Evaluating Double Descent in Machine Learning

Supervised by Professor Zamin Iqbal

April 2nd, 2025

SL50188 – Research Project 1B

MSc Bioinformatics

Department of Life Sciences

Faculty of Science

University of Bath

Total Wordcount: 3562

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 and ensemble size—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.

1.0 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 panel) as model complexity increases [Domingos, 2000; James et al. 2021]. 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) [Rajnarayan and Wolpert, 2025; James et al. 2021]. The optimal predictive performance is thought to lie at an intermediate point of complexity, where bias and variance are minimised [Briscoe and Feldman, 2011]. 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 [Fieldsend and Everson, 2008].

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 [German et al. 1992; Vapnik, 2000]. 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 [Goodfellow et al., 2016; Belkin, 2021; Bartlett et al., 2020]. To account for this observation, Belkin et al. (2019) proposed the double descent phenomenon (**Figure 1**, right panel), wherein test error initially decreases with complexity, rises near the interpolation threshold, and then decreases again as complexity increases further. This results in a non-monotonic, "double descent" curve that extends the classical U-shaped paradigm.

The double descent framework suggests that increasing model complexity can, under certain conditions, lead to improved generalisation even in highly overparameterised regimes [Lafon and Thomas, 2024]. 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 [Christensen, 2024; Belkin et al. 2019]. Nevertheless, its underlying theoretical basis remains a topic of ongoing debate [Sa-Couto et al. 2022]. Recent critiques argue that double descent may be a visual artefact of collapsing multidimensional model complexity into a single axis [Schaeffer et al. 2023; Sa-Couto et al. 2022]. Curth et al. (2023) 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 [Curth et al. 2023] (**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. (2023) 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* (*M. tuberculosis*) from whole-genome sequencing data. *M. tuberculosis* remains the leading cause of death from a single bacterial pathogen, with over 1.25 million deaths in 2024 alone [WHO, 2024]. 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 [Waller et al., 2023; Nimmo et al., 2022].

Building on this clinical relevance, we apply the experimental frameworks of Belkin et al. (2019) and Curth et al. (2023) to investigate whether double descent arises in classical machine learning models trained on genomic data from the Comprehensive Resistance Prediction for Tuberculosis (CRyPTIC) consortium [CRyPTIC, 2022]. 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. (2023), 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.

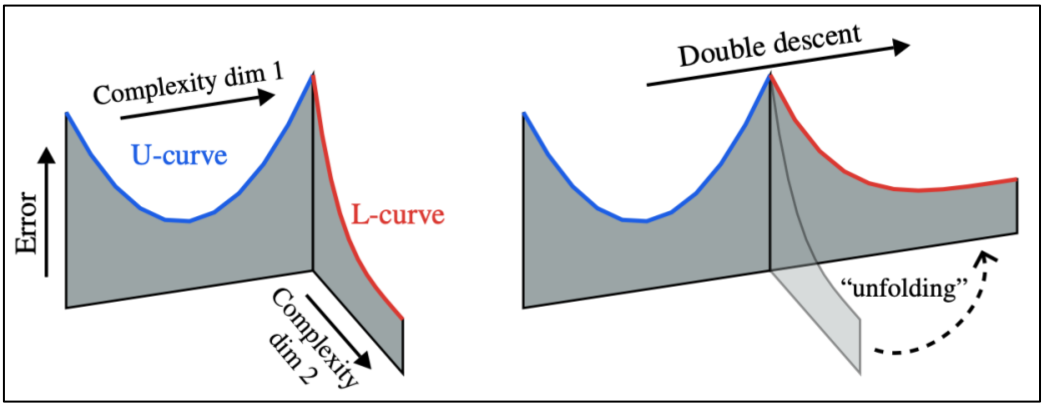

Figure 1 | Double descent illustration emerging from two complexity axes. Left*:* Error varies across two model complexity dimensions, forming a U-curve (blue) along one axis and an L-curve (red) along the other. Right: Collapsing these dimensions produces the double descent curve, suggesting it may arise from merging distinct generalisation behaviours.
Figure taken from Curth et al. (2023).

2.0 Methods

2.1 Data Sources

Whole-genome sequencing data were sourced from the June 2022 public release of the CRyPTIC consortium, comprising 12,289 *M. 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**.

2.2 Sample Selection and Pre-Processing

To ensure computational tractability and class balance, we selected a stratified subsample of 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 follows: 0 (homozygous reference), 1 (heterozygous), and 2 (homozygous alternate). The final feature matrix had dimensions of 765,413 SNPs × 4 features. Although the feature-to-sample ratio was high, no additional dimensionality reduction was applied. This decision follows the conventions of Belkin et al. (2019) and Curth et al. (2023), who recommend preserving high-dimensional structure when evaluating double descent, as it facilitates overfitting, thus ensuring a more rapid reach to the interpolation threshold [Ningyuan et al. 2022]. Pre-processing scripts have been numbered chronologically (scripts 01–10), and are available on GitHub: https://github.com/guillermocomesanacimadevila/RP1B/tree/main/Pre-processing.

2.3 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. (2019) and Curth et al. (2023), who demonstrated double descent under this loss function. Phenotypic labels were binarised as 0 (susceptible) and 1 (resistant) 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, P_{leaf} was varied from 2 to 500 with P_{ens} fixed at values of 1, 5, 10, and 50. Second, P_{ens} was varied from 1 to 50 with P_{leaf} fixed at 20, 50, 100, and 500. Third, a composite scaling experiment was conducted, wherein

P_{leaf} was increased within a single decision tree before subsequently varying P_{ens} in a random forest, using P_{leaf} values of 50, 100, 200, and 500. This composite design simulated the sequential growth of model capacity beyond the interpolation threshold as proposed by Curth et al. (2023).

Gradient boosting models were evaluated using the same general framework but with specific constraints. Base learners were limited to a maximum of 10 leaf nodes to remain within the regime studied by Belkin et al. (2019) and Curth et al. (2023). A high learning rate of γ = 0.85 was used to encourage rapid overfitting and thus increase the likelihood of interpolation [Babier et al. 2025]. In the first experiment, the number of boosting rounds (P_{boost}) was varied from 10 to 200 while P_{ens} was fixed at 1, 5, 10, and 50. In the second experiment, P_{boost} was held constant at 20, 50, 100, and 200 while P_{ens} was varied from 1 to 50. The third experiment adopted a composite approach: P_{boost} was fixed at 200, after which P_{ens} was gradually scaled from 1 to 50. This allowed the interaction between the number of boosting rounds and ensemble size to be evaluated beyond the interpolation threshold, as observed by Curth et al. (2023).

All model evaluations were performed using a consistent experimental grid and the same 70:30 train-test split. CRyPTIC-based machine learning experiments are available in full on GitHub: https://github.com/guillermocomesanacimadevila/RP1B/tree/main/ML CRyPTIC.

2.4 Synthetic Baseline

To validate observed generalisation dynamics in a controlled environment, we reproduced all experiments on a synthetic dataset proposed by Friedman (1991) and used in contemporary double descent literature [Curth et al. 2023]. The dataset contained 500 samples, and 50 independent features sampled from a uniform distribution U(0, 1). The regression target y was generated by:

$$y = \sin(\pi X_1 X_2) + 2(X_3 - 0.5)^2 + X_4 + 0.5X_5 + \varepsilon, \quad \text{where } \varepsilon \sim \mathcal{N}(0, 1).$$

This benchmark is widely used to evaluate how models handle non-linear interactions, sparse dependencies, and random noise [Friedman, 1991]. These properties make it especially valuable for studying how model performance evolves with increasing complexity—particularly in scenarios where overfitting becomes more pronounced, such as the double descent [Curth et al. 2023; Belkin et al. 2019]. All tree-based and boosting models were evaluated using the same experimental grid and 70:30 train-test split as in the CRyPTIC setting. All tree-based synthetic experiments can be accessed on GitHub: https://github.com/guillermocomesanacimadevila/RP1B/tree/main/ML Synthetic.

2.5 Code Availability

All preprocessing, feature extraction, and machine learning experiments were conducted using the Cloud Infrastructure for Microbial Bioinformatics [Connor et al., 2016]. Variant filtering was carried out using Bash scripts, while all downstream modelling was implemented in Python 3.12 using sci-kit-learn version 1.6.1 [Pedregosa et al., 2011], with supporting libraries including numpy 2.2.3 [Harris et al., 2020], pandas 2.2.3 [McKinney, 2010], matplotlib 3.10.1 [Hunter, 2007], and scipy 1.15.2 [Virtanen et al., 2020]. A fixed random seed was used across all experiments to ensure reproducibility. All code, and documentation are available at: https://github.com/guillermocomesanacimadevila/RP1B/tree/main.

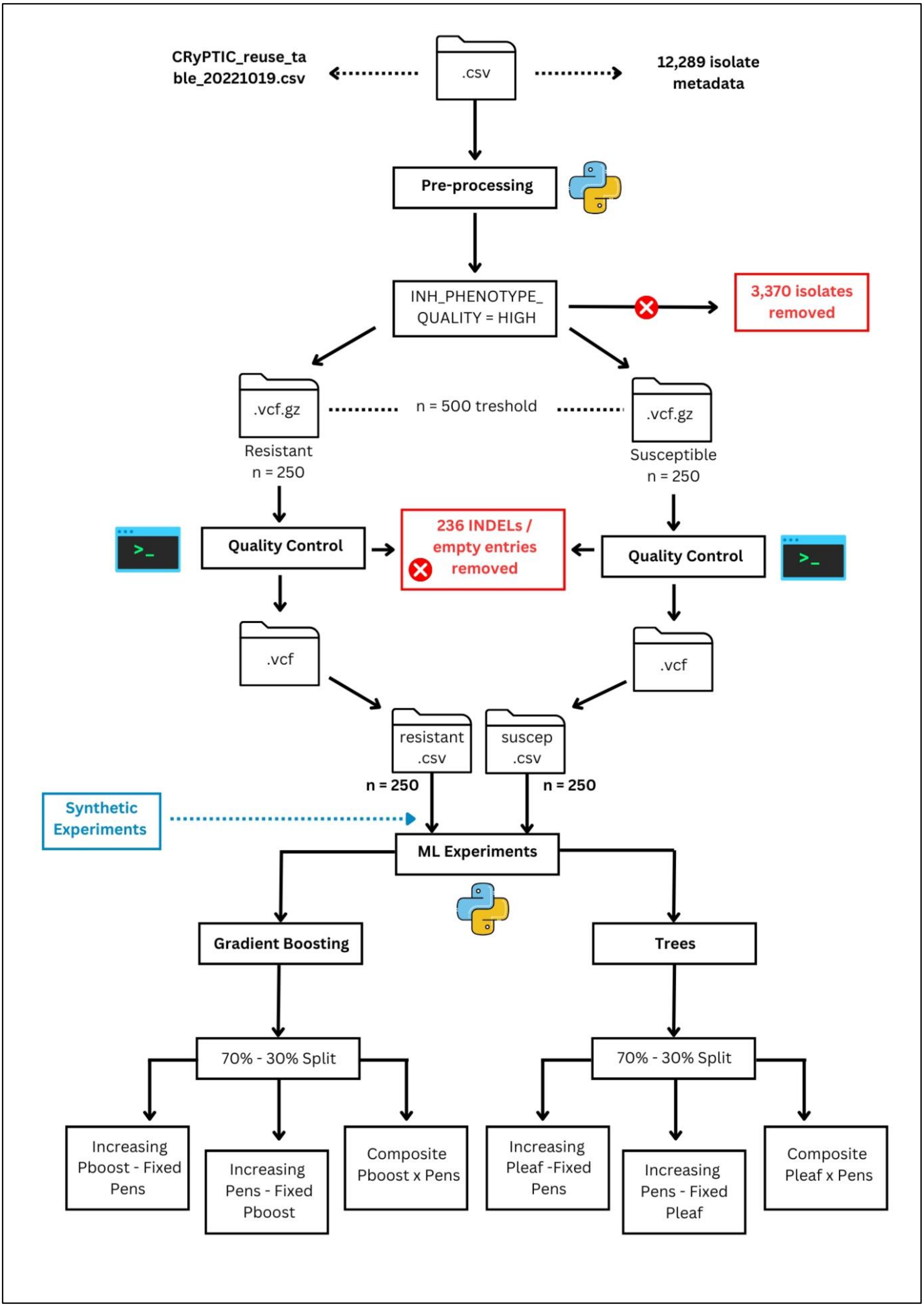

Figure 2 | Methodological pipeline. The blue dashed line indicates the branch where synthetic data experiments were conducted, following the same structure as the pipeline used for CRyPTIC data.

3.0 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 [Belkin et al., 2019], or that it is a projection artefact arising from collapsing separate complexity dimensions [Curth et al., 2023]. Across all experiments, our results consistently support the latter.

3.1 Composite Complexity Induces Double Descent in Trees and Boosting

When model complexity was increased in a composite manner—first by scaling learner capacity (e.g., increasing P_{leaf} or P_{boost}), followed by expanding ensemble size (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) from L2 to L$_{\text{max}}$, reaching a minimum at L10 (e.g., from 0.135 at L2 to 0.115 at L10 for P_{leaf} = 50). It then rose sharply near the interpolation threshold (marked by the dotted vertical line), peaking at 0.135–0.145 depending on the configuration (e.g., 0.140 at L100 for P_{leaf} = 100, 0.140 at L200 for P_{leaf} = 200, and 0.145 at L500 for P_{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 (P_{leaf} = 50, 100, 200, 500). 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**), test error declined from 0.118 at P_{boost} = 10 to a minimum of 0.081 at P_{boost} = 200, then rose near the interpolation threshold, before falling again to 0.074 at P_{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.

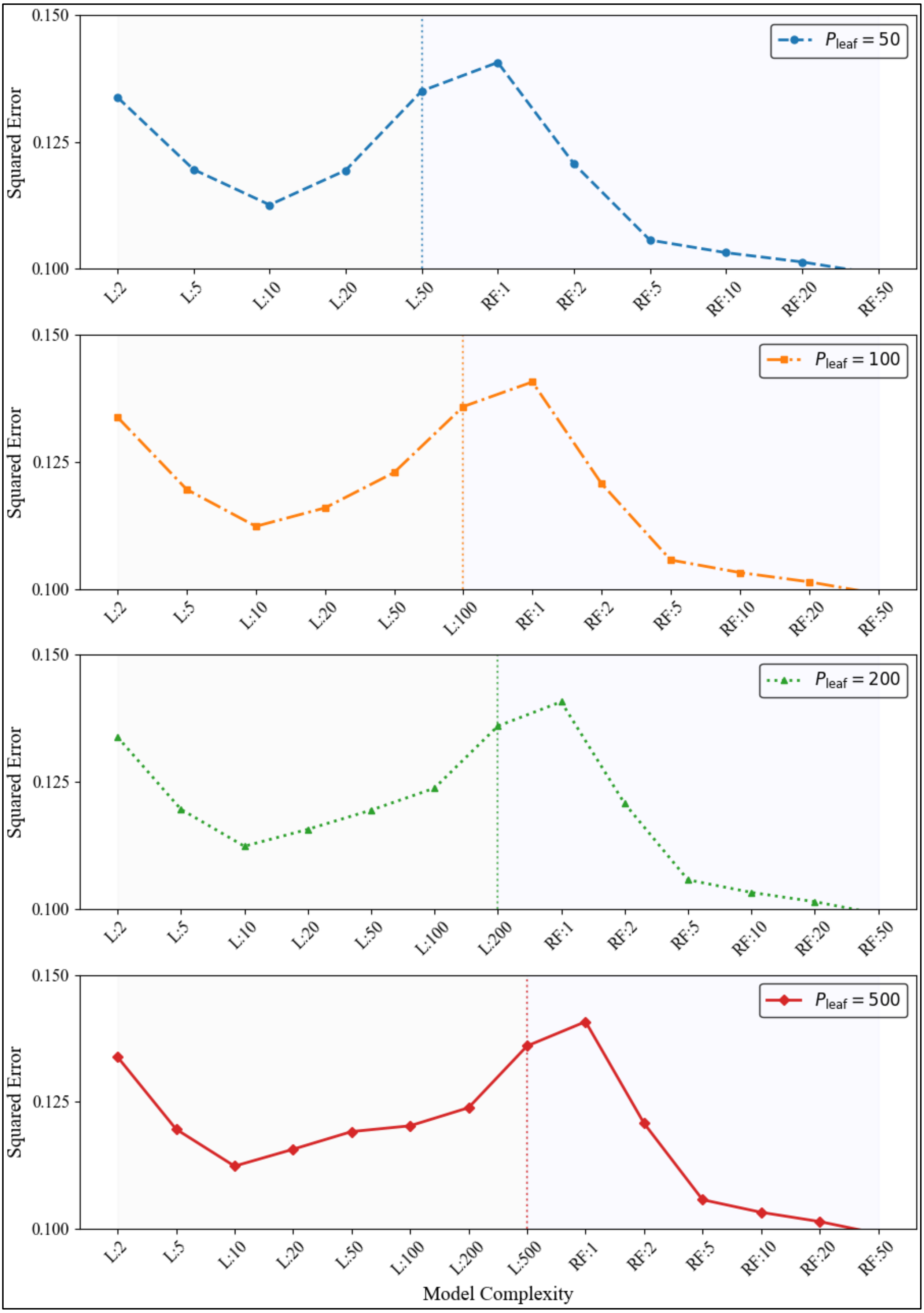

Figure 3 | Composite Complexity in Decision Trees and Random Forests on the CRyPTIC dataset. MSE is plotted against model complexity for four maximum leaf node settings: P_{leaf} = 50, 100, 200, 500. Within each subplot, complexity increases first by growing individual decision trees (L2 to L_{max}), followed by increasing ensemble size (P_{ens}) in random forests (RF1 to RF50). The vertical dotted line marks the interpolation threshold—the point at which single-tree capacity reaches its maximum and ensemble scaling begins.

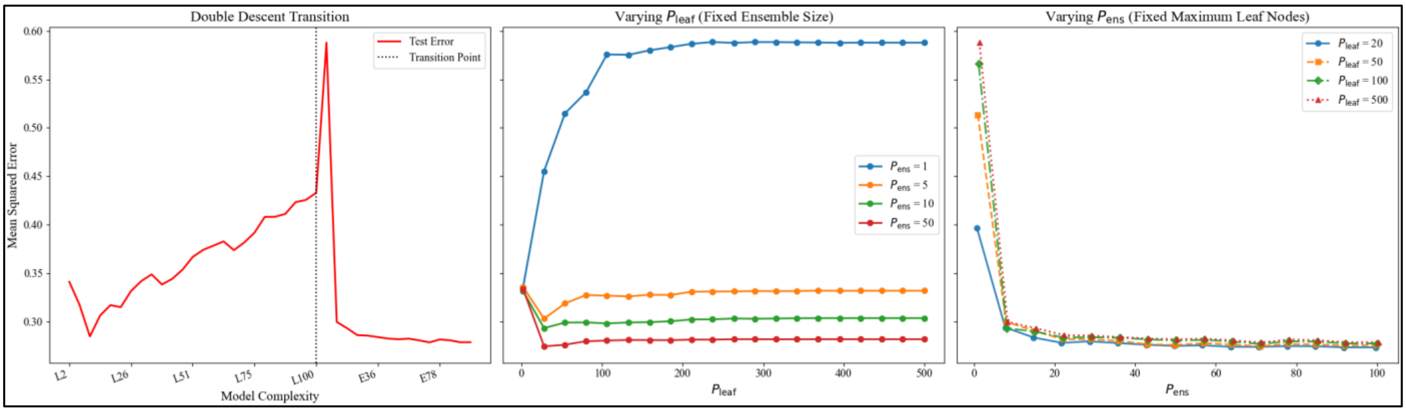

Figure 4 | Test set MSE for tree-based models on the synthetic dataset. Left: Composite complexity curve showing MSE across increasing P_{leaf}, followed by increasing P_{ens}. Middle: MSE as a function of P_{leaf} at fixed ensemble sizes. Right: MSE as a function of P_{ens}, at fixed tree depths. Each curve reflects a distinct fixed value of the non-varied parameter.

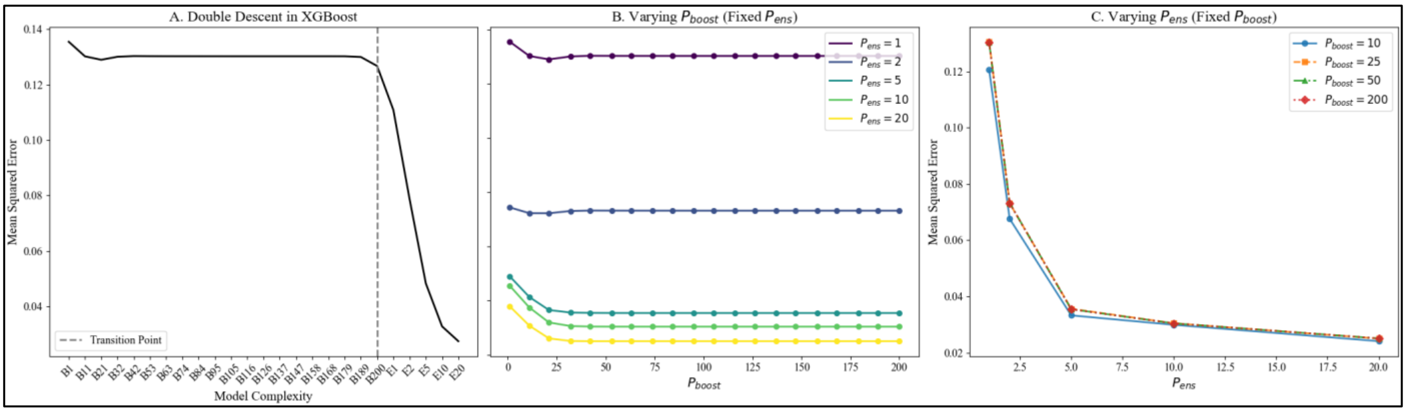

Figure 5 | Gradient boosting results on the CRyPTIC dataset. Panel A: MSE across composite complexity, beginning with increasing P_{boost}, followed by increasing P_{ens}. The vertical dotted line marks the transition point (i.e. interpolation threshold) between boosting depth and ensemble size. Panel B: MSE as a function of P_{boost} at fixed P_{ens}. Panel C: MSE as a function of P_{ens} at fixed P_{boost}. Each curve reflects a unique fixed value of the non-varied parameter.

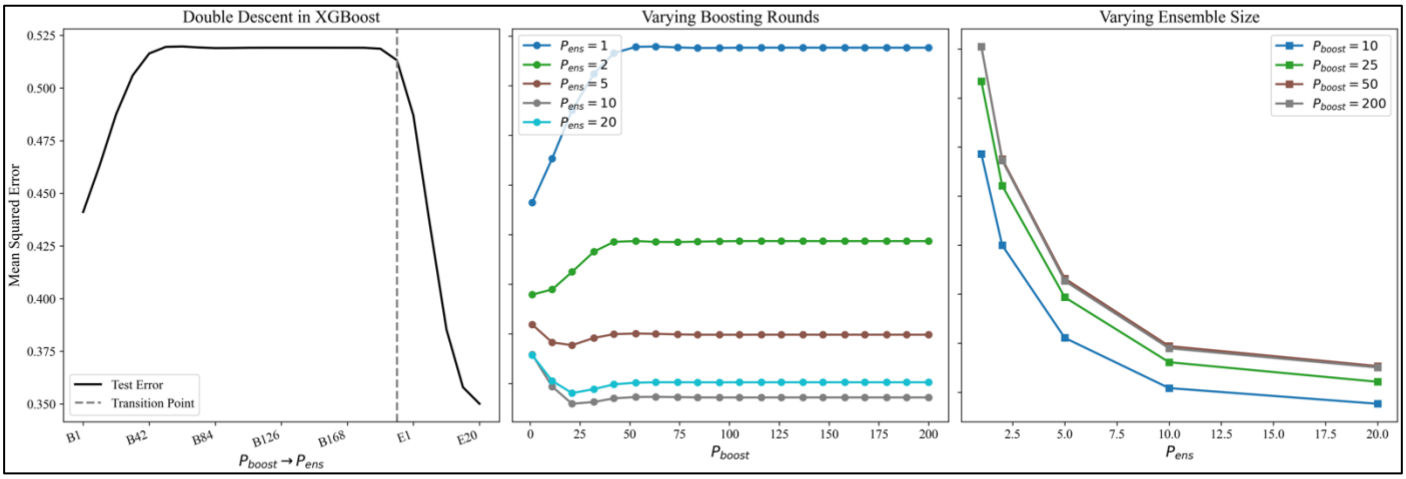

Figure 6 | Gradient boosting results on the synthetic dataset. Panel A: Composite complexity curve showing MSE with increasing P_{boost}, followed by increasing P_{ens}. The vertical dotted line marks the transition (i.e. interpolation threshold) between phases. Panel B: MSE as a function of P_{boost} at fixed ensemble sizes. Panel C: MSE as a function of P_{ens} at fixed boosting rounds. Each curve reflects a distinct fixed value of the non-varied parameter.

3.2 Axis-Specific Scaling Reveals Bias-Variance Tradeoff

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 the CRyPTIC dataset (**Figure 7**), increasing the P_{leaf} at fixed P_{ens} resulted in a characteristic U-shaped error curve. For instance, with P_{ens} = 1, MSE decreased from 0.137 at P_{leaf} = 2 to a minimum of 0.107 at P_{leaf} = 20 but rose sharply to 0.194 by P_{leaf} = 500. This sharp increase highlights overfitting in high-capacity learners without variance control, as described by Rocks and Mehta (2022). In contrast, holding P_{leaf} constant and increasing P_{ens}, reduced MSE smoothly and monotonically, denoting an L-shaped curve. For example, at P_{leaf} = 100, the test error dropped from 0.136 at P_{ens} = 1 to 0.097 at P_{ens} = 50. These trends were replicated in the synthetic dataset (**Figure 5**), reinforcing the generality of the effect.

Gradient boosting models showed an analogous pattern under axis-specific scaling (**Figures 5B, 5C; 6B, 6C**). At low ensemble sizes, increasing the P_{boost} introduced overfitting. In the CRyPTIC dataset (**Figure 5B**), test MSE rose from 0.123 at P_{boost} = 10 to 0.134 at 200 with P_{ens} = 1, reflects the traditional high-variance behaviour of overparameterised learners [James et al. 2021]. 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 ensemble size while keeping P_{boost} fixed consistently reduced test error. At P_{boost} = 50, MSE on the CRyPTIC dataset (**Figure 5C**) fell from 0.108 at P_{ens} = 1 to 0.042 at P_{ens} = 20. The synthetic dataset (**Figure 6C**) mirrored this L-shaped descent, with MSE dropping from 0.099 to 0.048. These results highlight ensemble size as a critical regulariser and show that double descent emerges only when model capacity and variance are scaled jointly. When these axes are varied independently, generalisation remains stable and consistent with classical bias-variance behaviour.

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 (ensemble size) 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.

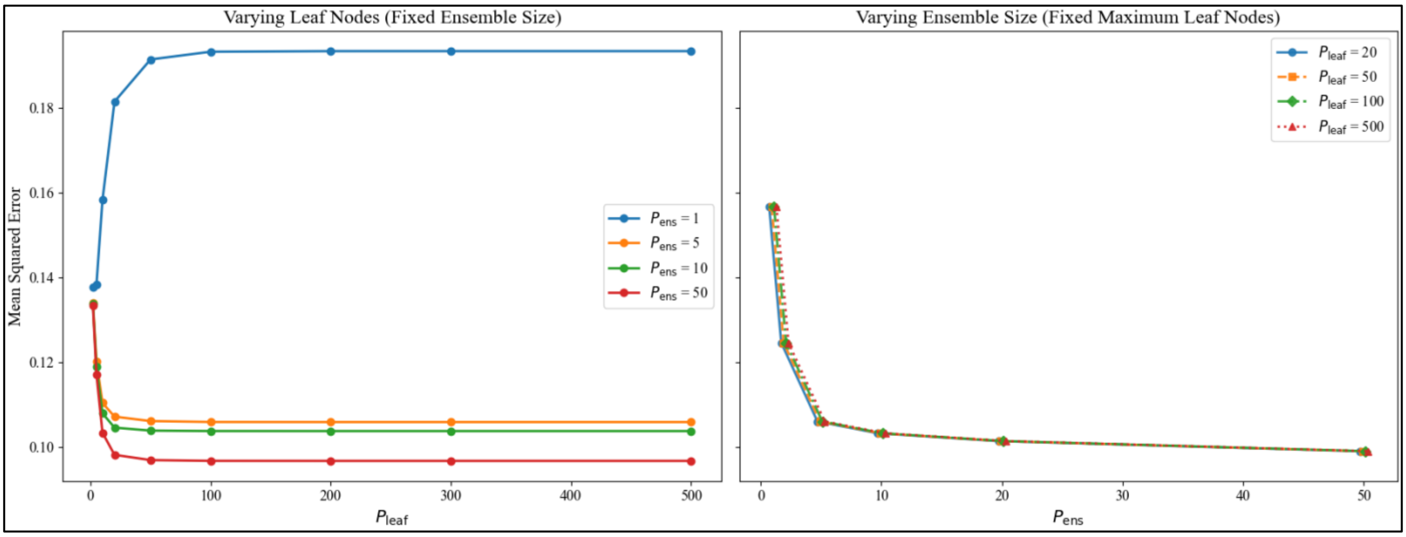

Figure 7 | Independent hyperparameter sweeps in tree-based models on the CRyPTIC dataset. The left panel shows MSE as a function of P_{leaf} for fixed ensemble sizes, while the right panel shows MSE as a function of P_{ens} for fixed values of P_{leaf}. Each curve corresponds to a distinct value of the non-varied hyperparameter.

3.3 Comparison with Existing Literature

While our error curves reflect the double descent dynamics originally described by Belkin et al. (2019), they align more closely with the "unfolding" hypothesis proposed by Curth et al. (2023). 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 [Schonlau et al., 2020; Park and Ho, 2020]. 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 [Babier et al., 2025; Raschka, 2018].

This interpretation helps reconcile conflicting findings in the literature. For example, Buschjäger and Morik (2022) 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. However, when complexity is scaled jointly across both dimensions, the double descent curve reliably re-emerges. These findings suggest that double descent is not a property of algorithm type, but of the trajectory through complexity space during training [Schaeffer et al., 2023; Curth et al., 2023].

3.4 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 [James et al., 2021; Domingos, 2000], but only when model complexity is treated as a multidimensional concept. As highlighted by Curth et al. (2023) and Schaeffer et al. (2023), 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 [Belkin et al., 2019]—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 [Babier et al., 2025; Buschjäger and Morik, 2022; Curth et al. 2023], are frequently misinterpreted as poor model performance, when they are actually artefacts of composite scaling. Therefore, as previously noted, our results support the unfolding hypothesis.

Moreover, as Schaeffer et al. (2023) 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 [Schaeffer et al. 2023; Curth et al. 2023].

3.5 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 Curth et al. (2023) 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 [Lee and Cherkassy, 2020]. Second, the scope of our study is restricted to classical (non-deep) learners. Whether the unfolding hypothesis, as formulated by Curth et al. (2023), 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., Belkin et al., 2019; Curth et al., 2023). However, biological data is often inherently noisy [Li et al. 2016], so this choice may have further amplified variance by retaining irrelevant or redundant features [Chizi and Maimon, 2009]. 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 *M. tuberculosis* [Marney et al. 2018].

4.0 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 and ensemble size—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 proposed by Curth et al. (2023), 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, ensemble size 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.

5.0Acknowledgements

I would like to thank Professor Zamin Iqbal for his patience and constant supervision throughout the completion of this project, and for granting access to the Cloud Infrastructure for Microbial Bioinformatics. I have learnt so much, and really enjoyed the project.

6.0 References

Barbier, J., Camilli, F., Nguyen, M.-T., Pastore, M. and Skerk, R., 2025. *Optimal generalisation and learning transition in extensive-width shallow neural networks near interpolation* [Online]. *arXiv.org*. Available from: https://arxiv.org/abs/2501.18530 [Accessed 29 March 2025].

Bartlett, P.L., Long, P.M., Lugosi, G. and Tsigler, A., 2020. Benign overfitting in linear regression [Online]. *Proceedings of the National Academy of Sciences*, 117(48), pp.30063–30070. Available from: https://doi.org/10.1073/pnas.1907378117 [Accessed 11 May 2021].

Belkin, M., 2021. *Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation* [Online]. *arXiv.org*. Available from: https://arxiv.org/abs/2105.14368 [Accessed 23 March 2025].

Briscoe, E. and Feldman, J., 2011. Conceptual complexity and the bias/variance tradeoff [Online]. *Cognition*, 118(1), pp.2–16. Available from: https://doi.org/10.1016/j.cognition.2010.10.004.

Chizi, B. and Maimon, O., 2009. Dimension Reduction and Feature Selection [Online]. *Data Mining and Knowledge Discovery Handbook*, pp.83–100. Available from: https://doi.org/10.1007/978-0-387-09823-4_5 [Accessed 31 March 2025].

Christensen, R., 2024. *Double Descent: Understanding Linear Model Estimation of Nonidentifiable Parameters and a Model for Overfitting* [Online]. *arXiv.org*. Available from: https://arxiv.org/abs/2408.13235 [Accessed 28 March 2025].

Connor, T.R., Loman, N.J., Thompson, S., Smith, A., Southgate, J., Poplawski, R., Bull, M.J., Richardson, E., Ismail, M., Thompson, S.E. -, Kitchen, C., Guest, M., Bakke, M., Sheppard, S.K. and Pallen, M.J., 2016. CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community [Online]. *Microbial Genomics*, 2(9). Available from: https://doi.org/10.1099/mgen.0.000086 [Accessed 7 September 2020].

CRyPTIC, 2022. A data compendium associating the genomes of 12,289 Mycobacterium tuberculosis isolates with quantitative resistance phenotypes to 13 antibiotics [Online]. *PLOS Biology*, 20(8), p.e3001721. Available from: https://doi.org/10.1371/journal.pbio.3001721.

Curth, A., Jeffares, A. and van, 2023. *A U-turn on Double Descent: Rethinking Parameter Counting in Statistical Learning* [Online]. *arXiv.org*. Available from: https://arxiv.org/abs/2310.18988 [Accessed 16 March 2025].

Domingos, P., 2000. *A Unified Bias-Variance Decomposition* [Online]. Available from: https://homes.cs.washington.edu/~pedrod/bvd.pdf.

Fieldsend, J.E. and Everson, R.M., 2008. Multiobjective Supervised Learning [Online]. *Natural Computing Series*, pp.155–176. Available from: https://doi.org/10.1007/978-3-540-72964-8_8 [Accessed 16 March 2025].

Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep Learning* [Online]. *www.deeplearningbook.org*. Available from: https://www.deeplearningbook.org.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P. and Gérard-Marchant, P., 2020. Array Programming with NumPy [Online]. *Nature*, 585(7825), pp.357–362. Available from: https://doi.org/10.1038/s41586-020-2649-2.

Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), pp.90–95.
James, G., Witten, D., Hastie, T. and Tibshirani, R., 2021. *An Introduction to Statistical Learning* [Online]. *Springer Texts in Statistics*. New York, NY: Springer US. Available from: https://doi.org/10.1007/978-1-0716-1418-1.

Lafon, M. and Thomas, A., 2024. *Understanding the Double Descent Phenomenon in Deep Learning*[Online]. *arXiv.org*. Available from: https://arxiv.org/abs/2403.10459 [Accessed 23 March 2025].

Lee, E.H. and Cherkassky, V., 2022. *VC Theoretical Explanation of Double Descent* [Online]. *arXiv.org*. Available from: https://arxiv.org/abs/2205.15549 [Accessed 31 March 2025].

Li, Y., Wu, F.-X. and Ngom, A., 2016. A review on machine learning principles for multi-view biological data integration [Online]. *Briefings in Bioinformatics*, p.bbw113. Available from: https://doi.org/10.1093/bib/bbw113.

Marney, M.W., Metzger, R.P., Hecht, D. and Valafar, F., 2018. Modeling the structural origins of drug resistance to isoniazid via key mutations in Mycobacterium tuberculosis catalase-peroxidase, KatG [Online]. *Tuberculosis*, 108, pp.155–162. Available from: https://doi.org/10.1016/j.tube.2017.11.007 [Accessed 19 September 2019].

McKinney, W., 2010. *Data Structures for Statistical Computing in Python* [Online]. *ResearchGate*. unknown. Available from: https://www.researchgate.net/publication/340177686_Data_Structures_for_Statistical_Computing_in_Python.

Nimmo, C., Millard, J., Faulkner, V., Monteserin, J., Pugh, H. and Johnson, E.O., 2022. Evolution of Mycobacterium tuberculosis drug resistance in the genomic era [Online]. *Frontiers in Cellular and Infection Microbiology*, 12, p.954074. Available from: https://doi.org/10.3389/fcimb.2022.954074 [Accessed 8 September 2023].

Park, Y. and Ho, J., 2020. Tackling Overfitting in Boosting for Noisy Healthcare Data [Online]. *IEEE Transactions on Knowledge and Data Engineering*, pp.1–1. Available from: https://doi.org/10.1109/tkde.2019.2959988.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É., 2018. Scikit-learn: Machine Learning in Python [Online]. *arXiv:1201.0490 [cs]*. Available from: https://arxiv.org/abs/1201.0490.

Rajnarayan, D. and Wolpert, D., 2025. *Bias-Variance Tradeoffs: Novel Applications* [Online]. *arXiv.org*. Available from: https://arxiv.org/abs/0810.0879 [Accessed 16 March 2025].

Raschka, S., 2018. *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*[Online]. *arXiv.org*. Available from: https://arxiv.org/abs/1811.12808.

Rocks, J.W. and Mehta, P., 2022. Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models [Online]. *Physical Review Research*, 4(1). Available from: https://doi.org/10.1103/physrevresearch.4.013201.

Sa-Couto, L., Ramos, J.M., Almeida, M. and Wichert, A., 2022. *Understanding the double descent curve in Machine Learning* [Online]. *arXiv.org*. Available from: https://arxiv.org/abs/2211.10322.

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

Schonlau, M. and Zou, R.Y., 2020. The random forest algorithm for statistical learning [Online]. *The Stata Journal: Promoting Communications on Statistics and Stata*, 20(1), pp.3–29. *Sagepub*. Available from: https://doi.org/10.1177/1536867x20909688.

Vapnik, V.N., 2000. *The Nature of Statistical Learning Theory* [Online]. New York, NY: Springer New York. Available from: https://doi.org/10.1007/978-1-4757-3264-1.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E. and Carey, C.J., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), pp.261–272.

Waller, N.J.E., Cheung, C.-Y., Cook, G.M. and McNeil, M.B., 2023. The evolution of antibiotic resistance is associated with collateral drug phenotypes in Mycobacterium tuberculosis [Online]. *Nature Communications*, 14(1). Available from: https://doi.org/10.1038/s41467-023-37184-7.

WHO, 2024. *Global Tuberculosis Report 2024* [Online]. *Who.int*. Available from: https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2024.