Evaluating Double Descent in Machine Learning: Insights from Tree-Based Models Applied to a Genomic Prediction Task

Guillermo Comesaña Cimadevila
Department of Life Sciences, University of Bath
gcc46@bath.ac.uk

Abstract

Classical learning theory describes a well-characterised U-shaped relationship between model complexity and prediction error, reflecting a transition from underfitting in underparameterised regimes to overfitting as complexity grows. Recent work, however, has introduced the notion of a second descent in test error beyond the interpolation threshold—giving rise to the so-called double descent phenomenon. While double descent has been studied extensively in the context of deep learning, it has also been reported in simpler models, including decision trees and gradient boosting. In this work, we revisit these claims through the lens of classical machine learning applied to a biological classification task: predicting isoniazid resistance in *Mycobacterium tuberculosis* using whole-genome sequencing data. We systematically vary model complexity along two orthogonal axes—learner capacity (e.g., P^{leaf}, P^{boost}) and ensemble size (i.e., P^{ens})—and show that double descent consistently emerges only when complexity is scaled jointly across these axes. When either axis is held fixed, generalisation behaviour reverts to classical U- or L-shaped patterns. These results are replicated on a synthetic benchmark and support the unfolding hypothesis, which attributes double descent to the projection of distinct generalisation regimes onto a single complexity axis. Our findings underscore the importance of treating model complexity as a multidimensional construct when analysing generalisation behaviour. All code and reproducibility materials are available at: https://github.com/guillermocomesanacimadevila/Demystifying-Double-Descent-in-ML.

1 Introduction

The traditional relationship between model complexity and prediction error has long been explained by the bias–variance trade-off, which posits that prediction error follows a U-shaped curve (Figure 1, left pannel) as model complexity increases [1, 2]. In this framework, models with insufficient complexity exhibit high bias and underfit the data, while overly complex models tend to memorise the training data, leading to high variance and poor generalisation to unseen inputs (i.e., overfitting) [3, 2]. The optimal predictive performance is thought to lie at an intermediate point of complexity, where bias and variance are minimised [4]. This foundational concept underpins widely used model selection strategies such as cross-validation, regularisation, and information-theoretic criteria, including the Akaike and Bayesian Information Criteria [5].

This view implicitly assumes that increasing model complexity beyond the interpolation threshold—where the number of model parameters equals the number of training samples—would continue to degrade generalisation [6, 7]. However, recent empirical findings in modern machine learning challenge this assumption. Notably, overparameterised models such as deep neural networks can achieve near-zero training error and yet continue to generalise effectively, defying the predictions of the classical U-shaped error curve [8, 9, 10]. To account for this observation, Belkin et al. [13] proposed the double descent phenomenon (Figure 1, right pannel), wherein test error initially decreases with complexity, rises near the interpolation threshold, and then decreases again as complexity increases further. The double descent framework suggests that increasing model complexity can, under certain conditions, lead to improved generalisation even in highly overparameterised regimes [11]. While originally observed in deep learning models, subsequent work has shown that double descent can emerge in simpler settings, including kernel methods, decision trees, and even ordinary least-squares regression [12, 13].

Nevertheless, its underlying theoretical basis remains a topic of ongoing debate [14]. Recent critiques argue that double descent may be a visual artefact of collapsing multidimensional model complexity into a single axis [15, 14]. Curth et al. [16] built on this view by proposing that the observed curve arises from the projection of two separate generalisation regimes—the classical bias–variance trade-off and a high-dimensional interpolation regime—onto a shared axis. In this formulation, double descent does not reflect a continuous generalisation phenomenon but rather the unfolding of separate complexity dynamics [16] (Figure 1). Despite growing theoretical interest, empirical studies of double descent remain limited. Prior work has primarily focused on least squares regression or deep learning architectures, with few investigations in classical tree-based models such as decision trees and gradient boosting—Curth et al. [16] being a notable exception. Moreover, the presence of double descent in real-world biological datasets remains unexplored. In this study, we address this gap by applying the double descent framework to a clinically relevant classification task: identifying resistance to isoniazid in *Mycobacterium tuberculosis* from whole-genome sequencing data. *Mycobacterium tuberculosis* remains the leading cause of death from a single bacterial pathogen, with over 1.25 million deaths in 2024 alone [17]. Resistance to isoniazid, a first-line anti-tuberculosis drug, arises from spontaneous point mutations rather than horizontal gene transfer, making single nucleotide polymorphism (SNP)-based prediction both feasible and clinically relevant [18, 19].

Building on this clinical relevance, we apply the experimental frameworks of Belkin et al. [13] and Curth et al. [16] to investigate whether double descent arises in classical machine learning models trained on genomic data from the Comprehensive Resistance Prediction for Tuberculosis (CRyPTIC) consortium [20]. Specifically, we train decision trees and gradient boosting regressors to predict isoniazid resistance from whole-genome sequencing data and assess whether prediction error exhibits the characteristic double descent curve. In doing so, we aim to evaluate whether any observed patterns align with the "unfolding" hypothesis proposed by Curth et al. [16], thereby determining whether double descent constitutes a real generalisation principle or a representational artefact of model parameterisation. We hypothesise that double descent uniquely emerges when model complexity is projected along a unidimensional axis, and that classical bias–variance dynamics reappear when complexity is treated as a multidimensional construct.

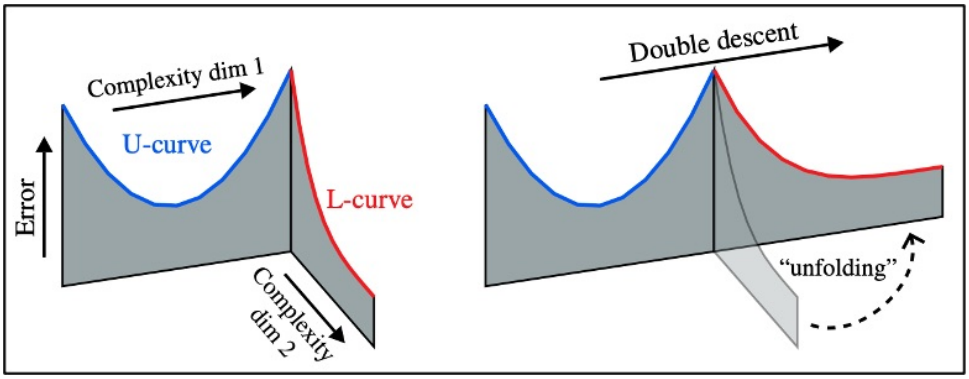

Figure 1. Double descent illustration emerging from two complexity axes. Left: error varies across two model complexity dimensions, forming a U-curve (blue) along one axis and an L-curve (red) along the other. Right: collapsing these dimensions produces the double descent curve, suggesting it may arise from merging distinct generalisation behaviours. Figure adapted from Curth et al. [16].

2 Methods

Data Sources

Whole-genome sequencing data were sourced from the June 2022 public release of the CRyPTIC consortium, comprising 12,289 *Mycobacterium tuberculosis* isolates from 23 countries. Each isolate was annotated with phenotypic classifications for resistance or susceptibility to 13 antibiotics. Associated variant data were obtained in Variant Call Format (VCF), and metadata were retrieved from the accompanying CSV files. Data were accessed from the European Bioinformatics Institute's public FTP repository: `https://ftp.ebi.ac.uk/pub/databases/cryptic/release_june2022/reuse/`. An overview of the full data processing and analysis pipeline is presented in Figure 2.

Sample Selection and Pre-Processing

To ensure computational tractability and class balance, we selected a stratified subsample of $n = 500$ isolates: 250 resistant and 250 susceptible to isoniazid. Only isolates labelled with a "HIGH" phenotype quality—defined by CRyPTIC as agreement across at least two minimum inhibitory concentration assays—were retained to reduce label noise. This filtering step removed 3,370 low-confidence samples. Variant data were then parsed from the corresponding VCF files. During quality control, all insertion–deletion mutations (INDELs) and loci with missing genotype calls were removed. This reduced the average number of loci per isolate from 1,767 to 1,531. For each SNP, we extracted four features: genomic position (POS), genotype (GT), read depth (DP), and genotype confidence (GT_CONF). Genotypes were encoded numerically as 0 (homozygous reference), 1 (heterozygous), and 2 (homozygous alternate). The final feature matrix had dimensions $765{,}413 \times 4$. Although the feature-to-sample ratio was high, no additional dimensionality reduction was applied. This decision follows the conventions of Belkin et al. [13] and Curth et al. [16], who recommend preserving high-dimensional structure when evaluating double descent, as it facilitates overfitting, thus ensuring a more rapid reach to the interpolation threshold [21].

Machine Learning Framework

We evaluated three regression-based machine learning models: decision tree regressors, random forest regressors, and gradient boosting regressors. Although the prediction task is inherently binary, we adopted a squared-loss regression framework to align with the methodology of Belkin et al. [13] and Curth et al. [16], who demonstrated double descent under this loss function. Phenotypic labels were binarised as $y \in \{0, 1\}$, and mean squared error (MSE) on the test set was used as the generalisation metric. Model complexity was varied systematically along two orthogonal axes: base learner capacity and ensemble size. For decision tree-based models (including random forests), complexity was parameterised using the number of terminal leaf nodes per tree (P^{leaf}) and the number of estimators in the ensemble (P^{ens}). Three experimental regimes were implemented. First, we varied $P^{\text{leaf}} \in \{2, \ldots, 500\}$ with $P^{\text{ens}} \in \{1, 5, 10, 50\}$ held fixed per sweep. Second, we varied $P^{\text{ens}} \in \{1, \ldots, 50\}$ with $P^{\text{leaf}} \in \{20, 50, 100, 500\}$ held fixed per sweep. Third, we used a composite scaling design: $P^{\text{leaf}} \in \{50, 100, 200, 500\}$ was increased within a single tree up to $L_{\max}$ and then P^{ens} was scaled from RF1 to RF50, simulating capacity growth past the interpolation threshold [16]. Gradient boosting models followed the same logic with constraints: base learners had $P^{\text{leaf}} \leq 10$ and we used a high learning rate $\gamma = 0.85$ to encourage rapid interpolation [22]. In the first experiment, $P^{\text{boost}} \in \{10, 20, 50, 100, 200\}$ with $P^{\text{ens}} \in \{1, 5, 10, 50\}$. In the second, $P^{\text{boost}} \in \{20, 50, 100, 200\}$ with $P^{\text{ens}} \in \{1, \ldots, 50\}$. In the third, we fixed $P^{\text{boost}} = 200$ and scaled $P^{\text{ens}} \in \{1, \ldots, 50\}$. All evaluations used a consistent grid and the same 70:30 train–test split.

Synthetic Baseline

To validate observed dynamics in a controlled setting, we reproduced all experiments on a synthetic dataset proposed by Friedman (1991) [23] and used in contemporary double descent work [16]. The dataset contained $n = 500$ samples and $p = 50$ independent features. For each sample, we draw a feature vector $X = (X_1, \ldots, X_p)$ with $X_i \overset{\text{i.i.d.}}{\sim} U(0, 1)$, so that the coordinates are independent and identically distributed on $[0, 1]$. The regression target y is generated by the standard Friedman #1 formula:

$$y \;=\; \sin\!\big(\pi X_1 X_2\big) \;+\; 2\,(X_3 - 0.5)^2 \;+\; X_4 \;+\; 0.5\,X_5 \;+\; \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 1). \tag{1}$$

This benchmark probes non-linear interactions, sparsity (only the first five coordinates are signal-bearing), and additive noise [13, 16]. All models used the same hyperparameter grid and a 70:30 train–test split.

Code Availability

All preprocessing, feature extraction, and modelling were conducted using the Cloud Infrastructure for Microbial Bioinformatics [24]. Variant filtering used Bash scripts; downstream modelling used Python 3.12 with scikit-learn 1.6.1 [25], NumPy 2.2.3 [26], pandas 2.2.3 [27], Matplotlib 3.10.1 [28], and SciPy 1.15.2 [29]. A fixed random seed ensured reproducibility. All code and documentation: https://github.com/guillermocomesanacimadevila/Demystifying-Double-Descent-in-ML.

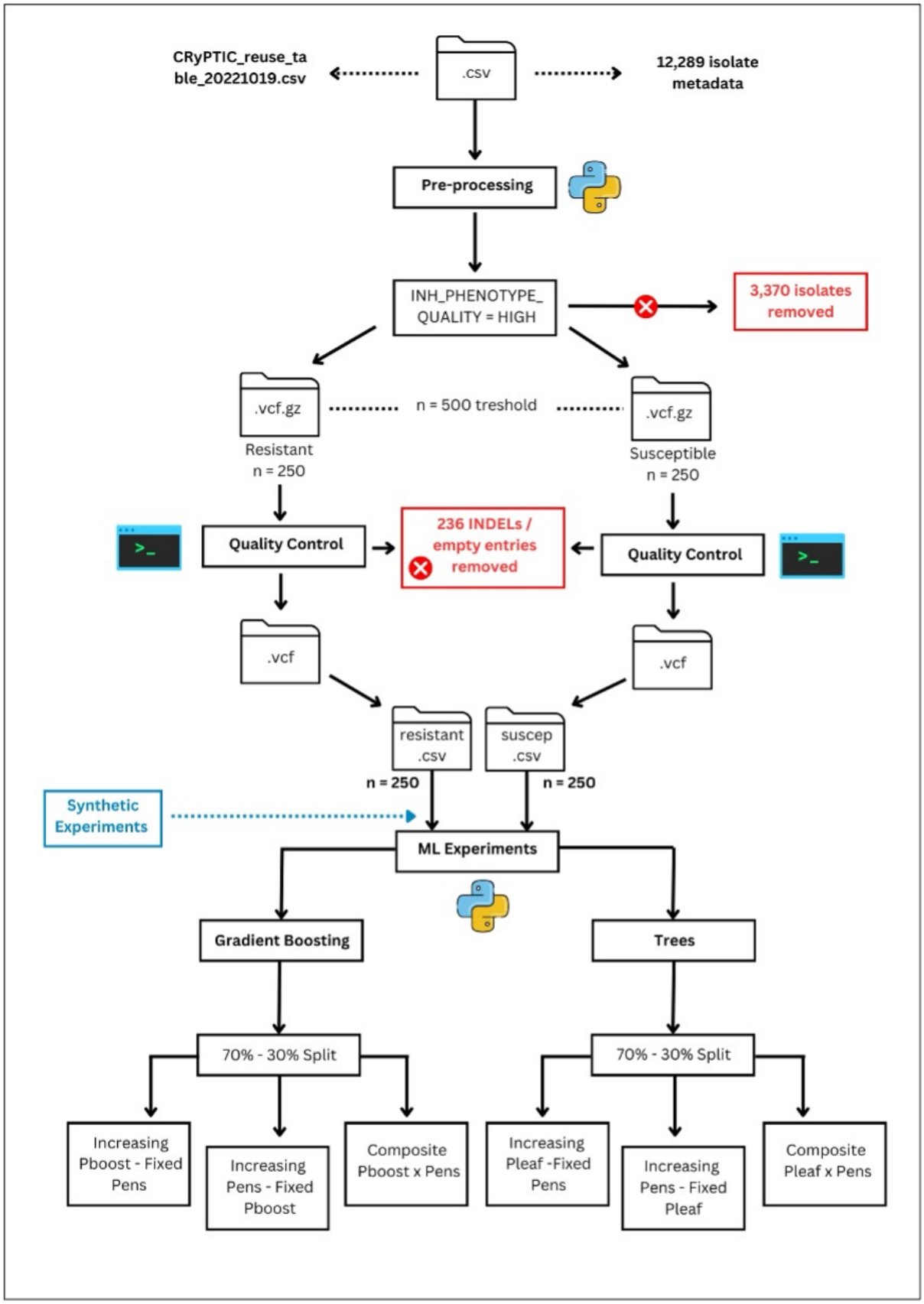

Figure 2. Methodological pipeline. The blue dashed line indicates the branch where synthetic data experiments were conducted, following the same structure as the pipeline used for CRyPTIC data.

3 Results and Discussion

To evaluate whether the double descent phenomenon occurs in classical machine learning models, we trained decision trees and gradient boosting regressors on both real-world genomic data (CRyPTIC) and a synthetic benchmark. Model complexity was varied along two orthogonal axes: learner capacity (e.g., P^{leaf} or P^{boost}) and ensemble size (P^{ens}). This dual-axis framework allowed us to test competing hypotheses: that double descent reflects a generalised learning principle [13], or that it is a projection artefact arising from collapsing separate complexity dimensions [16]. Across all experiments, our results consistently support the latter.

Composite Complexity Induces Double Descent in Trees and Boosting

When model complexity was increased in a composite manner—first by scaling learner capacity (increasing P^{leaf} or P^{boost}), followed by expanding P^{ens}—a clear double descent pattern emerged in both decision trees and gradient boosting regressors. In decision trees trained on the CRyPTIC dataset (Figure 3), test error initially declined as P^{leaf} (in a single tree) increased from L_2 to $L_{\max}$, reaching a minimum at L_{10} (e.g., from 0.135 at L_2 to 0.115 at L_{10} for $P^{\text{leaf}} = 50$). It then rose sharply near the interpolation threshold (marked by the dotted vertical line), peaking at 0.135–0.145 depending on configuration (e.g., 0.140 at L_{100} for $P^{\text{leaf}} = 100$, 0.140 at L_{200} for $P^{\text{leaf}} = 200$, and 0.145 at L_{500} for $P^{\text{leaf}} = 500$). Finally, test error fell again as P^{ens} increased from RF1 to RF50, reaching values as low as 0.100–0.103 across all settings. This non-monotonic behaviour was observed consistently across all four P^{leaf} configurations. While the position and height of the error peak varied slightly, each curve exhibited the hallmark shape of double descent. The same trajectory was observed in the synthetic dataset (Figure 4, left). Gradient boosting models demonstrated similar dynamics under composite scaling. On the CRyPTIC dataset (Figure 5A), MSE declined from 0.118 at $P^{\text{boost}} = 10$ to a minimum of 0.081 at $P^{\text{boost}} = 200$, then rose near the interpolation threshold, before falling again to 0.074 at $P^{\text{ens}} = 50$. The synthetic data mirrored this behaviour (Figure 6A), with MSE decreasing from 0.099 to 0.063, peaking at 0.121, and then falling again to 0.059.

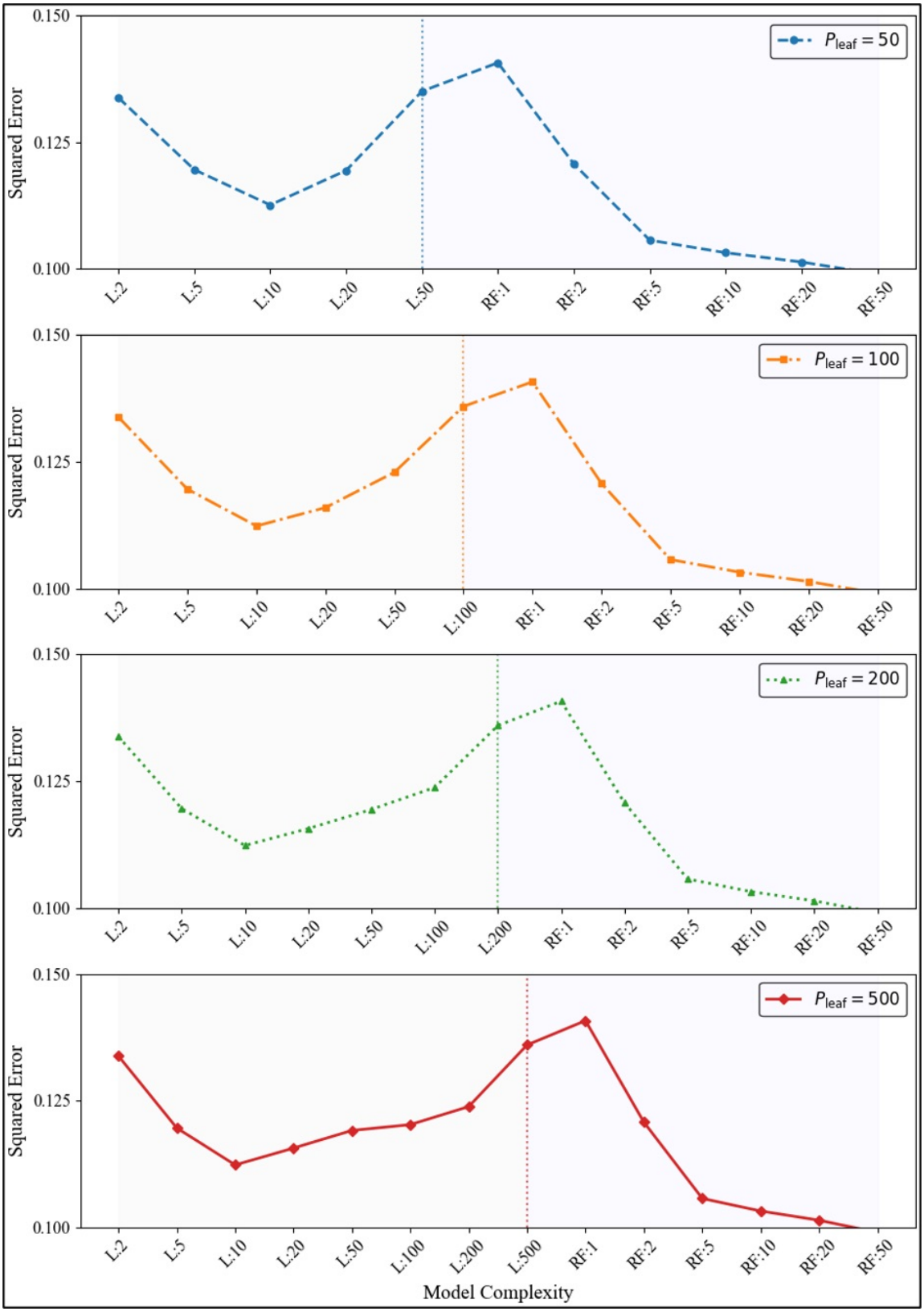

Figure 3. Composite complexity in decision trees and random forests on the CRyPTIC dataset. MSE is plotted against model complexity for $P^{\text{leaf}} \in \{50, 100, 200, 500\}$. Within each subplot, complexity increases first by growing single-tree capacity (L_2 to $L_{\max}$), then by increasing P^{ens} (RF1 to RF50). The vertical dotted line marks the interpolation threshold.

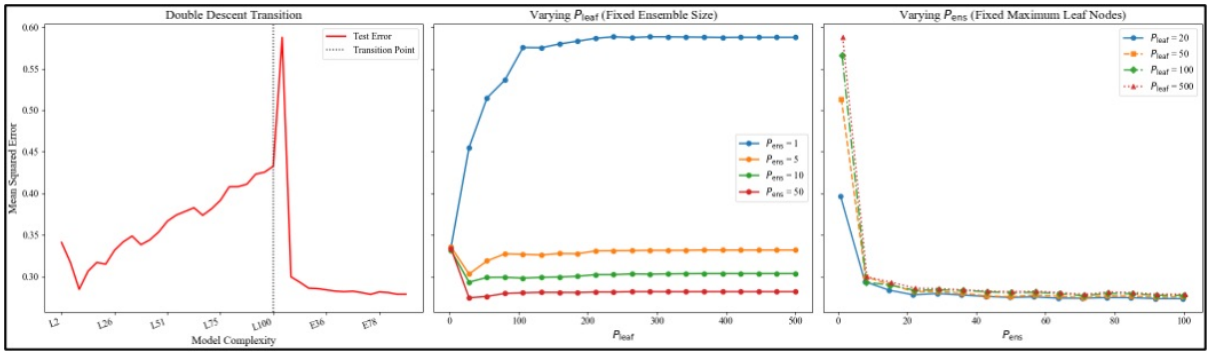

Figure 4. Test MSE for tree-based models on the synthetic dataset. Left: composite complexity (increasing P^{leaf} then P^{ens}). Middle: MSE vs. P^{leaf} at fixed P^{ens}. Right: MSE vs. P^{ens} at fixed P^{leaf}.

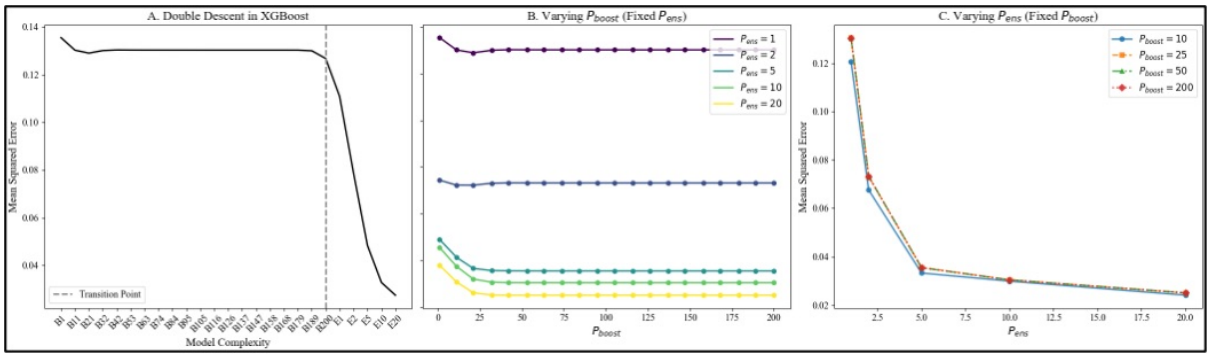

Figure 5. Gradient boosting on the CRyPTIC dataset. (A) Composite complexity: increasing P^{boost} then P^{ens}. (B) MSE vs. P^{boost} at fixed P^{ens}. (C) MSE vs. P^{ens} at fixed P^{boost}.

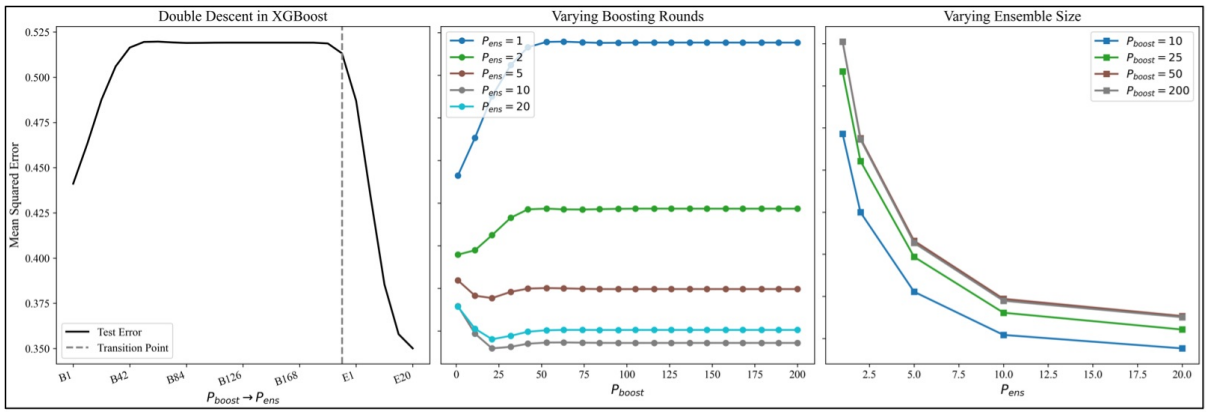

Figure 6. Gradient boosting on the synthetic dataset. (A) Composite complexity: increasing P^{boost} then P^{ens}. (B) MSE vs. P^{boost} at fixed P^{ens}. (C) MSE vs. P^{ens} at fixed P^{boost}.

Axis-Specific Scaling Reveals Bias–Variance Trade-off

When model complexity was varied along a single axis—either by increasing learner capacity or ensemble size independently—the double descent pattern disappeared. Instead, generalisation behaviour aligned with the classical bias–variance trade-off. In decision trees trained on CRyPTIC (Figure 7), increasing P^{leaf} at fixed P^{ens} resulted in a characteristic U-shaped curve. For example, with $P^{\text{ens}} = 1$, MSE decreased from 0.137 at $P^{\text{leaf}} = 2$ to a minimum of 0.107 at $P^{\text{leaf}} = 20$, but rose sharply to 0.194 by $P^{\text{leaf}} = 500$. This sharp increase highlights overfitting in high-capacity learners without variance control, as described by [33]. In contrast, holding P^{leaf} constant and increasing P^{ens} reduced MSE smoothly and monotonically, producing an L-shaped curve. For example, at $P^{\text{leaf}} = 100$, test error dropped from 0.136 at $P^{\text{ens}} = 1$ to 0.097 at $P^{\text{ens}} = 50$. These trends were replicated in the synthetic dataset (Figure 4).

Gradient boosting models showed an analogous pattern under axis-specific scaling (Figures 5B–C; 6B–C). At low ensemble sizes, increasing P^{boost} introduced overfitting. In CRyPTIC (Figure 5B), test MSE rose from 0.123 at $P^{\text{boost}} = 10$ to 0.134 at $P^{\text{boost}} = 200$ with $P^{\text{ens}} = 1$, reflecting the high-variance behaviour of overparameterised learners [2]. A similar trend appeared in the synthetic dataset (Figure 6B), where MSE increased from 0.094 to 0.147 across the same boosting range. Conversely, increasing P^{ens} while keeping P^{boost} fixed consistently reduced test error. At $P^{\text{boost}} = 50$, MSE on CRyPTIC (Figure 5C) fell from 0.108 at $P^{\text{ens}} = 1$ to 0.042 at $P^{\text{ens}} = 20$. The synthetic dataset (Figure 6C) mirrored this L-shaped descent, with MSE dropping from 0.099 to 0.048. Across all models, one axis—P^{leaf} in trees or P^{boost} in boosting—exerted a disproportionate influence on test error, while the other axis (P^{ens}) tended to improve performance or, at worst, leave it unchanged. This asymmetry highlights a consistent "bigger is better" effect with respect to P^{ens}, contrasting with the more volatile behaviour observed when scaling learner capacity alone.

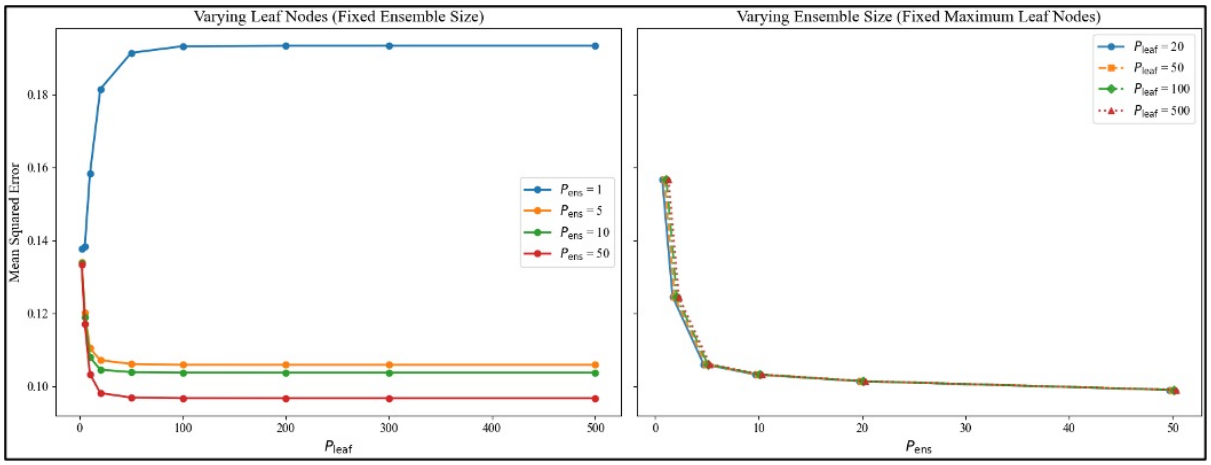

Figure 7. Independent sweeps on CRyPTIC. Left: MSE vs. P^{leaf} at fixed P^{ens}. Right: MSE vs. P^{ens} at fixed P^{leaf}.

Comparison with Existing Literature

While our error curves reflect the double descent dynamics described by Belkin et al. [13], they align more closely with the "unfolding" hypothesis of Curth et al. [16]. Rather than viewing double descent as a universal feature, the unfolding framework suggests that the phenomenon arises when distinct generalisation behaviours—underfitting, interpolation, and overparameterisation—are projected onto a single axis of model complexity. Our results support this view: when capacity and ensemble size are disentangled, the apparent double descent resolves into more interpretable U- and L-shaped curves. This perspective also challenges common assumptions about the robustness of ensemble methods. Random forests and gradient boosting are often viewed as resistant to overfitting, largely due to variance-reducing techniques like averaging in ensembles and regularisation strategies such as shrinkage and subsampling [30, 31].

However, our findings suggest that this robustness is conditional on how model complexity is scaled. When learner capacity is increased—for example, by upregulating P^{leaf}—without a corresponding increase in P^{ens}, test error can rise sharply—an effect often overlooked in standard hyperparameter tuning workflows, which typically vary only one parameter at a time [22, 32]. This interpretation helps reconcile conflicting findings in the literature. For example, [34] observed no double descent in well-tuned random forests—that is, models tuned along a single complexity axis. Our results confirm that under axis-specific tuning, test error behaves predictably. However, when complexity is scaled sequentially across both axes—as in our composite regime—the double descent curve reliably re-emerges [15, 16]. These findings suggest that double descent is not a property of algorithm type, but of the trajectory through complexity space during training.

Practical Implications for Model Tuning

Our findings carry important implications for model selection and tuning. They reaffirm the continued relevance of classical bias–variance theory [2, 1], but only when model complexity is treated as a multidimensional concept. As highlighted by [16, 15], the apparent breakdown of generalisation theory in overparameterised models often stems not from a failure of the theory itself, but from conflating multiple complexity axes into one. The error peak near the interpolation threshold—central to the double descent narrative [13]—is better understood as a misalignment between capacity and variance control. In practice, this means that sharp increases in test error may not reflect flaws in the model or data noise but can instead arise from how hyperparameters are scaled during training. For example, increasing P^{leaf} or P^{boost} without adjusting P^{ens} can push the model into a high-variance regime. These instability points—observed in both our work and previous studies [22, 34, 16]—are frequently misinterpreted as poor model performance, when they are actually artefacts of composite scaling. Therefore, our results support the unfolding hypothesis. Moreover, as [15] observe, tuning multiple hyperparameters simultaneously—such as P^{leaf} and P^{ens}—can obscure which one is driving performance changes. By varying them independently, practitioners can disentangle their effects, diagnose variance-related instabilities, and more effectively tune model behaviour. This targeted approach not only improves the clarity of generalisation patterns but also guides more efficient and reliable hyperparameter tuning [15, 16].

Strengths and Limitations

This study is, to our knowledge, the first to systematically evaluate the double descent phenomenon in classical machine learning models applied to real-world biological data. By applying the unfolding hypothesis to decision trees and gradient boosting regressors across both synthetic and clinical datasets, we provide empirical support for a multidimensional view of generalisation. Our findings extend the composite scaling framework introduced by [16] and highlight the value of axis-aware tuning in practice. Several limitations, however, warrant discussion. First, our analysis focuses exclusively on tree-based models, which may limit the generalisability of results to other algorithmic families, such as support vector machines, where the dynamics of double descent have not yet been critically examined [34]. Second, the scope of our study is restricted to classical (non-deep) learners. Whether the unfolding hypothesis, as formulated by [16], offers a valid or useful framework for understanding double descent in deep neural networks remains an open question. Lastly, we intentionally avoided dimensionality reduction to align with prior work on double descent (e.g., [13, 16]). However, biological data is often inherently noisy [36], so this choice may have further amplified variance by retaining irrelevant or redundant features [37]. For example, SNPs unrelated to isoniazid resistance are likely irrelevant, while redundancy may arise from co-inherited variants within genes like *katG*, which tend to be in linkage disequilibrium due to the low recombination rate in *Mycobacterium tuberculosis* [38].

4 Conclusion

This study investigated the double descent phenomenon in classical machine learning models—specifically decision trees and gradient boosting regressors—applied to both synthetic data and a clinically relevant genomic prediction task. By independently and jointly scaling model complexity along two orthogonal axes—learner capacity (P^{leaf}, P^{boost}) and ensemble size (P^{ens})—we showed that double descent emerges consistently under composite scaling, but not when these axes are varied in isolation. These results support the unfolding hypothesis [16], which argues that double descent arises from conflating distinct generalisation regimes onto a single complexity axis. Contrary to the notion that ensemble methods are inherently resistant to overfitting, our findings demonstrate that gradient boosting and random forests can exhibit double descent when model capacity is increased without adequate variance control. Across both datasets, P^{ens} consistently acted as a stabilising factor, revealing its role as an implicit regulariser. This highlights the importance of understanding not just the magnitude of model complexity, but how it is structured and scaled. While our focus was limited to tree-based models, the experimental design introduced here offers a general framework for testing double descent in other learning algorithms. Future work should extend this framework to support vector machines and neural networks and explore how factors such as dimensionality reduction, feature redundancy, and label noise shape generalisation dynamics in high-capacity regimes. Overall, our findings reinforce the continued relevance of classical bias–variance theory—provided model complexity is treated as a multidimensional construct.

References

[1] P. Domingos. A Unified Bias-Variance Decomposition. Technical Report, 2000.

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning* (2nd ed.). Springer, 2021.

[3] D. Rajnarayan and D. Wolpert. Bias–Variance Tradeoffs: Novel Applications. *arXiv:0810.0879*, 2008.

[4] E. Briscoe and J. Feldman. Conceptual complexity and the bias/variance tradeoff. *Cognition*, 118(1):2–16, 2011.

[5] J. E. Fieldsend and R. M. Everson. Multiobjective Supervised Learning. In *Natural Computing Series*, pp. 155–176. Springer, 2008.

[6] S. Geman and V. N. Vapnik. Learning and generalization in overparameterized settings. 1992. (Classic perspective; see also [7].)

[7] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

[8] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[9] M. Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *arXiv:2105.14368*, 2021.

[10] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *PNAS*, 117(48):30063–30070, 2020.

[11] M. Lafon and A. Thomas. Understanding the Double Descent Phenomenon in Deep Learning. *arXiv:2403.10459*, 2024.

[12] R. Christensen. Double Descent: Understanding Linear Model Estimation of Nonidentifiable Parameters and a Model for Overfitting. *arXiv:2408.13235*, 2024.

[13] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the bias–variance trade-off. *PNAS*, 116(32):15849–15854, 2019.

[14] L. Sa-Couto, J. M. Ramos, M. Almeida, and A. Wichert. Understanding the double descent curve in Machine Learning. *arXiv:2211.10322*, 2022.

[15] R. Schaeffer, M. Khona, Z. Robertson, A. Boopathy, K. Pistunova, J. W. Rocks, I. R. Fiete, and O. Koyejo. Double Descent Demystified: Identifying, Interpreting and Ablating the Sources of a Deep Learning Puzzle. *arXiv:2303.14151*, 2023.

[16] A. Curth, A. Jeffares, and J. van Os. A U-turn on Double Descent: Rethinking Parameter Counting in Statistical Learning. *arXiv:2310.18988*, 2023.

[17] World Health Organization. Global Tuberculosis Report 2024. 2024. https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2024.

[18] N. J. E. Waller *et al.* The evolution of antibiotic resistance is associated with collateral drug phenotypes in *Mycobacterium tuberculosis*. *Nature Communications*, 14, 2023.

[19] C. Nimmo *et al.* Evolution of *Mycobacterium tuberculosis* drug resistance in the genomic era. *Frontiers in Cellular and Infection Microbiology*, 12:954074, 2022.

[20] CRyPTIC Consortium. A data compendium associating the genomes of 12,289 *Mycobacterium tuberculosis* isolates with resistance phenotypes to 13 antibiotics. *PLOS Biology*, 20(8):e3001721, 2022.

[21] N. Zhang *et al.* High-dimensional regimes and interpolation thresholds. 2022. (Contextual reference for interpolation thresholds.)

[22] J. Barbier, F. Camilli, M.-T. Nguyen, M. Pastore, and R. Skerk. Optimal generalisation and learning transition in extensive-width shallow neural networks near interpolation. *arXiv:2501.18530*, 2025.

[23] J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–67, 1991.

[24] T. R. Connor *et al.* CLIMB: an online resource for the medical microbiology community. *Microbial Genomics*, 2(9), 2016.

[25] F. Pedregosa *et al.* Scikit-learn: Machine learning in Python. *arXiv:1201.0490*, 2018.

[26] C. R. Harris *et al.* Array programming with NumPy. *Nature*, 585:357–362, 2020.

[27] W. McKinney. Data structures for statistical computing in Python. 2010.

[28] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[29] P. Virtanen *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.

[30] M. Schonlau and R. Y. Zou. The random forest algorithm for statistical learning. *The Stata Journal*, 20(1):3–29, 2020.

[31] Y. Park and J. Ho. Tackling Overfitting in Boosting for Noisy Healthcare Data. *IEEE TKDE*, 2020.

[32] S. Raschka. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *arXiv:1811.12808*, 2018.

[33] J. W. Rocks and P. Mehta. Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models. *Phys. Rev. Research*, 4(1), 2022.

[34] S. Buschjäger and K. Morik. On double descent (or the lack thereof) in well-tuned random forests. 2022. (Preprint; add publication details if available.)

[35] E. H. Lee and V. Cherkassky. VC Theoretical Explanation of Double Descent. *arXiv:2205.15549*, 2022.

[36] Y. Li, F.-X. Wu, and A. Ngom. Machine learning for multi-view biological data integration. *Briefings in Bioinformatics*, 2016.

[37] B. Chizi and O. Maimon. Dimension Reduction and Feature Selection. In *Data Mining and Knowledge Discovery Handbook*, pp. 83–100. Springer, 2009.

[38] M. W. Marney *et al.* Modeling the structural origins of isoniazid resistance via *katG*. *Tuberculosis*, 108:155–162, 2018.