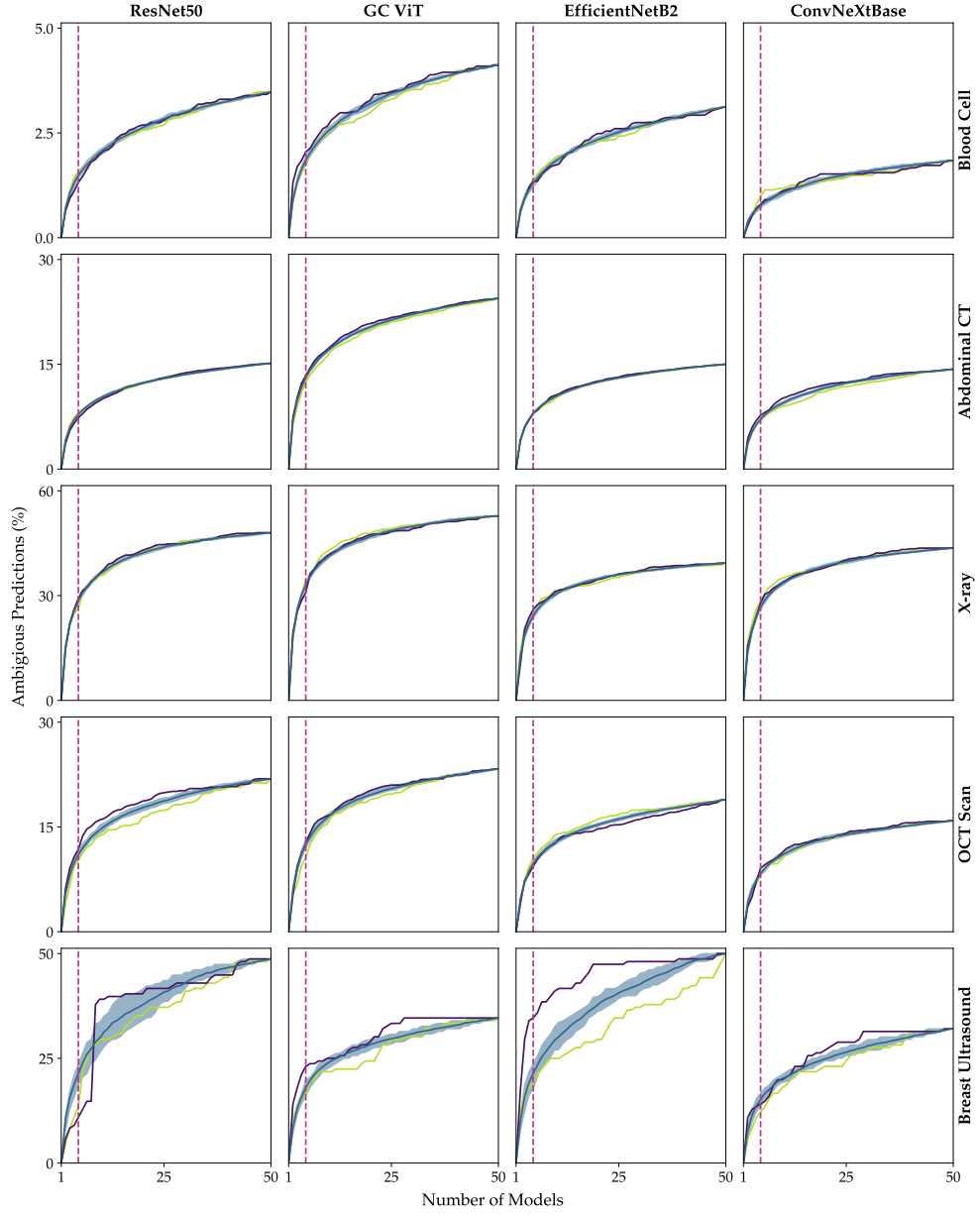


Fig. 3 A considerable amount of predictions depend on the initial model weights. As expected, the percentage of predictions that receive ambiguous predictions (y-axis) increases as more models are considered (x-axis) across all datasets (rows) and model architectures (columns). However, when considering up to five models (dashed line ●), the order in which we add models has little or no effect in most cases. We add models (●) *randomly* (shaded area: 25–75% IQR), by (●) *ascending* and (●) *descending* validation performance.

error rate (and the more models we consider), the more disagreement we might expect. Take the highly accurate Blood Cell/ EfficientNetB2 case where models in the empirical Rashomon set achieve a mean accuracy of 99.1% with a standard deviation of 0.1% on the validation and test sets. Under the assumption of *random, independent* errors, we would expect approximately 4.5% ambiguous predictions when considering five models, and 36.2% for 50 models.⁷ However, this dramatically overestimates the ambiguous predictions we actually observe: 1.3% for five models and 3.1% for 50 models. This discrepancy highlights a key insight: the errors made by the models in the empirical Rashomon set are *not independent*, but correlated and systematic. We conjecture that predictive multiplicity reveals a part of epistemic uncertainty: the models in the empirical Rashomon set are equally plausible explanations of the data; the models’ disagreement reflects their uncertainty.

2.3 Prediction reliability requires more than one model

In the following, we analyze the *stability of individual predictions* of single models and ensembles, which consist of multiple models from the empirical Rashomon set. Using an ensemble – rather than a single model – offers a key advantage: it enables the system to flag ambiguous and potentially harmful predictions. In such cases, decisions can be deferred to human experts, such as clinicians, instead of relying on an automated system. While this improves reliability and robustness, it comes at the cost of requiring additional models and potentially reducing sample coverage, as predictions are not made on all samples. In contrast, a single-model setup lacks this capacity to express uncertainty and abstain from potentially unreliable predictions.

We define the *expected pairwise agreement* as the agreement on held-out data in expectation between two distinct single models or ensembles of the same size randomly drawn from the empirical Rashomon set. Intuitively, the expected pairwise agreement captures how often two equally plausible models (or ensembles of the same size) are expected to make the same prediction on the same data points. Our definition aligns closely with prior notions of disagreement – its conceptual inverse – particularly as the complement of pairwise disagreement and discrepancy [9, 16, 39].⁸

Concretely, we calculate the expected pairwise agreement for single models and ensembles consisting of two and five models over 100 repetitions (and additionally report the corresponding standard deviation). We uniformly sample a baseline model or ensemble of the respective size from the empirical Rashomon set to represent plausible outcomes of standard model or ensemble training procedures. To estimate the pairwise agreement, we sample a competing model or ensemble (of the same size) from the *remaining* empirical Rashomon set and compute the percentage of matching predictions on the test dataset. Note that this approach underestimates predictive multiplicity, as

⁷We estimate ambiguity as $1 - p^M$, where p is the mean accuracy and M the number of models. While this assumes identical accuracies, it remains a close approximation when variance is small. In reality, the expected joint accuracy $\mathbb{E}[\prod_{i=1}^M p_i]$ is slightly lower than p^M , resulting in a slightly higher actual ambiguity.

⁸(Expected) pairwise disagreement [9] is defined over a distribution of models; in contrast, we compute it empirically using a constrained Rashomon set of 50 models and sample distinct model/ensemble pairs. Discrepancy [39] requires a baseline classifier and a threshold for model inclusion – neither of which can be meaningfully determined in our setting. As outlined in Section 2.1, the initial model for which we performed a hyperparameter search is neither superior to the other models in the empirical Rashomon set nor uniquely representative of the broader solution space; using a threshold may further exclude models with superior generalization performance.

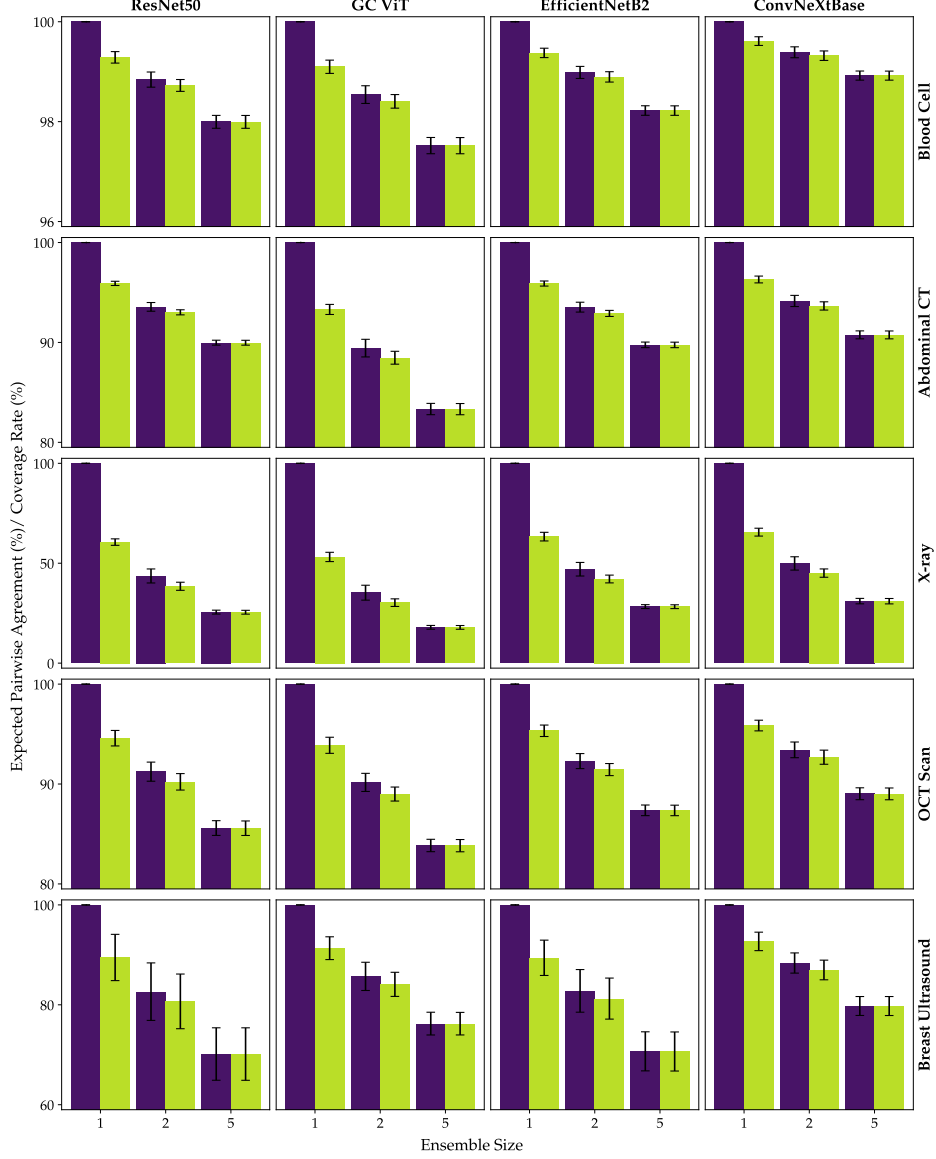


Fig. 4 Ensembles substantially reduce predictive multiplicity. We report coverage rate (●) for unanimous agreement among ensemble members and predictive multiplicity as measured by expected pairwise agreement (●) (y-axis) for ensembles of size 1 (i.e., single models), 2, and 5 (x-axis). Note: Expected pairwise agreement is computed over the *entire test set*, while coverage reflects the proportion of samples where all ensemble members agreed (i.e., a prediction was made). When pairwise agreement equals coverage, it implies that two independently drawn ensembles produce identical predictions on all covered samples – indicating no predictive multiplicity. Error bars represent the corresponding standard deviation.

we compare ensembles with ensembles and only measure coverage/agreement where both ensembles make predictions. This can lead to overly optimistic results due to abstentions in ensemble predictions. We use sampling without replacement to avoid the zero-inflation issue. For the ensembles, we adopt a conservative decision rule: a prediction is made *only if all constituent models agree*; otherwise, the ensemble abstains.

In Figure 4, we report the expected pairwise agreement (●) and the coverage (●) defined as the proportion of test samples on which the ensemble reaches full agreement and thus produces a prediction. Importantly, we present the percentage of expected pairwise agreement as the *percentage on all test samples*, regardless of abstention. This enables direct visual comparison across ensemble sizes, including single models that, by construction, never abstain. For example, if an ensemble of five models exhibits the same level of expected pairwise agreement and coverage, this means that on every sample where a prediction was made, two independently drawn ensembles of the same size produced the same output in expectation – implying no observed pairwise disagreement on this subset of test samples. Ensembles of size five achieve near-perfect expected pairwise agreement on the subset of test samples for which all members predict the same label, across all evaluated datasets and model architectures. We do not report accuracy in this figure, as it is well established that model ensembles improve predictive performance (see, for instance, [18] for a review). Our focus lies instead on predictive multiplicity – a distinct concern to ensure consistent predictions to guide treatment decisions appropriately. Nonetheless, predictive multiplicity must be considered in relation to accuracy: eliminating predictive multiplicity can lead to consistent yet incorrect predictions. Thus, consistency is only desirable when it leads to correct predictions. Notably, deep ensembles combined with an abstention mechanism can achieve both high consistency and near-perfect accuracy (see Appendix Figure 7 and Figure 9).

To illustrate the effect of ensembles with abstention on predictive stability, we contrast two cases: Blood Cell/ConvNeXtBase, where single models already show low predictive multiplicity, and Breast Ultrasound/ResNet50, where multiplicity is more pronounced. We report normalized expected agreement as the percentage of consistent predictions among *covered samples* – those for which the ensemble produces a prediction – and additionally provide the corresponding coverage rate. For Blood-Cell/ConvNeXtBase, single-model agreement is already high (mean = 99.60%, std = 0.10). A two-member ensemble increases agreement to 99.95% (std = 0.02) with 99.38% coverage (std = 0.06). In contrast, Breast Ultrasound/ResNet50 starts with lower single-model agreement (mean = 88.53%, std = 3.61). A two-member ensemble improves agreement to 98.43% (std = 1.61) with 87.92% coverage (std = 1.96). For both cases, a five-member ensemble achieves perfect expected pairwise agreement (100.00%, std = 0.00) at the cost of reduced coverage (98.97% coverage for Blood Cell/ConvNeXtBase and 71.34%, std = 3.76 for Breast Ultrasound/ResNet50). In summary, while larger ensembles can eliminate predictive multiplicity, they may do so at the expense of coverage, especially in settings with higher initial variability.

2.4 Multiplicity Dynamics in Overparameterized Models

Highly overparameterized models have driven major advances in machine learning and are the dominant approach across a range of tasks, including image classification [12, 28, 33, 45]. These models often have more trainable parameters than training samples, allowing them to converge to a variety of functionally different solutions. This enhanced capacity, however, also entails the risk of greater model multiplicity. In other words, while overparameterization can improve predictive performance, it allows for more equally accurate solutions, amplifying the potential for predictive multiplicity and, thus, the arbitrariness of individual predictions. Black et al. [9] theoretically demonstrate that improving accuracy by increasing variance – such as through the use of more complex model classes – can also lead to increased predictive multiplicity, a finding that aligns with their empirical results [6, 7]. Their analysis compares low- and high-complexity models (e.g., linear and deep learning models). In contrast, we examine predictive multiplicity *within* the already overparameterized regime, focusing on high-complexity networks of varying sizes. Specifically, we study how *model complexity*, *accuracy*, and *predictive multiplicity* are related using a controlled family of EfficientNet variants⁹: EfficientNetB0 (with 5.3M parameters), EfficientNetB2 (9.2M), and EfficientNetB4 (19.5M). Across all evaluated datasets (consisting of 546 - 97,477 training samples), these models operate in the overparameterized regime, enabling us to systematically explore how increased capacity affects prediction stability.

We begin by analyzing how *model capacity* affects predictive multiplicity. To quantify predictive multiplicity, we first measure the percentage of ambiguous predictions on the test set by incrementally adding models from the empirical Rashomon set (randomly selected, following the procedure in Section 2.2; we report the mean and the 25-75% IQR over 50 runs). Our results (first column in Figure 5) show that ambiguity remains relatively stable across model capacities for the Blood Cells and Abdominal CT datasets. In contrast, the OCT Scan dataset exhibits higher ambiguity for smaller models, which decreases as model capacity increases. A similar trend appears in the Breast Ultrasound dataset, though ambiguity levels plateau for EfficientNetB2 and B4. In summary, increased model capacity does *not* lead to higher ambiguity in the overparameterized regime. On the contrary, larger models tend to produce equal or lower predictive multiplicity.

The relationship between model capacity and predictive multiplicity may be influenced by *accuracy*: higher-capacity models can learn more accurate solutions, potentially reducing multiplicity. To test this, we assess whether accuracy moderates the relationship between model capacity and predictive multiplicity. We quantify predictive multiplicity using pairwise agreement (as defined in Section 2.3), computed as empirical averages – by measuring the percentage of matching predictions between two randomly selected models (or ensembles) from the Rashomon set on the test set. Our results (second column in Figure 5) show that changes in pairwise agreement are largely explained by changes in accuracy. Specifically, (1) when higher model capacity leads to higher accuracy, we observe higher pairwise agreement. On the OCT Scan and Breast Ultrasound datasets higher model capacity leads to improvements in accuracy and higher pairwise agreement. (2) When accuracy remains unchanged, so does

⁹We chose the EfficientNet architecture for its balance of accuracy and computational efficiency.

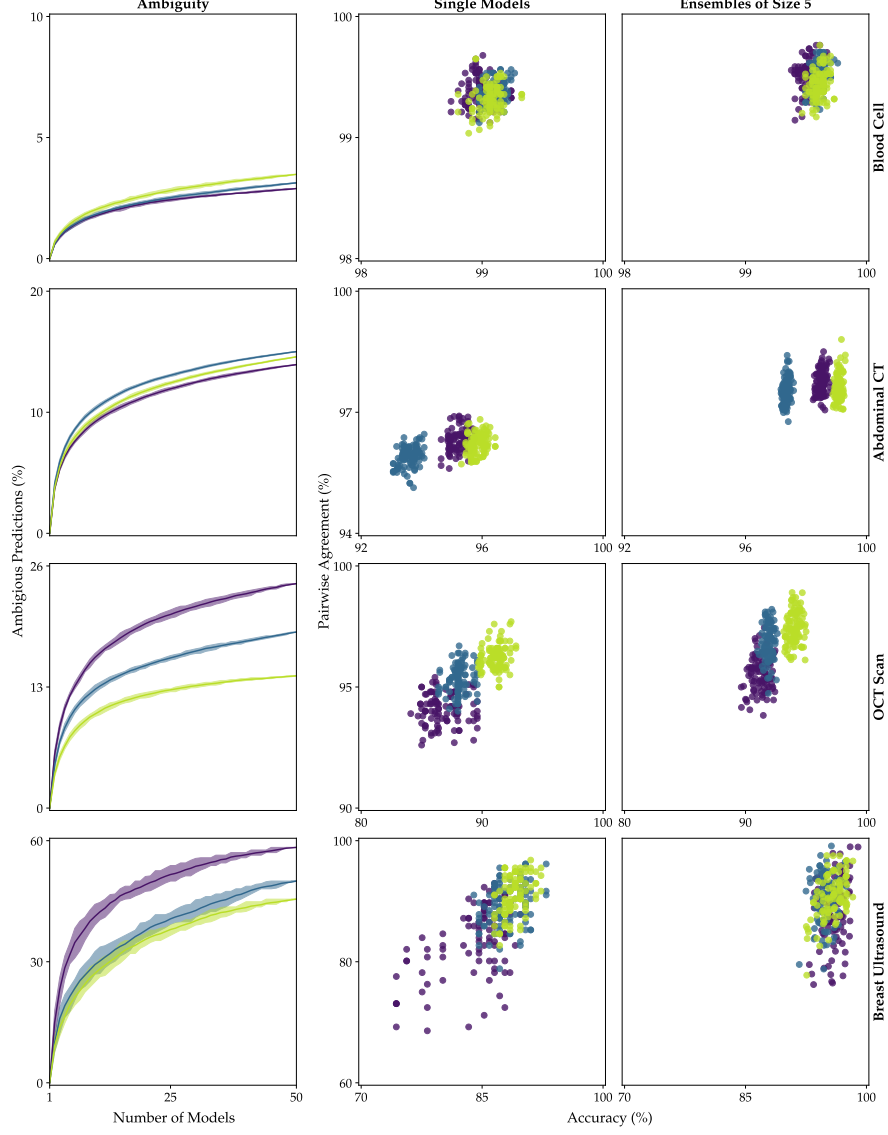


Fig. 5 Models that achieve higher accuracy tend to exhibit lower predictive multiplicity – characterized by fewer ambiguous predictions and greater consistency in outputs. The role of model complexity in this relationship is largely indirect by influencing model accuracy. We evaluated this relationship across multiple datasets (rows) and three model capacities: EfficientNetB0 (low ●), B2 (medium ●), B4 (high ●). First column: Percentage of ambiguous predictions (y-axis) when considering more and more models (x-axis); the solid line shows the mean, the shaded areas the 25-75% IQR. Second column: Pairwise agreement (y-axis) vs accuracy (x-axis), computed between two randomly selected, distinct *single models* from the Rashomon set. Third column: Same as the second column, but computed between two randomly selected, distinct *ensembles of size 5*.

agreement irrespective of model capacity. For example, on the Blood Cell dataset, the

EfficientNetB0, the model with the smallest capacity, already achieves high accuracy, increasing model capacity neither affects accuracy nor agreement. (3) When accuracy decreases, so does agreement. This notable case is observable on the abdominal CT dataset: EfficientNetB2 performs worse than both B0 and B4 and also shows slightly lower agreement. Overall, these results suggest that model capacity alone does not determine pairwise agreement – accuracy plays a key moderating role.

Finally, we extend our analysis to *ensembles of five models* to examine whether the patterns of predictive multiplicity observed at the individual model level persist in ensemble settings (see third row in Figure 5). On the Blood Cell, abdominal CT and Breast Ultrasound dataset ensembling with abstention leads to similar pairwise agreement irrespective of the model capacity. Interestingly, while the ensembles with smaller capacity on abdominal CT perform worse in terms of accuracy, pairwise agreement is largely similar across model capacities. For the Breast Ultrasound dataset some B0-based ensembles show lower stability despite high accuracy. This is likely due to greater variability among individual B0 models within the empirical Rashomon set, which affects ensemble consistency. Note that the spread in agreement is large on this dataset and that the higher accuracy achieved through ensembling might “mask” predictive multiplicity. The OCT dataset is the only dataset in which highest model capacity (B4) results in slightly less predictive multiplicity. To summarize, ensembling with abstention seems to “compensate” for performance gains resulting from increased model capacity – however, gains in accuracy through ensembling might mask persistently high predictive multiplicity as seen in the Breast Ultrasound dataset.

In summary, for single models higher accuracy is consistently associated with lower predictive multiplicity, reflected in both reduced ambiguity and increased pairwise agreement. Model complexity influences this relationship primarily through its effect on accuracy: when increased complexity improves accuracy, predictive multiplicity decreases; when it reduces accuracy (e.g., EfficientNetB2 on the Abdominal CT dataset), predictive multiplicity increases. When accuracy remains unchanged, model complexity has no clear impact on ambiguity and stability. Furthermore, ensembles consistently improve pairwise agreement – even when individual models perform poorly.

3 Discussion

In this work, we conducted a comprehensive evaluation of the empirical Rashomon set across a range of medical datasets and deep learning architectures. We found that validation set performance is not a reliable predictor of test set performance for models within the empirical Rashomon set. These models differ solely in the random seed used to initialize the weights of the final layer – a “choice” typically made haphazardly [41] without careful consideration. As a result, the models comprising the empirical Rashomon set represent alternative, equally plausible solutions that could have been obtained from slightly different yet reasonable training choices. When analyzing the predictions of these models on individual samples, we observed that some samples receive conflicting predictions when we consider more than one model from the empirical Rashomon set. Consequently, the predictions – and, by extension, the diagnoses and treatments – of some patients depend on an arbitrary decision. Notably,

this phenomenon occurs even among models with very high predictive performance, including models exceeding 99% accuracy. Therefore, *relying on any single model leads to arbitrary and potentially harmful predictions for some patients*: the ones, who would have received a different outcome from an equally valid alternative model. Taken together, these findings *challenge the prevailing paradigm of selecting a single model* and underscore the need for alternative strategies that *account for the existence of multiple models that fit the data equally well*.

In our experiments, we adopt a slightly different definition of the empirical Rashomon set. Rather than restricting it to models within a narrow validation performance range (from a baseline model or each other) – as is common in prior work (e.g., [25, 39, 43, 47]) – we define it as the set of models obtained by small perturbations to the training pipeline, specifically by varying the random seed used to initialize the final layer weights, regardless of validation performance. This broader definition is better suited for our goal: to empirically characterize the extent, sources, and consequences of predictive multiplicity. Traditional definitions may overlook models that generalize better to unseen data despite having lower validation scores (see Figure 2), leading to an incomplete or misleading characterization of the Rashomon set. Moreover, defining the empirical Rashomon set relative to a specific baseline model may introduce bias. As we show, the initial model – a seemingly natural choice for a reference – performs not (systematically) better than other models in the Rashomon set. Crucially, the exploration of the empirical Rashomon set is not an end in itself, it serves a purpose. In our case, it serves as a tool to uncover alternative, equally plausible models to reveal predictions which are arbitrary under a single model. Thus, considering a broader set of models allows us to gain a more holistic view of predictive multiplicity.

Further, considering all models obtained through random seed initialization offers an additional advantage: it can reveal configurations that may be ill-suited for automated classification, such as Breast Ultrasound/ResNet50. This becomes evident in two ways: (1) some models within the empirical Rashomon set show a correlation between average validation and test performance; (2) the order in which models are evaluated affects the perceived ambiguity of individual predictions. Moreover, using only a subset of these models in ensembles can obscure predictive multiplicity, as the ensemble reduce predictive multiplicity by classifying samples consistently incorrect.

When analyzing the stability of individual predictions, we find that using any single model – without considering other equally valid choices – leads to arbitrary and unjustified predictions for some patients, questioning the validity of using a single model. Even in high-accuracy regimes, a non-trivial proportion of patients receive conflicting predictions from the models within the empirical Rashomon set. Our findings highlight the limitations of relying on a single “best” model and instead advocate for an alternative strategy: leveraging ensembles – i.e., the predictions of multiple models – to achieve more robust and justifiable outcomes. Our analysis shows that ensembling with selective abstention – deferring on samples with model disagreement – effectively eliminates measurable predictive multiplicity in ensembles of five. Consistently predicted samples show no/little predictive multiplicity and thus may be suitable for automated decision-making. Conversely, samples where models from the ensemble disagree should

be relegated to manual review by clinicians.¹⁰ Note that our consensus-driven approach – requiring unanimous agreement across all models for a prediction – is intentionally naive, and while more sophisticated strategies (e.g., the method proposed by Black et al. [8] to use a hypothesis test to determine whether a consistent prediction exists at a predefined confidence level) may be better suited for real-world deployment, our aim is to demonstrate the viability of the underlying concept. The agreement criteria for accepting a prediction should be tailored to the specific application domain (this might also help with the aforementioned training with pathological instabilities). Further, the number of models to consider is a domain-specific trade-off between resource constraints and the desired confidence in prediction stability. Techniques such as ensemble distillation [38] may be suitable to reduce/eliminate predictive multiplicity while substantially reducing computational overhead at inference time. In summary, we foresee the requirement for interdisciplinary discourse to define appropriate settings for the aforementioned criteria.

As with most prior work, our approach to quantifying predictive multiplicity is based on thresholded scores, i.e., discrete classification decisions. This approach can exaggerate the apparent instability of model predictions. For instance, consider a binary classification scenario in which two models predict and assign the scores $[0.49, 0.51]$ and $[0.51, 0.49]$ to a sample. When predictions are thresholded at 0.5, this sample is flagged as unstable because it is classified differently by the two models (despite the two predictions being similar). Conversely, this approach can also underestimate the stability of predictions: The scores $[0.1, 0.9]$ and $[0.49, 0.51]$ (despite being quite different) would result in consistent predictions, thereby potentially masking the presence of predictive multiplicity. In fact, predictive multiplicity metrics for a fixed single model based on predicted classes may result in predictive multiplicity when the threshold criteria for output scores are changed. An alternative to considering the threshold scores would be to consider the probability distributions over all classes (such as the softmax outputs of a neural network) rather than relying on hard classification decisions; as, for example, proposed by Hsu and Calmon [25]. Our aim in this paper is not to propose/improve universal or task-specific solutions but rather to highlight that model multiplicity, and especially predictive multiplicity, is a pervasive and fundamental problem in machine learning. Any useful characterization of the empirical Rashomon set, as well as the choice of decision thresholds, must ultimately be tailored to the specifics of the task at hand.

Note that unless all plausible models are considered – that is unless the Rashomon set is fully characterized – any estimate of multiplicity metrics will necessarily be an underestimate based on the approximation of the true Rashomon set. Recall that for complex models, an exhaustive exploration of the Rashomon set is infeasible; finding a generally meaningful empirical characterization of the Rashomon set remains an open challenge and (likely) depends heavily on the particular application. In our approach, we limit the number of models per configuration to 50 models (which only differ in the random seed which informs the initialization of the last layer). While this constraint reflects the computational cost of training large numbers of models

¹⁰We intend to investigate in future works whether there are regularities in the samples that have high or low agreement.

and we see saturation trends before we consider 50 models, it risks underestimating the true variability between plausible models and the ambiguity of their predictions. The appropriate number of models should be tailored to the specific scenarios, and depending on the case, a larger number of models may be necessary to adequately capture the diversity within the empirical Rashomon set. In addition, we have only considered models from one model family in the empirical Rashomon set. If the goal is to achieve the best possible coverage of the empirical Rashomon set, considering different model architectures might be a useful alternative.

We further explored model-specific factors, concretely the relationship between model capacity, accuracy, and predictive multiplicity in the overparameterized regime. We find that the effect of model capacity is mainly indirect through accuracy: models that achieve higher accuracy tend to exhibit lower predictive multiplicity. Black et al. [9] showed that more expressive and sophisticated machine learning models (e.g., deep learning) have less bias and thus can more accurately approximate the Bayes optimal predictor than less expressive techniques (e.g., linear regression). They argue that this increased expressivity comes at the cost of high variance, which makes any particular model more sensitive to random choices in the development pipeline. We analyze model capacity in the already highly overparameterized regime and find that higher model capacity does not adversely affect the stability of individual predictions, on the contrary, if higher-capacity models achieve more accuracy, they show less predictive multiplicity.

To conclude, we agree that “accuracy is not an antidote to multiplicity, and model selection cannot simply be reduced to accuracy-maximization” [9] – 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: larger models can produce more accurate and consistent predictions, thereby reducing multiplicity.

4 Conclusion

In this study, we investigated the implications of *model multiplicity* – the existence of multiple models with similar average performance but differing individual predictions – in the medical domain. We showed that validation performance is an unreliable indicator of generalization behavior for the models in the empirical Rashomon set, rendering the selection of *any single model* based on performance effectively arbitrary. More critically, even highly accurate models produce conflicting predictions on a non-trivial fraction of samples. In the context of medical applications, this finding implies that predictions for some patients – and, ultimately, treatment decisions – depend on development choices, raising *serious concerns about the justification for deploying any one particular model*.

To address this problem, we find that ensemble-based approaches paired with abstention – making automated predictions only when ensemble members agree – improve the reliability of individual predictions. Even more, sufficiently large ensembles can eliminate measurable predictive multiplicity altogether. In cases of low model consensus, practitioners should default to abstaining from automated prediction and refer the case to manual clinician review.

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 (excluding, for example, segmentation or regression tasks), thereby omitting the full diversity of clinical scenarios and modeling paradigms. Moreover, we do not directly assess how predictive multiplicity influences downstream clinical decision-making or patient outcomes – an important area for future research. Our aim is to highlight the significance of model multiplicity and especially predictive multiplicity in medical application domains and to motivate further investigation into mitigating the risks that model multiplicity poses to the trustworthiness of machine learning systems in high-stakes domains.

Ultimately, model multiplicity matters most in real-world medical settings, where treatment decisions depend on individual-level 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.

5 Methods

5.1 Datasets

We use five medical imaging datasets spanning diverse modalities and classification tasks. All experiments rely on the official training, validation, and test splits to ensure comparability with prior work (see Irvin et al. [26] for X-ray and Yang et al. [49] for all other datasets).

The *Breast Ultrasound* dataset [2] contains images labeled for binary classification into malignant (73.1%) and benign (26.9%) cases. This class distribution is consistent across splits, which comprise 546 training, 78 validation, and 156 test samples.

The *Blood Cell* dataset [1] consists of microscopic images classified into eight categories, with training class proportions of 7.1%, 18.2%, 9.1%, 16.9%, 7.1%, 8.3%, 19.5%, and 13.7%. These distributions remain approximately stable across splits, which include 11,959 training, 1,712 validation, and 3,421 test samples.

The *X-ray* dataset [26] includes chest radiographs labeled for 14 conditions. We evaluate performance on five: Cardiomegaly, Edema, Consolidation, Atelectasis, and Pleural Effusion. Class prevalence varies across splits: Cardiomegaly: 12.1% (train), 29.1% (val), 30.2% (test); Edema: 29.2%, 19.2%, 15.6%; Consolidation: 6.6%, 14.1%, 5.8%; Atelectasis: 30.0%, 34.2%, 30.6%; Pleural Effusion: 43.8%, 28.6%, 20.8%. The dataset comprises 224,316 training, 234 validation, and 500 test images.

The *OCT Scan* dataset [30] contains retinal optical coherence tomography images categorized into four classes: choroidal neovascularization, diabetic macular edema, drusen, and normal. The class distribution in the training and validation sets are 34.4%, 10.5%, 8.0%, and 47.2%, respectively, while the test set is uniformly balanced (25% per class). The dataset includes 97,477 training, 10,832 validation, and 1,000 test images.

The *Abdominal CT* dataset, based on the Liver Tumor Segmentation Benchmark [5], consists of coronal abdominal CT scans labeled with eleven anatomical classes. The training set class distribution is: bladder (8.8%), femur-left (4.8%), femur-right (4.6%), heart (4.6%), kidney-left (8.4%), kidney-right (9.0%), liver (23.0%), lung-left

(7.7%), lung-right (7.9%), pancreas (9.0%), and spleen (12.1%). These proportions remain approximately stable across all splits. The dataset comprises 12,975 training, 2,392 validation, and 8,216 test samples.

5.2 Model training

We evaluated four model architectures in our base experiments: ResNet50 [24], GC ViT [23], EfficientNetB2 [44], and ConvNeXtBase [35], all pretrained on ImageNet [17]. In follow-up experiments, we additionally tested EfficientNetB0 and B4 [44] (also pretrained on ImageNet [17]). For each model, 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 [21].

For X-ray, we adopted the protocol of Irvin et al. [26], 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.

X-ray images were resized to $1 \times 224 \times 224$, all other datasets followed MedMNISTv2 preprocessing [49], resulting in $C \times 224 \times 224$ 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.

All models were trained with a batch size of 64 using the AdamW optimizer [37], 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 [36]. For X-ray, the decay steps equaled the total number of training iterations (epochs \times steps per epoch), while for all other datasets, 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, and one for X-ray. We used binary cross-entropy loss for X-ray and sparse categorical cross-entropy for all other datasets.

For the multi-label X-ray dataset, performance was evaluated using the mean ROC-AUC across the five target classes. ROC-AUC was computed separately for each of the five target classes and then averaged. For all other datasets, classification accuracy served as the primary performance metric. Table 1 reports both the performance of the initial model and the mean performance across the 50 models from the empirical Rashomon set.

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 a given empirical Rashomon set were trained on the same GPU to ensure consistency. Specifically, we used NVIDIA RTX A6000 (for all configurations with X-ray), NVIDIA RTX 6000 Ada Generation (for Breast Ultrasound/EfficientNetB4, and OCT Scan/ConvNeXtBase), NVIDIA H100 PCIe (for Blood Cell/EfficientNetB4,

Table 1 For each dataset (columns) and model architecture (rows), we report the performance of the initial model; the mean performance across the 50 models from the empirical Rashomon set is in parentheses. Test performance is measured using ROC-AUC for X-ray and accuracy for all other datasets. Best-performing baseline models per dataset are highlighted in bold. Benchmark results are from Irvin et al. [26] for X-ray (best mean ROC-AUC across five uncertainty methods) and from Yang et al. [49] for all other datasets (best accuracy across seven architectures).

	Benchmark	ResNet50	GC ViT	EfficientNetB2	ConvNeXtBase
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)
X-ray	89.5	88.9 (88.6)	87.2 (87.6)	89.8 (89.9)	89.7 (89.7)
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)

OCT Scan/ResNet50, OCT Scan/EfficientNetB2, OCT Scan/EfficientNetB4, and Abdominal CT/EfficientNetB4), NVIDIA A40 (for Breast Ultrasound/ResNet50, Breast Ultrasound/EfficientNetB0, Breast Ultrasound/EfficientNetB2, Blood Cell/EfficientNetB2, and Blood Cell/ConvNeXtBase), NVIDIA A100-SXM4-80GB (for Breast Ultrasound/GC ViT, Breast Ultrasound/ConvNeXtBase, Blood Cell/ResNet50, Blood Cell/GC ViT, Blood Cell/EfficientNetB0, Abdominal CT/ResNet50, Abdominal CT/GC ViT, Abdominal CT/EfficientNetB2, Abdominal CT/ConvNeXtBase, and OCT Scan/EfficientNetB0). We used NVIDIA H100 PCIe for the experiments with mislabeled samples; for the experiments where we removed samples, we used NVIDIA RTX 6000 Ada Generation.

5.3 Experiments

In our main experiments, we systematically evaluated the risks of predictive multiplicity across the medical imaging modalities and model architectures described above. To this end, we explored the empirical Rashomon set by training 50 models per dataset/architecture pair, resulting in a total of 1,000 models. 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 changing 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.

To evaluate how well validation performance predicts test performance across models in the empirical Rashomon set (Figure 2), we visualized their joint distribution using kernel density estimation plots implemented with Seaborn [46]. The KDE was smoothed using a bandwidth adjustment factor of 0.5, rendered with 100 contour levels, and thresholded at a density of 0.05 to suppress low-density regions.

To assess *ambiguity* across multiple models we considered up to 50 model instances per configuration (Figure 3 and Figure 5). Model instances were added by drawing randomly without replacement from the remaining models, in ascending and descending order of their validation performance. To take into account that the selection of a

specific model instance can lead to variability in the results, we report mean and IQR (25 and 75%) for the experiments with random order across 50 repetitions of this experiment.

To compute *pairwise agreement* (Figure 5), we randomly sampled 100 single model or ensemble pairs (of equal size) without replacement from the empirical Rashomon set. For each pair, we report the accuracy of the first model or ensemble in the pair. We considered ensembles of size two and five. To compute *expected pairwise agreement* (Figure 4), we averaged the results across these experiments, normalized by the coverage rate for improved visual comparability. Coverage is defined as 100% for single models and as the proportion of test samples on which all ensemble members agree for ensembles of sizes two and five. For the X-ray dataset, probability outputs were thresholded at 0.5 to obtain binary predictions. We defined agreement as both models correctly classifying all five target labels for a given sample.

To assess *architecture-specific characteristics* (Figure 5), we additionally trained EfficientNetB0 and EfficientNetB4 following the same protocol as described above.

Acknowledgements. 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).

A Independent errors

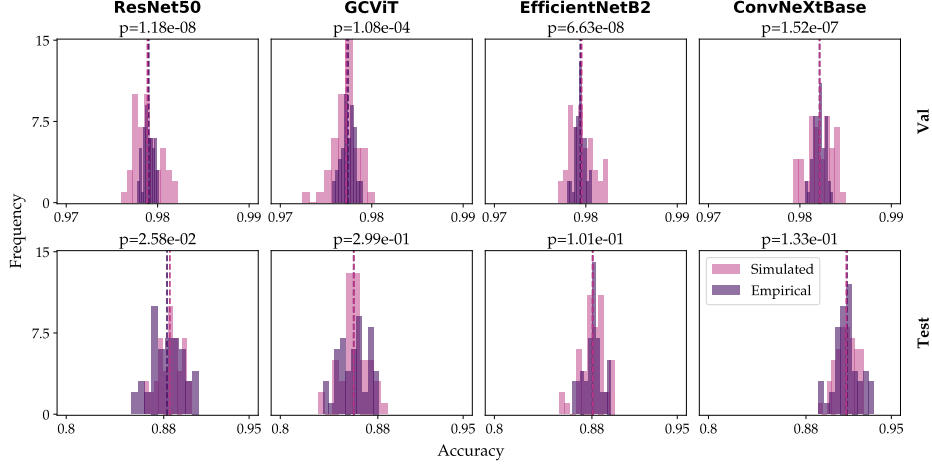


Fig. 6 Comparison of empirical and simulated distributions of model accuracies on the validation and test sets for the OCT Scan dataset. The empirical distributions are obtained from the respective 50 models from our main experiments (the individual models differ only in the random seed which informs the initialization of the last layer). The simulated distributions are constructed under the assumption that individual prediction errors are independent: for each model, the number of correct predictions is sampled from a binomial distribution with the dataset size as the number of trials and the mean accuracy as the success probability. These counts are then converted into accuracies. On the validation set, the empirical distribution is narrower than the simulated one, suggesting more consistency across different model initializations than would be expected under independent errors. On the test set, both distributions have comparable variance. To formally compare the variances of the empirical and simulated distributions, we use a two-sided F-test. The test statistic is calculated as the ratio of the empirical variance to the simulated variance. The corresponding p-value reflects the probability of observing such a ratio (or a more extreme one) under the null hypothesis that both distributions have the same variance.

B Predictive Performance

We have seen in the main manuscript of the paper that using ensembles reduces or even eliminates measurable predictive multiplicity. An additional (expected) effect of ensembles is improved performance.

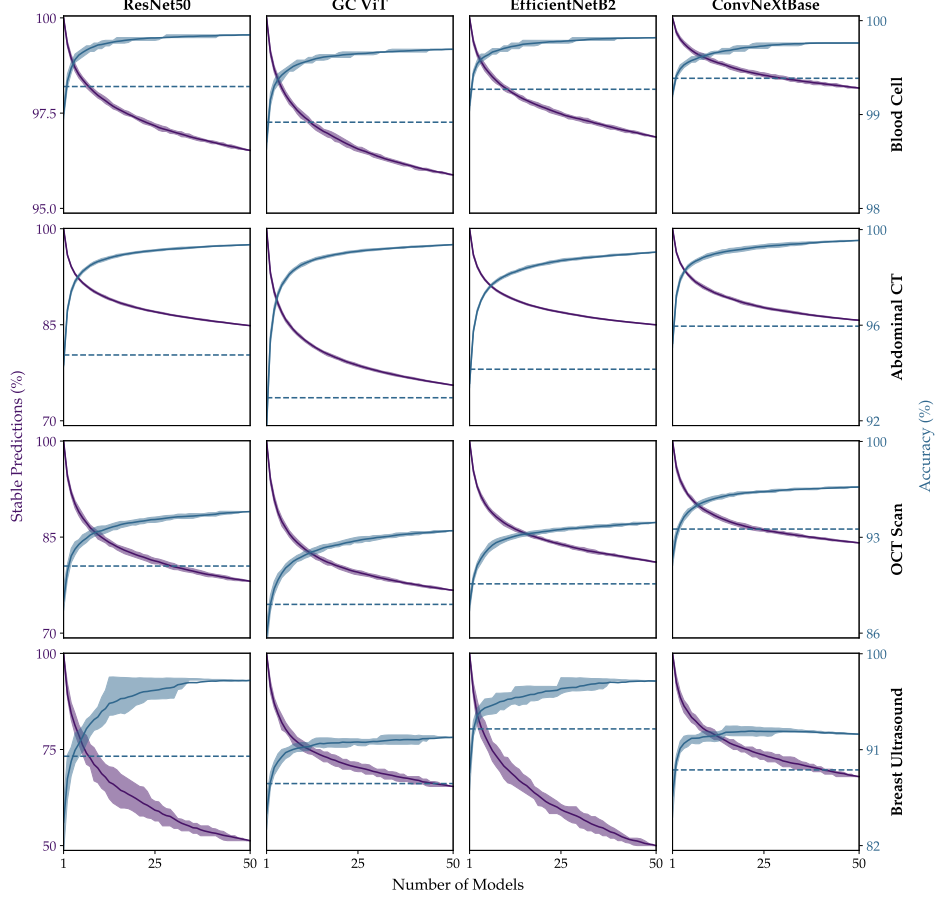


Fig. 7 By leveraging consensus among multiple models from the Rashomon set, we can abstain when predictions are uncertain and achieve higher accuracy on stable predictions than even the best individual model. When considering the Rashomon set, instead of relying on a single model, we can leverage multiple models, making predictions only when there is sufficient consensus; otherwise, we can abstain from automated predictions. This “consensus-driven” approach improves test performance on the samples where the considered models from the Rashomon set agree. With as few as two to 15 models, we outperform the best-performing model among 50 models from the empirical Rashomon set on the stable samples. The left y-axis represents the percentage of stable predictions (●) when we consider multiple models (x-axis). The right y-axis shows the percentage of correct classifications (●) for the proportion of stable predictions over the number of models (x-axis). In both curves, the solid line represents the mean value, and the shaded shows the IQR (25-75%) of 50 runs for a random order of the models. The dashed line (●) indicates the performance of the best of the 50 individual models on the whole test set.

Figure 7 shows the accuracy (●) when predictions are made exclusively on samples where *all considered models agree*, abstaining in cases of disagreement. We further plot the percentage of stable predictions (●), i.e., the percentage of samples on which we make predictions (which is the inverse of ambiguity). Remarkably, across all dataset/architecture combinations, this “consensus-driven” approach outperforms the

best single model among the 50 candidates (dashed line in \bullet) – even when relying on only a small number of models. Note that, unlike the ensemble, the best single model cannot abstain and must predict on the entire test set. We provide additional results in [Appendix C](#), where we evaluate predictive performance using a relaxed consensus criterion, requiring agreement from a predefined percentage of models instead of full unanimity.

In the following, we illustrate the benefits of using a small number of models with the consensus-driven approach on two contrasting cases. For the Blood Cell/ConvNeXtBase combination, predictive accuracy and stability are high even with a single model: Test accuracy of single models ranges from 99.0% to 99.4% across the Rashomon set, with only 1.8% of samples being subject to conflicting predictions. Selecting *two models* at random and predicting only on samples where their outputs agree matches the performance of the best individual model among all 50 models in expectation over all possible model combinations (mean = 99.42%, IQR = [99.38%, 99.44%]) while abstaining on just 0.4% of the samples. In contrast, model performance varies substantially for the Breast Ultrasound/ResNet50 combination: Accuracies differ by up to 17.8% (from 74.4% to 92.2%) across the empirical Rashomon set, with 48% of predictions being ambiguous across all 50 models. By randomly selecting *two models* and predicting only under consensus, we achieve an accuracy of 88.1% in expectation (IQR = [85.9%, 90.4%]) with coverage of 89.1%. Relying instead on the consensus of *six models*, we achieve the same performance as the best individual model among the 50 (mean = 92.4%, IQR = [90.8%, 93.8%]) in expectation on 75.3% of the samples. These examples show that consensus, even among only a small number of models, can dramatically stabilize predictive performance.

In summary, leveraging multiple models through a consensus-driven strategy can substantially improve the reliability and safety of automated decision-making. Even small ensembles – comprised of as few as two to six models – can match or exceed the performance of the best individual model, while providing the added benefit of abstention mechanisms that allow the system to defer uncertain or conflicting cases to clinical experts. This is particularly important in high-stakes medical settings, where erroneous predictions from a single model could have harmful consequences.

C Relaxed Consensus Thresholds

To further investigate the relationship between model agreement and predictive ambiguity, we conduct additional analyses in which a sample is deemed ambiguous only if a minimum fraction of classifiers disagree on the predicted label. Specifically, rather than requiring full unanimity to classify a sample as unambiguous, we allow for relaxed consensus thresholds, wherein partial agreement among the ensemble suffices. Results for this relaxed criterion are presented in [Figure 8](#).

We observe that under relaxed consensus criteria, the qualitative patterns evident under the strict unanimity requirement are preserved when using smaller ensembles. Notably, when the number of models increases, the proportion of samples classified as ambiguous does not continue to rise in the same way it does under the unanimity condition. In fact, the relaxed criterion often results in a plateau or even a decline

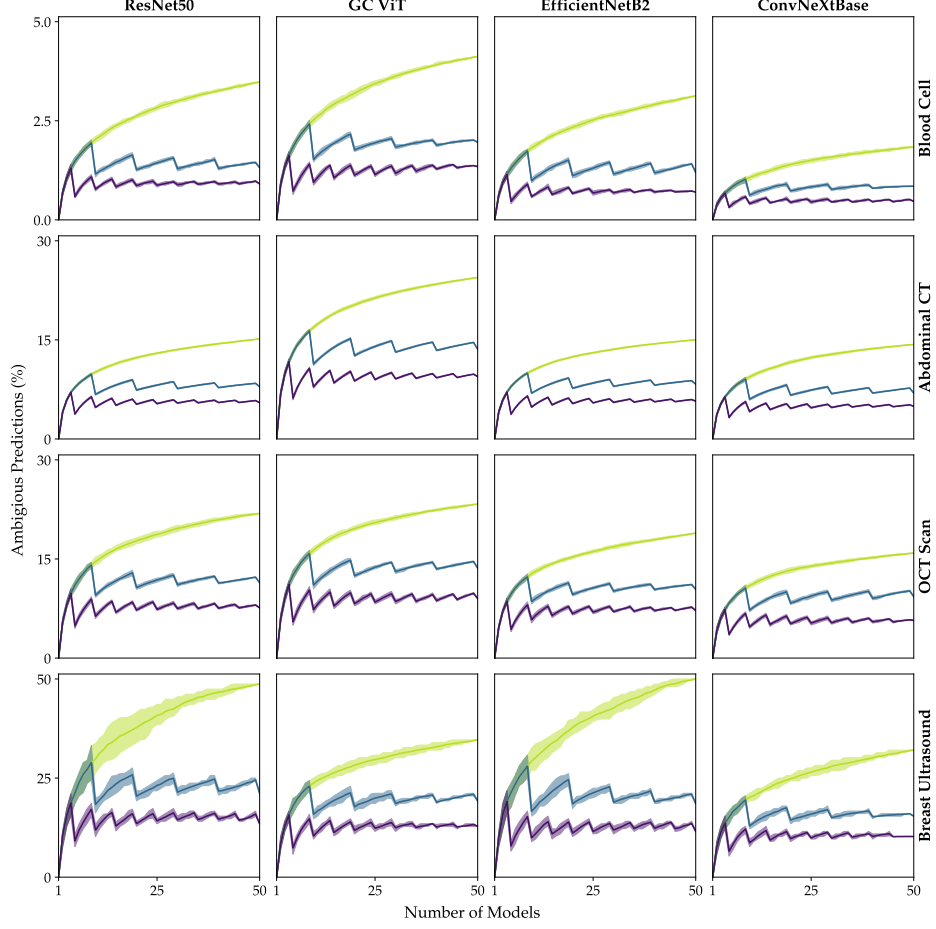


Fig. 8 When applying relaxed consensus criteria, we observe similar behavior as for unanimous agreement for the first models; for more models, we see substantially less ambiguity. We plot the percentage of ambiguous predictions (y-axis) as a function of the number of models considered (x-axis), across various datasets (rows) and model architectures (columns). We compare three levels of agreement: full consensus (●), 90% agreement (●), and 80% agreement (●). The jumps in the curves occur when the addition of a new model causes the consensus threshold to be crossed.

in ambiguity rates as ensemble size grows. This indicates that relaxing the consensus threshold can substantially improve coverage – defined as the fraction of data points for which the ensemble issues a non-ambiguous prediction – particularly when we consider a larger set of models from the empirical Rashomon set.

Importantly, any increase in coverage resulting from a relaxed agreement criterion is only practically meaningful if the additional predictions are also accurate. In other words, the utility of relaxed consensus thresholds hinges not only on their capacity to expand predictive coverage but also on their ability to maintain or improve predictive performance. As shown in Figure 9, this trade-off is partially realized: on the OCT Scan

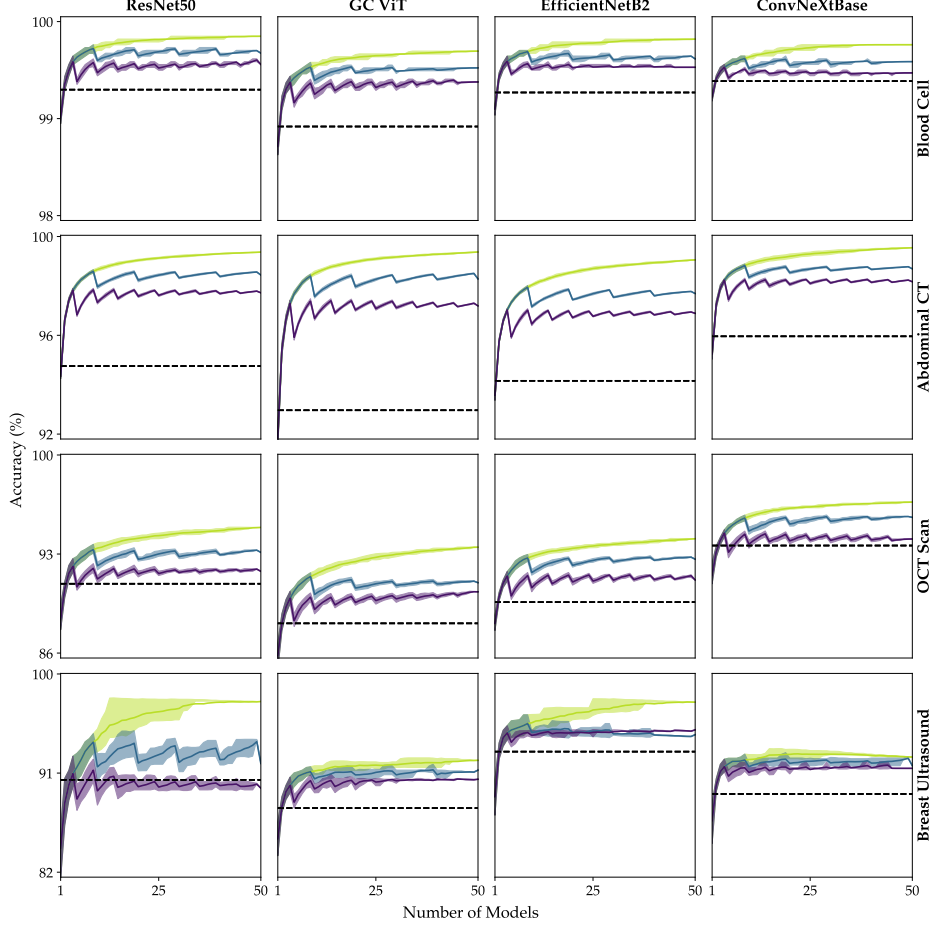


Fig. 9 Relaxed consensus can outperform individual models, with stricter agreement thresholds resulting in higher predictive accuracy. We report accuracy (y-axis) as a function of the number of models considered (x-axis), making predictions only when a specified proportion of models agree. Three consensus thresholds are shown: unanimity (●), 90% agreement (●), and 80% agreement (●). Discontinuities in the curves arise when the addition of a model causes the agreement threshold (e.g., 80% or 90%) to be newly satisfied, enabling predictions that were previously withheld. The dashed line (●) indicates the performance of the best of the 50 individual models on the whole test set.

dataset (with ResNet50 and ConvNeXtBase), an 80% agreement threshold achieves performance comparable to the best single model, while unanimity yields substantially higher predictive accuracy. For the Breast Ultrasound dataset, the 80% threshold results in performance that is similar to or worse than the best of 50 individual models; importantly, the single, best model predicts on the *whole test dataset*. In all other datasets, relaxed consensus thresholds lead to improved performance relative to the best individual model. Overall, as expected, increasing the strictness of the agreement threshold consistently results in higher predictive performance (albeit at the expense

of coverage). To summarize, this trade-off between coverage and reliability must be balanced based on application-specific constraints and priorities.

References

- [1] Acevedo A, Merino A, Alf  rez S, et al (2020) A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in brief* 30:105474
- [2] Al-Dhabyani W, Gomaa M, Khaled H, et al (2020) Dataset of breast ultrasound images. *Data in brief* 28:104863
- [3] Anders C, Pasliev P, Dombrowski AK, et al (2020) Fairwashing explanations with off-manifold detergent. In: *International Conference on Machine Learning*, PMLR, pp 314–323
- [4] Auer P, Herbster M, Warmuth MK (1995) Exponentially many local minima for single neurons. *Advances in neural information processing systems* 8
- [5] Bilic P, Christ P, Li HB, et al (2023) The liver tumor segmentation benchmark (lits). *Medical image analysis* 84:102680
- [6] Black E, Fredrikson M (2021) Leave-one-out unfairness. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp 285–295
- [7] Black E, Wang Z, Fredrikson M, et al (2021) Consistent counterfactuals for deep models. *arXiv preprint arXiv:211003109*
- [8] Black E, Leino K, Fredrikson M (2022) Selective ensembles for consistent predictions. In: *10th International Conference on Learning Representations, ICLR 2022*
- [9] Black E, Raghavan M, Barocas S (2022) Model multiplicity: Opportunities, concerns, and solutions. In: *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp 850–863
- [10] Black E, Gillis T, Hall ZY (2024) D-hacking. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp 602–615
- [11] Breiman L (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16(3):199–231
- [12] Brown T, Mann B, Ryder N, et al (2020) Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901
- [13] Chen C, Lin K, Rudin C, et al (2018) An interpretable model with globally consistent explanations for credit risk. *arXiv preprint arXiv:181112615*

- [14] Citron DK, Pasquale F (2014) The scored society: Due process for automated predictions. *Wash L Rev* 89:1
- [15] Creel K, Hellman D (2022) The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy* 52(1):26–43
- [16] D’Amour A, Heller K, Moldovan D, et al (2022) Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research* 23(226):1–61
- [17] Deng J, Dong W, Socher R, et al (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee, pp 248–255
- [18] Dietterich TG (2000) Ensemble methods in machine learning. In: *International workshop on multiple classifier systems*, Springer, pp 1–15
- [19] Dutta S, Wei D, Yueksel H, et al (2020) Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In: *International conference on machine learning*, PMLR, pp 2803–2813
- [20] Ganesh P, Chang H, Strobel M, et al (2023) On the impact of machine learning randomness on group fairness. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp 1789–1800
- [21] Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings*, pp 249–256
- [22] Hancox-Li L (2020) Robustness in machine learning explanations: does it matter? In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp 640–647
- [23] Hatamizadeh A, Yin H, Heinrich G, et al (2023) Global context vision transformers. In: *International Conference on Machine Learning*, PMLR, pp 12633–12646
- [24] He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 770–778
- [25] Hsu H, Calmon F (2022) Rashomon capacity: A metric for predictive multiplicity in classification. *Advances in Neural Information Processing Systems* 35:28988–29000
- [26] Irvin J, Rajpurkar P, Ko M, et al (2019) Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the*

- [27] Jordan K (2024) On the variance of neural network training with respect to test sets and distributions. ICLR
- [28] Jumper J, Evans R, Pritzel A, et al (2021) Highly accurate protein structure prediction with alphafold. *nature* 596(7873):583–589
- [29] Kaissis GA, Makowski MR, Rückert D, et al (2020) Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* 2(6):305–311
- [30] Kermany DS, Goldbaum M, Cai W, et al (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell* 172(5):1122–1131
- [31] Kobylńska K, Krzyżiński M, Machowicz R, et al (2024) Exploration of the rashomon set assists trustworthy explanations for medical data. *IEEE Journal of Biomedical and Health Informatics* 28(11):6454–6465
- [32] Krizhevsky A (2009) Learning multiple layers of features from tiny images. Tech. Rep. TR-2009, University of Toronto, technical report
- [33] Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25
- [34] Kulynych B, Hsu H, Troncoso C, et al (2023) Arbitrary decisions are a hidden cost of differentially private training. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp 1609–1623
- [35] Liu Z, Mao H, Wu CY, et al (2022) A convnet for the 2020s. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 11976–11986
- [36] Loshchilov I, Hutter F (2016) Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:160803983*
- [37] Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. *arXiv preprint arXiv:171105101*
- [38] Malinin A, Mlodozieniec B, Gales M (2019) Ensemble distribution distillation. *arXiv preprint arXiv:190500076*
- [39] Marx C, Calmon F, Ustun B (2020) Predictive multiplicity in classification. In: *International Conference on Machine Learning*, PMLR, pp 6765–6774
- [40] Pawelczyk M, Broelemann K, Kasneci G (2020) On counterfactual explanations under predictive multiplicity. In: *Conference on Uncertainty in Artificial*

Intelligence, PMLR, pp 809–818

- [41] Picard D (2021) Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. arXiv preprint arXiv:210908203
- [42] Rudin C, Zhong C, Semenova L, et al (2024) Position: amazing things come from having many good models. In: Proceedings of the 41st International Conference on Machine Learning, pp 42783–42795
- [43] Semenova L, Rudin C, Parr R (2022) On the existence of simpler machine learning models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp 1827–1858
- [44] Tan M, Le Q (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning, PMLR, pp 6105–6114
- [45] Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Advances in neural information processing systems 30
- [46] Waskom ML (2021) Seaborn: statistical data visualization. Journal of Open Source Software 6(60):3021
- [47] Watson-Daniels J, Parkes DC, Ustun B (2023) Predictive multiplicity in probabilistic classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 10306–10314
- [48] Wick M, Tristan JB, et al (2019) Unlocking fairness: a trade-off revisited. Advances in neural information processing systems 32
- [49] Yang J, Shi R, Wei D, et al (2023) Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data 10(1):41