Patterns

Article

Unraveling overoptimism and publication bias in ML-driven science

Highlights

- Accuracy is negatively associated with sample size in ML-driven science publications
- This counterintuitive relationship is attributed to overfitting and publication bias
- We introduce a model for observed accuracy based on parametric learning curves
- We propose a solution to compute realistic estimates of model performance

Authors

Pouria Saidi, Gautam Dasarathy,
Visar Berisha

Correspondence

psaidi@asu.edu

In brief

Machine learning (ML) is widely used, yet recent results suggest that reported model accuracies are overoptimistic. This research introduces a model to recover true model performances from inflated accuracy estimates, accounting for two overoptimism sources: overfitting and publication bias. The model estimates the true learning curve from overly optimistic accuracy observations, enabling a more realistic model performance evaluation. Its findings indicate that this approach can estimate the true model performance from a set of overly optimistic reports.

Saidi et al., 2025, Patterns *6*, 101185
April 11, 2025 © 2025 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.patter.2025.101185

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Unraveling overoptimism and publication bias in ML-driven science

Pouria Saidi,[1,3,4,*] Gautam Dasarathy,[1] and Visar Berisha[1,2]
[1]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281, USA
[2]College of Health Solutions, Arizona State University, Tempe, AZ 85281, USA
[3]Present address: Mayo Clinic, Scottsdale, AZ, USA
[4]Lead contact
*Correspondence: psaidi@asu.edu

https://doi.org/10.1016/j.patter.2025.101185

THE BIGGER PICTURE Advancements in machine learning (ML) have led to its widespread adoption across many scientific fields. However, published results are often overly optimistic due to publication bias, where less convincing findings are omitted, and overfitting, which is when the model retrieves extremely accurate results with the training or "learning" data but fails when encountering new data. As a result, reported performances in publications may not accurately reflect the true performance of ML models. Our research introduces a method to estimate true learning curves from overoptimistic results, offering a clearer understanding of actual model performance. This approach can support cost-benefit analyses for study design, enhance the transparency of ML-driven research, and complement existing evaluation methods. In addition, our work underscores the need to reform publishing standards by encouraging the inclusion of low-accuracy results. In the longer term, addressing these biases could accelerate the adoption of ML in critical areas like personalized medicine, climate modeling, and public policy.

SUMMARY

Machine learning (ML) is increasingly used across many disciplines with impressive reported results. However, recent studies suggest that the published performances of ML models are often overoptimistic. Validity concerns are underscored by findings of an inverse relationship between sample size and reported accuracy in published ML models, contrasting with the theory of learning curves where accuracy should improve or remain stable with increasing sample size. This paper investigates factors contributing to overoptimism in ML-driven science, focusing on overfitting and publication bias. We introduce a stochastic model for observed accuracy, integrating parametric learning curves and the aforementioned biases. We construct an estimator that corrects for these biases in observed data. Theoretical and empirical results show that our framework can estimate the underlying learning curve, providing realistic performance assessments from published results. By applying the model to meta-analyses of classifications of neurological conditions, we estimate the inherent limits of ML-driven prediction in each domain.

INTRODUCTION

Recent advancements in machine learning (ML) have opened new avenues for research across many disciplines, giving rise to the field of ML-driven science (e.g., sociology,[1] medicine,[2] education,[3] and digital health[4]). The rapid adoption of ML in these fields is driven in large part by high reported accuracies in academic publications.

Despite impressive reported results, several recent studies have raised questions about their validity.[5–7] For instance, a collection of results from a survey of predictions of brain disorders[6] reveals an unexpected negative association between sample size and reported accuracy in these studies. In Figures 1A and 1B, we illustrate this negative relationship for ML-driven studies focused on the prediction of Alzheimer's disease (AD) and schizophrenia. A similar trend has been reported in Berisha et al.,[8] where the authors analyzed published accuracies of speech-based ML models to predict AD and other forms of cognitive impairment (CI),[9–11] and in Vabalas et al.,[12] where the authors analyzed the performance of ML models for the detection of autism spectrum

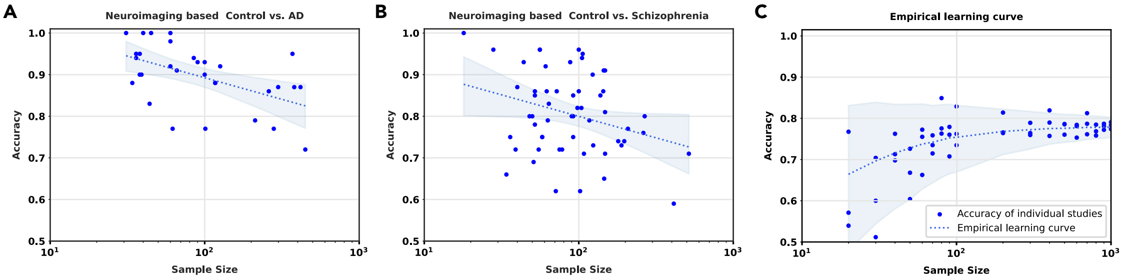

Figure 1. Reported classification accuracy vs. sample size

(A) Results from a meta-analysis[6] of neuroimaging-based classification models between a control group and a cohort of patients with AD.

(B) Results from a meta-analysis[6] of neuroimaging-based classification models between a control group and a cohort of patients with schizophrenia.

(C) An empirical learning curve that is obtained by solving a binary classification problem using different ML models and sample sizes.

The uncertainty bands represent 95% confidence intervals. The y axis is in linear scale, and the x axis is in log scale. See also Figure S1.

disorder (ASD). This is in contrast with our theoretical understanding of ML, where increasing sample sizes should not decrease the accuracy of a properly trained model.[8,13] Figure 1C illustrates a well-behaved learning curve that follows this intuition and is obtained by solving a binary classification problem using properly trained and evaluated ML models with randomly sampled datasets of different sample sizes.

Berisha et al.[8] postulate that overfitting and publication bias give rise to the negative association between reported accuracy and sample size. Overfitting occurs when a model captures not only the underlying patterns in the training data but also their noise and idiosyncrasies, leading to poor generalization on new, unseen data. This issue is especially pronounced in situations with limited data. It can become challenging to identify overfitting to the test set. This can happen due to unintended interdependencies between training and test datasets that can arise during model development.[7,14] Model development is inherently an iterative, adaptive process[15] where researchers reuse test sets to refine models. However, repeated use of the same test set can lead to inadvertently learning specific patterns or noise unique to that set, exacerbating overfitting issues. This problem has gained considerable attention recently in ML-driven science,[7] with the recognition that it likely results in the overestimation of model performance and unrealistic reported accuracy, particularly at small sample sizes.[16]

A less-studied cause of overoptimism in ML-driven papers is publication bias. It is known that training an ML model with limited data typically results in high-variance accuracy estimates.[17] This increased variability can lead to instances where the accuracy of the model is both underestimated and overestimated. However, models with higher estimated accuracy are more likely to be published,[18] a phenomenon known as publication bias or the file drawer effect in the social and medical sciences.[19] As a result, models with inflated estimates of the performance are more likely to be published. Meanwhile, models trained with larger sample sizes produce results that are more accurate and exhibit less variability, making them a more reliable measure of accuracy. Therefore, both overfitting and publication bias have a greater impact on models when the sample size is small, and this effect diminishes as the sample size increases. Hence, we posit that the negative association observed in Figure 1 is attributable to overfitting and publication bias. Both causes of overoptimism are likely compounded by the incentives created by the academic community's outsized emphasis on high accuracy as a primary reason for the publication of new methodologies in ML-driven science and the increased analytical flexibility of ML methods.[20]

Overoptimistic accuracy results reported in the literature create challenges for true scientific progress and the responsible deployment of ML models. They create a skewed perception of the state of knowledge in the field and inflate expectations for the practical application of research. These inflated expectations can lead to sensationalized stories in the press and premature deployment.[21,22] When these models fail after deployment, these expectations are not met and can negatively impact the public's trust in this technology.[23] We posit that this problem may be amplified for some fields by the new federal public access policy that mandates access to the results of federally funded research.[24,25] Ioannidis[20] discussed how rapidly evolving scientific fields with more scientific teams involved and greater flexibility in design and analytical methods have higher chances of reporting false and overoptimistic research findings. The availability of new data in ML-driven science will likely attract more scientific teams, and consequently, this may lead to an increase in the number of publications. This has the potential to set the stage for a perfect storm where deciding whether the new results are trustworthy or not becomes more difficult, even for the experts in the field.

In this paper, we present an observation model for the *published* classification accuracy of ML models based on the notion of parametric learning curves, taking into account both overfitting and publication bias. We leverage this model and further propose a solution to alleviate the overoptimism and determine realistic estimates of model performance by correcting the bias due to both causes.

Proposed model

We introduce the main idea of our modeling approach here. A more detailed explanation appears in the methods section. The relationship between training dataset size n and a ML model's performance $y(n)$ (say classification accuracy) is known to be effectively modeled using equations of the form $y(n) = A + \alpha n^{\beta}$.[26–30] Such models, known as power laws, have found applications in a variety of areas, such as network science,[31] social science,[32] and infection disease surveillance.[33]

Consider a scenario where a research team has access to a dataset of sample size *n* for the development of a binary classification model. During ML model development, the team iterates over the dataset multiple times, adjusting parameters and algorithms based on insights gained from previous runs, resulting in the final estimate of model accuracy. We propose an observation model for such estimates of the accuracy as

$$\mathbf{y}_n^\star = A + \alpha n^\beta + \mathbf{w}_n, \qquad \text{(Equation 1)}$$

where $\mathbf{y}_n^\star$ is the classification accuracy of an ML model given the sample size *n*, *A* is the limiting performance, $\beta < 0$ is the learning rate, $\alpha < 0$ is the power-law index, and $\mathbf{w}_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$ is a random variable with a Gaussian distribution. The idea behind adding $\mathbf{w}_n$ is 2-fold. The random variable $\mathbf{w}_n$ accounts for the inherent variations in the estimates of the model accuracy and the effects of overfitting on estimated accuracy. Variations in model accuracy arise due to differences in data selection and data splitting for model training and testing, and they scale with $n^{-0.5}$.[34,35] We incorporate this information and use $\sigma_n = c_1 n^{-0.5}$, where c_1 is a constant, to quantitatively model the variations in model accuracy. Similarly, previous studies have shown that overfitting to the test set inflates estimated model accuracies with $n^{-0.5}$.[34,36] As a result, we adjust the mean of the Gaussian distribution to $\mu_n = \zeta n^{-0.5}$ to model this. Here, ζ denotes the overfitting index or the bias in the average accuracy estimate due to overfitting.

The model proposed in Equation 1 posits that the average performance of the model can be represented as $A + \alpha n^\beta$ in the absence of overfitting. By applying this model across a range of sample sizes, the parametric model uncovers the true estimate of the learning trajectory, or learning curve, of the ML model.

A research team may elect to publish the observed results or not, depending on whether they deem the resulting accuracy sufficiently high for publication. Such self-selection has been shown to cause a bias in the published literature.[37] We model publication bias in the *reported* classification accuracies using a selection mechanism.[38] Under this model, a research team decides to publish their result if the estimated classification accuracy $\mathbf{y}_n^\star > \gamma_n$, where γ_n is a threshold that depends on the sample size. We can express this mechanism as follows:

$$\begin{cases} \mathbf{y}_n = \mathbf{y}_n^\star & \text{if} \quad \mathbf{y}_n^\star \geq \gamma_n, \\ \mathbf{y}_n^\star \text{ not observed} & \text{if} \quad \mathbf{y}_n^\star < \gamma_n \end{cases}. \qquad \text{(Equation 2)}$$

This operation models, for instance, self-selection by the authors and peer review. We posit that this process is carried out independently by multiple research teams at various levels of data availability. The selection model, therefore, implies that consumers of this literature only observe a biased sample of accuracies that are greater than some threshold.

In this work, we aim to estimate the parameters of the parametric learning curve in Equation 1 from the observed classification accuracies. Due to the censoring of observations from publication bias, conventional methods, such as ordinary least squares, lead to the unrealistic negative association observed in Figure 1. Therefore, we propose a new solution based on truncated regression to estimate the parameter values and provide theoretical and empirical results that demonstrate we can reliably estimate the true learning curve from a series of overoptimistic observations.

Contributions

The following points summarize the main contributions of this paper.

- We propose an observation model for the *published* classification accuracy of ML models. This model is based on the notion of parametric learning curves that can be represented using power-law models, and it accounts for overfitting and publication bias as two influencing factors for overoptimism in the ML literature.
- We propose a solution based on truncated regression and show theoretically that it is possible to identify the learning trajectory of ML models even without prior information about the selection model. Furthermore, we use the observation model to devise a cost function for estimating the true learning curve from overoptimistic accuracy results.
- We apply the model to different meta-analyses in the digital health literature. Particularly, we consider meta-studies of brain disorders, including AD and other forms of CI, schizophrenia, attention-deficit hyperactivity disorder (ADHD), and ASD; the reported ML model results are based on multiple modalities, such as neuroimaging and speech data. Our analysis highlights the prevalence of overoptimism in these fields and provides realistic estimates of ML model performance in each field.

RESULTS AND DISCUSSION

We evaluate the observation model and the accuracy of the solution in estimating ML model performance under overoptimism. The evaluation comprises three interconnected experiments, each highlighting a key component of the model's performance and relevance.

Experiment 1 serves as a foundational test, where we generate data directly from the observation model and estimate the known underlying learning curve. This experiment establishes the basic efficacy of our model, demonstrating its capability to accurately recover learning curve parameters in a controlled environment. This sets the stage for more complex scenarios in subsequent experiments.

Experiment 2 is a realistic simulation of the real-world process of ML model development, integrating elements of overfitting and selective reporting based on a minimal accuracy threshold. We consider binary classification problems and empirically derive the true learning curve by progressively increasing the sample sizes in ML model training. Next, we model overfitting to the test set by performing feature selection on all data and then splitting them into a training and a test set for model training and evaluation. Finally, we only "report" accuracies greater than a pre-established threshold (unknown to the recovery algorithm). Our model uses the reported accuracies to estimate the learning curve. This experiment is pivotal, as it showcases the model's robustness and validity in simulating real-world ML development scenarios, where overfitting and publication bias are prevalent.

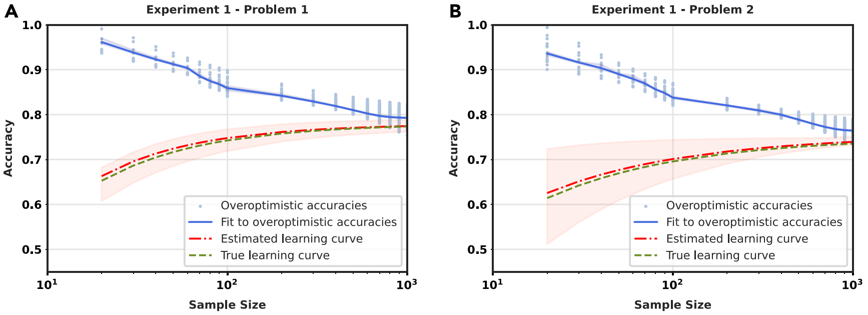

Figure 2. Evaluation of the proposed method under experiment 1 based on sampled observation models

(A) Problem 1: $A = 0.78$, $\alpha = -1.24$, $\beta = -0.76$, $\zeta = 0.45$, and $c_1 = 0.50$.

(B) Problem 2: $A = 0.75$, $\alpha = -0.75$, $\beta = -0.57$, $\zeta = 0.85$, and $c_1 = 0.40$.

The results show the overoptimistic accuracies (blue circles), the fit to the overoptimistic results (blue line), and the new estimates of the learning curve (red line) along the true learning curve (green line). The uncertainty bands represent 95% confidence intervals. The y axis is in linear scale, and the x axis is in log scale. See Figure S2 for additional cases.

Finally, experiment 3 uses reported accuracies from published meta-analyses across different digital health applications to estimate the underlying learning curve. Then, for some of the fields where data are available, we identify studies published after the meta-analysis publication date (to ensure they are not included in our algorithm) with relatively large sample sizes to compare our estimates of accuracy with those published from large data. This experiment allows us to evaluate the model on real data and estimate the amount of overoptimism across different fields of study. It is important to note that in this experiment, we extend the concept of learning curves beyond a single model and architecture[39] to encompass a specific field. Our underlying assumption is that the average performance of all ML models within this field can, in expectation, be expressed using the proposed model outlined in Equation 1.

Experiment 1 (sampling from the observation model)

In this experiment, we directly sample from the observation model in Equation 1. We simulate $K = 100$ researcher teams independently developing ML models for a binary classification problem, given a dataset with a sample size of n. The attained classification accuracies are governed by the parametric model in Equation 1. We consider the case where there is a pre-defined threshold γ_n (unknown to the recovery algorithm) that governs the decision of each of the teams to publish.

Herein, we highlight the results from two sets of model parameters, with additional cases detailed in the supplemental information.

- Problem 1: $A = 0.78$, $\alpha = -1.24$, $\beta = -0.76$, $\zeta = 0.45$, and $c_1 = 0.50$.
- Problem 2: $A = 0.75$, $\alpha = -0.75$, $\beta = -0.57$, $\zeta = 0.85$ and $c_1 = 0.40$.

We simulate our models using these two sets of parameters to generate 100 sample classification accuracies for various sample sizes n, ranging from 20 to 1,000, accepting only those above the set threshold γ_n. We aim to recover the true learning curve only from the reported accuracies.

Figures 2A and 2B show the results for problems 1 and 2, respectively. The blue points represent reported accuracies that are inflated due to overfitting and publication bias. The plots further show that the estimated learning curves (red dash-dot lines) and the true learning curves (green dashed lines) are in close agreement.

Experiment 2 (simulating overfitting to the test set and publication bias in binary classification)

We consider two binary classification problems with high-dimensional features **x** and corresponding labels **z**. A detailed description of the feature vectors and their corresponding labels is provided in the supplemental information. We consider the scenario where $K = 20$ research teams train ML models, with each team performing feature selection on the complete dataset (prior to the train-test split) to simulate a common form of overfitting.[12] Then, each team randomly selects a training set (70% of the data) and trains a different ML model. Finally, each team produces an estimate of the accuracy of their classifier on the remaining (30%) of the data. To account for publication bias, only those accuracies greater than a pre-specified threshold γ_n are reported. To establish a baseline for comparing the learning curve estimated from only the overoptimistic samples, we attain estimates of the true learning curve by iteratively increasing the sample size and training the model without overfitting to the test set and without publication bias. We repeat this process 100 times and average over these values.

Figure 3 shows the overoptimistic classification accuracies (blue circles) and the fit to these published results (solid blue line). The estimated learning curves (red dash-dot lines) are in close agreement with the true learning curves (green dashed lines). These results provide further backing to our theoretical results in the supplemental information that the true learning curve is estimable from the overoptimistic estimates using the proposed framework.

Experiment 3 (estimating the limits of prediction from meta-analyses)

Next, we turn our attention to data from published studies that use ML in a binary classification context. Our aim is to estimate the limits of using ML in a particular field from published results. We focus on published accuracies from meta-analyses of binary classification tasks for the prediction of brain disorders.[6,9–12] Several studies have reported on the overoptimism in this literature, as evidenced by the negative association in Figure 1. We consider several cases of interest.

- Case 1: classification accuracy of ML models developed using neuroimaging data[6] for classifying between patients with AD and healthy controls.
- Case 2: classification accuracy of ML models developed using speech data[9–11] for classifying between patients with AD and healthy controls.

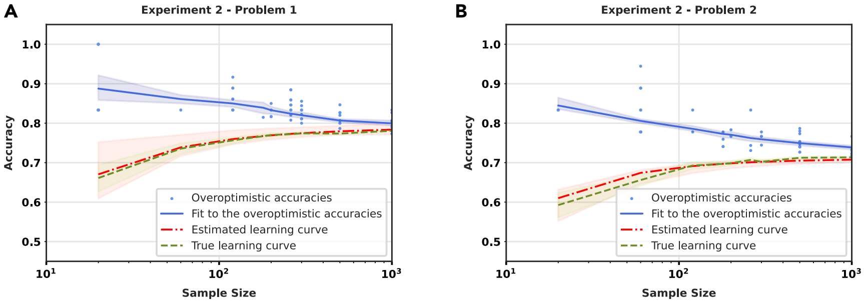

Figure 3. Evaluation of the proposed method under experiment 2 based on simulated overfitting and publication bias in binary classification

(A) Problem 1.

(B) Problem 2.

The results show the overoptimistic accuracies (blue circles), the fit to the overoptimistic results (blue line), the new estimates of the learning curve (red line), and the true learning curve (green line). The uncertainty bands represent 95% confidence intervals. The y axis is in linear scale, and the x axis is in log scale.

- Case 3: classification accuracy of ML models developed using neuroimaging data[6] for classifying between patients with ASD and healthy controls.
- Case 4: classification accuracy of ML models developed using neuroimaging data[6] for classifying between patients with ADHD and healthy controls.
- Case 5: classification accuracy of ML models developed using neuroimaging data[6] for classifying between patients with schizophrenia and healthy controls.
- Case 6: classification accuracy of ML models developed using speech data[9–11] for classifying between patients with forms of CI that are not AD (CI) and healthy controls.
- Case 7: classification accuracy of ML models developed using multi-modality data[12] for classifying between patients with ASD and healthy controls.

Here, we present the first two cases and provide the remainder in the supplemental information. Figures 4A and 4B illustrate the published classification accuracies (blue circles) and the fit to the overoptimistic results (blue), along with the estimated, de-biased learning curves (red), for the neuroimaging-based and speech-based classifications of AD, respectively. As depicted in Figure 4, the vast majority of the individual studies from the meta-analyses (blue circles) fall outside the upper confidence band of the corrected learning curves (red dash-dot lines); these are identified as overoptimistic published results by the model. Our extended results in the supplemental information show a similar trend; however, the extent of overoptimism varies across domains.

We compare the estimated limiting performance of the parametric model (i.e., parameter $\tilde{A}$) with large-scale studies published after the publication of the meta-analyses above in Table 1. In some cases, large-scale studies are not available, as collecting large datasets in these fields can be challenging. Nevertheless, we postulate that the reported performance of ML models trained with larger sample sizes should be more realistic, as the impacts of publication bias and overfitting diminish with increasing sample size. Table 1 lists the estimated limiting performance across different fields in digital health, including 95% confidence intervals, alongside the reported accuracy of recent large-scale individual studies with their corresponding sample sizes. The results indicate that reported classification accuracies from larger studies[40,41] fall within the confidence intervals of the limiting performance and the estimated learning trajectories for the prediction of AD and schizophrenia. On the other hand, the reported accuracy for the prediction of ASD by Prasad et al.[42] is below the anticipated value, estimated by the new, recovered learning curve. This suggests the possibility of further improvement in model development for the large sample-size study published in Prasad et al.[42]

Limitations of the study

As demonstrated in the supplemental information, there is no fundamental identifiability issue in estimating the trajectory of ML models from overoptimistic observations when many observation are available. However, the theoretical results provide limited insight into the sample size of observations required to recover the parameters. Furthermore, when applying the model to results from meta-analyses, evaluating the resulting estimates is difficult due to the absence of a ground truth. Although we recommend comparisons with individual, large-scale studies, such studies are often unavailable. Even when such studies are available, the complete model development process may remain opaque, and it is uncertain whether these models have achieved their maximum performance potential.

Highlighting another limitation of our work, the proposed model employs a hard threshold for the selection mechanism to simulate publication bias. However, this assumption may not fully capture the nuanced decision-making processes observed in real-world scenarios. To align the real-world observations with the assumptions of the model, we consider the published studies below the 0.1 quantile line as outliers. The exploration of a soft threshold remains an avenue for future work.

Additionally, the assumption about $n^{-0.5}$ dependency in our estimator, adopted from adaptive data analysis,[15,34] primarily addresses test set reuse. However, overfitting and data leakage could occur in several ways that our model does not comprehensively account for.[7] For example, in the presence of proxy variables, this assumption does not hold, as the sample size becomes largely irrelevant. Investigating these issues and extending the model to better capture broader forms of overfitting and publication bias is an important direction for future work.

Finally, our theoretical analysis in the supplemental information assumes that β is bounded away from 0.5 to ensure that the two exponential components are decoupled and linearly independent. However, we do not explicitly enforce this constraint in the optimization problem. Future extensions of the theoretical analysis should include a loosening of this constraint.

METHODS

In this section, we first discuss the causes behind overoptimism in ML-based science literature and the reasoning behind the

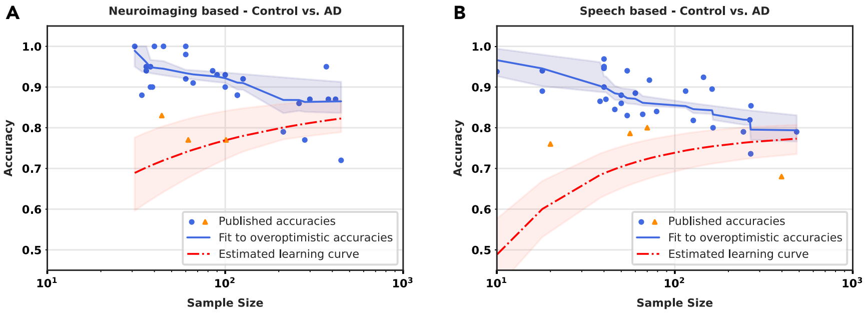

Figure 4. Estimates of learning trajectories under experiment 3 and from meta-analyses

(A) Meta-analysis of neuroimaging-based prediction of AD.[6]

(B) Meta-analyses of speech-based prediction of AD.[9–11]

The results show the new estimates of the empirical learning curve (red), the published results (blue circles), and the fit to the observations (blue). Orange triangles represent outliers excluded from the analysis. The uncertainty bands represent 95% confidence intervals. The y axis is in linear scale, and the x axis is in log scale. See also Figures S3 and S4.

proposed model. Then, we provide the details of the proposed solution toward alleviating the overoptimism in published results.

Observation model for reported classification accuracy

Learning curves are powerful tools that provide insight into the capabilities of ML models in a specific domain. In the context of supervised learning, the learning curves[43] describe the predictive performance of an ML model by illustrating an estimate of its performance for different sample sizes. One can construct an empirical learning curve by analyzing the classification accuracy of an ML model across a range of sample sizes. A true learning curve accurately represents an ML model's performance for a specific problem, offering insights into how the model's effectiveness changes with the addition of more data samples. Predicting an ML model's capabilities accurately is crucial for guiding future investments and conserving resources by avoiding inefficiencies. However, as demonstrated in Figure 1, the pervasive overoptimism in ML-based science research can obscure the identification of genuine learning curves. This issue of overoptimism, as suggested by Berisha et al.,[8] is likely driven by overfitting and publication bias within the ML scientific community.

Overfitting

The literature on ML in scientific research highlights the issue of overfitting to the test set, which can result in overly optimistic estimates of classification accuracy. One prevalent practice contributing to this issue is adaptive data analysis,[15,36] where researchers repeatedly use the same test set to evaluate various models, features, or hyperparameters in an effort to improve model performance. This approach is widespread in the ML community, where maximizing performance metrics is the primary goal. However, this practice becomes problematic in small-sample-size regimes, as models are more susceptible to learning from noise and unique dataset characteristics, thus increasing the risk of overfitting.[16] Evidence suggests that reusing a test set multiple times during model development introduces significant overfitting errors, which optimistically scales as $\sqrt{t/n}$, with t representing the test set's number of reuses.[15,34] Our proposed model, outlined in Equation 1, integrates this understanding by incorporating Gaussian noise with an elevated mean of $\zeta n^{-0.5}$ to account for overfitting effects.

Publication bias

Inspired by the work on selection bias in econometrics, we model publication bias in a similar fashion.[38] We assume that K research groups are independently working on a dataset of sample size n, with each group building an ML model using these data without influence from other groups and estimating its accuracy. Let $\mathbf{y}^{\star}_{n_k}$ denote the classification accuracy obtained by the k-th research team, where $k = 1, 2, \dots, K$ for datasets of size n. Then, we can write

$$\begin{cases} \mathbf{y}_{n_k} = \mathbf{y}^{\star}_{n_k} \text{ for } k = 1, \dots, K & \text{if} \quad \mathbf{y}^{\star}_{n_k} \geq \gamma_n, \\ \quad \mathbf{y}^{\star}_{n_k} \text{ not observed} & \text{if} \quad \mathbf{y}^{\star}_{n_k} < \gamma_n \end{cases}, \qquad \text{(Equation 3)}$$

where $\mathbf{y}_{n_k}$ denotes the published classification accuracy for team k using sample size n to build and test their model. Equation 3 indicates that accuracy $\mathbf{y}^{\star}_{n_k}$ is observed (published) only if it surpasses a threshold γ_n; otherwise, it remains unobserved (unpublished). Consequently, the count of published accuracies, M, is less than the total number of conducted studies and their corresponding accuracies, implying $M \leq K$. Acknowledging that publication bias more significantly affects studies with smaller sample sizes and diminishes as sample size increases, we model γ_n as a decreasing function of sample size. In addition, models trained with larger sample sizes generally exhibit higher reliability, and consequently, researchers are more inclined to publish studies with large sample sizes. This further supports the rationale that the threshold γ_n should be modeled as a decreasing function of n.

The proposed selection model (Equation 3) highlights publication bias, indicating that some studies' results may remain unpublished. To mitigate this bias and accurately estimate the true learning trajectory of ML models, we employ regression with truncated samples. This approach aims to estimate the statistics of Gaussian distributions from the observed data. We outline a solution for estimating these statistics and the parameters of the learning curve. Additionally, a theoretical analysis on the identifiability of this learning trajectory, through the lens of identifying the mean and variance of truncated Gaussian distributions, is presented in the supplemental information.

Proposed solution

Assume $x \sim \mathcal{N}(\mu, \sigma^2)$; then, we can write[44]

$$\mathbb{E}[x|x\rangle\gamma] = \mu + \sigma \frac{\phi\left(\frac{\gamma - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{\gamma - \mu}{\sigma}\right)}, \qquad \text{(Equation 4)}$$

and

$$\mathbf{Var}[x|x\rangle\gamma] = \sigma^2\left[1+\frac{\gamma-\mu}{\sigma}\left[\frac{\phi\left(\frac{\gamma-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\gamma-\mu}{\sigma}\right)}\right]-\left[\frac{\phi\left(\frac{\gamma-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\gamma-\mu}{\sigma}\right)}\right]^2\right], \quad \text{(Equation 5)}$$

where $\phi(.)$ denotes the probability density function (PDF) of a standard normal distribution, $\Phi(.)$ is the cumulative density function (CDF) of a standard normal distribution, and $\mathbf{Var}[x]$ denotes the variance of x. The ratio in Equation 4 is known as the inverse of the Mills ratio.[45] Given the parametric model in Equation 1 with $\mathbf{y}_n^\star \sim \mathcal{N}(A+\alpha n^\beta+\zeta n^{-0.5},\sigma_n^2)$, from Equation 4, we can write

$$\mathbb{E}[\mathbf{y}_n|\mathbf{y}_n\rangle\gamma_n] = A+\alpha n^\beta+\zeta n^{-0.5}+\sigma_n\psi(n;A,\gamma_n,\alpha,\beta,\zeta,c_1), \quad \text{(Equation 6)}$$

where $\sigma_n = c_1/\sqrt{n}$ and

$$\psi(n;A,\gamma_n,\alpha,\beta,\zeta,c_1) = \frac{\phi\left(\frac{\gamma_n-A-\alpha n^\beta-\zeta n^{-0.5}}{\sigma_n}\right)}{1-\Phi\left(\frac{\gamma_n-A-\alpha n^\beta-\zeta n^{-0.5}}{\sigma_n}\right)}. \quad \text{(Equation 7)}$$

Table 1. Estimated limiting performance and 95% confidence intervals from meta-analyses of prediction of brain disorders

Problem	Limiting Performance	Reference	Sample size	accuracy
NI HC vs. AD	0.88, [0.84 – 0.96]	Lu et al.[40]	85,721	0.91
NI HC vs. SZ	0.73, [0.70 – 0.81]	Yan et al.[41]	1,100	0.80
NI HC vs. ASD	0.87, [0.80 – 0.93]	Prasad et al.[42]	1,026	0.75

This table also includes the published classification accuracy and sample size from individual studies. HC, healthy control group; NI, neuroimaging; AD, Alzheimer's disease; SZ, schizophrenia; ASD, autism spectrum disorder. See also Table S1.

For brevity, we use the notation $\psi(n)$ for $\psi(n;A,\gamma_n,\alpha,\beta,\zeta,c_1)$. The average *published* classification accuracy can be expressed as in Equation 6, where the summation of the first two terms models the true learning curve, i.e., $A+\alpha n^\beta$, the term $\zeta n^{-0.5}$ models overfitting, and $\psi(n)$ in Equation 7 models the impact of publication bias. From Equation 5, the variance of this estimate is

$$\mathbf{Var}[\mathbf{y}_n|\mathbf{y}_n\rangle\gamma_n] = \sigma_n^2\left(1+\frac{\gamma_n-A-\alpha n^\beta-\zeta n^{-0.5}}{\sigma_n}\psi(n)-\psi^2(n)\right). \quad \text{(Equation 8)}$$

Cohen demonstrated that the method of moments can be used to estimate the parameters of a truncated Gaussian distribution by matching the empirical mean and variance of the observed samples to those expected from the truncated distribution.[44] Building on this concept, we propose the following multi-objective optimization problem, which aims to find the best match between the observed data's statistical properties and those of a theoretical truncated normal distribution:

$$\begin{aligned}&\min_{A,\alpha,\beta,\zeta,c_1} f_m(A,\alpha,\beta,\zeta,c_1)\\ &\text{subject to}\\ &f_1(A,\alpha,\beta,\zeta,c_1) = \sum_{n\in \mathrm{N}}\left(\overline{\mathbf{y}}_n - A-\alpha n^\beta-\zeta n^{-0.5}-\sigma_n\psi(n)\right)^2\\ &f_2(A,\alpha,\beta,\zeta,c_1) = \sum_{n\in \mathrm{N}}\left(\overline{\sigma}_n^2-\sigma_n^2\left(1+\frac{\gamma_n-A-\alpha n^\beta-\zeta n^{-0.5}}{\sigma_n}\psi(n)-\psi^2(n)\right)\right)^2\\ &0.5\le A\le 1,\ -2\le\alpha\le -0.5,\ -1\le\beta\le 0, 0\le\zeta\le 1, 0<c_1\le 0.5,\end{aligned} \quad \text{(Equation 9)}$$

where $\overline{\mathbf{y}}_n$ and $\overline{\sigma}_n^2$ are the sample mean and sample variance of the reported accuracies given a sample size n, and set N contains all sample sizes used in published studies within a specific field. We use the non-dominated sorting genetic algorithm (NSGAII)[46] to solve the non-convex optimization program in Equation 9. To find the optimal solution, we first construct the pareto front of non-dominated solutions to this problem. Among this set, we select the optimal solution by using $f_1(A,\alpha,\beta,\zeta,c_1)$ as the augmented scalarization function (ASF).[47]

We denote the new estimated learning curve as $\tilde{A}+\tilde{\alpha}n^{\tilde{\beta}}$, where $\tilde{A}$, $\tilde{\alpha}$, and $\tilde{\beta}$ are the estimates of the parameters A, α, and β, respectively; we denote the fit to the overoptimistic accuracies using Equation 6. We build the confidence intervals around these estimates using bootstrapping[48,49] where we sampled from the reported accuracies randomly and with replacement to construct a bootstrap sample and repeat this process 10,000 times.

In contrast to traditional regression analyses with truncated samples, which often presuppose knowledge of a threshold γ_n,[38,44] our approach does not assume this threshold is known beforehand. We use the maximum likelihood estimate of the threshold, which is the smallest observed value among the data, known as the minimum order statistic.[50] To smooth this statistic, we apply a sliding window technique, with a window length of 2 and a stride of 1, across reported sample sizes, γ_n, taking the lowest reported accuracy within each window as our threshold.

CONCLUSION

There is evidence of a prevalence of overly optimistic results in ML-based science, including in digital health literature. We posit that this overoptimism stems from publication bias and overfitting, phenomena that distort our understanding of a model's true performance. In this paper, we proposed a novel model based on parametric learning curves to express the reported classification accuracy of ML models, taking into account both overfitting and publication bias. We further proposed a solution based on regression with truncated samples to alleviate overoptimism in the literature. This novel technique paves the way for a more accurate understanding of the capabilities of ML models in specific fields using existing, yet overoptimistic, published results. Our results on synthetic data, based on the observation model, demonstrate the success of the proposed solution. Meanwhile, results on real data from meta-analyses in digital health reveal a divergence between the reported accuracies and the newly estimated learning curves. The estimated performance limits and convergence rates of these curves can help reveal the true capabilities of ML models across these fields. Furthermore, they can substantiate trust in published results, especially those obtained with limited samples, should they fall within the confidence intervals of these newly estimated learning curves.

Finally, parametric learning curves provide valuable insights into the trade-offs between the cost of data acquisition and the potential performance of a model. They help identify how additional data might improve model capabilities, aiding in cost-benefit analyses for study design.[51],[52] Our approach is well suited to complement these methods by providing a more accurate estimation of true model performance and could directly inform guidelines for appropriate sample sizes in ML studies. While our approach provides a means to correct for overoptimism in published results, it is complementary to broader systemic changes, such as encouraging the publication of low-accuracy results or mandating larger sample sizes for socially impactful models.[53] By quantifying the limitations of current practices, our method highlights the need for such reforms and can serve as a tool to support their adoption.

RESOURCE AVAILABILITY

Lead contact

Requests for resources should be directed to the lead contact, Pouria Saidi (psaidi@asu.edu).

Materials availability

This study did not generate new materials.

Data and code availability

All data and code have been uploaded to a Figshare repository. Computer code is available at Figshare: https://doi.org/10.6084/m9.figshare.27093958.[54] Data are available at Figshare: https://doi.org/10.6084/m9.figshare.27080794.[55]

ACKNOWLEDGMENTS

This work was supported in part by ONR grant N00014-21-1-2615, the John and Tami Marick Foundation, and the National Science Foundation under awards CCF-2007688 and CCF-2048223.

AUTHOR CONTRIBUTIONS

Conceptualization, G.D. and V.B.; methodology, P.S., G.D., and V.B.; software, P.S.; formal analysis, P.S., G.D., and V.B.; writing – original draft, P.S.; writing – review & editing, G.D. and V.B.; supervision, G.D. and V.B.; funding acquisition, G.D. and V.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT in order to improve the readability and language of the work. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2025.101185.

Received: September 24, 2024
Revised: December 23, 2024
Accepted: January 30, 2025
Published: February 25, 2025

REFERENCES

1. Rosenbusch, H., Soldner, F., Evans, A.M., and Zeelenberg, M. (2021). Supervised machine learning methods in Psychology: A practical introduction with annotated R code. Soc. Personal. Psychol. Compass *15*, e12579.
2. Vandewiele, G., Dehaene, I., Kovács, G., Sterckx, L., Janssens, O., Ongenae, F., De Backere, F., De Turck, F., Roelens, K., Decruyenaere, J., et al. (2021). Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. Artif. Intell. Med. *111*, 101987.
3. Alenezi, H.S., and Faisal, M.H. (2020). Utilizing crowdsourcing and machine learning in education: Literature review. Educ. Inf. Technol. *25*, 2971–2986.
4. Mathews, S.C., McShea, M.J., Hanley, C.L., Ravitz, A., Labrique, A.B., and Cohen, A.B. (2019). Digital health: a path to validation. NPJ Digit. Med. *2*, 38.
5. Berisha, V., Krantsevich, C., Stegmann, G., Hahn, S., and Liss, J. (2022). Are Reported Accuracies in the Clinical Speech Machine Learning Literature Overoptimistic? (Interspeech), pp. 2453–2457.
6. Arbabshirani, M.R., Plis, S., Sui, J., and Calhoun, V.D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. Neuroimage *145*, 137–165.
7. Kapoor, S., and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns *4*, 100804.
8. Berisha, V., Krantsevich, C., Hahn, P.R., Hahn, S., Dasarathy, G., Turaga, P., and Liss, J. (2021). Digital medicine and the curse of dimensionality. NPJ. Digit. Med. *4*, 153.
9. de la Fuente Garcia, S., Ritchie, C.W., and Luz, S. (2020). Artificial intelligence, speech, and language processing approaches to monitoring

Alzheimer's disease: a systematic review. J. Alzheimers Dis. *78*, 1547–1574.

10. Petti, U., Baker, S., and Korhonen, A. (2020). A systematic literature review of automatic Alzheimer's disease detection from speech and language. J. Am. Med. Inf. Assoc. *27*, 1784–1797.

11. Martínez-Nicolás, I., Llorente, T.E., Martínez-Sánchez, F., and Meilán, J.J.G. (2021). Ten years of research on automatic voice and speech analysis of people with Alzheimer's disease and mild cognitive impairment: a systematic review article. Front. Psychol. *12*, 620251.

12. Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A.J. (2019). Machine learning algorithm validation with a limited sample size. PLoS One *14*, e0224365.

13. Viering, T., and Loog, M. (2023). The shape of learning curves: A review. IEEE Trans. Pattern Anal. Mach. Intell. *45*, 7799–7819.

14. Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Trans. Knowl. Discov. Data *6*, 1–21.

15. Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A.L. (2015). Preserving statistical validity in adaptive data analysis. In Proceedings of the forty-seventh annual ACM symposium on Theory of computing (Association for Computing Machinery), pp. 117–126.

16. Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., and Schmidt, L. (2019). A meta-analysis of overfitting in machine learning. Adv. Neural Inf. Process. Syst. *32*.

17. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning, *112* (Springer).

18. Serra-Garcia, M., and Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. Sci. Adv. *7*, eabd1705.

19. Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychol. Bull. *86*, 638.

20. Ioannidis, J.P.A. (2005). Why most published research findings are false. PLoS Med. *2*, e124.

21. Raji, I.D., Kumar, I.E., Horowitz, A., and Selbst, A. (2022). The fallacy of AI functionality. In 2022 ACM Conference on Fairness, Accountability, and Transparency (Association for Computing Machinery), pp. 959–972.

22. Yawer, B.A., Liss, J., and Berisha, V. (2023). Reliability and validity of a widely-available AI tool for assessment of stress based on speech. Sci. Rep. *13*, 20224.

23. Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. Nat. Mach. Intell. *1*, 389–399.

24. Memorandum on increasing access to the results of federally funded scientific research (2013). https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

25. Memorandum on ensuring free, immediate, and equitable access to federally funded research (2022). https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf.

26. Hoiem, D., Gupta, T., Li, Z., and Shlapentokh-Rothman, M. (2021). Learning curves for analysis of deep networks. In International conference on machine learning, M. Meila and T. Zhang, eds. (PMLR), pp. 4287–4296.

27. Cortes, C., Jackel, L.D., Solla, S., Vapnik, V., and Denker, J. (1993). Learning curves: Asymptotic values and rate of convergence. Adv. Neural Inf. Process. Syst. *6*, 327–334.

28. John, G.H., and Langley, P. (1996). Static versus dynamic sampling for data mining. Kdd *96*, 367–370.

29. Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M.M.A., Yang, Y., and Zhou, Y. (2017). Deep learning scaling is predictable, empirically. Preprint at arXiv. https://doi.org/10.48550/arXiv.1712.00409.

30. Kolachina, P., Cancedda, N., Dymetman, M., and Venkatapathy, S. (2012). Prediction of learning curves in machine translation. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 22–30.

31. Artico, I., Smolyarenko, I., Vinciotti, V., and Wit, E.C. (2020). How rare are power-law networks really? Proc. Math. Phys. Eng. Sci. *476*, 20190742.

32. Kumamoto, S.-I., and Kamihigashi, T. (2018). Power laws in stochastic processes for social phenomena: An introductory review. Front. Physiol. *6*, 20.

33. Meyer, S., and Held, L. (2014). Power-law models for infectious disease spread. Ann. Appl. Stat. *8*, 1612–1639.

34. Arora, S., and Zhang, Y. (2021). Rip Van Winkle's razor: A simple estimate of overfit to test data. Preprint at arXiv. https://doi.org/10.48550/arXiv.2102.13189.

35. Vapnik, V. (2006). Estimation of Dependences Based on Empirical Data (Springer Science & Business Media).

36. Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015). Generalization in adaptive data analysis and holdout reuse. Adv. Neural Inf. Process. Syst. *28*. https://doi.org/10.48550/arXiv.1506.02629.

37. Easterbrook, P.J., Berlin, J.A., Gopalan, R., and Matthews, D.R. (1991). Publication bias in clinical research. Lancet *337*, 867–872.

38. Amemiya, T. (1984). Tobit models: A survey. J. Econom. *24*, 3–61.

39. Ruan, Y., Maddison, C.J., and Hashimoto, T. (2024). Observational scaling laws and the predictability of language model performance. Preprint at arXiv. https://doi.org/10.48550/arXiv.2405.10938.

40. Lu, B., Li, H.-X., Chang, Z.-K., Li, L., Chen, N.-X., Zhu, Z.-C., Zhou, H.-X., Li, X.-Y., Wang, Y.-W., Cui, S.-X., et al. (2022). A practical alzheimer's disease classifier via brain imaging-based deep learning on 85,721 samples. J. Big Data *9*, 101.

41. Yan, W., Calhoun, V., Song, M., Cui, Y., Yan, H., Liu, S., Fan, L., Zuo, N., Yang, Z., Xu, K., et al. (2019). Discriminating Schizophrenia using recurrent neural network applied on time courses of multi-site fMRI data. EBioMedicine *47*, 543–552.

42. Prasad, P.K.C., Khare, Y., Dadi, K., Vinod, P., and Surampudi, B.R. (2022). Deep learning approach for classification and interpretation of Autism spectrum disorder. In 2022 International Joint Conference on Neural Networks (IJCNN) (IEEE), pp. 1–8.

43. Mohr, F., and van Rijn, J.N. (2024). Learning curves for decision making in supervised machine learning: a survey. Mach. Learn. *113*, 8371–8425.

44. Cohen, A.C. (1991). Truncated and Censored Samples: Theory and Applications (CRC press).

45. Amemiya, T. (1985). Advanced Econometrics (Harvard university press).

46. Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA II. IEEE Trans. Evol. Comput. *6*, 182–197.

47. Wierzbicki, A.P. (1980). The use of reference objectives in multiobjective optimization. In Multiple Criteria Decision Making Theory and Application: Proceedings of the Third Conference Hagen/Königswinter, West Germany, August 20–24, 1979, G. Fandel and T. Gal, eds. (Springer), pp. 468–486.

48. Efron, B., and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat. Sci. *1*, 54–75.

49. Efron, B., and Tibshirani, R.J. (1994). An Introduction to the Bootstrap (CRC press).

50. Zuehlke, T.W. (2003). Estimation of a Tobit model with unknown censoring threshold. Appl. Econ. *35*, 1163–1169.

51. Figueroa, R.L., Zeng-Treitler, Q., Kandula, S., and Ngo, L.H. (2012). Predicting sample size required for classification performance. BMC Med. Inf. Decis. Making *12*, 8–10.

52. Larracy, R., Phinyomark, A., and Scheme, E. (2021). Machine learning model validation for early stage studies with small sample sizes. In 2021

43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (IEEE), pp. 2314–2319.

53. Kapoor, S., Cantrell, E., Peng, K., Pham, T.H., Bail, C.A., Gundersen, O.E., Hofman, J.M., Hullman, J., Lones, M.A., Malik, M.M., et al. (2023). Reforms: Reporting standards for machine learning based science. Preprint at arXiv. https://doi.org/10.48550/arXiv.2308.07832.

54. Saidi, P., Dasarathy, G., and Berisha, V. (2025). Unraveling overoptimism and publication bias in ML-driven science - Computer code for experiments 1,2, and 3. Figshare. Software. https://doi.org/10.6084/m9.figshare.27093958.

55. Saidi, P., Dasarathy, G., and Berisha, V. (2025). Unraveling Overoptimism and Publication Bias in ML-Driven Science - Data for Experiment 2 and 3 (Figshare). Data. https://doi.org/10.6084/m9.figshare.27080794.