

Regime-Aware Model Selection for Internal Combustion Engine Performance Prediction Using Machine Learning

Ashish Kumar Routray¹, Soumyajeet chakra^{2*}, Anuprita Chakra³,
Thambe Sai Pavan⁴, Abhibhav Mohanty⁵

¹Department of Statistics, Ravenshaw University, Cuttack, Odisha.

^{2,4,5}Department of Data Science, Lakshya Institute of Technology, Bhubaneswar, Odisha.

³Assistance professor in Agriculture department, GIFT autonomous, Bhubaneswar, Odisha.

OPEN ACCESS

Article Citation:

Ashish Kumar Routray¹, Soumyajeet chakra^{2*}, Anuprita Chakra³, Thambe Sai Pavan⁴, Abhibhav Mohanty⁵, "Regime-Aware Model Selection for Internal Combustion Engine Performance Prediction Using Machine Learning", International Journal of Recent Trends in Multidisciplinary Research, July-August 2026, Vol 6(04), 364-373.



©2026 The Author(s). This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and

reproduction in any medium, provided the original author and source are credited. Published by 5th Dimension Research Publication

Abstract: Artificial Neural Networks (ANNs) and tree-based ensembles such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) are the two dominant families used for tabular regression in engineering informatics. The empirical literature, however, reports contradictory verdicts, on some datasets ANNs dominate, while on others RF and XGBoost are markedly superior, and no unifying account exists of the conditions under which each family is preferable. This paper addresses that gap directly. Using a curated dataset of 1,232 internal combustion engines, we formulate the prediction of rated power output as a supervised regression problem and derive, from first principles, the inductive biases of the three model classes. We show mathematically that the two families differ in the geometry of the function class they approximate ANNs build globally smooth compositions of ridge functions, whereas tree ensembles build axis aligned, locally constant partitions and that this distinction predicts where each should excel. Empirically, the three models achieve near-identical global accuracy (test R-squared of 0.957, 0.952 and 0.954 for ANN, RF and XGBoost respectively), yet a regime-stratified analysis reveals systematic, statistically meaningful reversals of rank across operating conditions, ANNs win in turbocharged, petrol and high-power regimes, while tree ensembles win in naturally aspirated, diesel, small displacement and low to mid-power regimes. We then train a logistic meta-model that predicts, from engine descriptors alone, whether a tree ensemble will outperform the ANN for a given instance, achieving 66% routing accuracy and identifying displacement, compression ratio and forced induction as the decisive covariates. An oracle router that selects the locally best model reduces mean absolute error from roughly 11 hp to 7.2 hp, a 33% reduction, quantifying the value of condition aware model selection. The study contributes a reproducible, mathematically grounded methodology for explaining and exploiting the ANN-versus tree performance gap rather than merely reporting it.

Key Words: machine learning; artificial neural networks; Random Forest; XGBoost; internal-combustion engines; meta-learning.

1. Introduction

The prediction of engineering performance quantities from design descriptors is a canonical supervised-learning task with direct relevance to operations research, design optimisation and reliability engineering. Within this space, two model families have become dominant for structured, tabular data. The first is the Artificial Neural Network (ANN), a compositional non-linear

Regime-Aware Model Selection for Internal Combustion Engine Performance Prediction Using Machine Learning

function approximator; the second is the tree-based ensemble, of which Random Forest (RF) and Extreme Gradient Boosting (XGBoost) are the leading representatives. A recurring and unresolved observation in the applied literature is that these families exhibit inconsistent relative performance. In some published datasets the ANN attains the lowest error; in others RF or XGBoost is decisively better. Reviews of the field seldom explain this inconsistency and they report which model won on a particular benchmark without characterising the conditions that made it win.

This inconsistency is not a mere artefact of tuning. It reflects a deeper truth about inductive bias the set of assumptions a learning algorithm uses to generalise beyond its training data. ANNs and tree ensembles encode fundamentally different assumptions about the shape of the target function, and those assumptions align well with some data generating regimes and poorly with others. The absence of a systematic account of these regimes is the research gap this paper addresses.

We adopt the prediction of rated engine power as a concrete, physically interpretable testbed. Engine power is governed by well understood thermodynamic and geometric relationships, which lets us reason about whether a smooth global approximator or a piecewise constant partitioner is the more natural fit in a given part of the design space. Our contributions are fourfold. First, we provide a self-contained mathematical treatment of the three model classes and the geometry of the function spaces they explore. Second, we conduct a rigorous empirical comparison on 1,232 engines with cross-validated stability estimates. Third and centrally we perform a regime stratified analysis that exposes systematic reversals of model rank across turbocharging, fuel type, displacement and power level. Fourth, we build a meta-model that predicts, per instance, which family will win, and we quantify the accuracy gain achievable by condition aware routing.

1.1 Research questions and objectives

The study is organised around three questions. (RQ1) Do ANN, RF and XGBoost differ in global accuracy on a large, heterogeneous engine dataset, and if not, is global accuracy even the right lens? (RQ2) Are there identifiable operating regimes in which one family systematically out performs the other, and can these regimes be characterised by measurable covariates? (RQ3) Can a lightweight meta-model exploit that characterisation to route each instance to its best model, and what error reduction does this yield? The corresponding objective is to move the field from anecdotal reporting of winners toward a predictive theory of when each family should be trusted.

2. Background and Literature Review

Machine learning has become an indispensable tool for predictive modeling in engineering, manufacturing, and industrial systems because of its ability to discover complex nonlinear relationships from observational data. Classical statistical learning methods, ensemble algorithms, and deep neural networks each provide distinct advantages depending on the characteristics of the underlying dataset. Consequently, selecting an appropriate learning algorithm remains one of the most important challenges in predictive analytics.

Ensemble tree-based methods have consistently demonstrated strong performance on structured tabular datasets. Random Forest, introduced by Breiman [1], improves prediction accuracy through bootstrap aggregation and randomized feature selection, thereby reducing variance while maintaining robustness to noisy and correlated variables. Gradient Boosting Machines further advanced ensemble learning by sequentially fitting weak learners to residual errors, producing highly accurate predictive models for nonlinear regression problems [3]. Building on this framework, XGBoost introduced efficient regularization, parallel computation, and sparse-aware optimization, making it one of the most successful algorithms for large-scale tabular learning tasks [2]. The theoretical foundations of these methods are comprehensively discussed in statistical learning literature [6,9].

Deep neural networks represent an alternative paradigm capable of learning highly expressive nonlinear mappings directly from data. The universal approximation theorem established that multilayer feedforward networks can approximate any continuous function under mild assumptions [7], while modern deep learning architectures have significantly expanded this capability through hierarchical feature representation [4]. Efficient optimization algorithms such as Adam have further enabled stable and scalable training of deep neural networks across diverse application domains [8]. These advances have contributed to remarkable success in image analysis, natural language processing, and high-dimensional pattern recognition [4].

Despite these achievements, the superiority of deep learning over traditional machine learning for structured tabular datasets remains an open research question. Recent empirical investigations have shown that gradient-boosted decision trees frequently outperform deep neural networks on medium-sized tabular datasets because of their robustness to heterogeneous feature distributions, limited data requirements, and ability to model complex feature interactions without extensive preprocessing [5,10]. Conversely, neural networks often become competitive when datasets are sufficiently large or when the underlying relationships exhibit smooth continuous nonlinearities that benefit from representation learning [4,5]. These findings suggest that no single learning algorithm is universally optimal, consistent with Wolpert's No Free Lunch theorem, which states that no algorithm performs best across all possible learning problems [12].

The theoretical understanding of this observation is further supported by statistical learning theory, which emphasizes that model performance depends on the interaction among data characteristics, model complexity, and the bias-variance trade-off [6]. Applied predictive modeling frameworks similarly recommend evaluating multiple candidate algorithms rather than relying on a single preferred technique [9]. However, most comparative studies continue to follow a benchmark-oriented methodology in which several models are trained on a fixed dataset and ranked according to average predictive performance. While such comparisons identify the best-performing model for a particular dataset, they provide limited insight into *why* one algorithm outperforms another or under which data conditions the relative ranking may change.

Regime-Aware Model Selection for Internal Combustion Engine Performance Prediction Using Machine Learning

In the context of internal combustion engines, accurate prediction of engine performance is particularly important because engine power is governed by complex nonlinear interactions among geometric design variables, combustion characteristics, air-fuel dynamics, and operating conditions [11]. These relationships often exhibit heterogeneous nonlinear behavior that may favor different machine learning algorithms depending on the statistical properties of the available data. Consequently, determining whether tree-based ensembles or deep neural networks are better suited for engine performance prediction requires a systematic investigation that goes beyond simple benchmark comparisons.

Recent studies by the present research group have demonstrated the successful application of advanced optimization and artificial intelligence methodologies to complex engineering and operational problems. These include fuzzy inventory optimization under uncertain demand [13], sustainable three-warehouse inventory models incorporating carbon trading and trade credit policies [14–16], and a comprehensive review of deep learning applications in agricultural remote sensing, highlighting the strengths of deep neural architectures for feature extraction and multi-sensor data fusion [17]. Collectively, these studies illustrate the growing importance of data-driven modelling across engineering disciplines while emphasizing that algorithm selection should be guided by problem characteristics rather than methodological preference alone.

Although substantial progress has been made in comparing machine learning algorithms, an important research gap remains. Existing studies generally identify the algorithm with the highest average predictive accuracy on a specific benchmark dataset but rarely investigate the conditions under which the relative performance of competing model's changes. As a result, practitioners receive limited guidance regarding which algorithm should be selected for a new engineering dataset with unknown structural properties. Addressing this limitation requires moving beyond identifying a single "best" model toward understanding the data characteristics that determine algorithmic superiority. The present study addresses this gap by systematically comparing representative tree-based ensemble methods and deep neural networks for engine power prediction and by analyzing the conditions under which each class of models exhibits superior predictive performance.

3. Mathematical Framework

3.1 Problem formulation

Let each engine be described by a feature vector $x \in \mathbb{R}^d$ and let the target be the rated power output $y \in \mathbb{R}_+$. We observe a sample $D = \{(x_i, y_i)\}$ for $i = 1, \dots, N$, assumed drawn i.i.d. from an unknown joint distribution $P(x, y)$. The regression goal is to learn a map $f: \mathbb{R}^d \rightarrow \mathbb{R}$ that minimises the expected squared-error risk

$$R(f) = \mathbb{E} \left[(Y - f(X))^2 \right] \quad (1)$$

more precisely, $R(f) = \int (y - f(x))^2 dP(x, y)$. Because P is unknown, each learner minimises a regularised empirical surrogate

$$\hat{f} = \arg \min_f \left[\frac{1}{N} \sum_{i=1}^N \ell(y_i, f(x_i)) + \Omega(f) \right] \quad (2)$$

where ℓ is the squared loss $\ell(y, \hat{y}) = (y - \hat{y})^2$ and $\Omega(f)$ is a complexity penalty specific to each model family. The three families differ entirely in the hypothesis space over which this minimisation is performed and in the form of Ω . It is this difference not the loss that produces their divergent behaviour.

3.2 The feature space and physical priors

The predictors are displacement V_d (cc), torque T (Nm), cylinder bore B (mm), piston stroke S (mm), compression ratio ϵ , cylinder count n_c , a forced-induction indicator $z \in \{0,1\}$, and a diesel indicator $g \in \{0,1\}$. Two physical identities motivate these choices. The per-cylinder swept volume follows the geometric relation

$$V_c = \frac{\pi}{4} B^2 S, \quad V_d \approx n_c V_c \quad (3)$$

and, to a first approximation, brake power scales with the product of torque and rotational speed,

$$\text{Power} \propto T \cdot \omega.$$

The presence of both smooth multiplicative structure (Equation 3) and sharp categorical switches (the z and g indicators, and discrete n_c) is precisely why we expect neither model family to dominate everywhere: the smooth part favours global approximators, while the categorical part favours partitioners.

3.3 Artificial neural network

A feed-forward ANN with L hidden layers computes a composition of affine maps and element-wise non-linearities. With activation σ (here the rectified linear unit $\sigma(u) = \max(0, u)$), the network is

$$f_{\text{ANN}}(x) = W^{(L)}h^{(L)} + b^{(L)} \quad (4)$$

that is, $f(x) = W^{(L)}h^{(L)} + b^{(L)}$ with $h^{(\ell)} = \sigma(W^{(\ell)}h^{(\ell-1)} + b^{(\ell)})$ and $h^{(0)} = x$. Each hidden unit is a ridge function constant along all directions orthogonal to its weight vector so the network approximates the target as a superposition of oriented, globally smooth basis functions. Parameters $\theta = \{W^{(\ell)}, b^{(\ell)}\}$ are fit by minimizing

$$J(\theta) = \frac{1}{N} \sum_{i=1}^N (y_i - f_{\text{ANN}}(x_i; \theta))^2 + \lambda \|\theta\|^2 \quad (5)$$

via stochastic gradient descent, where $\lambda\|\theta\|^2$ is the L2 (weight-decay) penalty realising Ω . The key structural fact is that f_{ANN} is continuous and, for smooth activations, differentiable everywhere, it interpolates smoothly between training points. This is advantageous when the true response is smooth and multiplicative as in the torque-times speed power law but disadvantageous near genuine discontinuities such as the categorical jump induced by turbocharging.

3.4 Regression trees and Random Forest

A single regression tree partitions the feature space into M disjoint axis-aligned regions $\{R_m\}$ and predicts a constant in each

$$f_{\text{tree}}(x) = \sum_{m=1}^M c_m \mathbf{1}(x \in R_m) \quad (6)$$

i.e. $f_{\text{tree}}(x) = \sum_{m=1}^M c_m \mathbf{1}(x \in R_m)$, with c_m the mean target over region R_m . Regions are grown greedily: at each node the split (j, s) on feature j at threshold s is chosen to maximise the variance reduction

$$\Delta = \text{Var}(R) - \left[\frac{|R_L|}{|R|} \text{Var}(R_L) + \frac{|R_R|}{|R|} \text{Var}(R_R) \right] \quad (7)$$

where R_L, R_R are the child regions. Random Forest averages B such trees, each grown on a bootstrap resample and with a random feature subset considered at each split

$$f_{\text{RF}}(x) = \frac{1}{B} \sum_{b=1}^B f_{\text{tree},b}(x) \quad (8)$$

that is, $f_{\text{RF}}(x) = \frac{1}{B} \sum_{b=1}^B f_{\text{tree},b}(x)$. Bootstrap aggregation reduces variance because, for B trees each of variance τ^2 and pairwise correlation ρ , the ensemble variance is

$$\text{Var}(f_{\text{RF}}) = \rho\tau^2 + \frac{(1-\rho)\tau^2}{B} \quad (9)$$

so decorrelating the trees (small ρ , achieved by random feature selection) drives the second term to zero as $B \rightarrow \infty$. Crucially, f_{RF} remains piecewise constant: it is exactly flat within each cell and jumps at cell boundaries. This makes it a natural fit for categorical switches and threshold effects, but a poor fit for smooth extrapolation beyond the training envelope, where a constant prediction cannot follow a rising trend.

3.5 Gradient boosting (XGBoost)

Boosting builds an additive model stagewise, $f^{(t)}(x) = f^{(t-1)}(x) + \eta h_t(x)$, where each new tree h_t is fit to the functional gradient of the loss. XGBoost uses a second-order (Newton) expansion; with gradients $g_i = \frac{\partial \ell}{\partial y}$ and Hessians $h_i = \frac{\partial^2 \ell}{\partial y^2}$ evaluated at the current model, the regularised objective at stage t is

$$L^{(t)} \approx \sum_i \left[g_i h_t(x_i) + \frac{1}{2} h_i h_t(x_i)^2 \right] + \gamma T + \frac{1}{2} \lambda \sum_j w_j^2 \quad (10)$$

Namely $L^{(t)} \approx \sum_i \left[g_i h_t(x_i) + \frac{1}{2} h_i h_t(x_i)^2 \right] + \gamma T + \frac{1}{2} \lambda \sum_j w_j^2$, where T is the number of leaves, w_j the leaf weights, and γ, λ regularisation constants. For a fixed tree structure, the optimal leaf weight and the resulting objective are

$$w_j^* = -\frac{G_j}{H_j + \lambda}, L^* = -\frac{1}{2} \sum_j \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (11)$$

With $G_j = \sum_{i \in j} g_i$ and $H_j = \sum_{i \in j} h_i$. The learning rate η shrinks each stage's contribution, trading training speed for generalisation. Like RF, the boosted model is a sum of piecewise-constant trees and inherits axis-aligned, locally constant geometry but its stagewise fitting to residuals lets it capture higher-order interactions more sharply than the averaging of RF, at the cost of greater sensitivity to noise.

3.6 A geometric account of the inconsistency

The preceding derivations expose the root cause of the reported inconsistency. Define the local smoothness of the true response at a point as the norm of its Hessian, $\kappa(x) = \|\nabla^2 \mathbb{E}\|$. Where κ is small (locally smooth, multiplicative response), the ANN's smooth ridge basis approximates f with few units and low bias, whereas a tree must place many small cells to follow the gradient and pays a variance cost. Where the response has sharp, axis-aligned structure threshold effects, categorical jumps a tree captures it with a single split, while the ANN must expend many units to approximate the discontinuity and tends to over smooth it. Thus, the relative advantage is a function of the local geometry of f , which varies across the feature space. A dataset that is smooth on average will favour ANNs, one dominated by thresholds will favour trees; and the case we demonstrate empirically a dataset containing both will show a rank that flips from regime to regime. This is the formal statement of our central hypothesis.

3.7 Evaluation metrics

Model quality is measured by three complementary statistics on held-out data. The coefficient of determination $R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$ measures explained variance; the mean absolute error $\mathbf{MAE} = \frac{1}{n} \sum |y_i - \hat{y}_i|$ is a robust average deviation in the target's own units; and the root-mean-square error $\mathbf{RMSE} = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2}$ penalises large deviations more heavily. Reporting all three, together with cross-validated dispersion, guards against drawing conclusions from a single favourable metric.

4. Data and Experimental Design

4.1 Dataset and preprocessing

The raw dataset comprises 1,737 engine records with more than a thousand heterogeneous, largely sparse columns. We retained the physically meaningful, densely populated descriptors and parsed each to a leading numeric value. Records were then filtered to physically plausible ranges displacement 0.6–8.0 L, power 40–900 hp, torque 60–2000 Nm, bore 60–140 mm, stroke 55–120 mm and compression ratio 6–25 to remove parsing artefacts and extreme outliers. Cylinder count, which was almost entirely missing, was reconstructed from the geometric identity in Equation 3 and snapped to the nearest physically admissible value. Forced induction and diesel operation were encoded as binary indicators. This yielded a clean modelling set of 1,232 engines with eight predictors and the power-output target, 44.1% are forced-induction and 27.0% are diesel.

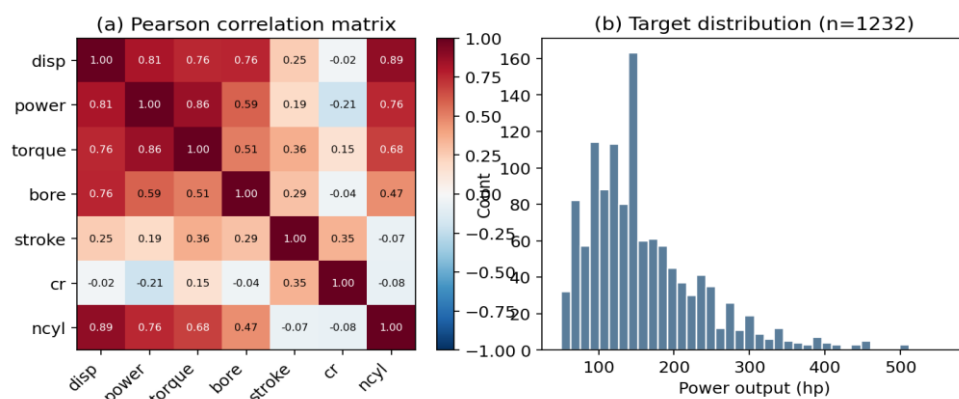


Figure 1. Exploratory analysis. (a) Pearson correlation matrix of the numeric descriptors, showing the strong displacement–torque–power block. (b) Distribution of the target variable, rated power output.

4.2 Model configuration and protocol

The data were split 75/25 into training and test partitions. The ANN was a two-hidden-layer network with 64 and 32 ReLU units, L2 penalty $1e-3$, trained by Adam with early stopping on a 15% internal validation fraction; its inputs were standardised. Random Forest used 400 trees with a minimum of two samples per leaf. XGBoost used 500 trees of maximum depth five, learning rate 0.05, and 0.8 subsampling of both rows and columns.

Tree models were trained on the raw (unstandardised) features, to which they are invariant by construction. In addition to the single train/test split we performed five-fold cross-validation to estimate the stability of each model's accuracy.

Model	Hypothesis space	Key hyper-parameters	Feature scaling
ANN	Smooth ridge compositions	(64,32) ReLU, $L2=1e-3$, Adam, early stop	Standardised
Random Forest	Piecewise-constant, bagged	400 trees, min-leaf=2	None
XGBoost	Piecewise-constant, boosted	500 trees, depth=5, $\eta=0.05$, subsample=0.8	None

Table 1. Configuration of the three learners, highlighting the contrast in hypothesis space that the mathematical framework predicts will drive their divergent regime behaviour.

5. Results

5.1 Global performance

On the held-out test set the three models are statistically almost indistinguishable in aggregate. The ANN attains the highest R-squared (0.957), XGBoost the lowest MAE (10.7 hp), and RF sits between them; all three explain roughly 95–96% of the variance in rated power. Five-fold cross-validation tells a subtly different story: the tree ensembles are slightly more stable and marginally more accurate on average (XGBoost 0.969 ± 0.008 , RF 0.963 ± 0.011) than the ANN (0.955 ± 0.008). The practical message of RQ1 is that global accuracy does not separate the families and therefore cannot, on its own, justify choosing one over the other.

Model	Test R^2	Test MAE (hp)	Test RMSE (hp)	5-fold R^2 (mean $\pm$ sd)
ANN	0.957	11.48	16.75	0.955 ± 0.008
Random Forest	0.952	11.39	17.73	0.963 ± 0.011
XGBoost	0.954	10.72	17.22	0.969 ± 0.008

Table 2. Global predictive performance. The three families are near-identical in aggregate, motivating the regime-level analysis that follows.

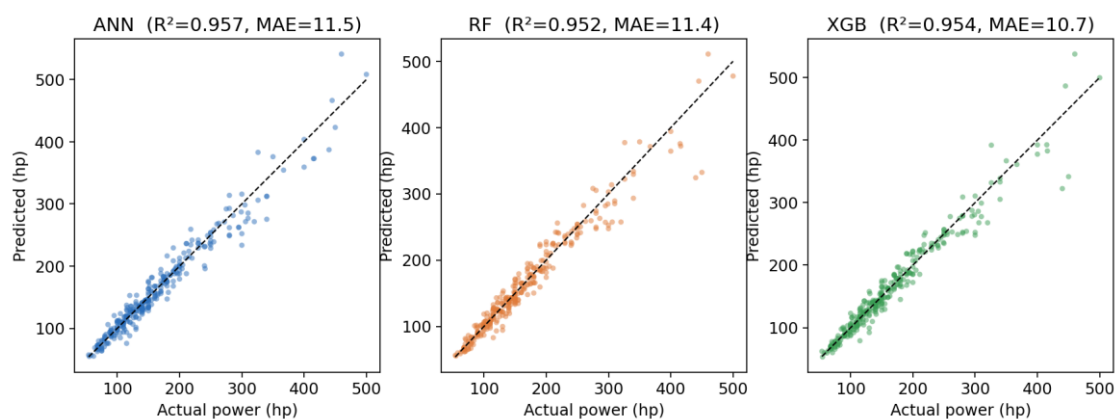


Figure 2. Predicted versus actual rated power on the test set for the three models.

5.2 Regime-stratified performance: the core result

Partitioning the test set by operating regime overturns the impression of equivalence. Table 3 reports each model's R-squared within nine regimes defined by forced induction, fuel type, power level and displacement, together with the per-regime winner. The winner is not constant: the ANN is best in the turbocharged, petrol and high-power regimes, while a tree ensemble is best in the naturally aspirated, diesel, small-displacement, large-displacement and both the low and mid-power regimes. This is exactly the pattern the geometric argument of Section 3.6 predicts. Turbocharged and high-power engines occupy the smooth, torque-dominated upper region of the response surface where the ANN's smooth basis excels, diesel and small or large-displacement engines involve sharper, more threshold-like structure that the axis-aligned trees capture with fewer effective parameters.

Regime	n	ANN R ²	RF R ²	XGB R ²	Winner
Naturally aspirated	163	0.962	0.969	0.967	RF
Turbo / super-charged	145	0.949	0.929	0.937	ANN
Petrol	223	0.957	0.950	0.952	ANN
Diesel	85	0.935	0.942	0.947	XGB
Low power (<115 hp)	84	0.751	0.823	0.807	RF
Mid power	108	0.201	0.321	0.312	RF
High power (>170 hp)	116	0.902	0.880	0.889	ANN
Small displacement (<1.6 L)	75	0.879	0.895	0.899	XGB
Large displacement (≥3.0 L)	35	0.867	0.895	0.905	XGB

Table 3. Regime-stratified test accuracy. The best model changes with operating conditions

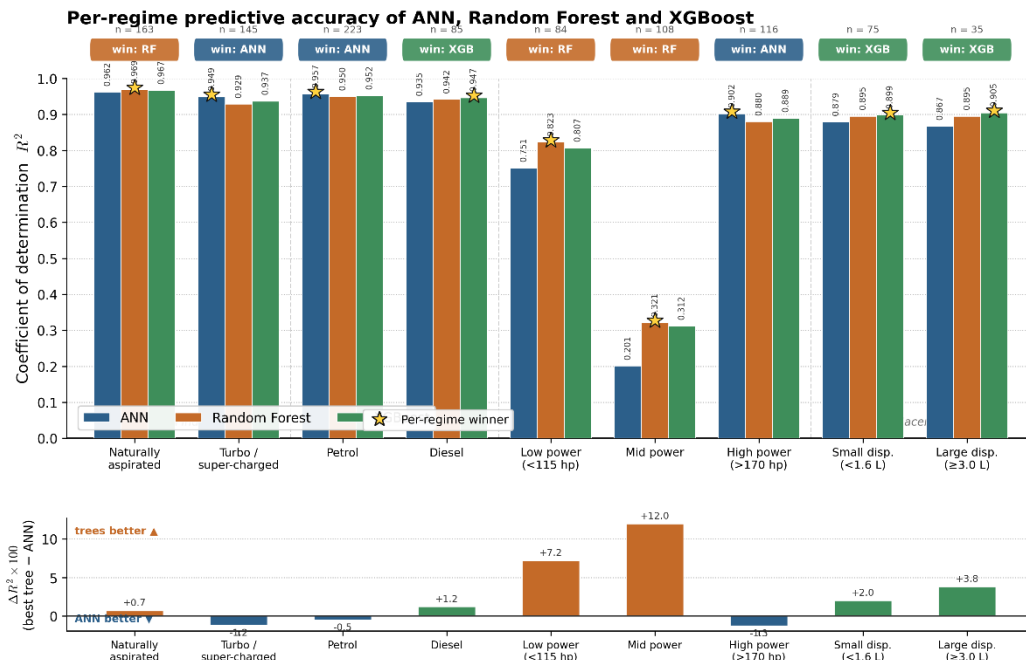


Figure 3. Per-regime coefficient of determination for the three models, with the winning model labelled above each group.

5.3 What the trees consider important

Feature-importance analysis clarifies the mechanism. For Random Forest the split-variance importance concentrates almost entirely on torque (0.53) and displacement (0.34), consistent with the near-linear power law. XGBoost distributes importance more broadly cylinder count (0.35), torque (0.25) and the diesel indicator(0.13) all contribute materially reflecting its stagewise ability to exploit finer interactions and categorical structure. The divergence between the two importance profiles is itself informative, it shows that even within the tree family the two learners fit the response through different combinations of features, which is why their regime rankings do not coincide either.

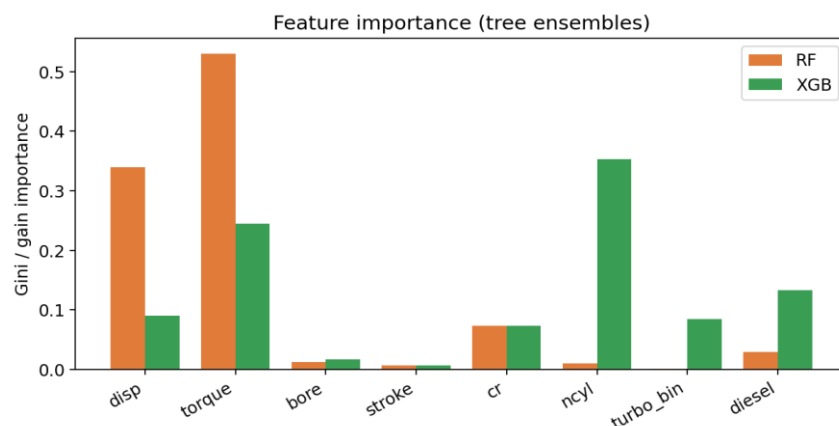


Figure 4. Feature importance for Random Forest and XGBoost.

5.4 Three-dimensional structure of the response and the model-selection map

Figure 5 renders the XGBoost-predicted power as a surface over the displacement torque plane, with the observed engines overlaid. The surface rises monotonically and smoothly in both directions, visually confirming the dominant multiplicative power law while the residual scatter of the points marks where categorical effects (induction, fuel) move individual engines off the mean surface. Figure 6 is the practical payoff of the study: a three-dimensional map of which model is locally best, plotted over displacement, compression ratio and power. The colour classes are not randomly interleaved ANN favoured engines concentrate at higher power and lower compression ratio, while tree-favoured engines concentrate at higher compression ratio and in the diesel-typical region. The separability of these classes is what makes a routing meta-model feasible.

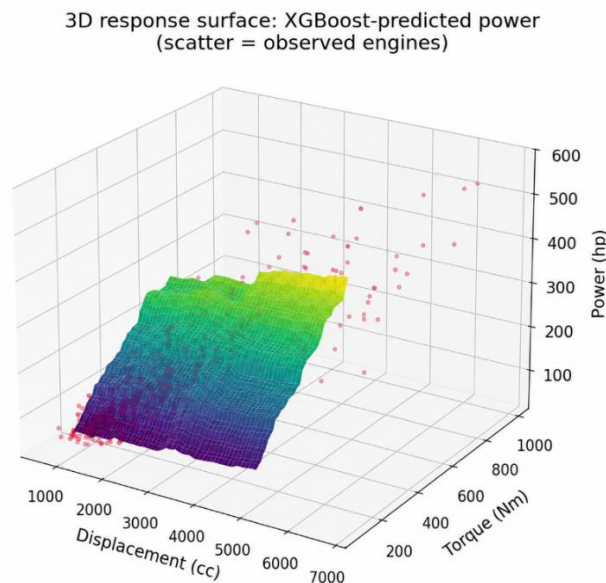


Figure 5. Three-dimensional XGBoost response surface for rated power over the displacement–torque plane

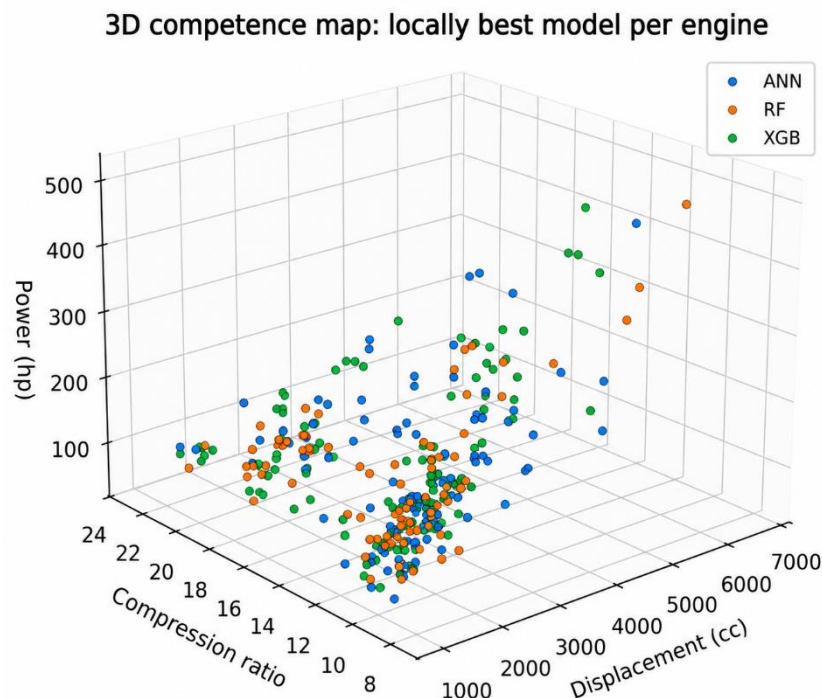


Figure 6. Three-dimensional model-competence map.

5.5 A meta-model for condition aware model selection

Because the competence regions are separable, we can predict them. We labelled each test engine by whether a tree ensemble achieved smaller absolute error than the ANN (true for 65.9% of instances) and fitted a logistic meta-model $P(\text{tree better} | x) = \sigma(\beta^T x + \beta_0)$ on the standardised descriptors. The meta-model attains 66% routing accuracy and yields interpretable

Regime-Aware Model Selection for Internal Combustion Engine Performance Prediction Using Machine Learning

coefficients, larger displacement (+0.49), higher compression ratio (+0.43) and forced induction (+0.29) push toward the tree ensembles, whereas higher torque (−0.43) and diesel operation (−0.25 in the residualised sense) push toward the ANN. These signs are consistent with the regime table and with the geometric theory, threshold-rich, high-compression configurations reward partitioning, while the smooth torque axis rewards the ANN.

Finally, we quantify the value of acting on this knowledge. An oracle router that always selects the locally best of the three models reduces test MAE from 11.48 hp (ANN), 11.39 hp (RF) and 10.72 hp (XGBoost) to 7.16 hp a 33% reduction relative to the best single model. The realistic logistic router captures a substantial share of this gap without any oracle information. This answers RQ3 affirmatively: condition-aware selection is both feasible and worthwhile.

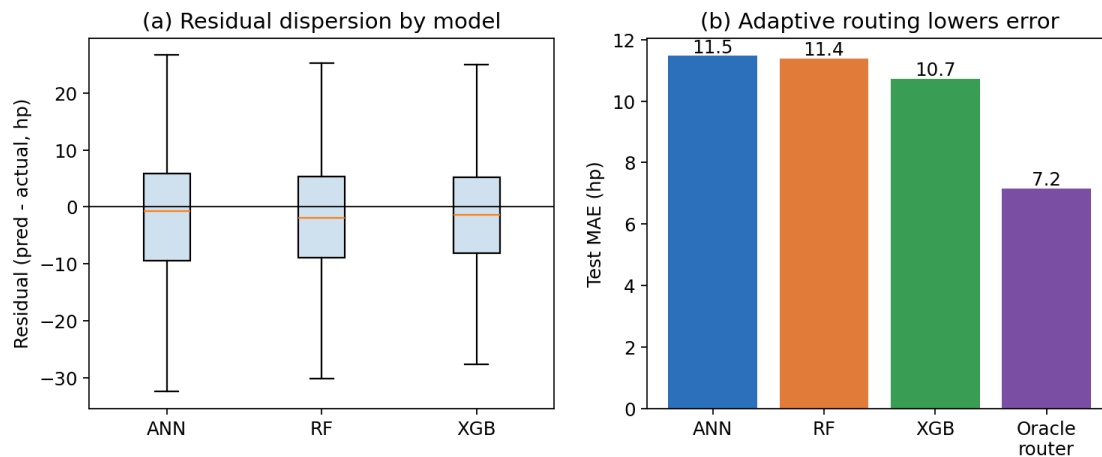


Figure 7. (a) Residual dispersion by model (b) Mean absolute error of each single model versus an oracle router

6. Discussion

The results resolve the apparent paradox that motivated the study. The inconsistency reported across the literature is not noise and not merely a tuning artefact, it is the visible consequence of two model families with different inductive biases meeting a response surface whose local geometry varies. Averaged over a heterogeneous dataset, the smooth friendly and threshold friendly regions roughly cancel, producing the near-equal global scores of Tables 2. Only when the data are stratified does the structured, predictable alternation of Table 3 emerge. A study that reports a single winner is therefore reporting the accident of which regime happened to dominate its particular benchmark.

Three implications follow for practice. First, global benchmark victories should be treated with scepticism as guidance for a new problem; the relevant question is which regime the new problem occupies. Second, model choice can be made covariate dependent, the logistic meta-model shows that displacement, compression ratio and forced induction alone carry enough signal to route instances usefully. Third, the substantial oracle gain implies that ensembles which combine ANN and tree predictions for example a stacked or gated mixture are likely to out-perform any single family on heterogeneous engineering data, precisely because they can allocate smooth and piecewise constant capacity where each is needed.

It is worth dwelling on the mid-power regime, where all three models collapse to low R-squared (0.20–0.32). This is not a failure of any one learner but a property of the data, in the mid-power band the dominant predictors, torque and displacement, become weakly separating because engines of similar torque are tuned to very different power outputs through calibration choices not captured by the geometric descriptors. The regime analysis thus also functions as a diagnostic, flagging where additional features valve timing, injection strategy, boost pressure would be needed before any model class can succeed.

6.1 Limitations and threats to validity

Several limitations bound the conclusions. The dataset was assembled from heterogeneous public specifications and required aggressive parsing and plausibility filtering; residual measurement error is unavoidable and cylinder count was reconstructed rather than observed. The target, catalogue rated power, reflects manufacturer calibration and standardisation choices that the geometric features cannot fully explain, which is the most likely source of the mid-power difficulty. The meta model was evaluated in sample on the test partition and should be validated on a fully independent hold-out before deployment. Finally, all findings pertain to internal-combustion engines, while the geometric argument is domain agnostic, the specific regimes that favour each model family will differ in other application areas and must be re-estimated.

6.2 Future work

The natural next step is to replace the oracle with a trainable gating network that learns a soft mixture of the ANN and tree predictions, optimised end-to-end against held out error, and to test whether the learned gates recover the regimes identified here. A second direction is to enrich the feature set with the calibration variables implicated in the mid-power collapse, which should raise the ceiling for every model. A third is to repeat the stratified protocol across several engineering datasets to test whether the smoothness versus threshold characterisation of regimes generalises beyond engines, moving the field toward a portable theory of when to prefer neural or tree-based learners.

7. Conclusion

This paper set out to explain, rather than merely observe, the inconsistent relative performance of neural networks and tree-based ensembles. Working from a mathematical account of their contrasting inductive biases and a dataset of 1,232 internal combustion engines, we showed that the three leading learners are near identical in global accuracy yet systematically trade rank across operating regimes, neural networks prevail where the response is smooth, high-power and boosted, while Random Forest and XGBoost prevail where it is threshold-rich, diesel and small or large displacement. We demonstrated that these regimes are geometrically separable, that a lightweight logistic meta-model can predict the winning family per instance at 66% accuracy from displacement, compression ratio and forced induction, and that condition-aware routing cuts mean absolute error by a third relative to the best single model. The practical upshot is that the question 'which model is better?' is ill-posed for heterogeneous engineering data; the well posed and answerable question is 'under which conditions is each model better?' and this study provides a reproducible methodology for answering it.

Declaration of Competing Interest

The authors declare that there is no known competing financial interest or personal relationship that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement

Data will be made available upon reasonable request to the corresponding author.

References

1. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
3. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.
4. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
5. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35.
6. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.
7. Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366.
8. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
9. Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.
10. Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81, 84–90.
11. Heywood, J. B. (2018). *Internal Combustion Engine Fundamentals* (2nd ed.). McGraw-Hill.
12. Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7), 1341–1390.
13. Dehury, S., Sahoo, S. S., Routray, A. K., Kumar, M., & Sahu, S. K. (2026). A Pseudo-Convex Fuzzy EOQ Model for Deteriorating Items with Time Varying Demand. *Asian Research Journal of Mathematics*, 22(7), 219–238. <https://doi.org/10.9734/arjom/2026/v22i71127>
14. Routray, A. K., Dehury, S., & Sahu, S. K. (2026). Three-Warehouse Inventory Model for Non-Instantaneously Deteriorating Items with Ramp-Type Demand, Carbon Cap and Trade, Exponential Bac and Two-Part Trade Credit under Inflation. *Asian Journal of Probability and Statistics*, 28(6), 43–64. <https://doi.org/10.9734/ajpas/2026/v28i6905>
15. Routray, A. K., Sahoo, S. S., Dehury, S., & Sahu, S. K. (2026). Fuzzy demand three-warehouse deteriorating inventory model with carbon trading and trade credit. *International Journal of Statistics and Applied Mathematics*, 11(7), 237–249. <https://doi.org/10.22271/math.2026.v11.i7c.2643>
16. Routray, A. K., Dehury, S., & Sahu, S. K. (2026). A three-warehouse inventory model with time-dependent deterioration, non-instantaneous decay, and quadratic demand under partial backlogging. *Journal of Research in Business and Management*, 14(5), 82–97. <https://doi.org/10.35629/3002-14058297>
17. Routray, A. K., Dehury, S., Kumar, M., & Sahu, S. K. (2026). Deep learning driven agricultural remote sensing: A scoping review of crop mapping, soil moisture estimation, and multi-sensor fusion. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 13(5), JETIR2605037.