Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Hybrid enzyme action optimizer CatBoost ensemble for drawbar load prediction with graphical user interface and comparative evaluation against nature inspired metaheuristic algorithms

**Frankline Mwiti**

franklinemwiti2@gmail.com

University of Nairobi

**Ayub Gitau**

University of Nairobi

**Duncan Mbuge**

University of Nairobi

**Research Article**

**Keywords:** Hybrid machine learning, CatBoost ensemble, Hyperparameter tuning, Enzyme action optimizer, Model generalization, SHapley Additive exPlanations, Graphical user interface

**Posted Date:** October 30th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11169776/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Hybrid enzyme action optimizer CatBoost ensemble for drawbar load prediction with graphical user interface and comparative evaluation against nature inspired metaheuristic algorithms

**Frankline Mwiti [a*], Ayub Gitau [a, b], Duncan Mbuge [c]**

[a] Faculty of Engineering, Mechanization, Power & Machinery Engineering Division, University of Nairobi, P.O. Box 30197, Nairobi GPO, 00100 Kenya. https://orcid.org/0000-0002-9662-9430
[b] Office of the Vice Chancellor, University of Nairobi, P.O. Box 30197, Nairobi GPO, 00100 Kenya. https://orcid.org/0000-0003-1050-5232. ayub.gitau@uonbi.ac.ke
[c] Faculty of Engineering, University of Nairobi, P.O. Box 30197, Nairobi GPO, 00100 Kenya. https://orcid.org/0000-0002-3943-2877. duncan.mbuge@uonbi.ac.ke
* franklinemwiti2@gmail.com

## Highlights

- Nature-inspired algorithms can optimize standalone Ensemble models for engineering
- 25 nature-inspired algorithms are evaluated to optimize the Base-CatBoost learner
- Enzyme action optimizer (EAO) hybrid with Base-CatBoost generalized most precisely
- SHAP explainable hybrid AI revealed tractive force dominates off-road drawbar load
- A user-friendly EOA-CatBoost ensemble GUI for off-road drawbar pull is presented

## Abstract

Cognitive intelligence of standalone machine learning models is constrained by target leakage, weak data mining, inadequate hyperparameter tuning, and premature convergence to local optima. This study integrated Base-CatBoost Ensemble learner with 25 recent nature-inspired metaheuristic algorithms and evolved new hybrid-CatBoost models for predicting drawbar load ($H_{Db}$). These included, Aurora Intelligent Metaheuristic-AIM algorithm, Artificial Lemming Algorithm-ALA, Bobcat Optimization Algorithm-BOA, Bacteria Phototaxis Optimizer-BPO, Chicken Swarm Optimization-CSO, Enzyme Action Optimizer-EAO, Grizzly Bear Fat Increase Optimizer-GBFIO, Gazelle Optimization Algorithm-GOA, Grey Wolf Optimizer-GWO, Harris hawks optimization-HHO, Hippopotamus Optimization Algorithm-HO, Jellyfish Search Optimizer-JSO, Mantis Shrimp Optimization Algorithm-MShOA, Osprey Optimization Algorithm-OOA, Parrot Optimizer-PO, Pufferfish Optimization Algorithm-POA, Reindeer Cyclone Optimization Algorithm-RCOA, Secretary bird optimization algorithm-SBOA, Swift Flight Optimizer-SFO, Snake Optimizer Algorithm-SO, Walrus Optimization-WAOA, Warthog Optimization Algorithm-WartOA, Wolf-Bird Optimizer-WBO, Whale Optimization Algorithm-WOA and White Shark Optimizer-WSO. A scikit-learn library for the Base-CatBoost learner was

created, and critical hyperparameters of hybrid ensemble models were heuristically optimized, targeting the experimental $H_{Db}$ obtained from 3360 values in the 14 soil-machine database. Total multi-metric accuracy (R, $R^2$, $R^2$adjusted, EVS, RAE, RRMSE, RMSRE, NRMSE, SI, PI, WMAPE, VAF, PA, CV) and noise robustness (MSE, RMSE, MAE, MedAE, Huber loss-Lδ, Log-Cosh loss-LCL) score ranking (980), regression error characteristic curve-REC (AOC=0.236), Taylor reference clustering (R-σ-cRMSE), and convergence stability (perturbation score, 0.005, generalization gap, 4.545e-06) confirmed hybrid EAO-CatBoot as the superior model. SHAP explainable hybrid AI-XHAI revealed that tractive thrust most significantly dominated $H_{Db}$ prediction dispersion and convergence decision path attribution in EAO-CatBoost. Python-based GUI for hybrid EAO-CatBoost is presented for user-friendly HDb prediction in engineering practice.

**Keywords:** Hybrid machine learning, CatBoost ensemble, Hyperparameter tuning, Enzyme action optimizer, Model generalization, SHapley Additive exPlanations, Graphical user interface.

## 1. Introduction

Cat boosting ensemble (CatBoost) is a Gradient Boosting Decision Tree-based algorithm, well-suited to machine learning tasks involving heterogeneous and complex variables. According to Prokhorenkova et al. [1], CatBoost was designed for computationally intensive tasks involving highly heterogeneous datasets. Since its debut in late 2018, CatBoost has been successfully deployed in machine learning studies to unravel complex interdependencies, achieving superior performance compared to other machine learning algorithms [2]. However, in the cognitive engineering of tillage machinery systems, standalone CatBoost models encounter specific limitations when predicting tractor-implement performance from dynamic soil-wheel and soil-tool interaction variables. These bottlenecks are primarily driven by the extreme nonlinearity of soil dynamics and by the sensitivity of gradient boosting to high-dimensional, noisy agricultural data. While CatBoost utilizes ordered target statistics to minimize leakage [3], it remains vulnerable to learning and modelling tillage variables due to sequential dependency of soil-machine variables in situ. This is because soil-tool and soil-wheel interactions often involve multiple passes and tooling disturbances whereby features related to previous passes inadvertently leak future state information into the model, leading to artificially high accuracy that fails during other live field operations. At the same time spatio-temporal autocorrelation of soil properties may cause standalone models to "learn" these spatial patterns as proxies for the target variable rather than the underlying dynamics, leading to poor generalization across different fields. Further, stand-alone CatBoost struggles to learn the "cause-and-effect" of dynamic load transfers in tillage

machinery systems because they lack the inherent physical-numerical constraints of soil-machine interactions that encompass complex geometric and material non-linearities. As such, a stand-alone CatBoost may fail to identify the physico-mechanical non-linearities in variable interactions on mechanical dynamics at critical thresholds, especially when soils shift from elastic to plastic deformation, which can introduce noise into the dataset. All these challenges put standalone CatBoost models at high risk of target leakage. On a different note, implementation of the standalone CatBoost algorithm often relies on manual hyperparameter tuning [4, 5], which frequently results in suboptimal predictions due to premature convergence to local optima, particularly when handling multi-modal search spaces in soil mechanics data. Dynamic fluctuations in heterogeneous soil-tire and soil-tool interface variables near the surface and at various depths in situ render standalone CatBoost models inferior for predicting highly instantaneous drawbar pull. Due to weak feature learning [6, 7], Base-CatBoost models struggle to predict the effects of dynamic interactions on critical mechanical nonlinearities during soil-wheel-and soil-tool interactions in tillage. As such, standalone Cat-Boost models may also face reliability challenges under heterogeneous conditions [8, 9], such as off-road drawbar pull ($H_{Db}$) estimation. This is because of heterogeneous physio-mechanical variabilities across heterogeneous agricultural soil matrix and the effects of dynamic soil deformation nonlinearities on the tire and tractor-implement configurations, which continuously alter traction and load transfer mechanisms. As such, due to weak learning of underlying relationships standalone Catboosting ensemble models can converge prematurely [10], especially when subjected to new operating environments and soil types with new unforeseen soil-machine datasets, leading to poor generalization of drawbar predictions. Hybridizing the Base-CatBoost algorithm with recent nature-inspired algorithms can be used for predictive modelling of such a dynamic target, which involves many variable interrelationships that require simultaneous optimization, superior data exploration, and variable exploitation. Literature suggests that boosting ensemble algorithms often outperforms several conventional machine learning approaches, including Artificial Neural Networks (ANN), in modelling heterogeneous datasets with complex nonlinearities and mixed data structures due to their robustness to feature variability and reduced sensitivity to extensive hyperparameter tuning [11, 2]. However, despite the native effectiveness of the Base-CatBoost learner in handling numerous categorical features within datasets, standalone CatBoost models are highly sensitive to feature learning, hyperparameter tuning, overfitting, and premature convergence, especially when applied to new datasets or unforeseen events. For instance, too high learning rate and learning depth lead to overshooting and overfitting, respectively, while the number and size of iterations can cause underfitting or overfitting [12]. At the same time

premature convergence can occur if the early stoppage parameters are not properly tuned, causing convergence to a local optimum [13]. All these challenges cause target leakage due to weakened computational efficiency, leading to poor generalization with unseen datasets, especially in highly complex and dynamic soil environments in tillage. Further, the Base-CatBoost model-parameter tuning, overfitting, and premature convergence are interconnected challenges that, when improperly managed, cause the models to memorize the training data rather than learn the complex underlying patterns in the tillage datasets. This further undermines cognitive robustness and often leads to target leakage, in which information and noise from outside the training variables corrupt the model, resulting in high training accuracy but poor generalization. This renders the predictive strength of standalone CatBoost models weak when applied to new, unseen tillage data, making their adoption inconclusive. As such, the modelling accuracy and predictive performance of CatBoost models critically depend on the algorithm type and hyperparameter tuning methodologies for robustness against overfitting. Therefore, the predictive performance of the standalone CatBoost models is limited by overfitting challenges related to local optima, slow convergence rates, difficulties with hyperparameter tuning, and weak computational power [14]. To resolve these challenges in agricultural tractors this study adopted 24 recently developed nature inspired machine learning algorithms to develop hybrid CatBoost models for off-load prediction of drawbar load of agricultural tractors using 14 highly dynamic and heterogeneous soil-machine variables in situ. These included; Aurora Intelligent Metaheuristic (AIM), Artificial lemming algorithm (ALA), Bobcat Optimization Algorithm (BOA), Bacteria phototaxis optimizer (BPO), Chicken Swarm Optimization (CSO), Enzyme Action Optimizer (EAO), Grizzly Bear Fat Increase Optimizer (GBFIO), Gazelle optimization algorithm (GOA), Grey Wolf Optimizer (GWO) Harris hawks optimization (HHO), Hippopotamus Optimization Algorithm (HO), Jellyfish Search Optimizer (JSO), Mantis Shrimp Optimization Algorithm (MShOA), Osprey Optimization Algorithm (OOA), Parrot Optimizer (PO), Pufferfish Optimization Algorithm (POA), reindeer cyclone optimization algorithm (RCOA), Secretary bird optimization algorithm (SBOA), Swift Flight Optimizer (SFO), Snake optimizer algorithm (SO), Walrus Optimization (WAOA), Warthog Optimization Algorithm (WartOA), Wolf-Bird Optimizer (WBO), Whale Optimization Algorithm (WOA) and White Shark Optimizer (WSO).

### 1.1 Novelty of the research

The novelty of this study is fourfold. First, the work presents one of the most rigorous machine learning evaluations of recent nature-inspired optimization algorithms for boosting the performance of CatBoost ensemble learners. The study developed a set of 25 new hybrid CatBoost ensemble learning models based on the latest nature-inspired metaheuristic

optimization algorithms to achieve intelligent engineering decision support in off-road machinery systems. Secondly the study builds and proposes a robust, efficient and high-performance hybrid EAO-CatBoost machine learning framework, with high predictive accuracy, computational efficiency, and exceptional generalization capacity in unseen data for predicting drawbar load. Thirdly, it embeds SHAP-based explainable hybrid artificial intelligence (XHAI) to enhance the transparency and engineering interpretability of the hybrid EAO-CatBoost machine-learning model. Lastly, it develops an intelligent Python-based user-friendly GUI for practical implementation and deployment of EAO-CatBoost model.

### 1.2 Significance of the study

Under highly dynamic soil-machine operating conditions, intelligent, real time drawbar load prediction is fundamental for accurate design, selection, and optimization of off-road machinery and traction systems. The study highlights the potential of nature-inspired optimization algorithms to improve the predictive intelligence, robustness, and explainability of ensemble machine learning models for complex engineering applications and established hybrid EAO-CatBoost as highly generalizable framework for HDb prediction. Integrated SHAP-based hybrid AI interpretability established tractive force as the dominant variable in drawbar load dispersion, thereby generating valuable engineering knowledge that enhances interpretability of AI-assisted decisions by the hybrid EAO-CatBoost model. Finally, deployment of the Python-based GUI provides a transparent and highly accurate decision-making support tool for intelligent off-road machinery systems and precision engineering. This enables practical implementation of the EAO-CatBoost model, translating the research into real-world practice for off-road engineering applications by engineers, researchers, and off-road agricultural machinery practitioners.

## 2. Materials and methods

### 2.1 Experimental database creation

A remotely monitored drawbar load dynamometer was logged into a laptop device and coupled to the drawbar of the research tractor and the front tow-end of the auxiliary tractor. A 2-tine deep strip tiller was then fitted with heavy-duty shear pins and limbered up on the three-point hitch of the auxiliary tractor. The engine was cranked at 1500 rpm, and the tillage locomotion engaged in a completely randomized combination of quintuplet (5) tire inflation pressures ($P_{Tire}$), quadruplet (4) rear wheel loads ($W_{Load}$), and (4) strip tillage depth ($T_{depth}$), and the drawbar dynamometer decoupled after sequential triplications (240). These included $P_{Tire}$ levels of 110.4, 151.8, 193.2, 234.6, 275.8kPa; $W_{Load}$ levels of 1.3, 11.8, 12.3, and 12.8kN; $T_{deth}$ levels of 1-10, 10-20, 20-30, and 30-40cm. The total tractive thrust (Ftr), implement draft force ($F_D$), rolling resistance force ($F_{Rr}$), soil-tire patch contact area ($A_{Stc}$), wheel rut depth ($R_{depth}$), wheel slippage

($S_{Wheel}$), and fuel injection rate ($Ø_{Fuel}$) were obtained for each combination of wheeling load, tillage depth, and tire inflation pressure levels. The technical specifications of the research tractor and tires are shown in Table 1. At the same time, soil parameters for each treatment were established using in situ and standard laboratory procedures to determine soil cone index ($CI_{soil}$), soil shear strength ($\tau_{Shear}$), soil moisture content ($\theta_{Soil}$), soil bulk density ($\gamma_{Soil}$), and soil plasticity index ($I_{PSoil}$) for every replica, yielding a database of 3360 values (240 sets of 14 soil-machine values).

**Table 1** Key technical specifications of the research tractor

| | Particulars/parameters | Specifications/dimensions |
|---|---|---|
| Tractor | Make and model | CASE IhJXM 90 HP, FPT S |
| | Model series and type | Farmall, 4WD |
| | Engine type/aspiration | 4-cylinder, turbocharged |
| | Engine rated power and r.p.m | 88hp/65.7kW @ 2300rpm |
| | Maximum engine torque | 4000Nm @1300 rpm |
| | Front Axle | 2WD/4WD |
| | Rear axle | Heavy-duty straight axle |
| | Static weight and standard ballast weight | 3600kg, 3250kg |
| | Max Lift capacity and ground clearance | 2500kg, 436mm |
| | Wheelbase 4WD,2WD | 2258, 2283mm |
| | Overall dimensions (LxWxH) | 4131x2087x2589mm |
| Rear tire | Tire size (width and rim diameter) | 18.4-30 |
| | Ply rating, Brand, and type | 14-ply, BKT, TT |
| | Pattern name and type construction | TR-136 Heavy-duty, Crossply |
| | Speed weight index and speed radius | 151A6, 725 |
| | Section width and overall diameter (mm) | 465, 1545 |
| | Static loaded radius (mm) | 712 |
| | Rolling circumference and tire weight | 4719mm, 94.67kg |

## 2.2 Database pre-processing

### *2.2.1 Normalization and variable scaling*

Since the database contained disparate units and ranges for various input variables, a data normalization procedure was implemented to achieve feature scaling. All numerical features were scaled using z-score normalization, which linearly transformed each feature to a standardized range (0-1), using Eq. 1 [15]. This method helped avoid exploding gradients and maintain directionality, thereby enhancing computational efficiency by ensuring algorithms sensitive to feature magnitudes perform effectively. This step was also critical for metaheuristic optimizers like the Bobcat Optimization Algorithm (BOA), distance-based algorithms, and gradient-descent-based optimizers, as it prevents features with larger numerical scales from dominating the search space, causing biased weight updates and slowed convergence while mitigating regression multicollinearity.

$$x' = \frac{x}{|x|_2} = \frac{x}{\sqrt{\sum_{i=1}^{n} x_i^2}} \quad (1)$$

where $(|x|_2)$ is the Euclidean norm of the vector ($x$).

### *2.2.2 Correlation and multi-statistical dependency*

Python-based Seaborn data visualization library was adopted to analyze and visualize the dynamics of correlations among variables, given its inherent graphical capabilities, thereby facilitating a more interpretative and transparent understanding of the modelling datasets. The quantitative measure of interrelationships among dataset variables and their levels of strength, directionality, and linear relationship among and between the variables and $H_{Db}$ was evaluated using Pearson's correlation ($r$) as shown in Eq. 2 [16].

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2.\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (2)$$

where n, $x_i$, $y_i$, $\bar{x}$, and $\bar{y}$ are the number of datapoints, individual datapoint variables, and their respective sample means of x and y. Multivariate assessment of dual input-output variable dependencies, tripartite and multipartite interdependencies, and dependencies were evaluated with the Partial Mantel test using the "fortify_mantel" function in the "ggcor" and "vegan" packages of the statistical software R. The distant dissimilarity or similarity relationships between two variables on the $H_{Db}$, while controlling the effect of a third variable, were evaluated by computing the partial correlation coefficient ($r_{AB.C}$) using Eq. 3 [17].

$$r_{AB.C} = \frac{r_{AB}-(r_{AC}.r_{BC})}{\sqrt{(1-r_{AC}^2)(1-r_{BC}^2)}} \quad (3)$$

where $r_{AB}$, $r_{AC}$, and $r_{BC}$ are the Mantel correlations between soil-machine interaction variables A and B, A and C, and B and C, respectively, while $r_{AB.C}$ defines the correlation remaining between A and B after the linear influence of C is discarded.

## 2.3 Hybrid CatBoost modelling

### *2.3.1 Base learner CatBoost model development*

A CatBoost library was created, and the base learner was implemented in a Visual Studio Code® environment using Python®, and the model was interpreted using the *Scikit-Learn*® framework. The Base-CatBoost learner was then executed by mathematically representing the given set of training datasets containing N samples with M features, where each sample was denoted as ($x_i$, $y_i$), with $x_i$ being the vector of M features and $y_i$ the corresponding target variable. CatBoost was then trained on the datasets to learn the function F(x) for predicting the target variable (y), the $H_{Db}$, by summing the initial guess with multiple products of summations over the ensemble trees and over the training samples using Eq. 4 [18].

$$F(x) = F_0(x) + \sum_{n=1}^{M}\eta f_m(x_i) \quad (4)$$

where $F_0(x)$ is the initial or baseline guess, M is the total number of trees in the ensemble, $\eta$ is the learning rate, while $f_m(x_i)$ is the prediction of the m$^{th}$ tree for the i$^{th}$ training sample. The final base learner CatBoost model was defined based on the gradient boosting algorithm and represented by a collection of decision trees, as shown in Eq. 5 [18]

$$F(x) = \sum_{i=1}^{n} \eta_i h_i(x) \quad (5)$$

where $\eta_i$ and $h_i(x)$ are the learning rate and decision trees for the i$^{th}$ sample, respectively. During modelling, the error gradients updated the learning algorithm using an unbiased gradient ($\delta_u$) computed using Eq. 6 [1]

$$\delta_u = \frac{\Delta L}{\Delta F(x_i)|_{F=\bar{F}}} \quad (6)$$

where $L$ and $F(xi)$ are the i$^{th}$ cost function and predicted values, respectively. To avoid data leakage, the target statistic method was used for numerical encoding in the CatBoost algorithm and for conversion of the category features (14 soil-machine variables), implemented using Eq. 7 [1].

$$\mu = \frac{class_{count} + prior\ x\ \mu_0}{count + prior} \quad (7)$$

where c$lass_count$ is the total number of samples in each class, *prior* is the smoothing parameter, and $\mu_0$ is the targeted mean. Further, during learning and predictive modelling, overfitting was optimized using tree depth and L2 regularization approaches defined by Eq. 8 and Eq. 9, respectively [19]

$$D_{Penalty} = \beta x d_{Curr} \quad (8)$$

$$\|w_{model}\|_2^2 = \sum_{i=1}^{n} {w_i}^2 \quad (9)$$

where $D_{penalty}$, $\beta$, and $d_{curr}$ are the depth penalty, penalty coefficient of depth, and current tree depth, while $w_{model}$, $n$, and $w_i$ are the weight vector, total model weights, and the weight of the i$^{th}$ variable in the leaves of the decision tree. The underlying patterns and interrelationships among experimental datasets ($W_{Load}$, $R_{depth}$, $P_{tire}$, $A_{Stc}$, $F_{Rr}$, $S_{Wheel}$, $T_{deth}$, $F_D$, $Ø_{Fuel}$, $CI_{soil}$, $T_{Shear}$, $\theta_{Soil}$, $\gamma_{Soil}$, and $I_{PSoil}$) were all employed to train and constitute drawbar load predictions of the base learner using algorithm implementation, model selection, fitting, and evaluation tools in the *Scikit-Learn*® library.

### *2.3.2 Integration and fine-tuning of Base-CatBoost with nature-inspired algorithms*

Algorithm selection and hyperparameter tuning influence the convergence behavior, model performance, and generalization ability of hybrid CatBoost machine learning approaches. In this study, the Base-CatBoost learner was hybridized using 24 Python-based nature-inspired metaheuristic algorithms. During modelling, auto-heuristic tuning was deployed to automatically establish the optimal hyperparameters for the developed hybrid ML algorithms. Randomized

search CV and the grid search CV techniques were sequentially executed in the modelling library for all ML algorithms to perform critical hyperparameter tuning and optimize model performance. The Scikit-Learn GridSearchCV function was deployed to perform exhaustive exploration search of estimators over specified grids with a range of combined hyperparameter values optimized by cross-validated grid-search over the parameters, and to evaluate model performance for different combinations of various hyperparameters using CV. This exhaustive search ensured that all possible combination approaches within the grid were considered. At the same time, a randomized search CV was used to randomly sample a specified number of hyperparameter combinations from a specified distribution defined by prior knowledge of the hyperparameter space. Random search hyperparameter combination sampling enabled the discovery of better performance configurations within the grid. During tuning, hybrid Catboosting iterations were determined by the number of ensemble decision trees, model complexity, and the model's learning capacity. The learning rate was determined by the step size upon which the weight update of each ensemble tree was based. Maximum tree depth controlled the complexity of individual decision trees and governed their capability to capture complex, intricate patterns within the soil-machine datasets. The strength of $L2$ regularization was adopted to determine the penalties applicable to the weights of the tree leaves to prevent overfitting for enhanced model generalization. Table 2 presents details of the various algorithms and the newly developed hybrid CatBoost models optimized in this study. All models were trained on $H_{Db}$ experimental data obtained from tillage datasets, and their predictions were compared using statistical accuracy and robustness metrics.

**Table 2** Base-CatBoost learner, nature-inspired algorithms, and the developed hybrid CatBoost models

| Algorithm | Year developed | Reference | New hybrid models developed in this study |
|---|---|---|---|
| Base-CatBoost | 2017 | Prokhorenkova et al. [20] | - |
| Aurora Intelligent Metaheuristic (AIM) | 2025 | Fakhouri et al. [21] | AIM- CatBoost |
| Artificial lemming algorithm (ALA) | 2025 | Xiao et al. [22] | ALA-CatBoost |
| Bobcat Optimization Algorithm (BOA) | 2024 | Benmamoun et al. [23] | BOA- CatBoost |
| Bacteria phototaxis optimizer (BPO) | 2023 | Pan et al. [24] | BPO-CatBoost |
| Chicken Swarm Optimization (CSO) | 2014 | Meng et al. [25] | CSO- CatBoost |
| Enzyme Action Optimizer (EAO) | 2025 | Rodan et al. [26] | EAO- CatBoost |
| Grizzly Bear Fat Increase Optimizer (GBFIO) | 2025 | Dehghani et al. [27] | GBFIO- CatBoost |
| Gazelle optimization algorithm (GOA) | 2023 | Agushaka et al. [28] | GOA-CatBoost |
| Grey Wolf Optimizer (GWO) | 2014 | Mirjalili et al. [29] | GWO-CatBoost |
| Harris hawks optimization (HHO) | 2019 | Heidari et al. [30] | HHO-CatBoost |
| Hippopotamus Optimization Algorithm (HO) | 2024 | Amiri et al. [31] | HO- CatBoost |

| | | | |
|---|---|---|---|
| Jellyfish Search Optimizer (JSO) | 2021 | Chou & Truong [32] | JSO- CatBoost |
| Mantis Shrimp Optimization Algorithm (MShOA) | 2025 | Sánchez Cortez et al. [33] | MShOA- CatBoost |
| Osprey Optimization Algorithm (OOA) | 2023 | Dehghani & Trojovský [34] | OOA- CatBoost |
| Parrot Optimizer (PO) | 2024 | Lian et al. [35] | PO-CatBoost |
| Pufferfish Optimization Algorithm (POA) | 2024 | Al-Baik et al. [36] | POA- CatBoost |
| reindeer cyclone optimization algorithm (RCOA) | 2025 | Chaudhary & Rawal [37] | RCOA-CatBoost |
| Secretary bird optimization algorithm (SBOA) | 2024 | Fu et al. [38] | SBOA-CatBoost |
| Swift Flight Optimizer (SFO) | 2025 | Kareem et al. [39] | SFO- CatBoost |
| Snake optimizer algorithm (SO) | 2022 | Hashim & Hussien [40] | SO-CatBoost |
| Walrus Optimization (WAOA) | 2023 | Trojovský & Dehghani [41] | WaOA-CatBoost |
| Warthog Optimization Algorithm (WartOA) | 2025 | Das et al. [42] | WartOA- CatBoost |
| Wolf-Bird Optimizer (WBO) | 2025 | Azizi et al. [43] | WBO-CatBoost |
| Whale Optimization Algorithm (WOA) | 2016 | Mirjalili & Lewis [44] | WOA- CatBoost |
| White Shark Optimizer (WSO) | 2022 | Braik et al. [45] | WSO-CatBoost |

## 2.4 Performance evaluation of hybrid CatBoost models

### *2.4.1 Statistical accuracy /error evaluation metrics*

Various statistical error metrics were adopted to evaluate the accuracy of hybrid CatBoost ensemble models in predicting drawbar pull, as shown in Eqs. 10-23 [46, 47, 48]. The coefficient of multiple correlation (R) and the coefficient of determination (R2) were used to assess the accuracy and strength of the relationships between the predicted and observed values for both the hybrid models and the base learner, using Eqs. 10 and 11. $R^2$adjusted modifies ($R^2$) to account for the number of predictors by penalizing the inclusion of irrelevant or redundant variables, making it more reliable for comparing hybrid machine learning models (Eq. 12). Further, explained variance score (EVS) was adopted to evaluate how best the hybrid models accounted for the variability in the actual $H_{Db}$ and assessed the extent to which the model's captured variance in the predicted $H_{Db}$ as shown in Eq. 13. Relative absolute error (RAE) was used to establish relative magnitude of absolute prediction errors of the hybrid models compared to the Base-CatBoost learner. In contrast, the Relative Root Mean Square Error (RRMSE) was used to quantify and evaluate the relative prediction accuracy of the hybrid models compared with the base learner (Eq. 14-15). Root Mean Square Relative Error (RMSRE) was adopted as a performance measure to penalize larger errors and determine the average magnitude of prediction errors relative to the true values and is an important metric in machine learning and regression modeling in agricultural engineering applications, especially where scale-independent evaluation is considered (Eq. 16). Normalized root mean square error (NRMSE)

was adopted as a standardized percentage metric for comparing $H_{Db}$ predictions accuracy of the models across training and testing datasets by normalizing the RMSE using the range, mean, or standard deviation (σ) of the actual $H_{Db}$ as shown in Eq. 17. The weighted mean absolute relative percentage Error (WMAPE) was used to assess how large the $H_{Db}$ prediction errors were relative to the actual $H_{Db}$ values (Eq. 18) while Variance accounted for (VAF) was adopted to quantify the proportion of variance in the actual $H_{Db}$ that is explicative in the model predictions, thereby indicating the model's accuracy in capturing the underlying dataset variabilities (Eq. 19). The coefficient of variation (CV) as shown in Eq. 20, was adopted as a statistical measure of relative variability of the precision, stability, and consistency of model predictions and performance across different datasets and simulation runs, as opposed to just the absolute error. The Scatter Index (SI) was adopted to normalize the RMSE by the mean of observed $H_{Db}$, providing a scale-independent performance measure that facilitated consistency in relative performance comparison of the models, across the training and testing datasets (Eq. 21). Further, Performance Index (PI) merges error evaluation metrics, $R^2$ and RMSE and constitutes a balanced ratio of combined modeling performance allowing comprehensive comparison of various models directly. Rather than a single error component, PI, as shown in Eq. 22, integrates several metrics to evaluate the overall quality, accuracy, and efficiency of machine learning models. Hybrid models with larger PI values exhibit greater accuracy, reliability, and overall predictive performance [49, 50]. Also, prediction accuracy (PA) is a statistical measure of data mining and machine learning models adopted in this study to assess how well the model would forecast outcomes with new, unseen or future data (Eq. 23). All the statistical metrics allowed for evaluation and comparison of hybrid Cat Boost models based on accuracy, efficiency, and performance robustness and superiority providing a decisive scientific basis for adopting one hybrid model over others.

$$\mathrm{R} = \frac{\sum_{i=1}^{n}(\mathrm{H_{DbP}} - \overline{\mathrm{H}}_{\mathrm{DbP}})(\mathrm{H_{DbA}} - \overline{\mathrm{H}}_{\mathrm{DbA}})}{\sqrt{\sum_{i=1}^{n}(\mathrm{H_{DbP}} - \overline{\mathrm{H}}_{\mathrm{DbP}})^2 \mathrm{x} \sum_{i=1}^{n}(\mathrm{H_{DbA}} - \overline{\mathrm{H}}_{\mathrm{DbA}})^2}} \quad (10)$$

$$R^2 = \frac{\sum_{i=1}^{n}(H_{DbP} - H_{DbA})^2}{\sum_{i=1}^{n}(H_{DbP} - \overline{H}_{DbA})^2} \quad (11)$$

$$Adjusted\ R^2 = 1 - \frac{(1-R^2)(s-1)}{S-F-1} \quad (12)$$

$$EVS = 1 - \frac{Var(H_{DbA} - H_{DbP})}{Var(H_{DbA})} \quad (13)$$

$$RAE = \frac{\sum_{i=1}^{n}|H_{DbP} - H_{DbA}|}{\sum_{i=1}^{n}|HDb_A - \overline{H}_{DbA}|} \quad (14)$$

$$\mathrm{RRMSE} = \frac{1}{|\overline{\mathrm{H}}_{\mathrm{DbP}}|}\sqrt{\frac{\sum_{i=1}^{n}(\mathrm{H_{DbA}} - \mathrm{H_{DbP}})}{\mathrm{n}}} \quad (15)$$

$$RMSRE = \sqrt{\frac{1}{n}\sum\left(\frac{H_{DbA}-H_{DbP}}{H_{DbA}}\right)^2} \quad (16)$$

$$\mathrm{NRMSE} = \frac{RMSE}{MaxH_{DbP}-MinH_{DbP}} \quad (17)$$

$$WMAPE\ (\%) = \frac{\sum_{i=1}^{m}|H_{DbA}-H_{DbP}|}{\sum_{i=1}^{m}|H_{DbA}|}x100 \quad (18)$$

$$VAF = \left(1-\frac{var(H_{DbA}-H_{DbP})}{var(H_{DbA})}\right)x100 \quad (19)$$

$$\mathrm{CV} = \frac{\sqrt{\sum_{i=1}^{n}\left(\frac{\left(H_{Dbp}-\overline{H}_{Dbp}\right)^2}{n-1}\right)}}{\left(\frac{\sum_{i=1}^{n}H_{DbA}}{n}\right)} \text{ or } \mathrm{CV} = \frac{\mathrm{RMSE}}{\mathrm{Mean}}\mathrm{x}100 \quad (20)$$

$$SI = \frac{RMSE}{\overline{H}_{DbA}} \quad (21)$$

$$\mathrm{PI} = \frac{\mathrm{RRMSE}}{1+\mathrm{R}} \text{ or } \mathrm{PI} = \frac{R^2}{\mathrm{RMSE}} \quad (22)$$

$$\mathrm{PA} = \left[1-\left(\frac{1}{n}\sum_{i=1}^{n}\frac{|H_{DbA}-H_{DbP}|}{H_{DbA}}\right)\right]\mathrm{x}100 \quad (23)$$

where $n$ is the number of $H_{Db}$ variables, $H_{DbP}$, and $\overline{H}_{Dbp}$ are the predicted $H_{Db}$ and its corresponding mean, while $H_{DbA}$ and $\overline{H}_{dbA}$ represent the actual experimental and mean of actual $H_{Db}$, respectively. $S$ and $F$ are the numbers of samples and features, respectively. The ideal values and accuracy metrics criteria are shown in Table 3.

**Table 3** Performance evaluation criterion of accuracy metrics

| Accuracy metrics | Ideal criteria/Perfect prediction value |
|---|---|
| R, $R^2$, $R^2$adjusted, EVS, | 1 |
| RAE, RRMSE, RMSRE, NRMSE, SI, PI, WMAPE | 0 |
| VAF, PA | 100 |
| CV | <10% (Excellent consistency/Low variation), 10-20% (Good/Acceptable/Moderate variation), 20-30% (Acceptable in some fields), |

### *2.4.2 Robustness metrics*

Robust regression loss functions were implemented to establish a hybrid CatBoost model that converges fast and has reliable resistance to outliers in handling $H_{Db}$ predictions. These included Mean Squared Error (MSE) and Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Median Absolute Error (MedAE), Huber loss ($L_\delta$), and Log-Cosh loss (LCL) as shown in Eqs 24-29. The MSE and RMSE were essential for assessing the quality of predictions, evaluating overall prediction accuracy, and measuring global goodness-of-fit of the models (Eq. 24-25). Further, MAE was adopted to capture absolute errors, quantify the average magnitude of prediction errors and evaluate overall prediction accuracy (Eq. 26). At the same time, MedAE

(Eq. 27), was adopted to measure the median of all absolute differences between predicted $H_{Db}$ and actual values making it a robust model evaluation metric against outliers and is used as the "gold standard" for post-training validation of models in noisy environments like soil-machine analysis due to its high invariant to extreme values [51]. Further, the Huber loss ($L_\delta$) function was adopted as a robust regression loss function that combines the sensitivity of MSE to small drawbar load residuals with the resilience of MAE to extreme values, making it particularly suitable for evaluating prediction robustness of CatBoost models, as shown in Eq. 28 [52]. Similarly, the Log-Cosh loss (LCL) approximated MSE for small residuals and behaved like MAE for large residuals, demonstrating robustness to extreme deviations as shown in Eq. 29 [53].

$$MSE = \sum_{i=1}^{n}(H_{DbA} - H_{DbP})^2 \quad (24)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(H_{DbP} - H_{DbA})^2}{n}} \quad (25)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|H_{DbP} - H_{DbA}| \quad (26)$$

$$MedAE = Median(|H_{DbA} - H_{DbP}|) \quad (27)$$

$$L_\delta(H_{dbA}, H_{dbP}) = \begin{cases} \frac{1}{2}(H_{DbA}, H_{DbP})^2, & for \quad |H_{DbA}, H_{DbP}| \le \delta \\ \delta \cdot \left(|H_{DbA}, H_{DbP}| - \frac{1}{2}\delta\right), for\ |H_{DbA}, H_{DbP}| \not\le \delta \end{cases} \quad (28)$$

$$LCL(H_{DbA}, H_{DbP}) = \frac{1}{2}\sum_{i=1}^{n} log\big(cosh(H_{DbP} - H_{DbA})\big) \quad (29)$$

where $n$ is the number of $H_{Db}$ variables, $H_{DbP}$, and $\bar{H}_{Dbp}$ are the predicted $H_{Db}$ and its corresponding mean, while $H_{DbA}$ and $\bar{H}_{DbA}$ represent the actual experimental and mean of actual $H_{Db}$, respectively. $S$ and $F$ are the number of samples and features, respectively. Small and large error boundaries are defined by the positive hyperparameter value (δ), while *log (cosh)* is the natural logarithm of hyperbolic cosine. The ideal values of MSE, RMSE, MAE, and medAE are zero [48], while low values of $L_\delta$ and LCL indicate minimal difference between predictions and actual $H_{Db}$ values, demonstrating that the model has effectively mitigated noise and outliers, and accurate predictions are preserved [54].

### *2.4.3 Score ranking and multicriteria selection of the best model*

Whenever the single best machine learning model cannot be conclusively established as the most superior model based on a set of statistical evaluation metrics, performance rank score analysis can be adopted as a decision-support approach. This method facilitates robust model selection in multi-criteria optimization problems, especially in heterogeneous and nonlinear systems such as soil–machine interaction modelling. In this study, each model was ranked across multiple statistical evaluation criteria (both accuracy and robustness statistical metrics) during training and testing phases. The aggregate ranking scores were then computed to obtain

a composite measure of overall performance for each model, enabling a more balanced assessment by accounting for trade-offs between competing performance metrics.

### *2.4.4 Modelling regression error-characteristic analysis*

Regression error characteristic curve (REC) analysis was adopted to evaluate diagnostic accuracy and robustness of the 26 models by characterizing the residual error distributions of the models. REC curves were plotted with error tolerance (ε) on the x-axis against the corresponding percentages of predictions within the tolerance on the y-axis. The resulting curve estimates the cumulative distribution function of the error as the proportion of predicted $H_{Db}$ whose absolute errors fall within the specified tolerance, as expressed using Eq. 30.

$$\varepsilon = \frac{1}{n}\sum_{1}^{n} 1\left(|H_{DbA} - H_{DbP}| \leq \varepsilon\right) \tag{30}$$

where ε is the error tolerance accuracy, n is the number of instances, $H_{DbP}$, and $H_{DbA}$ are the predicted $H_{Db}$ and the actual experimental $H_{Db}$, respectively. By varying ε from 0 to the maximum admissible error, the $H_{Db}$ REC curves capture the cumulative distribution function of absolute errors, thereby providing a complete depiction of model performance across different precision requirements and metrics. A key metric used to evaluate models in this study is the area over the REC curve (AOC) to quantify the prediction error, given by Eq. 31 [55].

$$AOC = \int_{0}^{\varepsilon_{max}} [1 - Accuracy(\varepsilon)] d\varepsilon \tag{31}$$

Lower AOC values indicate superior model performance, while an ideal model exhibits a steep REC curve that rapidly approaches unity accuracy at minimal error tolerance, corresponding to near-zero AOC. The technique is useful for high-level modelling, for rigorous comparison of competing models to support error-tolerance-based decision-making, and for providing an interpretable linkage between application-specific performance and statistical accuracy constraints. An ideal model should exhibit a very small area over the curve (AOC = 1-AUC) and a high area under the curve (AUC), with its error curve achieving high accuracy levels (y-axis) while remaining steep and approaching parallelism with the y-axis, at minimal error tolerance (x-axis), thus achieving maximum predictive power [56, 57].

### *2.4.5 Convergence generalization evaluation metrics*

Assessing generalization and identifying performance overfitting of hybrid models is critical, as they occur whenever there is excellent performance on the training data but underperform on unseen and new data. This is because the learning algorithm is exposed to noise and random and irrelevant patterns present in the new datasets, resulting in poor predictions. To assess model generalization ability and detect overfitting, two metrics were used, including generalization gap (G-gap) and perturbation stability score (PS). G-gap evaluates the difference

between the accuracy on the training versus test datasets and quantifies overfitting of machine learning models by measuring the performance deviation between the training and unseen data, as shown in Eq. 32 [58]. The perturbation score evaluates how slight changes in the training data affected predictions, whereby the most stable and robust models do not vary their outputs upon ingress of noise in the input data, as shown in Eq. 33 [59].

$$G_{gap} = M_{train} - M_{test} \tag{32}$$

$$PS = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{H}_{Dbi} - \widehat{H}_{Db}^{P}\right| \tag{33}$$

Where $M_{train}$ and $M_{test}$ represent the evaluation metric, such as MSE or $R^2$, while $\widehat{H}_{Dbi}$ and $\widehat{H}_{Db}^{P}$ are the original prediction and the prediction upon perturbation. High PS indicates high sensitivity to small input changes, suggesting poor robustness and the ingress of overfitting noise, while the lowest G-gap indicates better generalization ability without overfitting noise [60].

### *2.4.6 Taylor Analysis*

The Taylor diagram tool was used to graphically quantify, evaluate, and visually compare model performance by simultaneously integrating three statistical metrics, including the correlation coefficient (R), the standard deviation (σ), and the centered root mean square error (cRMSE) between model predictions and observed data. The diagram is based on the law of cosines, which links these three statistics via Eq. 34 [61]. In the diagram, the radial distance from the origin corresponds to the model's standard deviation (σ), the angular position with respect to the x-axis represents R, while the distance between the model point and the reference point (observations) indicates the centered RMSE. Plotting multiple models on the same Taylor diagram enables rapid visual assessment of how well each model reproduces the observed variability, pattern similarity, and overall accuracy, thereby facilitating comparative performance evaluation and uncertainty analysis. In a Taylor diagram, a model that is positioned closer to the reference point (usually on the x-axis at a distance equal to the σ of the observed data) is viewed as more precise, as it shows reduced prediction error, stronger correlation, and better overall agreement with observed data [62]

$$cRMSE^2 = \sigma_P^2 + \sigma_A^2 - 2\sigma_P\sigma_A R \tag{34}$$

Where *cRMSE* is the centered RMS difference, $\sigma_P$, $\sigma_A$ are the standard deviations of the model predictions and actual experimental observations, while R is the correlation coefficient.

## 2.5 SHAP analysis

Machine learning Shapley Additive exPlanations (SHAP) was deployed to assess the complex drawbar load-generating behavior and to obtain insights into the dynamic impact of variables on the predictions by the most effective model. Further, the SHAP method was adopted to establish

how model predictions were produced from specific input variables and to understand which variables (features) were driving the results either positively or negatively and by what magnitude. For each feature, SHAP values corresponding to the variable effect on the prediction were computed, and the sum of SHAP values for all the features of a particular prediction was computed to account for the difference between the predicted value and the mean baseline values of the entire dataset. SHAP summary provided insights into the behavior of the EAO-CatBoost model, showing the most important features globally, local prediction drivers, their impact, and the distribution of their effects. Positive and negative feature effects on the $H_{Db}$ are denoted by red and blue points, respectively [63]. The flowchart for the entire study is shown in Fig. 1.

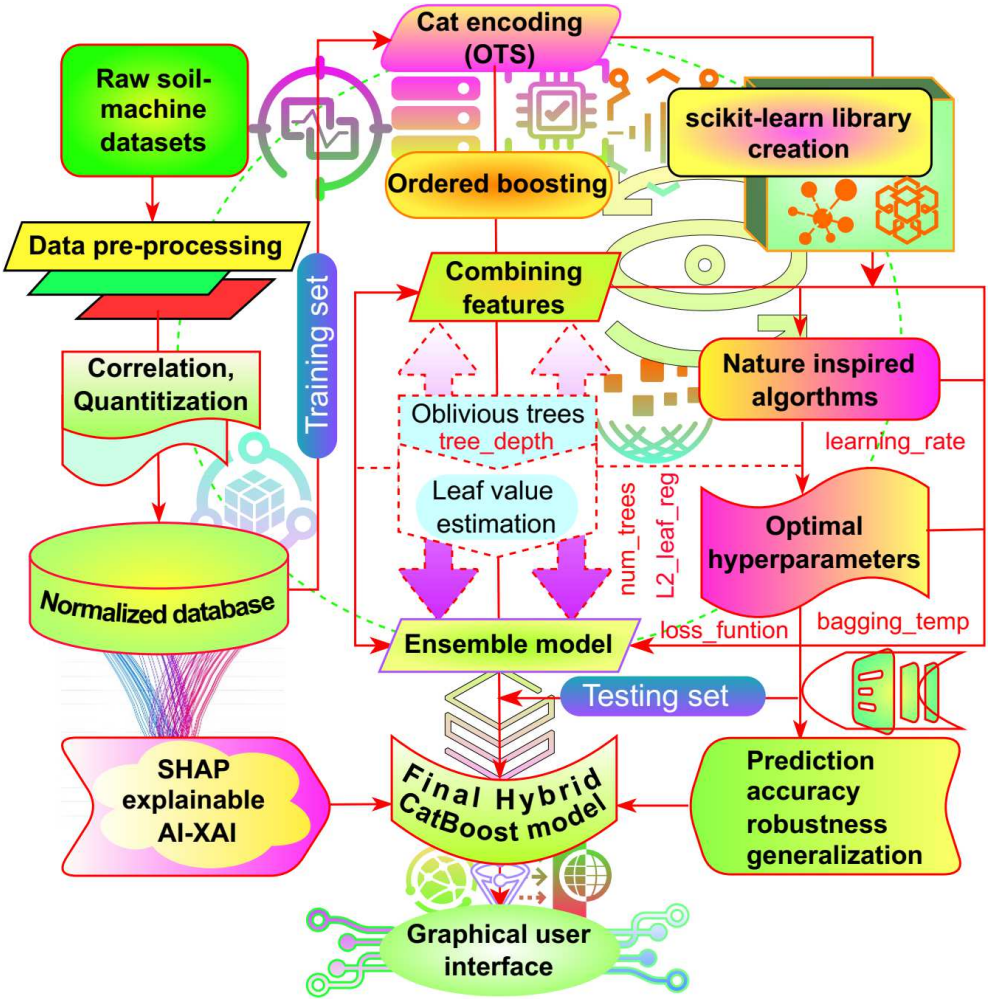


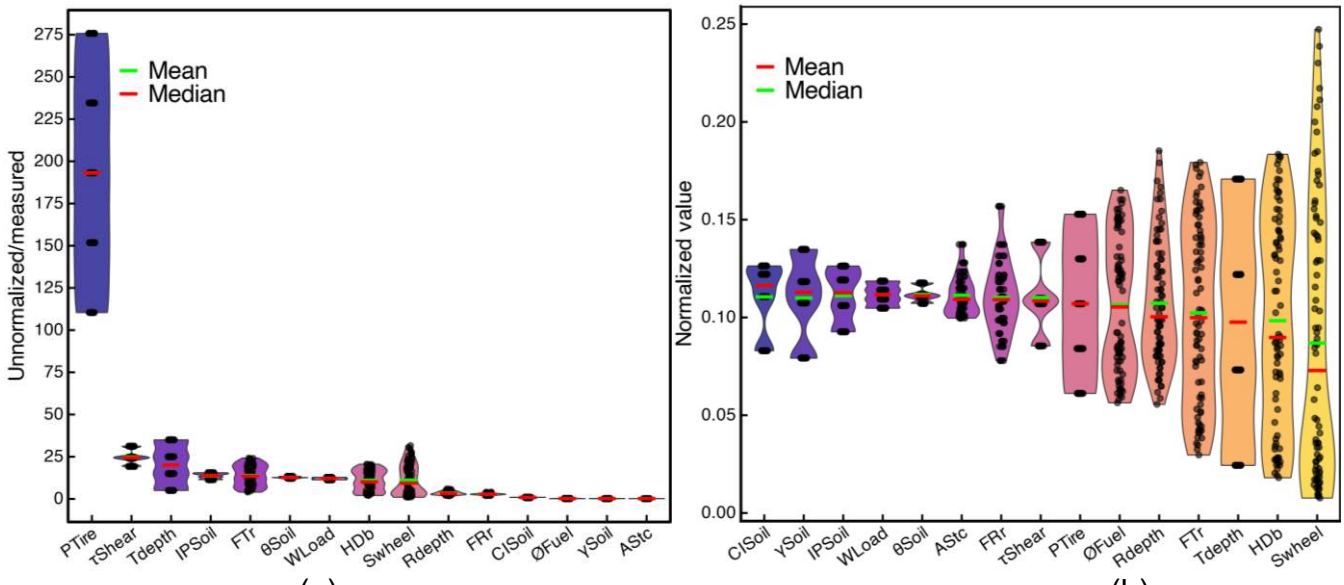

**Fig. 1.** Hybrid CatBoost model development and evaluation framework for the study

## 3. Results and discussions

### 3.1 Database normalization

A normalized database combining the features of a box plot and a kernel density plot using violin plots is shown in Fig. 2, allowing visualization of the median, quartiles, and the full distribution

shape for a comprehensive comparison with non-normalized data. The violin plot shows the dataset with raw data values in their original scale, with the y-axis showing the raw absolute values, and the median indicating the central values. In contrast, the violin density shape indicates how the values are distributed and where the raw measurements are concentrated, focusing on the scale for magnitude comparison. The differences appear large because of scale. Further, Fig. 2(b) shows the comparison distribution of the datasets after database normalization. All the data were normalized to a new range of (0, 1) to reduce dispersion among variables and improve the convergence of CatBoost models. The y-axis represents standardized values (unlike raw units), and the distributions from different groups become more directly comparable, while shape differences become clearer, focusing on patterns in their distributions. Further, the violins are aligned around a common centre, and the median lines are similar across samples. The normalized database thus enabled shapes to be more comparable, and the differences now reflect true experimental variations rather than scale effects. Since the input variables were intrinsically heterogeneous in units and scales, normalizing the entire database improved data quality for model training and prediction comparison.




(a) (b)

**Fig. 2.** Violin plot showing the distribution of multiple variables in the database before and after normalization for (a) magnitude comparison of raw non-normalized data and (b) distribution patterns of the normalized database

### 3.2 Data visualization in Pearson's and Partial Mantel test analysis

The normalized database showed considerable Pearson and partial Mantel correlations and multicollinearity ($r > 0.8$). Fig. 3 shows a comprehensive visual display of relationships and correlation coefficients among the modelling variables, providing a more in-depth evaluation and understanding of data point distribution patterns. There was a strong positive correlation

between $H_{Db}$ with $F_{Tr}$ (1.00), $T_{depth}$ (0.96), $Ip_{Soil}$ (0.96), $Ø_{Fuel}$ (0.93), and $S_{wheel}$ (0.81). Further, the total $H_{Db}$ generated at the drawbar was significantly affected by the tractive thrust force ($F_{Tr}$) generated by the tractive devices (dependent on wheel load), as greater $T_{depth}$ increased specific soil resistance, resulting in higher ground reaction forces, which increased the $S_{wheel}$ and $Ø_{Fuel}$ at higher $H_{Db}$. Pearson correlation matrix further revealed interdependence among soil physical parameters, including $CI_{Soil}$, $\theta_{Soil}$, $\gamma_{Soil}$, and $I_{PSoil}$, reflecting the inherent multicollinearity among soil strength and consistency indicators. There was a strong positive correlation between $T_{depth}$ and $Ip_{Soil}$ (0.99), $Ø_{Fuel}$ and $T_{depth}$ (0.98), $Ip_{Soil}$ and $Ø_{Fuel}$ (0.97), ØFuel and $S_{wheel}$ (0.93), $W_{load}$ and $F_{Rr}$ (0.76), and a negative correlation between $\theta_{Soil}$ and $\gamma_{Soil}$ (-1.00). This concludes that the variables will affect $H_{Db}$, either directly or indirectly, at various levels of significance, and agrees with soil–tire interaction models that highlight the coupled nature of soil deformation and strength indices in predicting traction outcomes [64,65].

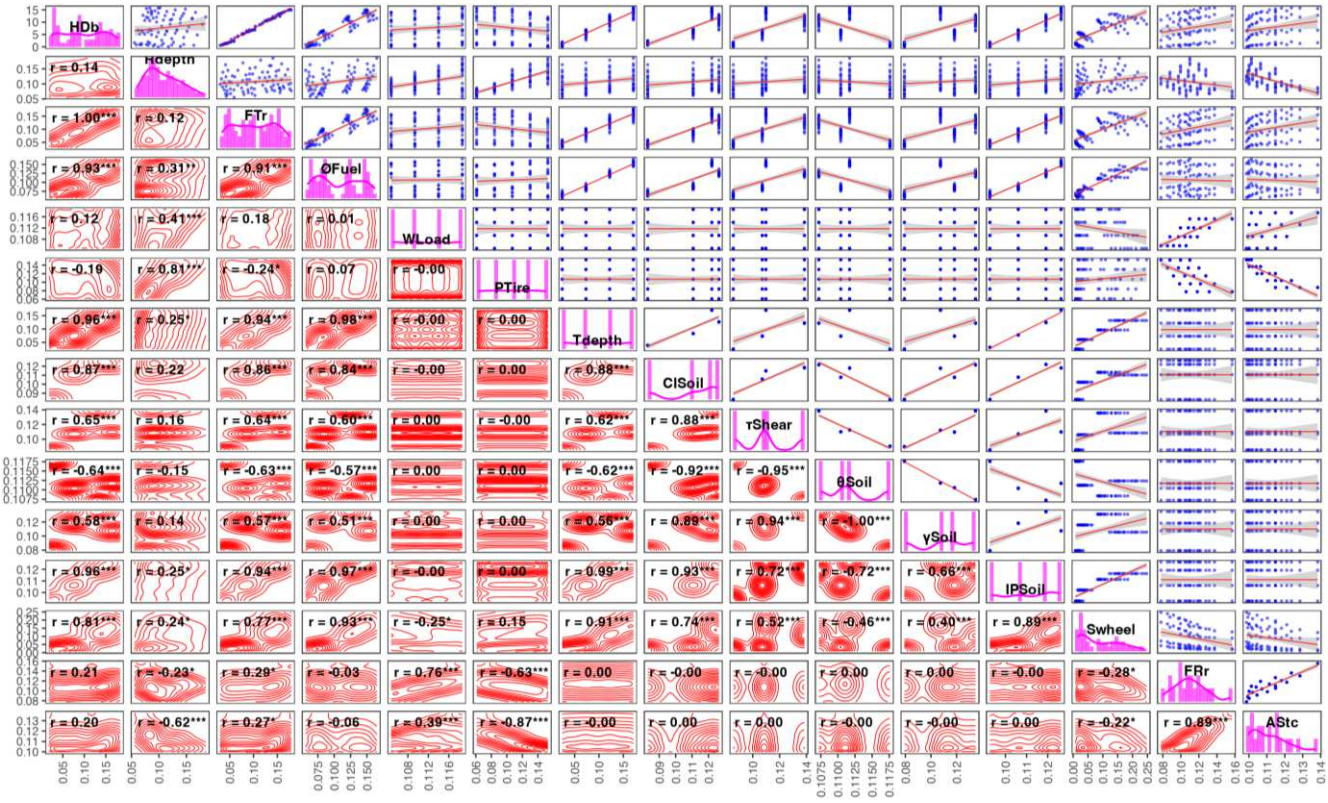



**Fig. 3.** Correlation among drawbar load prediction variables for CatBoost models. *denotes a statistically significant correlation ($p$<0.05), and the number of ∗ signs denotes the level of significance, i.e. *, ** and *** indicates $p$<0.05 at the 0.05 level, ($p$ <0.01) at the 0.01 level, and ($p$ <0.001) at the 0.001 level, respectively. The circular and elliptical contours with different numbers on the lines reflect the magnitude and direction of the correlations among the paired drawbar prediction variables. The confidence bounds of regression estimates are shown by the shaded

area around regressed lines of the scatter plot, indicating the variability of estimated regression scatter and the likely range of true regression lines.

Partial Mantel test revealed highly significant correlations ($r \geq 0.4, p < 0.01$) of $H_{Db}$ with $F_{Tr}$, $T_{depth}$, $\tau_{Shear}$, and $S_{wheel}$ as shown in Fig. 4. These findings are consistent with traction studies and advances in terramechanics that identify soil mechanical resistance and soil-wheel interface dynamics as the dominant factors governing traction and drawbar performance in off-road agricultural field operations [66, 67, 68, 69]. In contrast, $W_{Load}$ and $P_{Tire}$ displayed weak or statistically insignificant relationships ($p \geq 0.05$), suggesting minimal influence on $H_{Db}$ variations within the operational range of high-power range tractors tested. These findings emphasize that drawbar pull optimization under field conditions is predominantly determined by soil deformation behavior and soil-wheel contact mechanics**,** rather than by adjustments in load distribution or tire inflation pressure. This underscores the importance of observing soil strength and contact characteristics to enhance power transmission to the drawbar and minimize energy losses during field operations. Similar observations have been reported in experimental investigations, where inflation pressure and load adjustments contributed less to traction efficiency compared to soil strength and slip–sinkage effects [65, 69, 68,70]

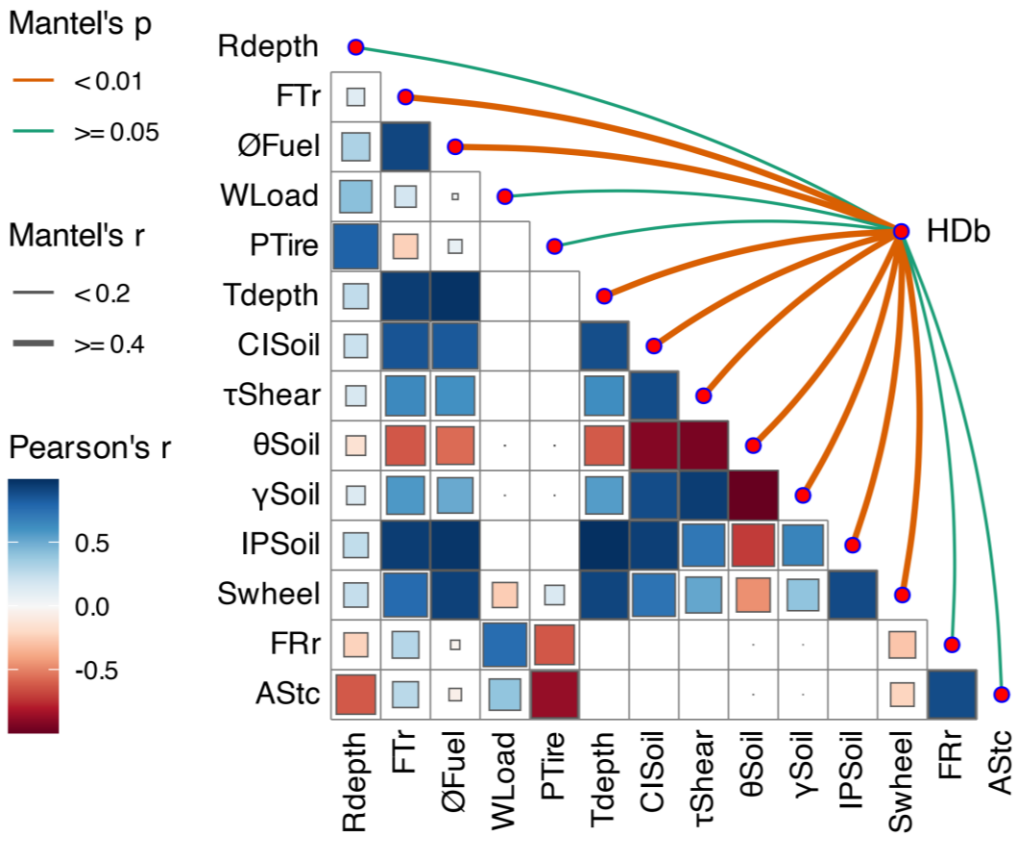



**Fig. 4.** Mantel test matrix illustrating and quantifying the correlations between $H_{Db}$ and dependent variables during tillage. Line thickness indicates the magnitude and strength of the correlation,

as well as the absolute value of Mantel's *r* statistic. While line colour reveals levels of statistical significance of the association and if the relationship is reliable (p-value), with orange (P<0.01) and green (P≥0.01)

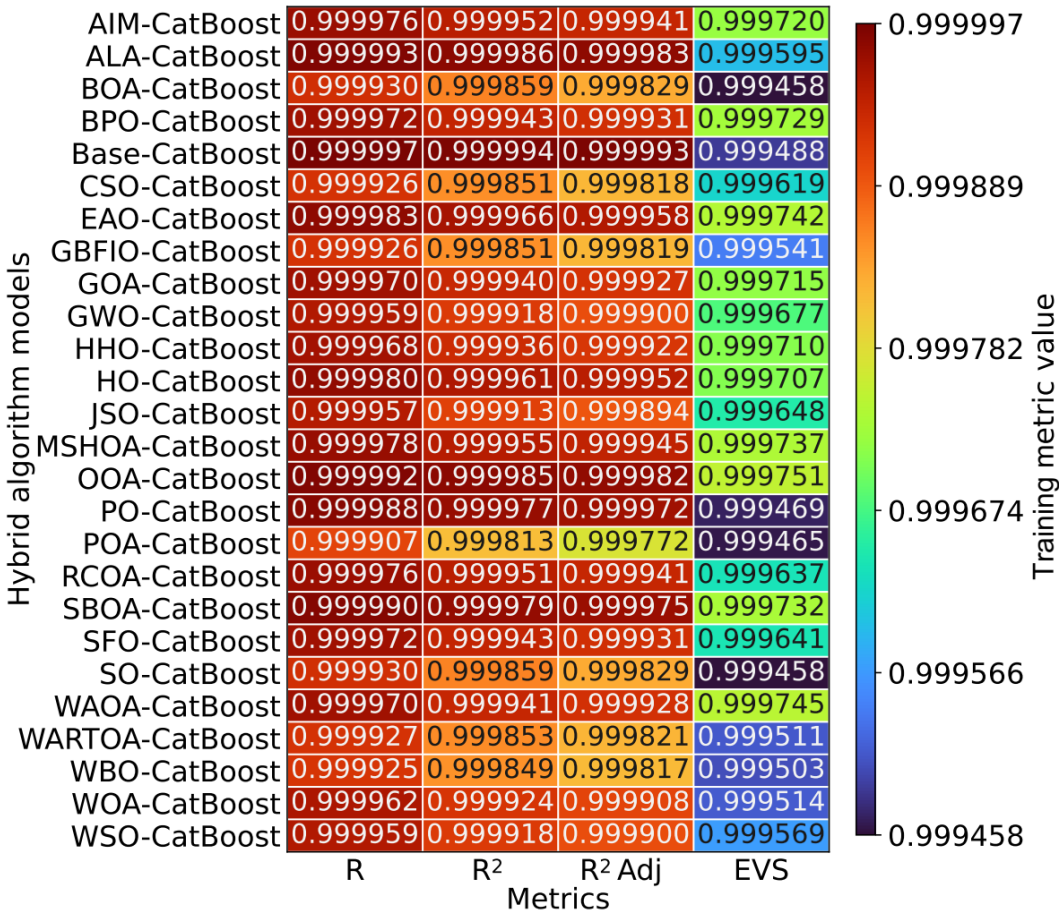
### 3.3 Cognitive performance and selection of the most effective models

After modelling the $H_{Db}$ with all the hybrid Cat Boost models, an important step was to identify the most effective model based on a robust criterion combining all evaluation metrics.

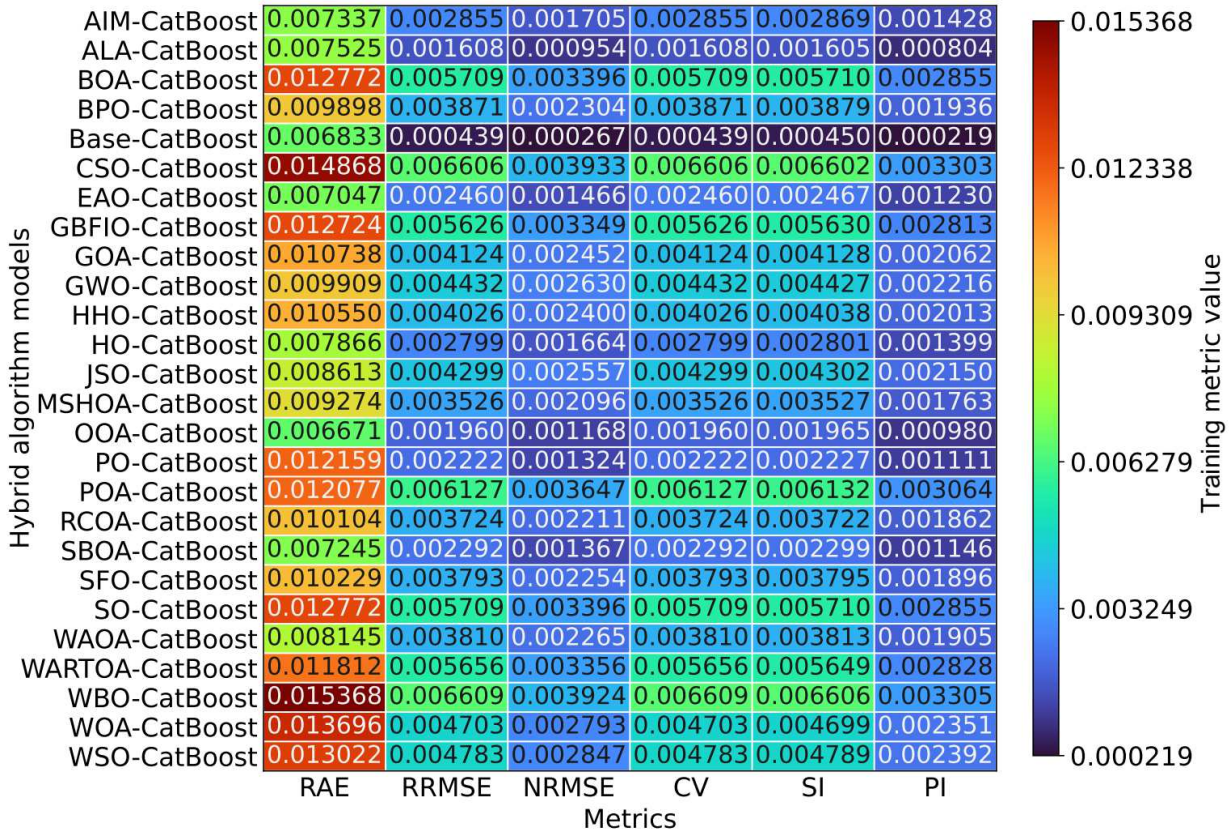
#### *3.3.1 Error /Accuracy evaluation matrix*

An accuracy matrix was adopted to facilitate a comprehensive assessment of accuracy levels achieved by all the models and their comparison with ideal metrics during training and testing. During training, the Base-CatBoost learner achieved the best metrics of R, $R^2$, $R^2$adj of 0.999997, 0.999994, 0.999993, and the best RRMSE, NRMSE, CV, SI, PI of, 0.000439,0.000267,0.000439, 0.000450, and 0.000219, respectively, as shown in Fig. 5a-d. However, OOA achieved the best EVS (0.999751) and RAE (0.006671) during training against the Base-CatBoost (EVS, 0.999488 and RAE, 0.006833). During testing, the base learner performed poorly while EAO-CatBoost achieved the best values of $R^2$ (0.998361), $R^2$adj (0.998008), EVS (0.998363), RAE (0.035851), RRMSE (0.021915), NRMSE (0.013083), CV(0.021915), SI (0.021915), and PI (0.10962). However, SBOA Cat Boost achieved the best R (0.999214) against the EAO-Cat Boost (0.999197) as shown in Fig. 5c-d. At the same time Fig. 5 e-f shows that Base-CatBoost achieved the best values of WMAPE (0.03508%) during training, while EAO CatBoost achieved the best during testing (1.72517%). However, all the models had WMAPE values under 5% and were all considered highly accurate in terms of WMAPE. Nonetheless, Base-CatBoost performed poorly in terms of VAF (99.94879%) compared to EAO Cat Boost (99.97418%), but both models lagged behind OOA-CatBoost (99.97514%) and WAOA-Cat Boost (99.97446%). In comparison, both SO-CatBoost and BOA-Cat Boost (99.94582%) had the lowest values during training, as shown in Fig. 5g. However, during training, EAO CatBoost was the best model (VAF, 99.83626%) while the Base-CatBoost learner was the worst (VAF-99.63180%), as shown in Fig. 5h. Although all the models had 100% PA during training, they declined during testing with CSO-CatBoost attaining the best values (98.75%) followed by EAO-CatBoost (97.50%). In comparison, the Base-CatBoost (91.25%) had the lowest values as shown in Fig. 5 i-j. This indicates that single accuracy metrics such as R and PA do not necessarily reflect modelling robustness and versatility in machine learning, especially when applied to unseen datasets. It can be observed that selecting a single best model across all accuracy metrics was not entirely conclusive, especially during training, and

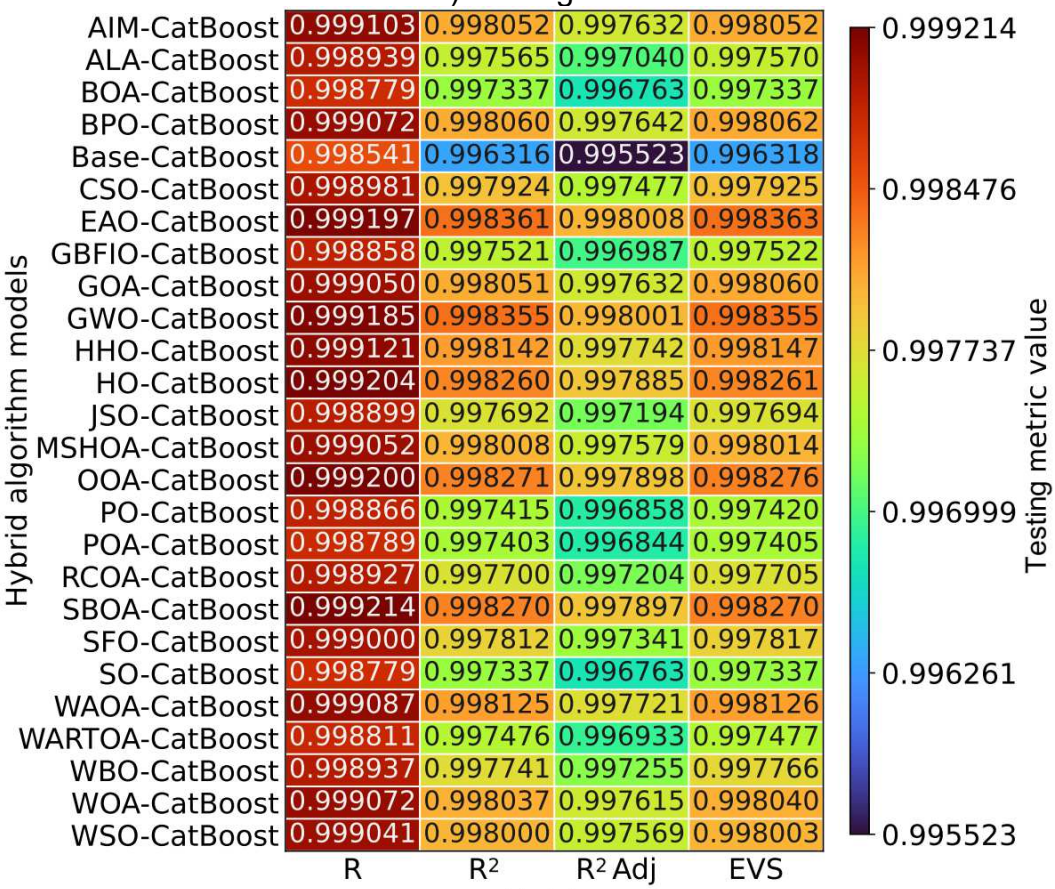
the models did not achieve the ideal zero or unity values. This could be due to limitations, such as high sensitivity to dataset noise, necessitating the adoption of robustness metrics to ensure the selection of a model that maintains consistent and reliable performance even when inputs are perturbed, noisy, or unexpected. While accuracy measures how often a model is correct on a specific dataset, it often fails to provide a complete picture of reliability, especially in real-world or high-stakes scenarios. This necessitated further evaluation of the models using robustness metrics. The models were thus evaluated for robustness to achieve a trade-off, as accuracy metrics alone were inconclusive and could mislead the choice of the best model.



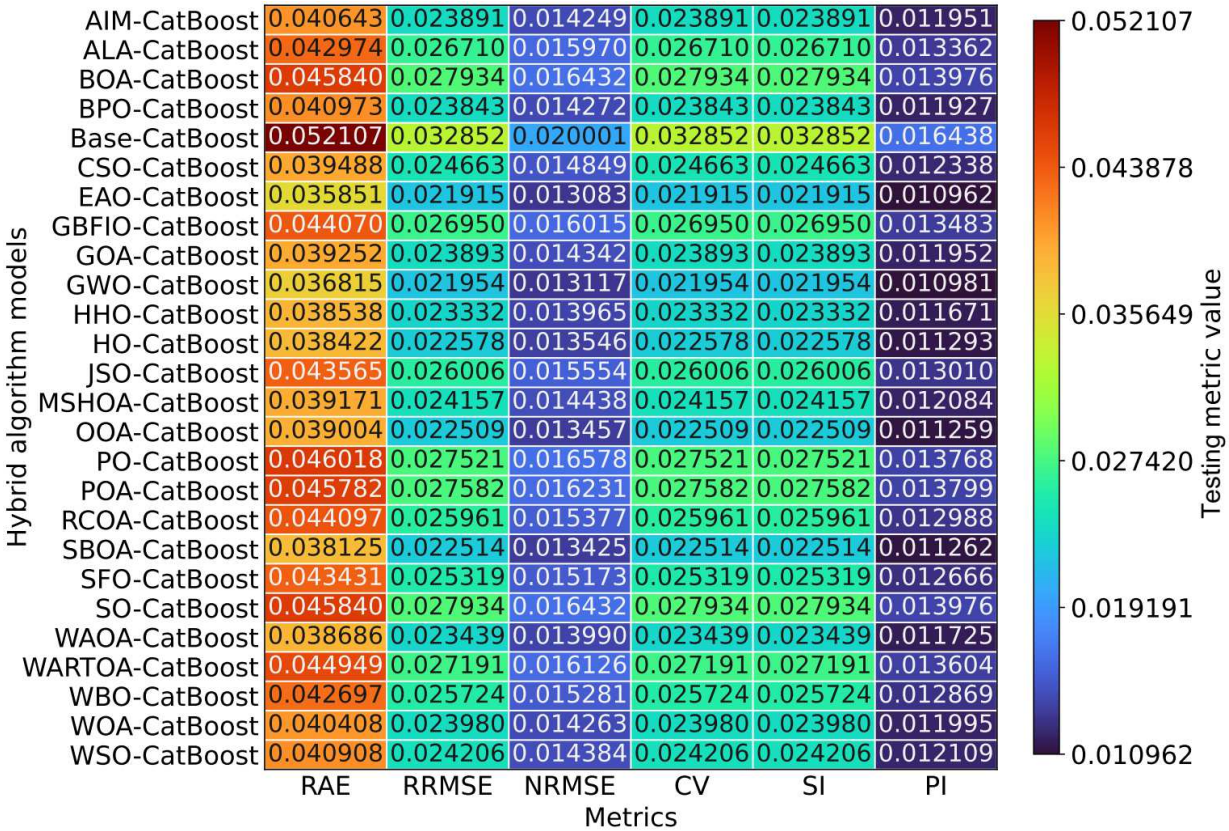

(a) training



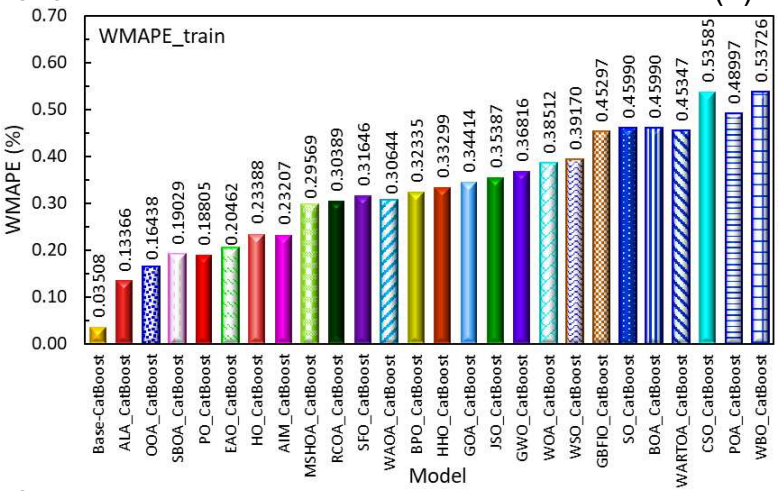

b) training

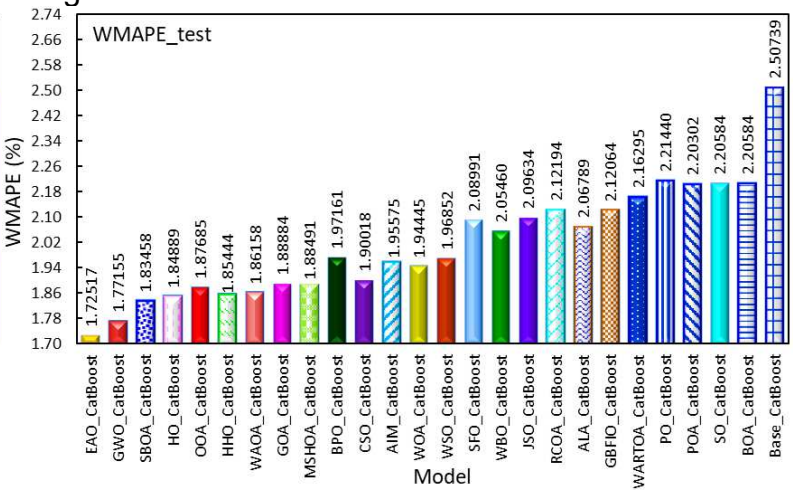



(c) testing

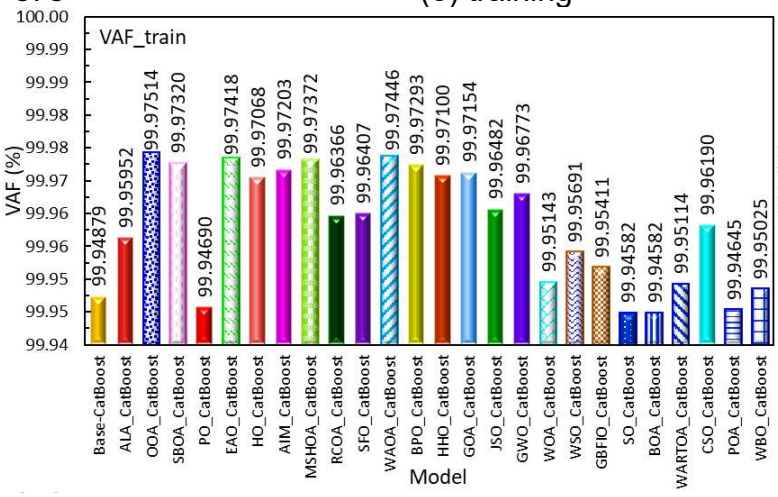


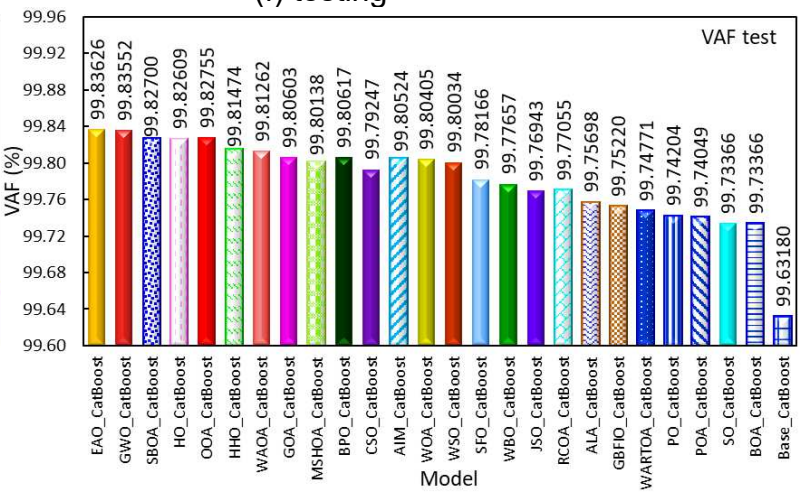

(d) testing



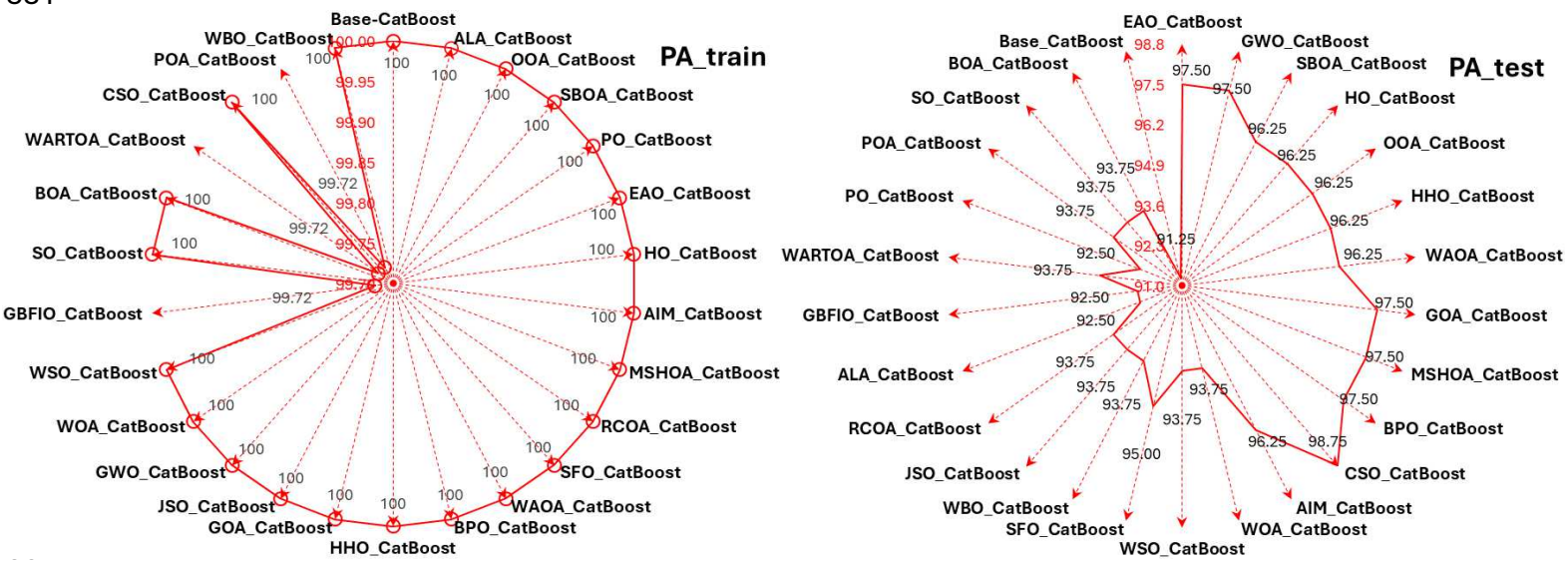




(e) training (f) testing



(g) training (h) testing

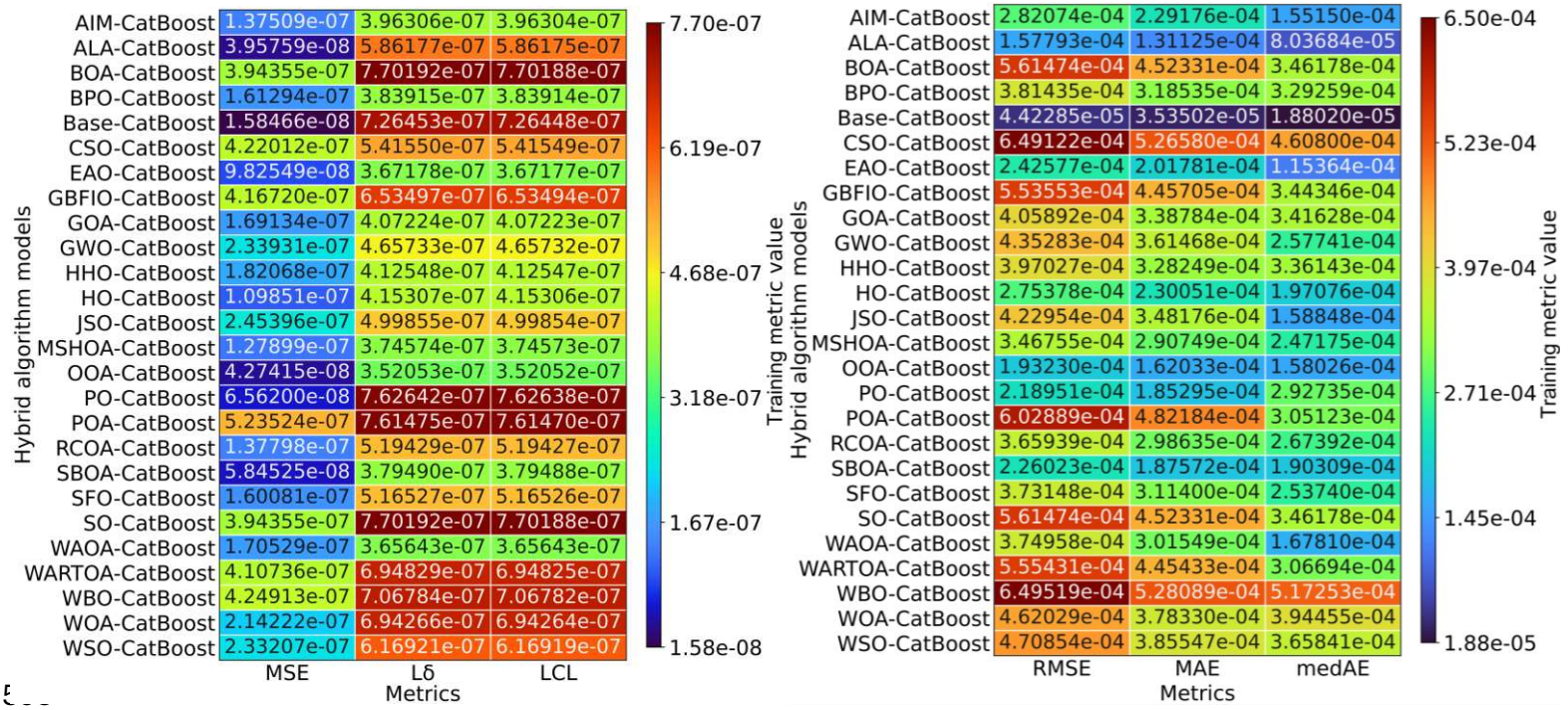



(i) training-prediction accuracy (j) testing-prediction accuracy

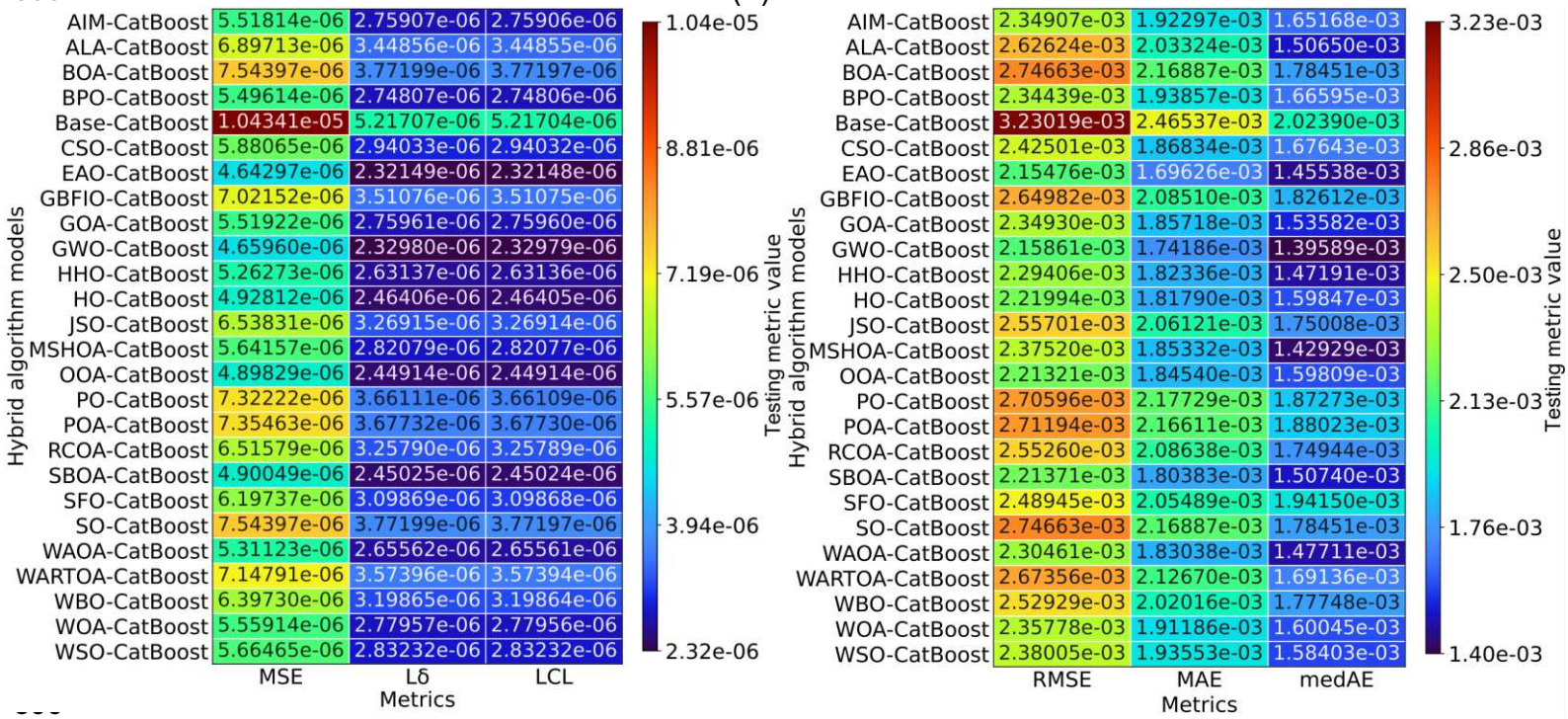
**Fig. 5.** Analysis of error/accuracy evaluation metrics of all the 25 models during training (a, c, e, g, i) and (b, d, f, h, j) testing phases.

### *3.3.2 Robustness metrics of hybrid models*

Fig. 6 shows the performance of all models across robustness metrics. During training the Base-CatBoost performed the best in terms of MSE (1.58466e-08), Lδ (7.26453e-07), LCL (7.26448e-07), RMSE (4.42285e-05), MAE (3.53502e-05), and MedAE (1.88020e-05). During testing, all 25 hybrid models performed better than the Base-CatBoost model across all evaluation metrics, thanks to enhanced robustness in hyperparameter tuning and bias-variance trade-off, resulting in improved stability and reduced error on unseen data. EAO-CatBoost was the best model in terms of MSE (4.64297e-06), Lδ (2.32149e-06), LCL (2.32148e-06), RMSE (2.15476e-03), MAE (1.69626e-03), but was third best in terms of MedAE (1.88020e-05), preceding MSHOA-CatBoost (1.42929e-03) and GWO-CatBoost (1.39589e-03). This deviation in medAE necessitated further evaluation of the model's accuracy and overfitting metrics, while focusing on performance score ranking.




(a) train




(b) test

**Fig. 6.** Prediction robustness metrics for the Base-CatBoost learner and 25 hybrid CatBoost models

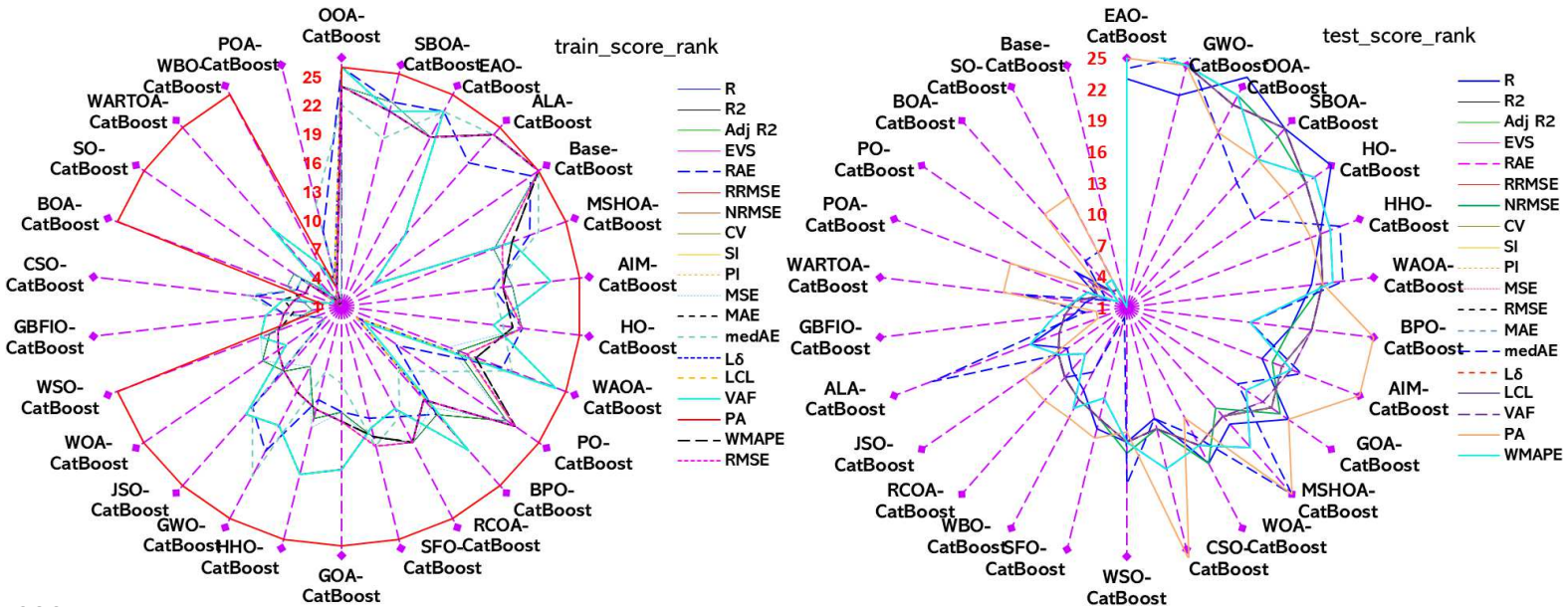
### *3.3.3 Multi-metric accuracy and robustness score ranking of the models*

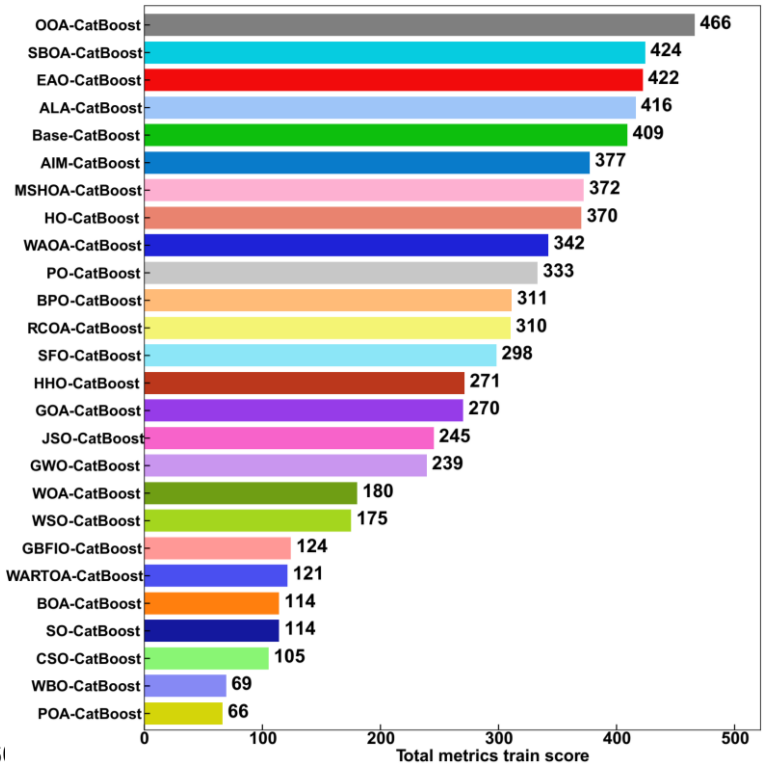
Achieving ideal metrics and models across all evaluated accuracy and robustness metrics, using the complete set of criteria in all modelling phases, was not feasible in practice. This necessitated selecting the model that yielded the most optimized values, taking into account all error and robustness evaluation metrics across all modelling phases without compromising any indices.

The most important aspect is the need for the best hybrid model to strive and achieve performance values closest to the ideal values for all the metrics evaluated in the criterion. As such, it was necessary to obtain the best model by metric scoring and performance ranking of all the models. All computed metrics were programmatically ranked using a statistical performance scoring technique to provide a comprehensive comparison during training and testing based on their proximity to the ideal value for each metric. Metrics for which larger values indicate better performance, such as R, $R^2$, and PA, were ranked in descending order, with the algorithm with the largest value receiving the highest ranking (in this case, 26 out of 26 models) and the algorithm with the smallest value receiving the lowest ranking (1). Metrics representing prediction error or dispersion, including RMSE, MSE, MAE, RRMSE, MARE, RMSRE, and CV, were ranked in ascending order, with smaller values indicating better performance. As such model metrics with an ideal target value were ranked based on the absolute deviation from the ideal values with those closest to the ideal reference point being assigned the best rank, while those with a monotonic optimum were ranked according to their optimization direction, with algorithms producing the most desirable lower or high values receiving the best ranks as they indicated reduced loss, reduced overfitting tendency, smaller train–test performance discrepancy, lower bias, and improved model reliability. The total score for each algorithm was computed separately by summing the individual error and robustness metric scores during the training and testing phases, along with the entire modelling diagnostics, as shown in Fig. 7. All the models were then ordered according to the scores from the training and testing performance metrics. Separate scores were calculated for training metrics and validation diagnostics, allowing the evaluation process to distinguish between algorithms that performed well on training data and those that generalized effectively to unseen test data. During training, the OOA-CatBoost performed best for all the metrics (train total score of 25) while the POA-CatBoost performed the least (train total score of 1). The Base-CatBoost learner model ranked 5th, preceded by SBOA, EAO, and ALA, as shown in Fig. 7a. During testing, the EAO-CatBoost performed best (test total score of 25).

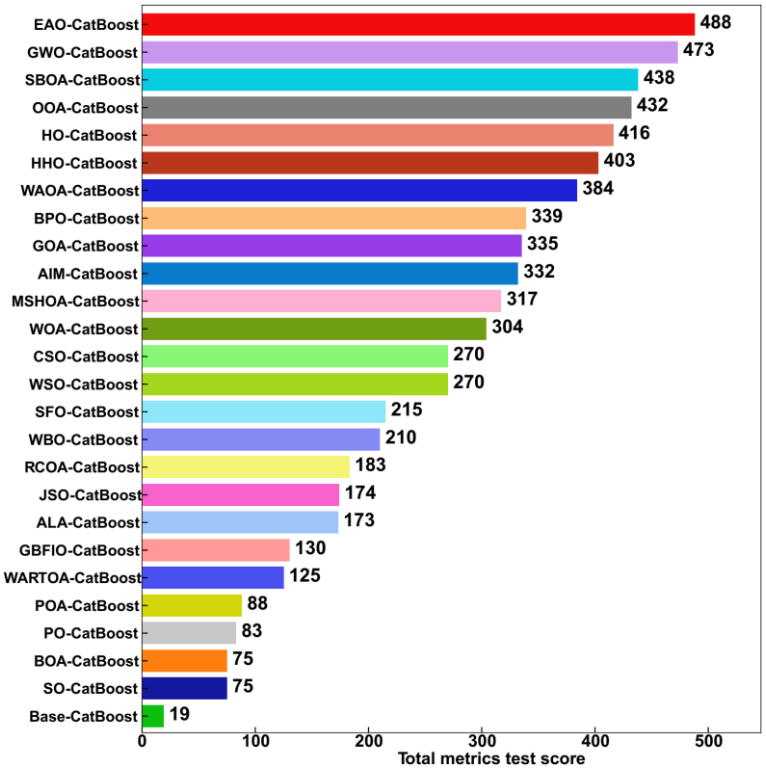
In contrast, Base-Cat-Boost ranked lowest (test total score of 1), as shown in Fig. 7b. Further, evaluating the models' metrics using a unified overall rank score across the entire modelling phases was necessary to establish algorithms that demonstrated stable behavior without excessive overfitting. To provide a comprehensive evaluation of model rank across the entire modelling cycle, a summary leaderboard was constructed by combining training and test scores for both accuracy and robustness metrics. This further enabled the selection of algorithms that

demonstrated strong predictive capability on unseen future data while maintaining stable behavior across the training and validation phases. The overall model ranking scores for the entire modelling phases were computed as the sum of the three stage-specific scores, as shown in Fig. 7c-e. Notably, OOA-CatBoost demonstrated the highest total metrics ranking during training (466), followed closely by SBOA-CatBoost (424) and EAO-CatBoost (422), ALA-CatBoost (416), and the Base-CatBoost learner (409). At the same time, EAO-CatBoost achieved the best rank scores during testing with unseen data (488) while Base-CatBoost was ranked the least (19). Although the Base-Cat Boost learner ranked fairly among models during training as the fifth-best, it is worth noting that all the hybrid models performed better than the stand-alone Base-CatBoost during testing. This shows that the Base-CatBoost learner suffers reduced performance on unseen data due to poor generalization. Further, EAO-CatBoost ranked as the best overall model across all modelling phases (980), while the Base-CatBoost model (465) ranked 17th, and POA-CatBoost (201) ranked worst overall, at position 26 for the entire modelling cycle. Thus, the recently developed EAO algorithm, hybridized with the CatBoost learner, was the most effective hybrid model (EAO-CatBoost) and most reliably generalized on the training dataset for predicting $H_{Db}$ in agricultural tractors.

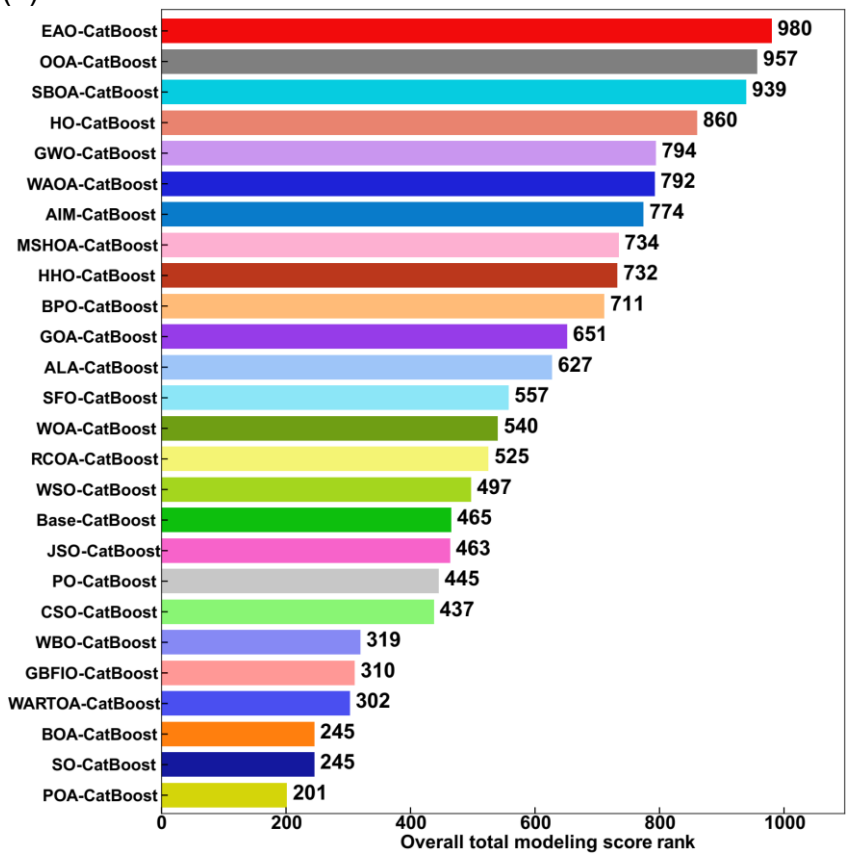





(c)




(d)




(e)

**Fig. 7.** Multi-metric scores for accuracy and robustness during (a) training, (b) testing phases, and the total model rank during (c) training and (d) testing, and the (e) overall final model ranking for the entire modelling cycle

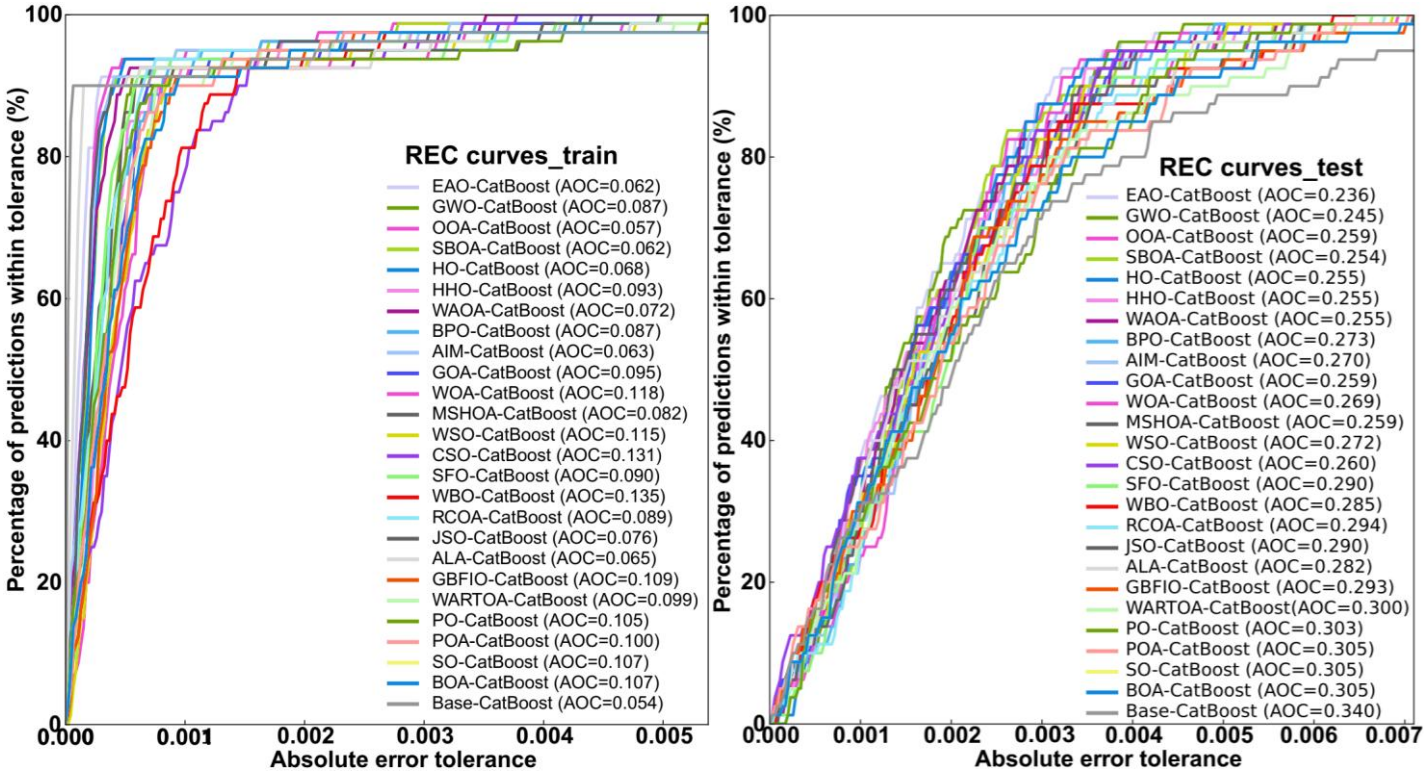
### *3.3.4 Regression error characteristic curve (REC)*

To further establish a valid credit for selecting the EAO-CatBoost as the best model, we analyzed all 26 models using the model regression error characteristic curve (REC) across all modelling phases. REC curve generalized the model regression receiver operating characteristic (RROC), as shown in Fig. 8, thereby aiding in the interpretation of model performance by plotting error tolerance against the proportion of predictions falling within the given error tolerance. This resulted in a curve that estimates the cumulative distribution function of the error, and served as a robust diagnostic tool for confirming model performance, robustness, and stability. The REC curve afformed the robustness of all the models by computing the area over the curve (AOC) and the area under the curve (AUC) as key metrics for overall assessment of model performance, as they reflected cumulative distribution of prediction errors during training and testing phases respectively, illustrating the cumulative prediction proportions within various levels of error tolerances, visually depicting model accuracy and robustness in training and validation. The REC graphs values for Hybrid models generally reveal that hybrid-Cat Boosting not only enhanced the prediction performance of the optimized EAO-CatBoost model but also improved the performance of other hybrid boosting models above the Base-CatBoost learner during testing (Fig. 8). Overall, Base-CatBoost achieved the lowest AOC (0.054) during training (highest AUC) but the highest AOC during testing (0.340), depicting the lowest AUC. Further, EAO-CatBoost achieved the lowest AOC (0.236), indicating the highest AUC during testing (Fig. 8b) and the best error tolerances and stability with a narrow error distribution.

In contrast, Base-CatBoost showed the highest AOC (0.34) during training, indicating the lowest AUC, the largest cumulative error, and the weakest predictive accuracy. The Base-CatBoost training curve indicates greater prediction uncertainty, lower accuracy, and stability. Compared to all hybrid models, EAO-CatBoost REC curves show the highest prediction accuracy at very low error tolerance, underscoring its reliability and robustness on unseen data. Further, the EAO-CatBoost produced the steepest and longest REC curve, achieving the highest cumulative accuracy and demonstrating superior performance and generalization. This suggests strong predictive accuracy with minimal overfitting, indicating that EAO-CatBoost models can achieve high predictive accuracy within very small error bounds, reflecting strong generalization capability and fast convergence.

In contrast, Base-CatBoost achieved low training errors, showing clear signs of overfitting, as evidenced by a lower AOC (0.034) during testing and a comparatively flatter REC curve at low AOC (0.054) during training. This further demonstrates the base learner model's memorization of the training data, highlighting its tendency to overfit and its limited ability to generalize to

unseen data. Fig. 8c compares the REC of the best model (EAO-CatBoost) with those of its top preceding models, with the Base-CatBoost learner showing the widest range of residual errors during testing, resulting in a slower increase in the cumulative distribution compared to the EAO-CatBoost, which showed steeper curves indicating fewer residual errors during testing. Overall, EAO-CatBoost maintained superior generalization, achieving higher accuracy and consistently minimizing residual errors across all phases (8d). The results confirm that the EAO-CatBoost model was the most accurate, robust, and reliable for $H_{Db}$ prediction among the 26 models (25 hybrid models and the base learner) evaluated in this study.




(a) (b)

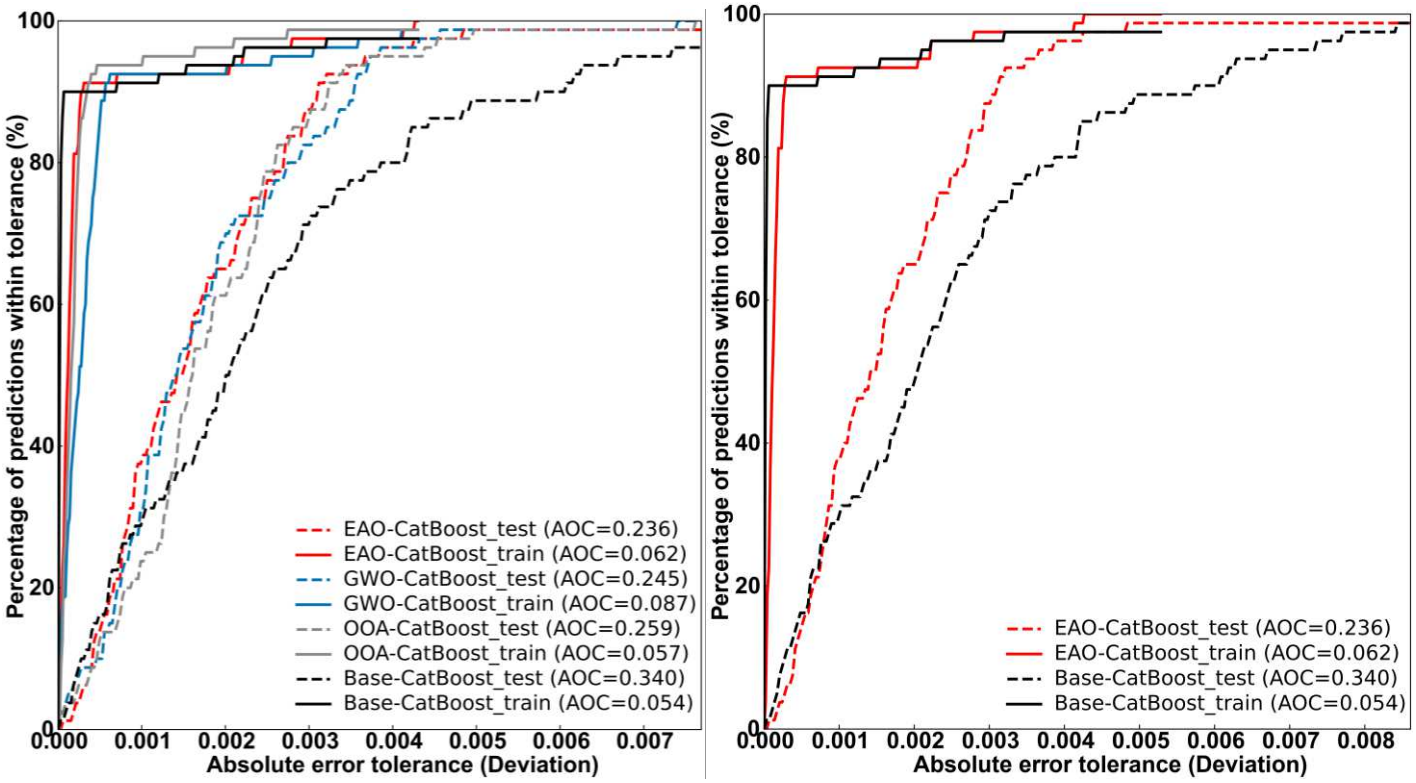



**Fig. 8.** Illustration of the regression error characteristic (REC) of all 26 models during (a) training, (b) testing, (c) comparison of the best three hybrid models, and (d) the selected model with the base learner.

### *3.3.5 Model generalization and overfitting score*

Model generalization was assessed using the perturbation score (PS) and the Generalization gap (G-gap). The perturbation metric evaluated model stability by quantifying how a slight change in the input database (perturbation) is likely to alter predictions and was adopted to measure how the $H_{Db}$ prediction changes when inputs are slightly modified, analogous to heterogeneous conditions and future unseen events in field datasets. A high PS indicates that predictions are highly sensitive to small input changes, suggesting poor robustness and a higher likelihood of overfitting. This is because the model tends to capture noise rather than underlying patterns within the datasets. Conversely, a low PS indicates stable predictions, indicating that the model has learned generalized patterns and is more robust to variations in input data. The hybrid EAO-CatBoost model achieved lower PS (0.005) than the top preceding models obtained in REC curves GWO(0.12), OOA (0.08) and the CatBoost base learner (0.017), affirming resilient to dataset variations, improved robustness and reduced overfitting (Fig. 9a). Consequently, the model is better suited for deployment in real-world scenarios where dynamic drawbar datasets distributions deviate from the training set. Further, the G-gap performance score evaluated the models' performance on unseen data in practical applications, assessed and measured as the

MSE gap between the training and testing phases. The hybrid EAO-CatBoost achieved the lowest G-gap (4.545e-06), indicating better generalization ability. At the same time, a larger G-gap in the Base-Cat Boost (1.042e-05) suggested ingress of overfitting (Fig. 9b). Further, OOA-CatBoost and GWO-CatBoost showed higher PS and G-gap closest to the Base-CatBoost and were eliminated for further evaluation. Nonetheless, the EAO-CatBoost was affirmed to be the best model across all evaluation metrics.

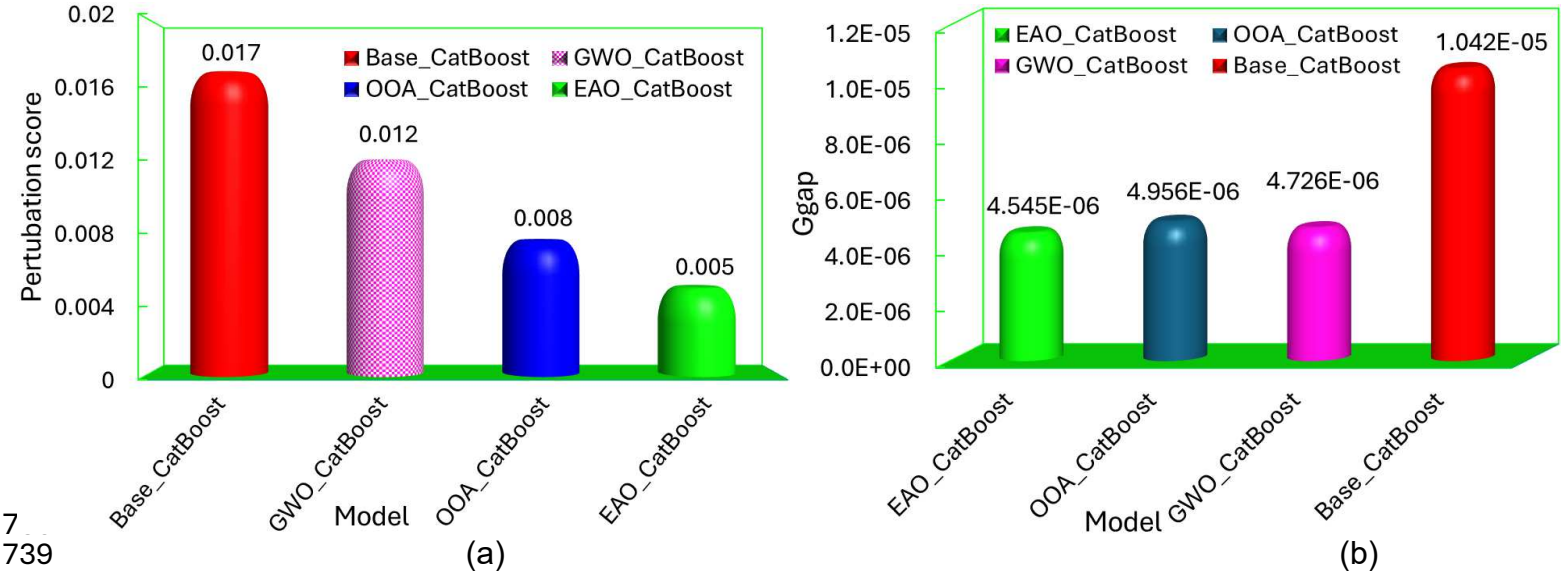



**Fig. 9.** Model performance overfitting evaluation using (a) Perturbation score and (b) generalization gap for the selected model (EAO-CatBoost) with the best of these models

### *3.3.6 Taylor analysis*

Taylor analysis was adopted to assess the prediction performance of EAO-CatBoost and the standalone Base-CatBoost learner. Taylor plot simultaneously constitutes multiple statistical metrics in terms of R, normalized standard deviation (σ), and the centered root-mean-square difference (cRMSE) in a single polar plot, to visually compare measured reference and predicted values. Fig. 10 shows that the EAO-CatBoost predicted $H_{Db}$ with σ = 0.05315 more closely to the reference σ of the measured database (0.05320) than the Base-CatBoost model σ = 0.05311 during training. This indicates better agreement of the EAO-CatBoost with the observed variance, whereas Base-CatBoost shows a slight underestimation of variability. Further, the EAO-CatBoost and Base-CatBoost predicted $H_{Db}$ with σ values of 0.0529 and 0.0517, respectively, during testing. It can be observed that the EAO-CatBoost achieved a σ closer (0.0529) to the actual database ( 0.0532) than the Base-CatBoost model (0.0517) during testing. It is concluded that the prediction capabilities of EAO-CatBoost improved the σ with unseen data by 0.0003 deviation difference, very close to the actual database, while the Base-CatBoost lagged by a larger margin (0.0015). At the same time, EAO-CatBoost achieved the strongest

correlations (R>0.9999) during training and testing (0.9992) compared to Base-CatBoost (0.9997 train and 0.9985 test). Further, a lower cRMSD positions EAO-CatBoost closest to the reference point on the Taylor diagram during training (0.0009) and testing (0.0022), indicating superior agreement with actual observed data and overall predictive accuracy on unseen data. Conversely, Base-CatBoost is farther from the reference, reflecting higher dispersion and reduced agreement with the observed data during training (0.0012) and testing (0.0032). High prediction clustering to the reference point indicated superior predictive stability of EAO-CatBoost in aligning its predictions closest to the measured values, suggesting a stronger capability in capturing the underlying variabilities in the $H_{Db}$ datasets.

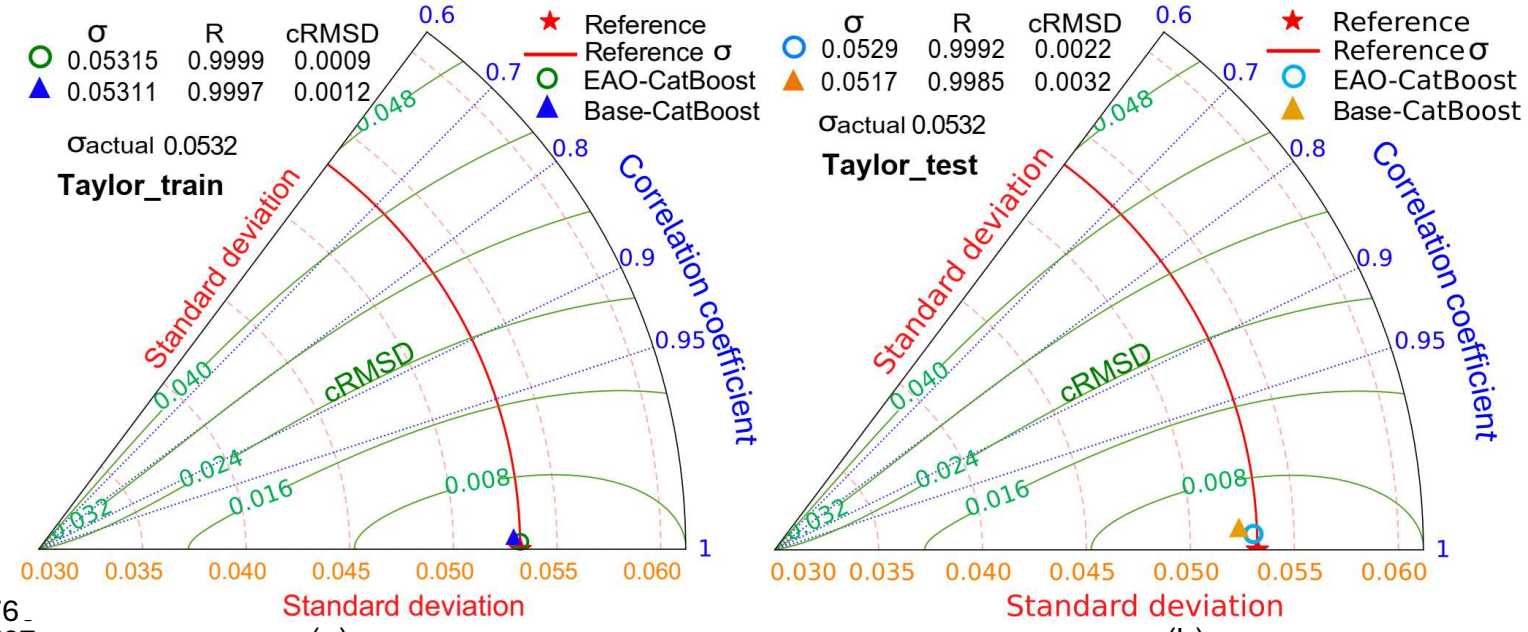


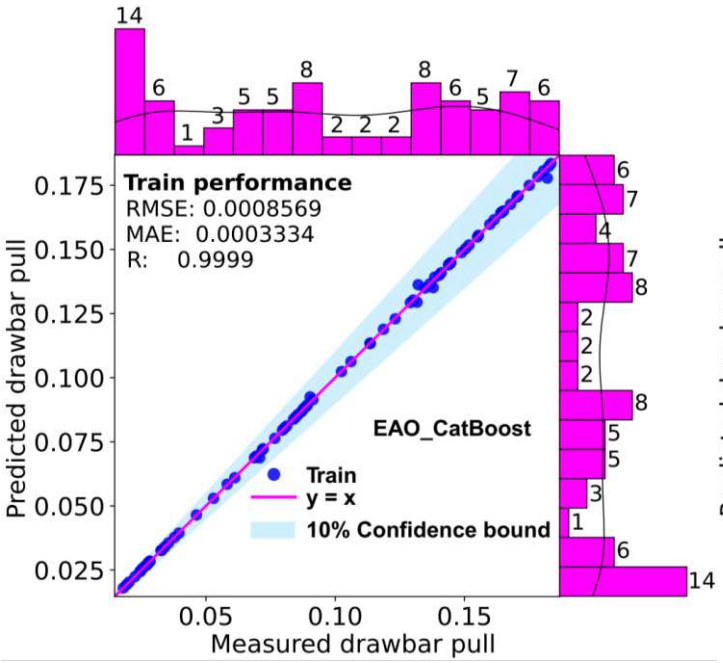

**Fig. 10.** Taylor plots showing comparison of Base-CatBoost and EAO-CatBoost models in (a) training and (b) testing phases. The points on the graph indicate the positions of the hybrid EAO-CatBoost model's predictions relative to the base learner. The red line represents the standard deviation of the actual data. It serves as the experimental reference line, whereby models closest to it have predictions closest to the actual data, indicating high prediction accuracy.

## 3.4 Comparative regression evaluation of hybrid EOA-CatBoost with the Base-CatBoost

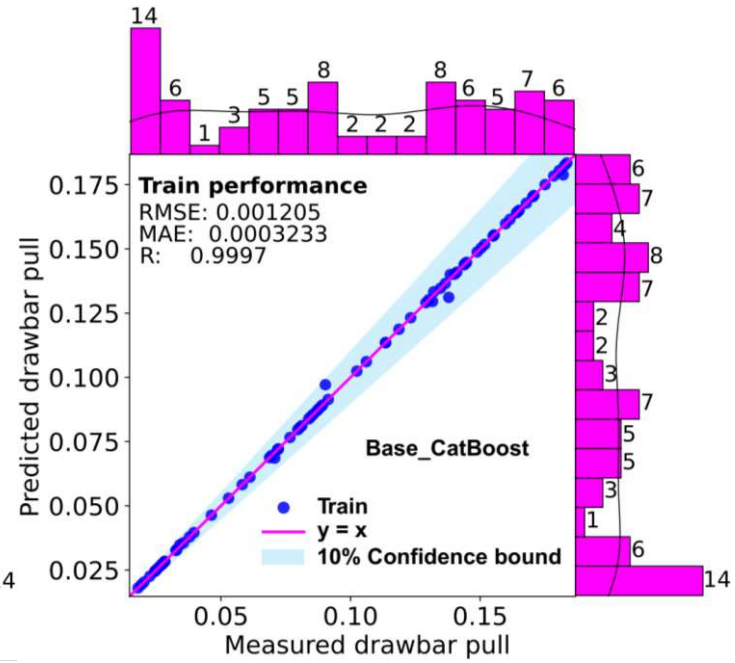
### *3.4.1 Performance scatter and regression deviation*

Fig. 11 shows the regression scatter plot and deviation histogram for EAO-CatBoost and the Base-CatBoost learner during training and testing on the entire database, comparing the measured and predicted $H_{Db}$. This detailed illustration enables evaluation of the accuracy and reliability of each model prediction, along with its associated actual, non-normalized prediction. EOA-CatBoost regressed its predictions closest to the experimental values, with a narrow

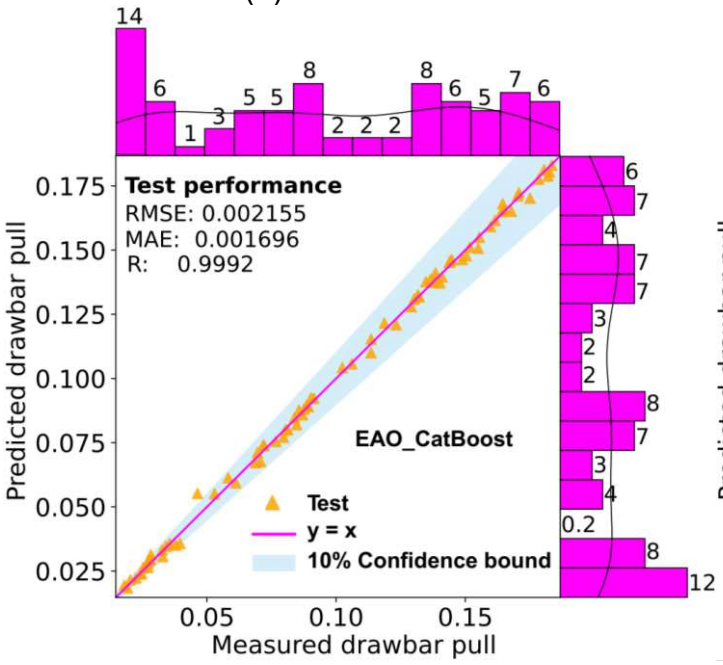
781 confidence interval and a 45-degree regression line, yielding the strongest agreement with
782 observed $H_{Db}$, the highest R (0.9999), and the lowest RMSE (0.0008569). Still, the Base-
783 CatBoost achieved the best MAE (0.0003233) during training. However, during testing, EAO-
784 CatBoost achieved the best values of R (0.9992), MAE (0.001696), and RMSE (0.002155) than
785 the Base-CatBoost, with R (0.9985), MAE (0.002465), and RMSE (0.00323). At the same time,
786 modelling with the entire database showed that the EAO-CatBoost performed best in both
787 training (R=0.9999) and RMSE (0.0008569) and testing (R=0.9992, RMSE=0.002155)
788 compared to Base-CatBoost during training (R=0.9997, RMSE=0.001205 and testing
789 (R=0.9985, RMSE=0.00323. Conversely, EAO-CatBoost achieved the worst MAE in training
790 (0.0003334) compared to Base-CatBoost (0.0003233) but achieved the best MAE in testing
791 (0.001696) compared to Base-CatBoost (0.002465). However, compared to EAO-CatBoost,
792 Base-CatBoost had the best MAE values during training cycles (Fig. 11a, b, e), but the worst in
793 testing and the entire validation database, as shown in Fig. 11c, d, and f, respectively. This shows
794 that the Base-CatBoost model learns the training data patterns (including noise) very well, but
795 those patterns do not generalize to unseen test or validation data. It optimizes to fit the training
796 data, inherently causing overfitting, and uses specialized categorical encodings and target
797 estimators that are more precise on known data, computed from labels, leading to lower bias
798 due to overfitting. However, when applied to an unseen test or validation dataset, where the
799 memorized targets are unknown, it does not generalize or eliminate generalization gaps, leading
800 to higher test MAE. Therefore, the model is inherently tailored to the unique structure of the
801 training dataset and becomes precise only when applied to familiar patterns, but exhibits target
802 leakage on unseen data. The model dispersion of EAO-CatBoost closely aligns strongly with
803 unseen data, as shown by the regression 1:1 line (dots in the scatter plots), indicating more
804 accurate predictions. Further, the slope values of the trendlines indicate that the predictions of
805 $H_{Db}$ are closer to the measured values. Thus, the error histograms clearly show that the
806 predictions are more accurate in EAO-CatBoost, potentially aligning more closely with the
807 measured HDb than in Base-CatBoost, due to lower correlations between the measured and
808 $H_{Db}$ predictions. At the same time, the relative error histograms of EOA-CatBoost predictions
809 showed minimal deviations from all the measured values, revealing that the models converged
810 the $H_{Db}$ predictions with consistent accuracy during validation tests.


Train performance
RMSE: 0.0008569
MAE: 0.0003334
R: 0.9999
EAO_CatBoost
Train
y = x
10% Confidence bound
Predicted drawbar pull
Measured drawbar pull

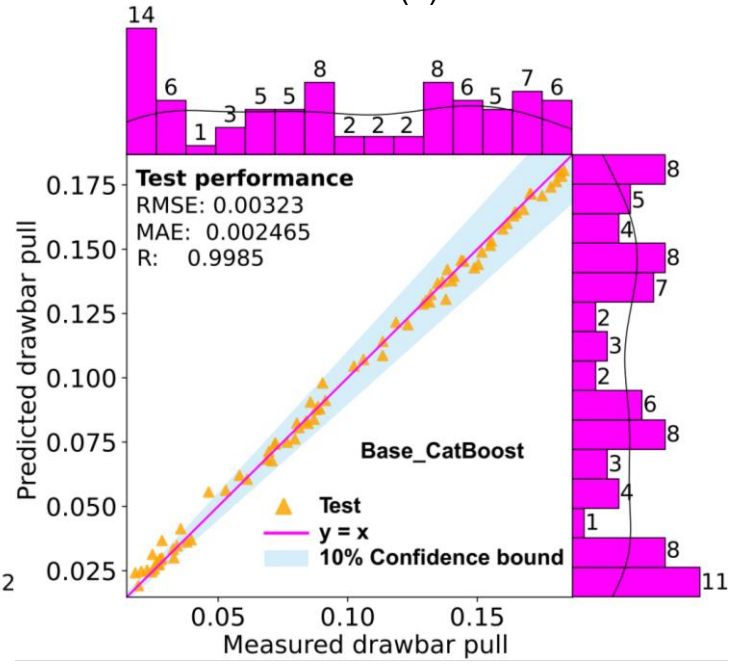
(a)


Train performance
RMSE: 0.001205
MAE: 0.0003233
R: 0.9997
Base_CatBoost
Train
y = x
10% Confidence bound
Predicted drawbar pull
Measured drawbar pull

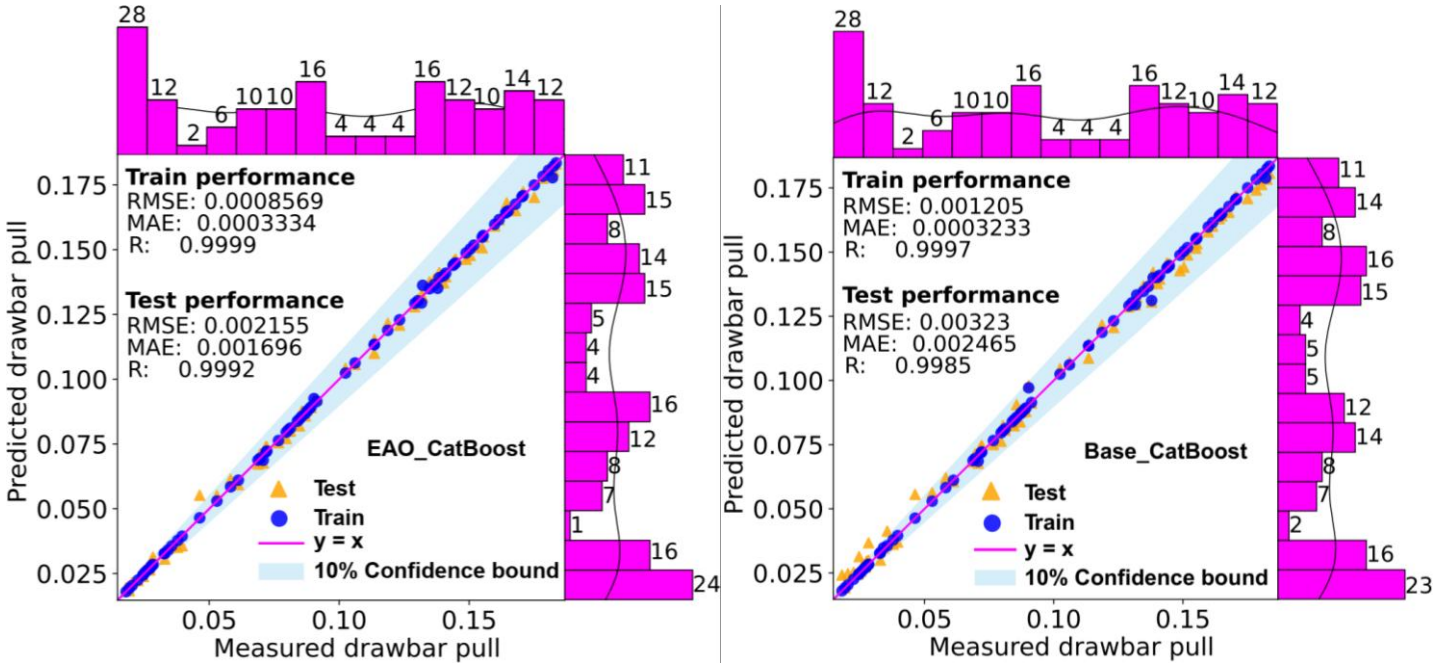
(b)


Test performance
RMSE: 0.002155
MAE: 0.001696
R: 0.9992
EAO_CatBoost
Test
y = x
10% Confidence bound
Predicted drawbar pull
Measured drawbar pull

(c)


Test performance
RMSE: 0.00323
MAE: 0.002465
R: 0.9985
Base_CatBoost
Test
y = x
10% Confidence bound
Predicted drawbar pull
Measured drawbar pull

(d)




(e) (f)

**Fig. 11.** Regression comparison of Hybrid EAO-CatBoost and the Base-CatBoost models during training (a-b), testing (c-d), and with the entire database ( e-f), respectively.

### *3.4.2 Model training and prediction error residuals*

Prediction residual error plots of Base-CatBoost and EAO- CatBoost showed differences between the actual $H_{Db}$ values and the respective model predictions (Fig. 12). The modelling domain of EAO- CatBoost produced narrower, more tightly packed training and testing curves that were closely clustered around the actual $H_{Db}$ line, confirming higher model precision, consistency, and reduced variance susceptibility. Furthermore, the pattern of the EAO-Cat Boost absolute error line showed smoother residual dispersion and more evenly distributed prediction residuals, indicating reduced target leakage, less overfitting, and greater prediction stability. Conversely, the Base-CatBoost showed comparatively wider and irregular residual fluctuations around the actual $H_{Db}$ profile line, depicting weak precision and low residual stability. In certain prediction regions, the Base-CatBoost residual distribution curve showed greater error concentration due to localized error variability, suggesting poor hyperparameter tuning. This indicates under-optimized learning leading to reduced robustness in capturing the complex interactions that govern $H_{Db}$ predictions within the database. Thus, hybrid optimization of the base learner with EAO captured the complex nonlinearities in the datasets while reducing prediction bias.

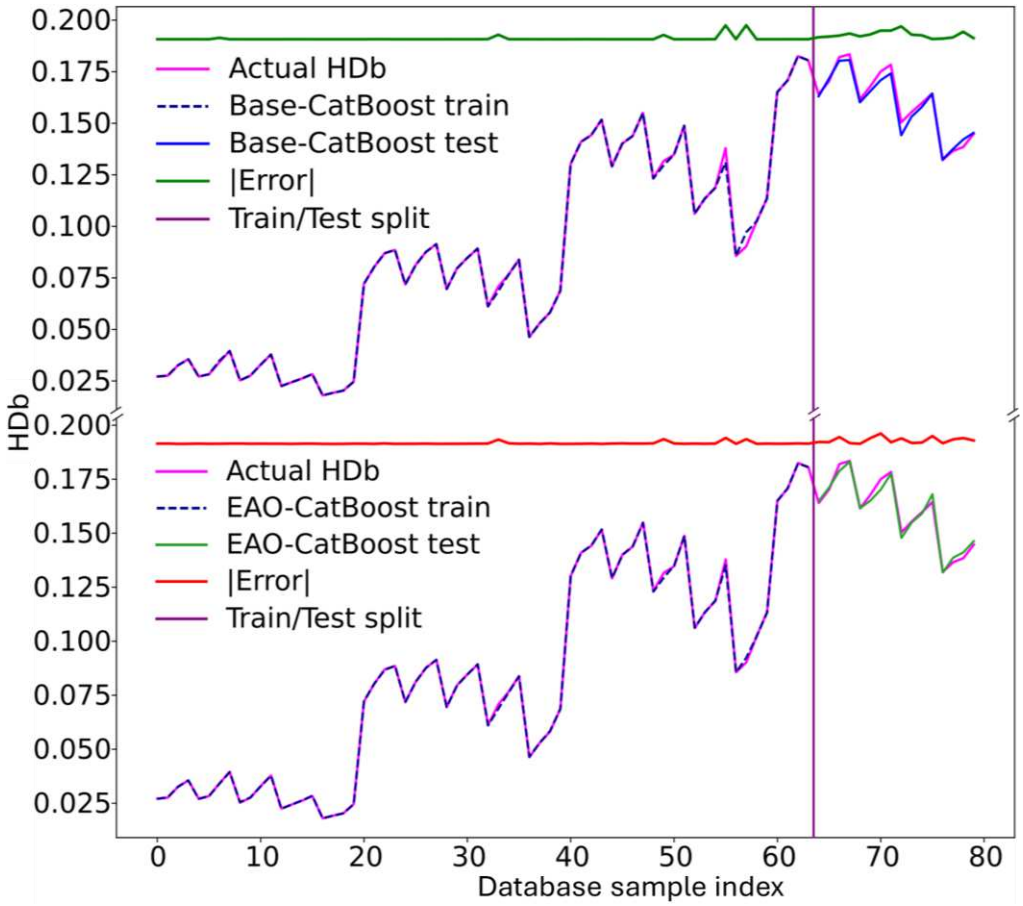



**Fig. 12.** Prediction domain profile and error residuals of (a) Base-CatBoost and (b) EAO-CatBoost

### 3.5 SHAP analysis based on the EAO-CatBoost model

#### *3.5.1 Explainable feature importance based on EAO-CatBoost*

Explainable hybrid artificial intelligence (XHAI) machine learning based on Shapley Additive exPlanations (SHAP) analysis revealed the dynamic impact of variables on the predictions by the EAO-CatBoost model. Bee swarm plots provided a complex information-rich display of qualitative measures of both local and global model behavior and interpretability by combining both feature relative importance and impact directionality, revealing each feature's contribution to the output $H_{Db}$, indicating whether it increases or decreases the final $H_{Db}$ prediction value and by how much and their global distribution across all predictions (Fig. 13a). Further SHAP framework decoder was applied to the entire dataset, to generate global Bee swam plots for extraction of feature importance as shown in Fig. 13b. Results visualize whether a feature increases (positive) or decreases (negative) the predicted value compared to the base rate (average prediction). SHAP values along the x-axis indicate a feature's relative impact, while color gradients reveal the variable's magnitude. A strong positive or negative influence on $H_{Db}$ prediction is indicated by the concentration of red color points on the positive or negative axis,

854 respectively, and vice versa. Further, the distribution of the feature point values on the x-axis
855 shows that high $F_{Tr}$ contributes to positive prediction impact and low $F_{Tr}$ contributes to negative
856 prediction outcomes of $H_{Db}$. Global Bee swarm SHAP summary revealed that $F_{Tr}$ achieved the
857 greatest distribution spread with positive pink points (highest values) and blue negative points
858 (lowest values), indicating that higher $F_{Tr}$ increases EAO-CatBoost drawbar pull predictions. In
859 contrast, lower $F_{Tr}$ values decrease drawbar predictions. However, both high $\theta_{Soil}$ and $P_{Tire}$
860 showed deviations, with negatively clustered pink points contributing to negative prediction
861 outcomes and lower $\theta_{Soil}$ and $P_{Tire}$ positively clustered blue points contributing to positive
862 prediction outcomes. This confirms that higher $\theta_{Soil}$ and $P_{Tire}$ increased predictions. Additionally,
863 each feature had a consistent impact as indicated by the vertical spread of points; $CI_{Soil}$, $T_{depth}$,
864 and $\tau_{shear}$ showed the greatest variability and thus the greatest influence on CatBoost model
865 predictions across samples, while $R_{depth}$ and $W_{Load}$ had the least. Further, Fig. 13(b) ranks the
866 variables by their predictive power as indicated by the hierarchy of their mean absolute SHAP
867 values. $F_{Tr}$ dominated the predictive hierarchy (0.0168), higher (>50%) than the preceding $CI_{Soil}$
868 (0.00732), indicating that tractive force is the single-most critical predictor of drawbar pull.
869 Ranging from 0.00506 to 0.0168, $T_{depth}$, $IP_{Soil}$, $CI_{Soil}$, and $F_{Tr}$ accounted for the greatest influence
870 on the model predictions. At the same time, $R_{depth}$ (0.000569) and $W_{Load}$ (0.0012) had minimal
871 impact on the predictions with the lowest SHAP values (Fig.13b). The model relied primarily on
872 tractive force, cone index, tillage depth, and soil shear strength to construe the drawbar loading
873 estimate.

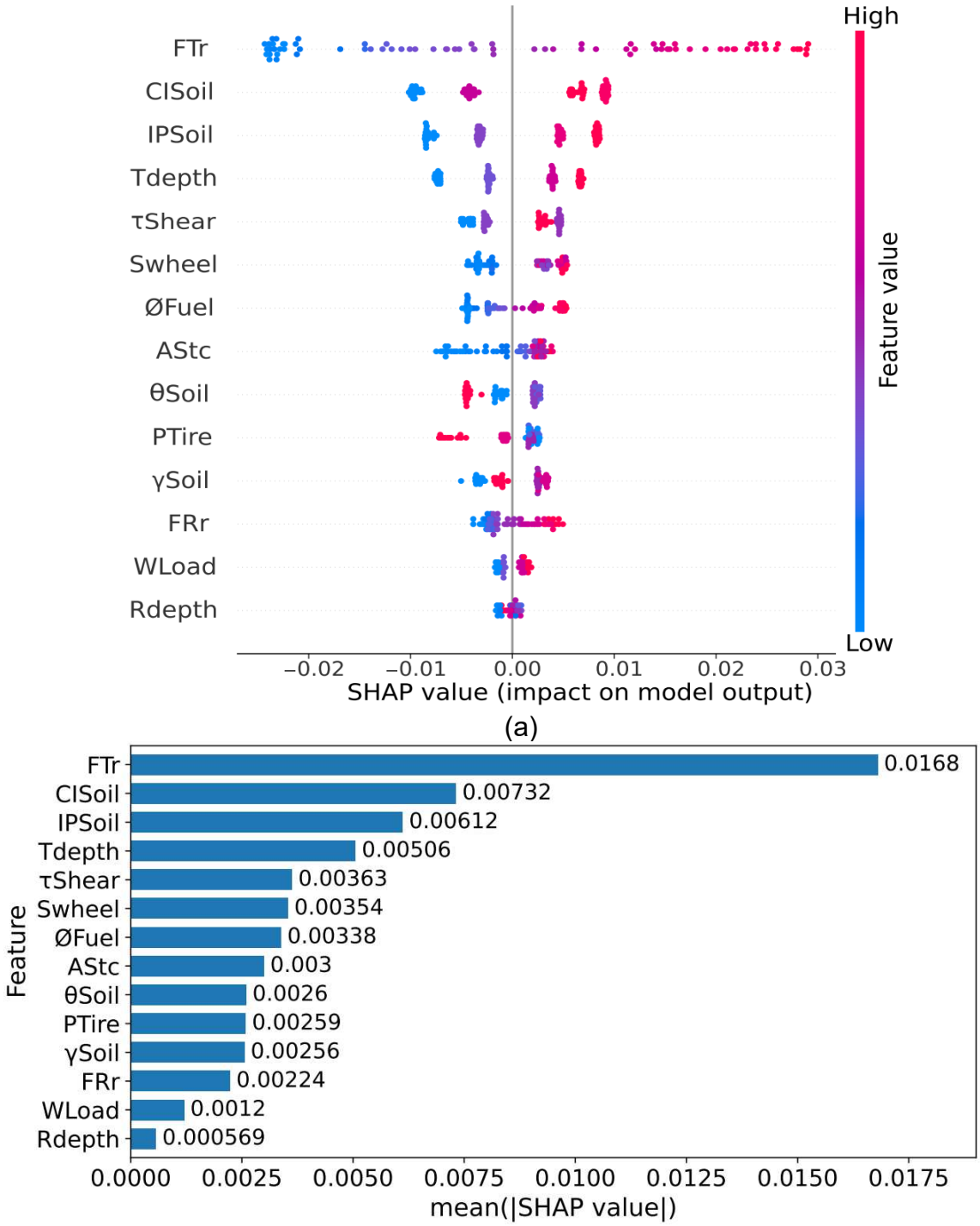



**Fig. 13.** SHAP summary of (a) features and (b) feature importance and contribution plot to the hybrid model predictions for tractor drawbar pull by EAO-CatBoost model. Each point represents a sample, and the point's color indicates the value of the corresponding instance; blue and pink denote low and high feature values, respectively. The X-axis shows the (SHAP) value of a specific feature, which signifies the feature's effect on the prediction. An absolute higher SHAP value indicates a greater impact on the predicted $H_{Db}$, and the sign indicates whether the average prediction increases or decreases. The Y-axis shows the arrangement of features in order of importance (high to low) from top to bottom, computed from average absolute SHAP values.

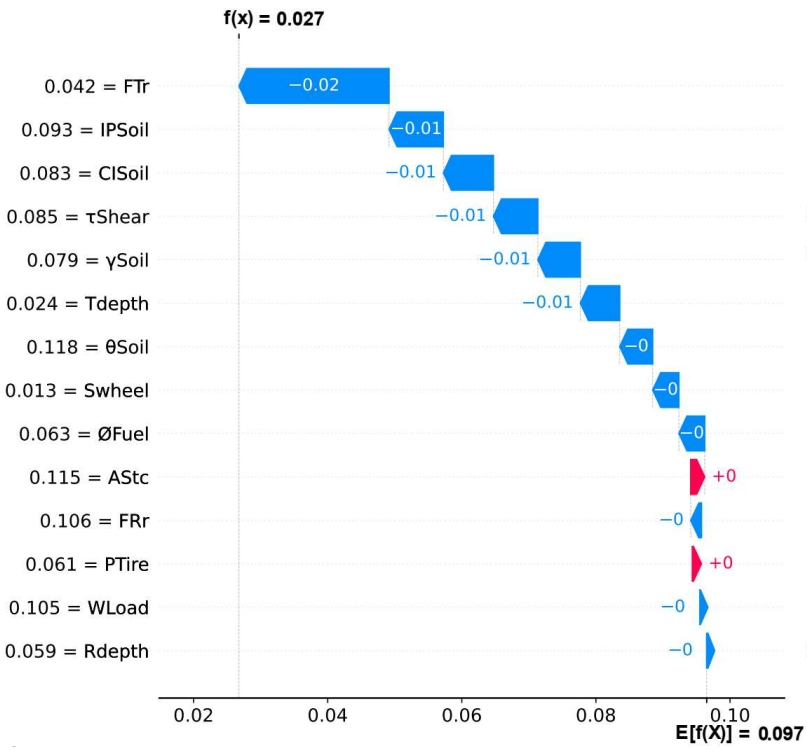
Opposing side feature values are shown as high (red) and low (blue) values. Pink points denote a higher positive or negative contribution, while blue points denote a lower positive or negative impact on the prediction. The vertical axis represents the prediction features of $H_{Db}$ ordered by their mean absolute SHAP values and their distribution rank according to their contribution to predictions.

### *3.5.2 Waterfall and individual force plots explanations*

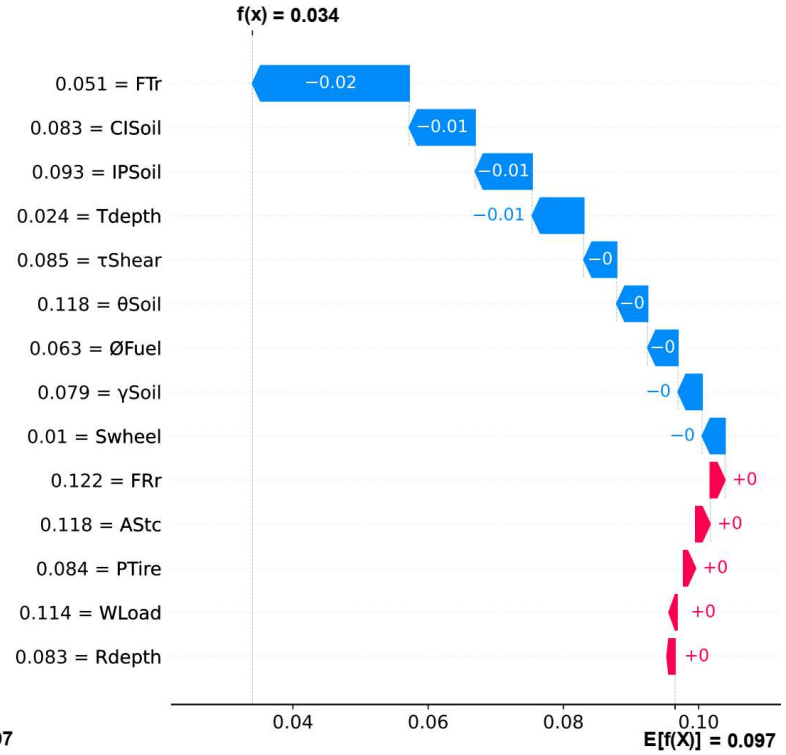
The SHAP-waterfall plot (Fig. 14) provides a visual interpretation of the local feature contributions to a single prediction instance. Each row indicates how the positive (pink) or negative (blue) contribution of each feature shifts the value from expected output (E[ f(x)]) to the model-predicted value (f (x)). f(x) = E( f(x))+ shapley value for each feature. The X-axis shows average SD, and the Y-axis the features arranged in descending order according to their SHAP values (from top to bottom) to visualize how much the features impacted either negatively (blue) or positively (red) and how that shifted the predictions from the expected outcome E[f(x)]. The values beside each feature row indicate the respective observed values. The SHAP force plots are also shown as horizontal projections of the waterfall plots, highlighting features that drive higher (pink) or lower (blue) drawbar pull predictions. For the first sample, the model output is f(x) = 0.027, and the expected final output is E[f(x)] = 0.097. Each feature influences this model's outcome deviation. The observed $F_{Tr}$ =0.042 (normalized) provided a low negative contribution (-0.02), while $IP_{Soil}$, $CI_{Soil}$, $\tau_{Shear}$, $\gamma_{Soil}$, and $T_{depth}$ all had negligible contributions (-0.01), cumulating into a low (-0.07) contribution, pointing to the negative predictions. In contrast, others had negligible (<0.005) negative contributions ($\theta_{Soil}$, $S_{wheel}$, $\emptyset_{Fuel}$, $F_{Rr}$, $W_{Load,}$ and $R_{depth}$) and positive contributions ($A_{Stc}$, $P_{Tire}$) to the prediction, respectively. This demonstrated that the features push the predictions towards higher values primarily due to the largest contribution from tractive force, moderate contributions from soil plasticity, cone index, shear strength, soil bulk density, and tillage depth, while other features played supportive parametric roles in the complex prediction process.

The corresponding SHAP force plot of Fig. 14 e explains the key information on the predictions but in a more condensed format than the expansive waterfall plots that spare no detailed explainable behavior in the generalized predictions (a) and the $20^{th}$ (b), $40^{th}$ (c),$60^{th}$ (d), and $80^{th}$ (e) percentile instances. Features that raised the $H_{Db}$ prediction above the base SHAP value (0.09651) have pink/red arrows pointing right, while those that pushed the $H_{Db}$ prediction lower from the base have blue arrows pointing left. The base value (0.09651) is the average $H_{Db}$ (normalized) in the database, and features with greater impact (larger SHAP values) have the longest arrows pushing the prediction instance from 0.09651(base) to 0.03, 0.08, 0.12, and 0.15

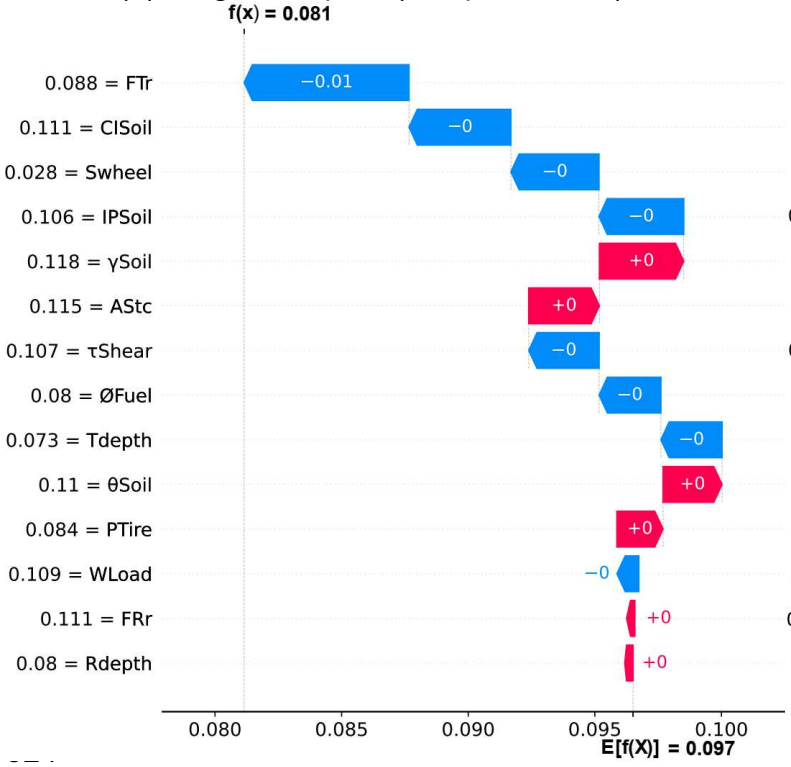
at the 80th percentile. The dominant blue section for $F_{Tr}$ = 0.04248 visually indicates a strong negative contribution towards lower prediction, followed by $IP_{Soil}$ (0.09266) and $CI_{Soil}$ (0.08299). The compact force plot shows $A_{Stc}$ (0.115) and $P_{Tire}$ (0.061) on the left edge, with minimal positive impact towards higher $H_{Db}$ predictions.



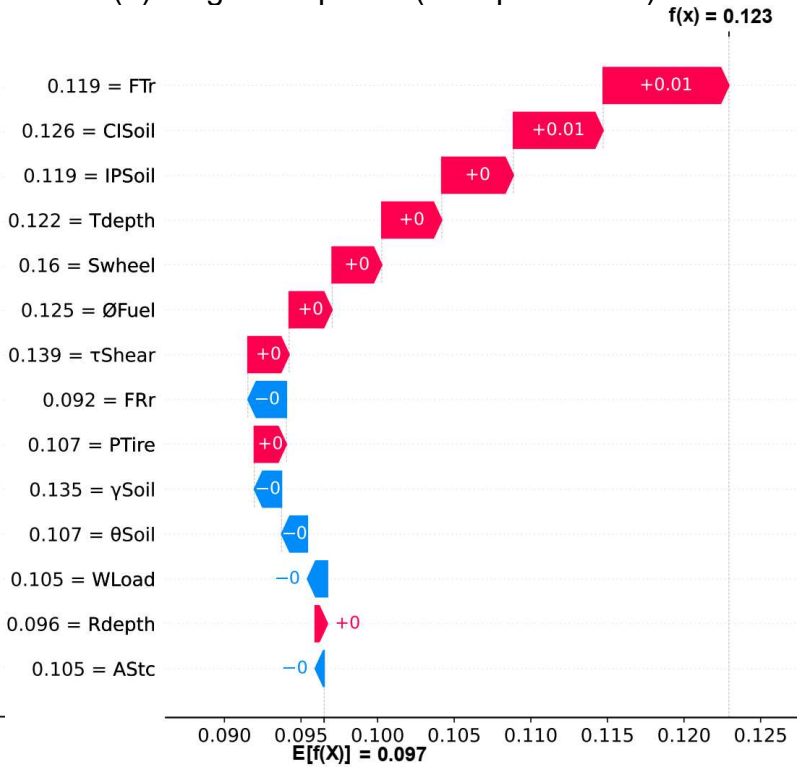

(a) Single sample 1(first per-centile)



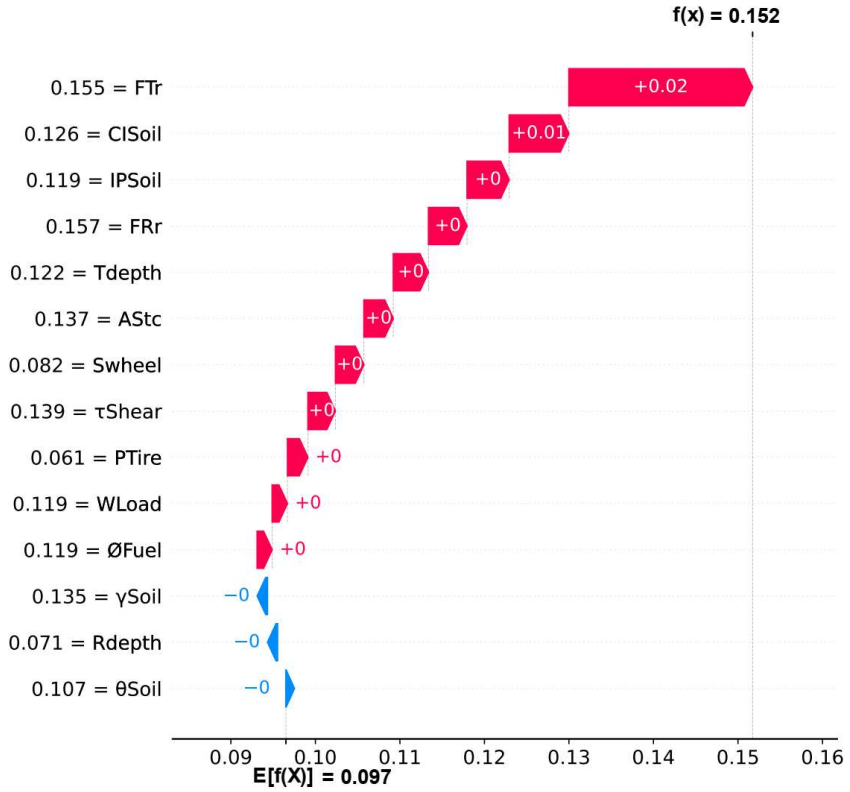

(b) Single sample 16 (20th per-centile)




(c) Single sample 32 (40th percentile)




(d) Single sample 48 (60th percentile)




(e) Single sample 64 (80th percentile)

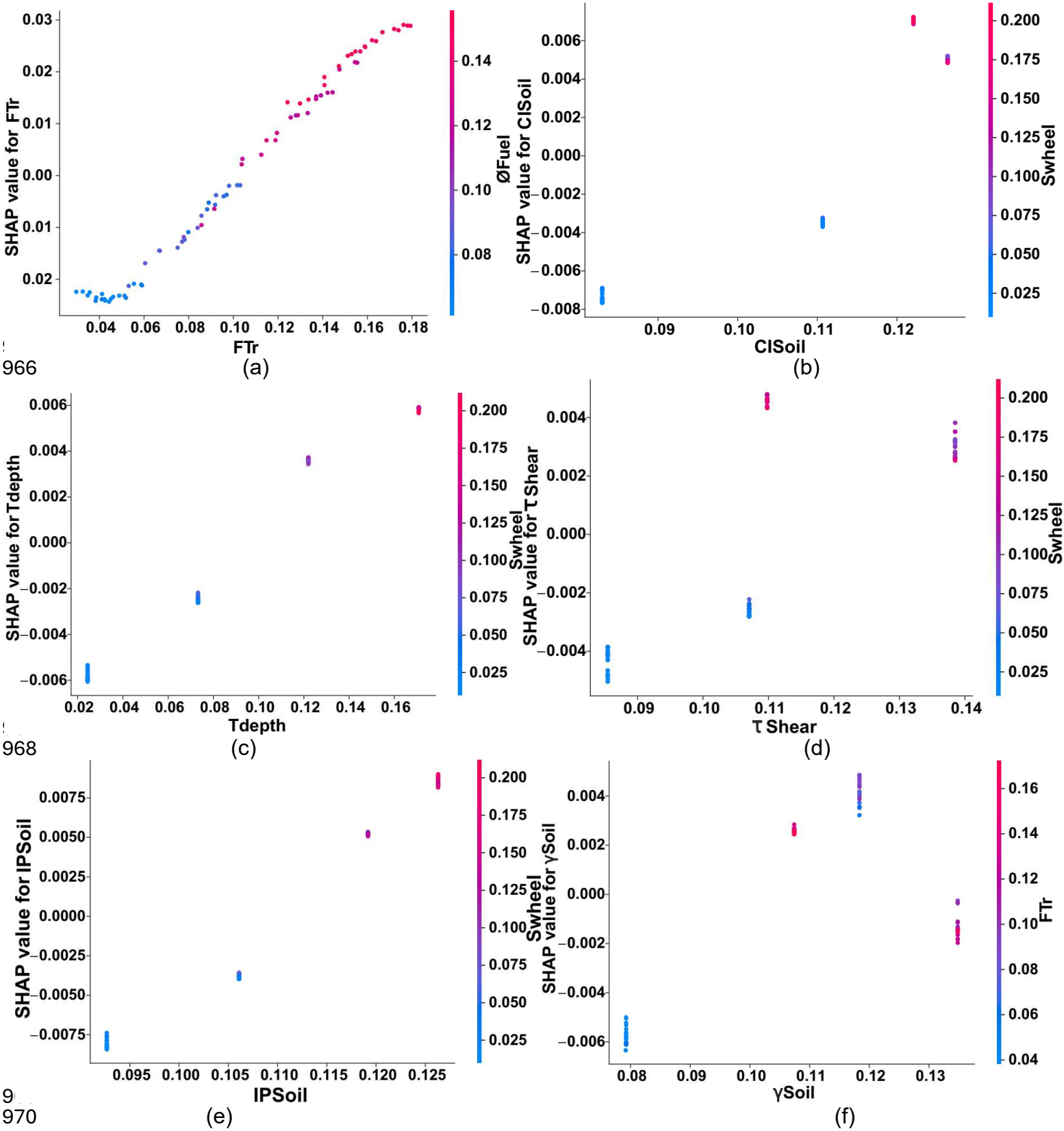



(f) SHAP force plot on selected samples

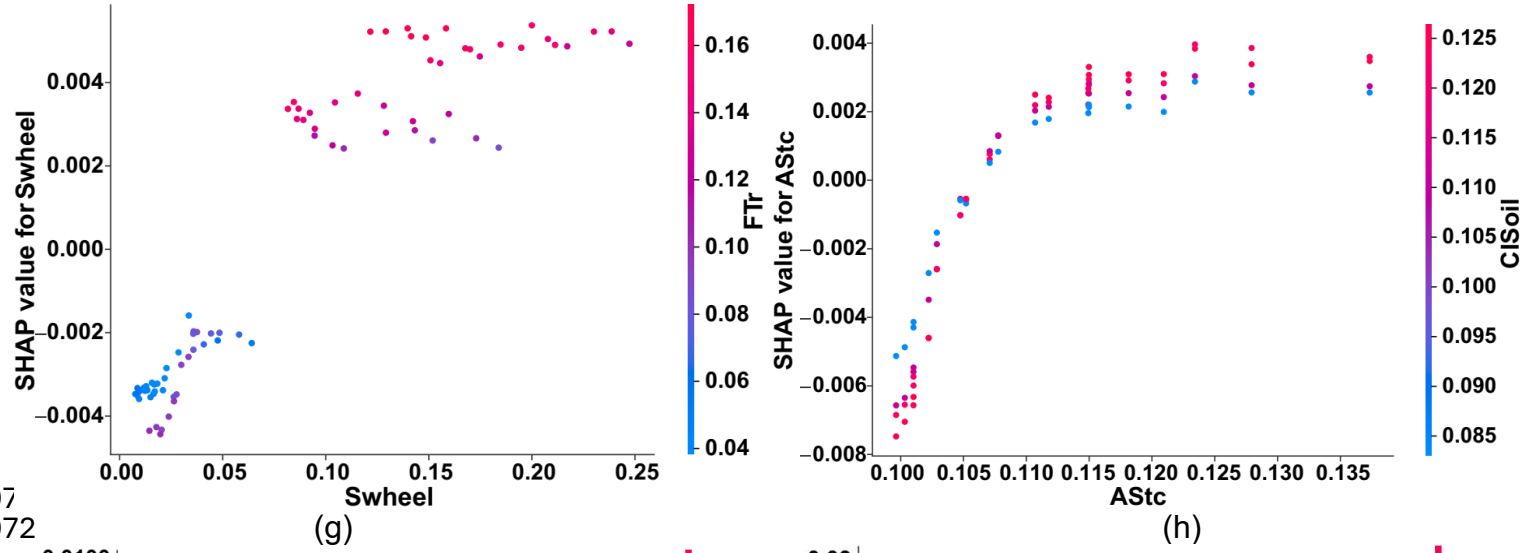
**Fig. 14.** SHAP analysis local explanations of waterfall plots (a-e) and the corresponding interpretation force plots (a-e) for dominant features showing how the features drive the EAO-CatBoost model predictions and the features' impact on local instances. The x-axis shows the individual $H_{Db}$ predictions from the CatBoost model. E[f(x)] (0.097) represents the baseline value, indicating the model output of the entire dataset, and f (x) represents the individual model output for each feature. The y-axis represents the input features of the models.

### *3.5.3 Prediction dependency scatter plot of EAO-CatBoost*

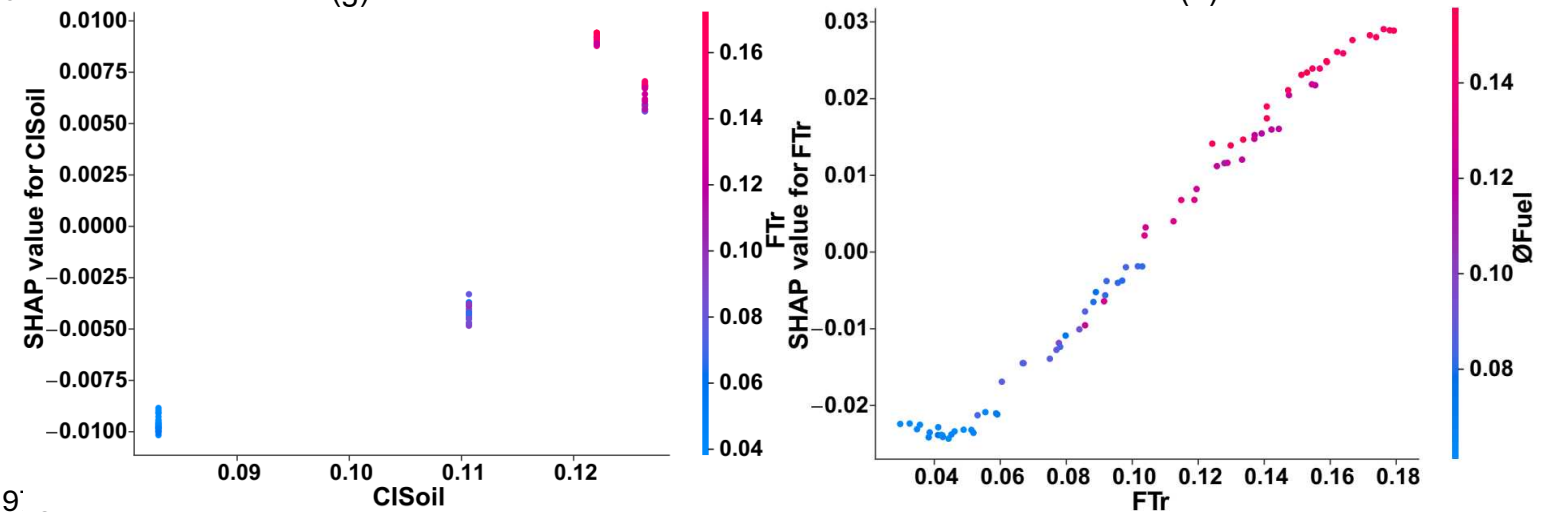
It was necessary to understand the true relationship between the model's dominant features and their prediction outcomes, because the Beeswarm plots were information-dense and provided broad SHAP values for many features at once. Fig. 15a-s shows the true dependence plots for the model features, revealing that the relationships between SHAP values and variable values were quite different for each feature. SHAP dependency plots showed inverse interaction effects and nonlinearities among features. SHAP values above the y=0 line (positive values) showed higher predictions of $H_{Db}$, and vice versa (Fig. 15a). Positive values increase the prediction, while negative values decrease it. Lower levels of $F_{Tr}$ (<0.1) favored low $Ø_{Fuel}$ and lower $H_{Db}$ predictions, while the combination of higher $F_{Tr}$ levels (>0.1) consumed more Fuel and produced higher $F_{Tr}$ predictions. At the same time, the $CI_{Soil}$-$S_{Wheel}$ interactions showed that $CI_{Soil}$ <0.12 strengthened $H_{Db}$ discrimination, leading to lower SWheel predictions, while $CI_{Soil}$ >0.12 at higher $S_{Wheel}$ led to higher $S_{Wheel}$ predictions. Fig 15c shows that $T_{depth}$ (<0.12) discriminated low (y<0) and high $H_{Db}$ predictions (y>0) with lower $T_{depth}$(<0.12) leading to lower $S_{Wheel}$ (y<0) while higher $T_{depth}$ (>0.12) leading to higher $S_{Wheel}$ at higher $H_{Db}$ predictions (y>0). Further, the $\tau_{Shear}$ – $S_{Wheel}$ interaction showed that $\tau_{Shear}$ < 0.11 and low $S_{Wheel}$ (<0.125) was an important discriminating factor for $H_{Db}$ (Y=0 line), as indicated in Fig.15d. $IP_{Soil}$- $S_{Wheel}$ responded in a similar trend to $CI_{Soil}$ with $S_{Wheel}$. However, Fig. 15f clearly shows that the $H_{Db}$ predictions dramatically grow (y>0 line) when $\gamma_{Soil}$ increases up to 0.11 at high $F_{Tr}$ (0.14) and further increases to higher values at higher $\gamma_{Soil}$ (0.12) even when $F_{Tr}$ declines, but thereafter declines as $F_{Tr}$ maintains moderately high values at higher $\gamma_{Soil}$ (>0.13). However, $H_{Db}$ predictions are lowest at low values of $F_{Tr}$ and $S_{Wheel}$, but increase as FTr and SWheel increase. This is more so as $CI_{Soil}$, $T_{depth}$, $\tau_{Shear}$, and $IP_{Soil}$ increase forcing increased $Ø_{Fuel}$ as deduced in figures 15a, b, c, d, and e. These interactive patterns are evidence of the high-precision discrimination aided by the hybrid EAO-CatBoost model, which captures complex nonlinearities across various combinations of model prediction features.


SHAP value for FTr
FTr
ØFuel
SHAP value for ClSoil
ClSoil
Swheel
SHAP value for Tdepth
Tdepth
SHAP value for τ Shear
τ Shear
SHAP value for IPSoil
IPSoil
SHAP value for γSoil
γSoil
FTr

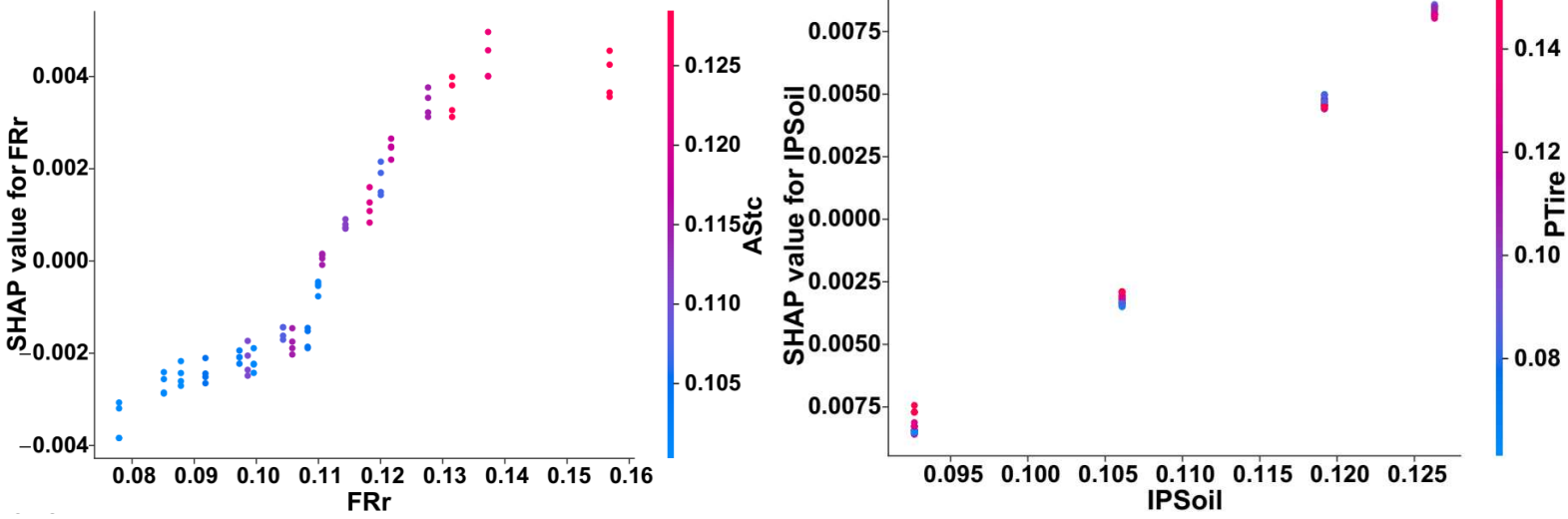
(a) (b) (c) (d) (e) (f)


SHAP value for Swheel
Swheel
FTr
SHAP value for AStc
AStc
ClSoil

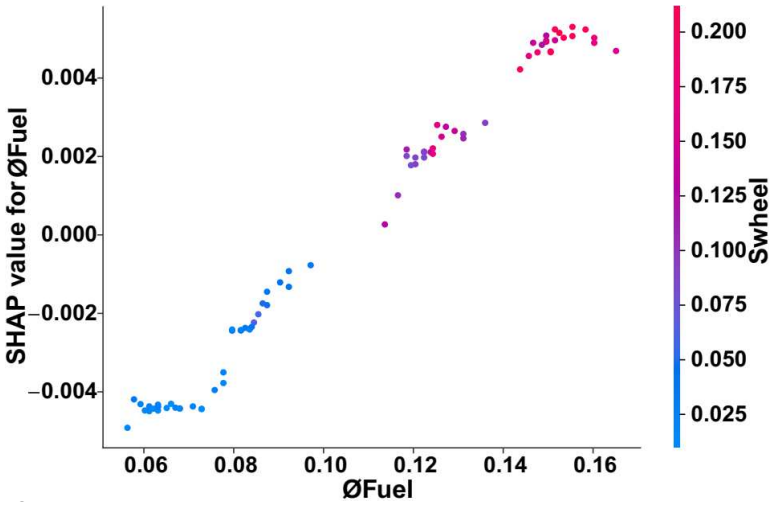
(g) (h)


SHAP value for ClSoil
ClSoil
FTr
SHAP value for FTr
FTr
ØFuel

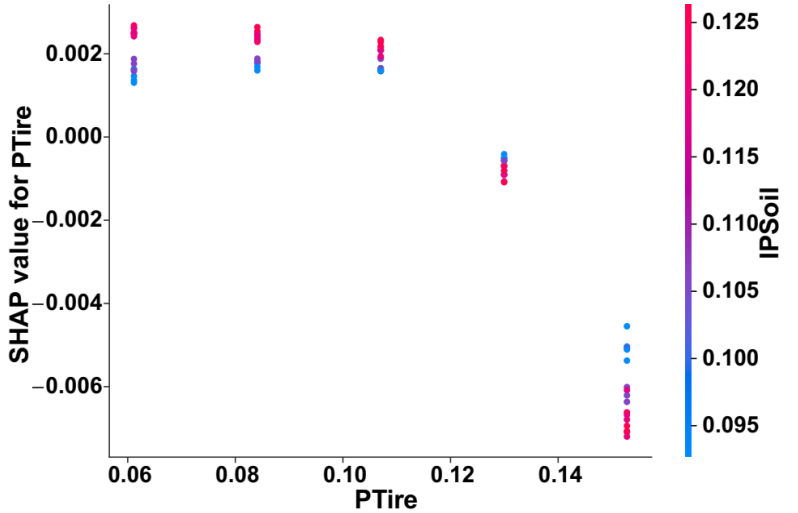
(i) (j)


SHAP value for FRr
FRr
AStc
SHAP value for IPSoil
IPSoil
PTire

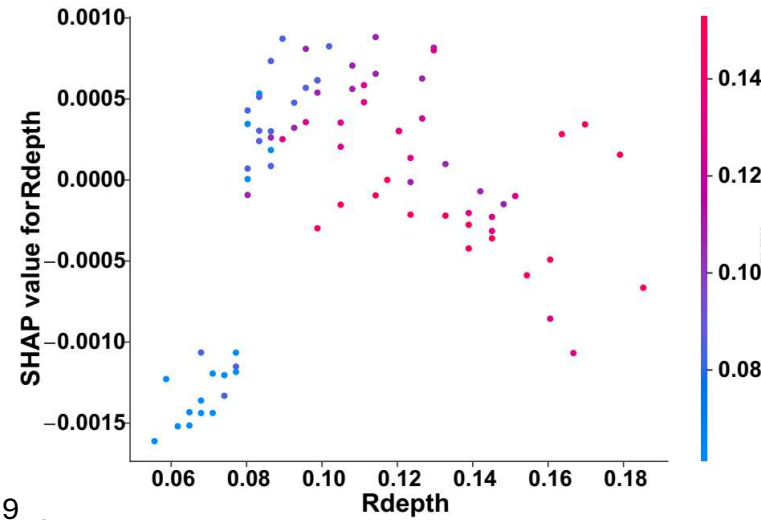
(k) (l)



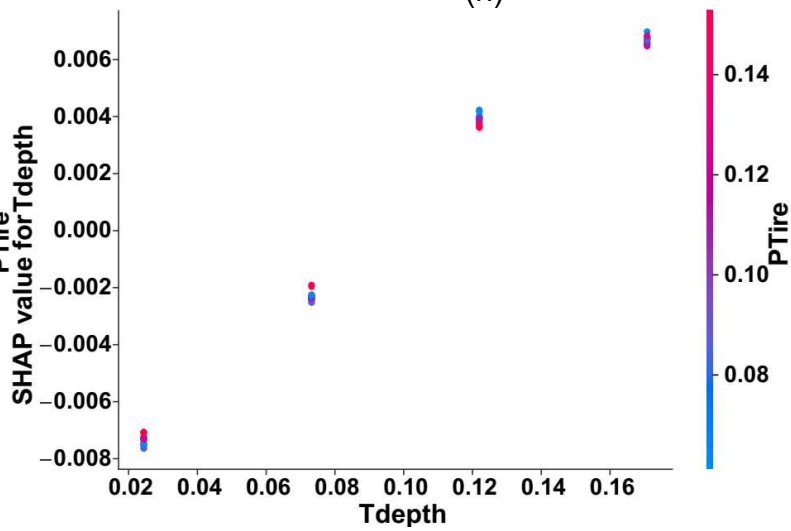

978 (m)



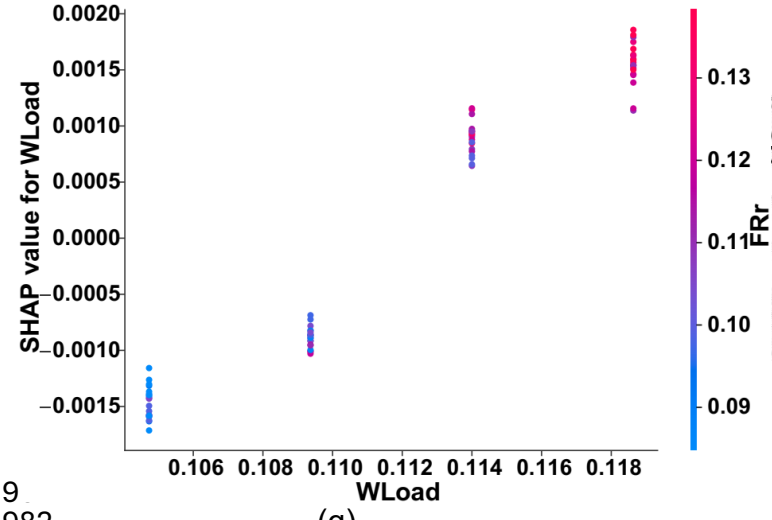

(n)



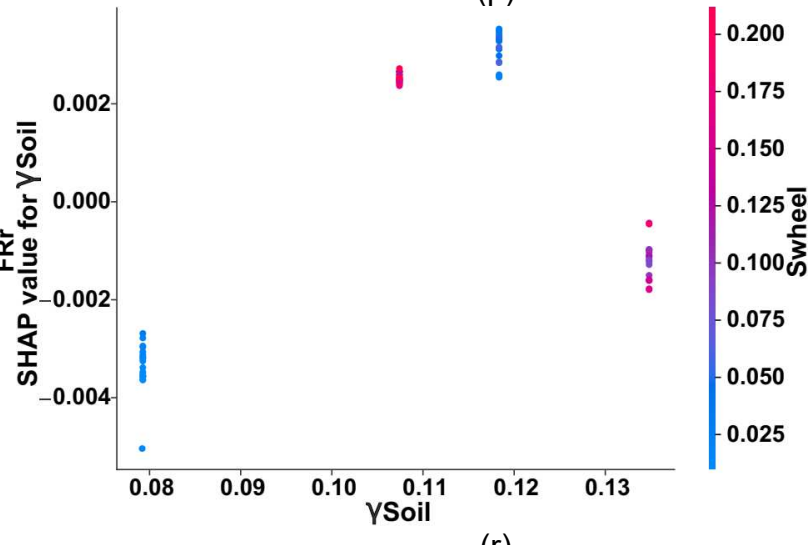

9
980 (o)



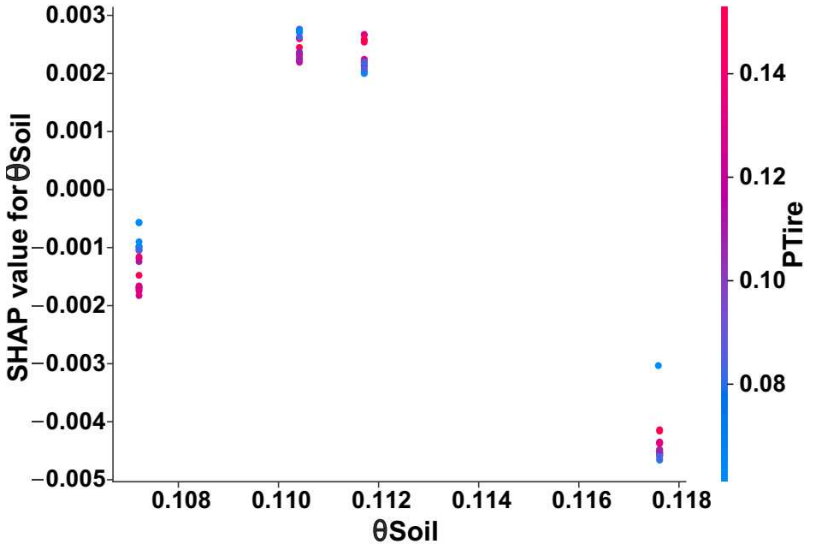

(p)




9
982 (q)




(r)




(s)

**Fig. 15.** SHAP dependency analysis for the six dominant features, coloured by the values of strongly interacting features. The plots reveal the influence of a feature value (x-axis) on the EOA-CatBoost prediction (y-axis: SHAP value). SHAP values (Y<0) and Y>0 indicate low and high predictions, respectively. Each point represents a single sample, and its color indicates the value of the second feature (secondary y-axis) to help visualize their multi-partite interaction. i.e., Effects of $H_{Db}$ prediction outcome by SHAP values (primary Y-axis) as affected by various model features.

### 3.6 Decision path and attribution clustering of the EAO-CatBoost model

The SHAP decision plot was adopted to visualize how a complex hybrid CatBoost model arrives at predictions by plotting the cumulative sum of SHAP values, showing how individual features move a prediction away from the model's base value (average prediction) towards the final prediction. SHAP decision path and attribution clustering pattern revealed heterogeneous decision making by the EAO-CatBoost model, providing more interpretable visualization of the modelling"path" that a data point takes through the drawbar predictions step by step, from the baseline (average prediction) to the final output (Fig. 16). The EAO-CatBoost decision path encoding showed consistent structural partitions defined by shared traversal paths. In contrast, attribution clustering revealed distinct patterns of feature attribution. Feature values are provided for reference next to the prediction lines. Each line represents one sample of the testing datasets. The prediction by a single sample depicted how the model arrived at the final prediction result from the baseline value through the contributions of each feature and the cumulative influence of each process variable on the output value of SHAP (predicted concentration value), providing typical insights into the cumulative effect of the interaction and prediction path. Fig. 16a shows how each feature shifted the prediction up or down relative to the average prediction, and the

decision plot shows the step-by-step cumulative contribution along the prediction path. Fig. 16b shows the feature-cluster heat map summarizing SHAP values across all prediction instances and illustrates the contribution of each feature to the prediction in each instance. Positive and negative prediction impacts are indicated in red and blue colors, respectively. The clustering heatmap (Fig.16b) revealed two distinct model-cluster attribution patterns. The first positively clustered samples (0-30) were associated with positive SHAP contributions from $F_{Tr}$, corresponding to positive $H_{Db}$ predictions ($f(x) > 0$), while samples 38–80 (second cluster) were characterized by negative SHAP contributions from $F_{Tr,}$ corresponding to negative $H_{Db}$ predictions ($f(x) < 0$). This means that whenever $F_{Tr}$ has a negative SHAP value, the hybrid Cat-Boost ensemble model relies on the positive SHAP values from $\gamma_{Soil}$, $A_{Stc}$, $Ø_{Fuel}$, $F_{Rr}$, $\theta_{Soil}$, $P_{Tire}$, to enhance predictions of $H_{Db}$. This shows their significant role in the model output. Another notable observation is that $W_{Load}$, when present, consistently contributes positively to the $H_{Db}$ prediction, albeit with low impact. In contrast, $R_{depth}$ has no significant positive or negative impact on the predictions. At the same time, whenever $\gamma_{Soil}$, $A_{Stc}$, $Ø_{Fuel}$, $F_{Rr}$, $\theta_{Soil}$, $P_{Tire}$ have a negative or positive contribution, the $F_{Tr}$, $IP_{Soil}$, $CI_{Soil}$, $\tau_{Shear}$, $T_{depth,}$ and $S_{wheel}$ always contribute positively with SHAP values changing accordingly. This pattern confirms the dynamic heterogeneity of soil and machine variables. However, $F_{Tr}$ shows the most pronounced, deep red and blue colour intensity variations, indicating the largest magnitudes and substantial differences across prediction instances and samples. At the same time, IPSoil, CISoil, τShear, Tdepth, and Swheel showed a mix of marginal pale pink on the left and light blue to blue on the right, consistent with their non-linear inverse relationships, where low instance values are associated with a positive contribution. The $\gamma_{Soil}$, $A_{Stc}$, $Ø_{Fuel}$, $F_{Rr}$, $\theta_{Soil}$, $P_{Tire}$, and $W_{Load}$ contributions are predominantly pale blue across all prediction instances, indicating minimal overall impact on model output. In contrast, $R_{depth}$ did not impact the predictions. Overall, $F_{Tr}$, $IP_{Soil}$, $CI_{Soil}$, $\tau_{Shear}$, $T_{depth,}$ and $S_{Wheel}$ exhibited distinct and strongly varying impacts across prediction instances. At the same time, $\gamma_{Soil}$, $A_{Stc}$, $Ø_{Fuel}$, $F_{Rr}$, $\theta_{Soil}$, and $P_{Tire}$ had a more moderate and consistent effect. $F_{Tr}$ was the single most significant feature contributing to the model's prediction output, while $R_{depth}$ had no significant contribution to the $H_{Db}$ prediction instances. Together, the decision path and attribution clustering visualizations enabled a comprehensive understanding of the dominant predictors that governed both the global and the sample-specific prediction behavior, thereby strengthening the interpretability of EAO-CatBoost modelling.

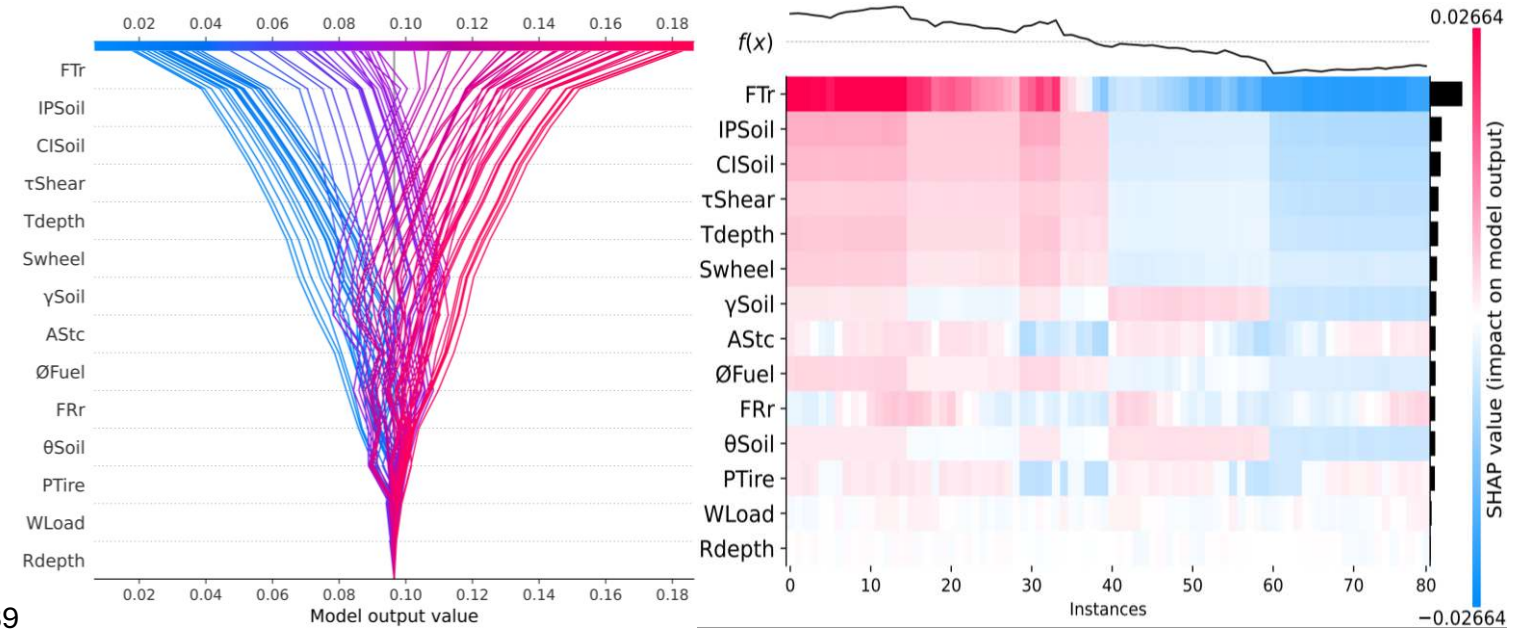



**Fig. 16.** Drawbar prediction decision paths for 80 representative samples centered, and attribution clustering heat map (a) X-axis represents the model's output. At the same time, the Y-axis lists the model's features ordered by descending importance computed over the plotted observations. Each observation's prediction is represented by a colored path starting from the base (E[f(x)]) value at the bottom of the x-axis to strike the model's output value f(x) at the top, with the colour bar line representing the magnitude and value of the final prediction. At the same time SHAP values for each feature are cumulatively added to the model's base value to show contribution of each feature to the overall prediction whilst the observations converge at expected value at the bottom of the plot, and (b) Feature clustering heatmap with colors representing the directional magnitude of SHAP values (red for positive, blue for negative), intuitively displaying attribution patterns and clearly defining distinct sample clusters that negatively of positively affect drawbar pull. The heat map curve shows the corresponding sample model raw output f(x).

### 3.7 Graphical user interface for drawbar load prediction using hybrid EAO-CatBoost

The hybrid EAO-CatBoost model demonstrated superior, robust predictive performance compared to the other evaluated models. To enhance the practical application of this study, EAO-CatBoost was coded into a standalone user-friendly application for predicting drawbar pull using PySide6 desktop graphical user interface (GUI) framework in Python. To optimize GUI performance for a user-friendly experience, the normalized database variables that facilitated model training were subjected to post-processing denormalization using Eq. 35. This allows users of the interactive GUI application to easily input variables using their easily measurable units to obtain drawbar pull in KN without transformation of variable units for rapid user-friendly

response and experience. At the same time, the EAO-CatBoost GUI enables standalone output visualization devoid of manual computations, as shown in Fig. 17.

$$H_{Db_{User}} = H_{normalised\ Pred} x \left\| H_{All_dabase} \right\|_2 \text{ and } \left\| H_{database} \right\|_2 = \sqrt{\sum_{i=1}^{n} H_{Dbi}^2} \quad (35)$$

The EAO-CatBoost graphical user interface exhibited strong predictive capability, with R = 0.999, $R^2$ = 0.998, RMSE = 0.0385, MAE = 0.0017, MAPE = 0.0026, RMSRE = 0.0385, and RRMSE = 0.0219 during testing with new unseen datasets, validating its effectiveness for field applications in practice.

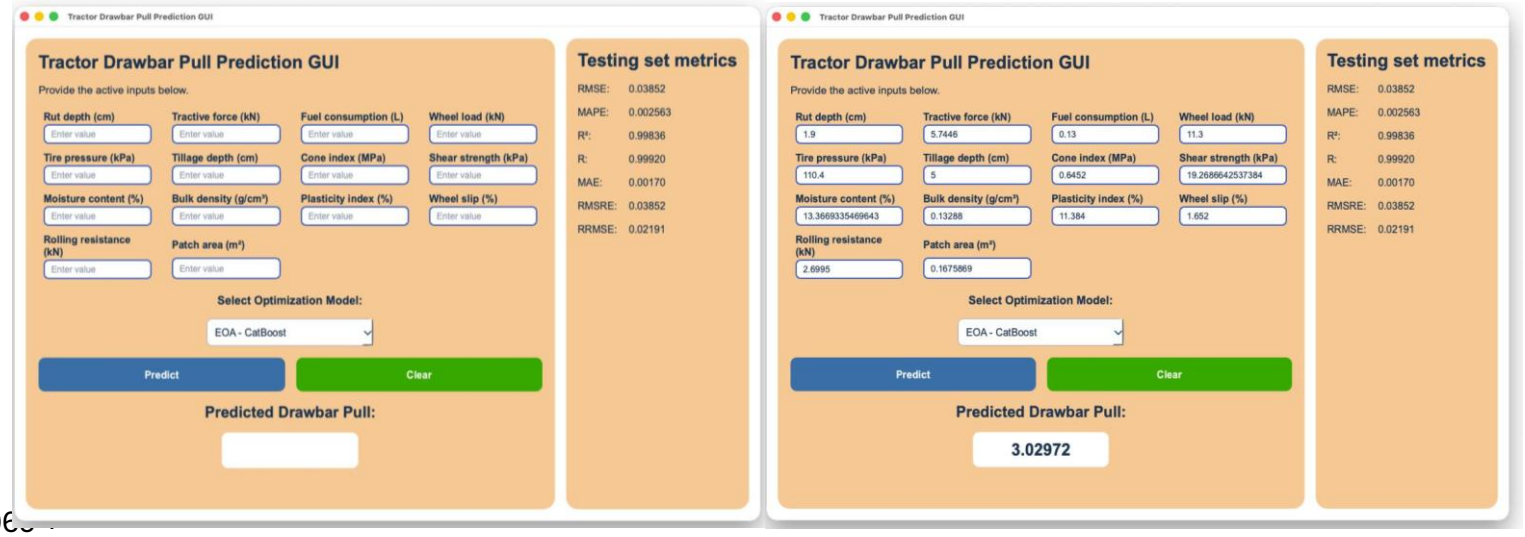



**Fig. 17**. User-friendly EAO-CatBoost graphical user interface for predicting drawbar pull.

### 3.8 Discussion

Among the 26 models, the Base-CatBoost model achieved high accuracy and lower deviation error metrics during training, but was outperformed by the Hybrid CatBoost models during testing. Specifically, EOA-CatBoost offered the best accuracy, robustness, stability, and generalization on the validation set. This outcome indicates that while Base-CatBoost models may fit the training data closely, adaptive hybridization with EAO methods performs better at handling variability in unseen data. Additionally, compared to the base learner, models using EAO nature-inspired algorithms to optimize the Base-CatBoost learner perform better at detecting, mapping, and capturing complex relationships and variabilities in unseen datasets across all instances, which is necessary for accurate predictions. This is because of EAO's highly adaptive iteration and fine-tuning mechanisms, its high-performance convergence, and its superior feature selection for dynamically simulating the catalytic behaviour of enzyme–substrate interactions, which balance exploration and exploitation (refining the best solution) and are suitable for machine learning and engineering optimization tasks. As such, the developed state-

of-the-art hybrid EAO-CatBoost model outperformed all the nature-inspired algorithm models in boosting the performance of the conventional Base-CatBoost learner and forms the best hybrid CatBoost model for drawbar predictions. The hybrid EAO-CatBoost model superiorly outperformed other nature-inspired algorithms by integrating the high-dimensional global search efficiency of the Enzyme Action Optimizer with the strong categorical feature handling framework of the CatBoost gradient boosting library, resulting in high predictive accuracy, robustness, and generalization. Thus, hybridizing the standalone Base-CatBoost learner with the recently developed EAO-nature-inspired algorithm significantly improved its cognitive performance for off-road engineering applications in drawbar load predictions. Based on the accuracy evaluation matrix, overfitting metrics, multi-metric score ranking, error characteristic curve, performance scatter, regression analysis, prediction deviation, robustness and generalization, as well as prediction error residuals and Taylor analysis, the EAO-CatBoost was best suited for drawbar predictions. The Hybrid EAO-CatBoost showed the best generalized cognitive ability among other models, on unseen validation datasets. Although Base-CatBoost performed almost perfectly during training, it produced lower validation metrics due to poor generalization and was a concern of overfitting. This highlights the need to evaluate hybrid neurocomputing models beyond training accuracy metrics using robustness and generalization metrics. Furthermore, cognitive interpretation of the EAO-CatBoost model predictions using Shapley Additive exPlanations revealed $F_{Tr}$ as the most significant feature contributing to $H_{Db}$ variability and dispersion, followed by $IP_{Soil}$, $CI_{Soil}$, $T_{Shear}$, $T_{depth,}$ and $S_{Wheel}$. These meaningful findings align well with off-road soil-machine dynamics and agricultural drawbar load mechanics in tillage and traction that fundamentally govern drawbar pull generation and control for practical applications in tillage operations and off-road locomotion.

## 4. Conclusion

This investigation evaluated the cognitive performance of hybrid CatBoost ensemble models and the Base-CatBoost learner in predicting drawbar load of agricultural tractors. Comparatively, hybrid-CatBoost models outperformed the standalone Base-CatBoost learner across evaluation measures on unseen data during testing, indicating that hyperparameter optimization using nature-inspired algorithms significantly improved the Base-CatBoost learner's predictive reliability. Further, hybrid EAO-CatBoost showed superior predictive performance when tested compared to the Base-CatBoost learner, highlighting the benefit of integrating CatBoost into a bio-inspired hyperparameter optimization framework. However, the Base-CatBoost learner algorithm performed best during training, demonstrating its powerful gradient-boosting

capabilities for handling nonlinear interactions and complex relationships between soil-machine features and the target $H_{Db}$ variable using the training data. The observed weakness of the base CatBoost learner during testing indicates that it uses default hyperparameters that memorize the training data rather than fully leverage heuristic tuning to exploit its predictive structure in an unseen dataset. Conversely, the superior robustness and generalization associated with EAO-CatBoost indicate that the hybrid optimization process heuristically improved the bias-variance trade-off by selecting appropriate combinations of learning rate, tree depth, regularization, and iteration-related parameters. Although the Base-CatBoost is known for excellent gradient-boosting performance, the EAO nature-inspired optimizer fine-tuned and improved the Base-CatBoost learner beyond the default hyperparameter settings for the training set. This improvement further underscores that the effective comparative analysis of machine learning models should be interpreted critically, not only to include accuracy but also to balance robustness-generalization trade-offs with overfitting risks. The superiority of the hybrid models was not only demonstrated by the lowest values of error-based metrics RMSE, MAE, MAPE, NRMSE, RRMSE, SI, SE, sMAPE, WMAPE, MARE, RMSRE, medAE, CV, and PI, but also they demonstrated better agreement based goodness of fit indicators R, $R^2$, VAF, score rank, lowest values of robustness and overfitting metrics including MSE, RMSE, MAE, medAE, Lδ, LCL, and lowest G-gap, PS, and cRMSD Taylor diagram proximity. Overall, the Hybrid EAO-CatBoost achieved the highest stable ranking across accuracy and robustness metrics, the best residual characteristic curves, the lowest G-gap, the best PS, and the closest Taylor diagram agreement compared to all the other 24 hybrid models and the Base-CatBoost. This study shows that improvements in the accuracy alone for hybrid machine learning models over the base learner are not sufficient to determine their superiority. Instead, the superiority gains need to be evaluated with respect to stable validation behaviour, decreased residual dispersion, reduced generalization gaps, and increased agreement with the observed data. Thus, the most defensible hybrid model (EAO-CatBoost) improved testing accuracy while maintaining numerical stability, model parsimony, and generalization reliability compared to the Base-CatBoost learner. EAO-CatBoost achieved a good trade-off between accuracy and robustness in estimating $H_{Db}$ during training and attained the highest accuracy, robustness, and generalization during validation on unseen datasets compared to the base learner (Base-CatBoost). This work integrated the stand-alone CatBoost learner with the enzyme action optimizer and evaluated the EAO-CatBoost model for both local (individual) and global interpretability using SHAP to gain clear insights into the main drivers of drawbar pull variability and their cognitive modelling mechanism. The validity of the insights and the explanatory reliability of the EAO-CatBoost

model are consistent with known soil-machine dynamics in off-road traction, and the model is capable of making accurate, robust, and explainable predictions. The developed hybrid CatBoost framework is recommended for integration into intelligent off-road decision-support systems and smart machinery platforms. Further, the deployment of the user-friendly EAO-CatBoost GUI will enable data-driven, intelligent operational adjustments of autonomous off-road machinery for adaptive control and optimized drawbar load utilization.

## CRediT authorship contribution statement

**F.M:** Research conceptualization, Methodology, Data collection and Curation, Formal analysis and Visualization, Software, modelling and graphics, Writing original manuscript, Literature analysis and synthesis, Validation, Comprehensive revision, Final editing & Proofreading. **A.G:** Supervision, Research administration, Proofreading & Research mentorship. **D.M:** Supervision, Proofreading & Research mentorship.


## Funding and acknowledgements

This research received no external funding. Machinery adjustments and experimental operations were supported by Muliro, B. A., & Haron, N. O. from the University of Nairobi.

The authors gratefully acknowledge the Editor and anonymous reviewers for their rigorous evaluation, insightful comments, and constructive feedback, which substantially enhanced the engineering quality, clarity, and scientific rigor of this work.


## Data availability

The datasets generated during and/or analysed during the current study can be obtained from the corresponding author upon reasonable request

## Declarations

## Ethics approval consent to participate

Not applicable. This study did not involve human participants, human data, or animal experiments The authors declare no conflicting interests

## Consent for publication

Not applicable.

## Competing interest

The authors declare no competing interests

## References

1. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst [Internet]. Curran Associates, Inc.; 2018 [cited 2025 Mar 6]. https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html. Accessed 6 Mar 2025

2. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data [Internet]. 2020 [cited 2025 Mar 7];7:94. https://doi.org/10.1186/s40537-020-00369-8

3. Dutta J, Roy S. OccupancySense: Context-based indoor occupancy detection & prediction using CatBoost model. Appl Soft Comput [Internet]. 2022 [cited 2026 Apr 1];119:108536. https://doi.org/10.1016/j.asoc.2022.108536

4. Bhih M, Elassad ZEA, Siti H, Meslouhi OE, Abouelmehdi K, Boustani AE. Temporal pattern-aware temperature forecasting using CatBoost: A hybrid machine learning approach. Results Eng [Internet]. 2026 [cited 2026 Apr 1];30:110212. https://doi.org/10.1016/j.rineng.2026.110212

5. Larionov A, Adams NM, Webster KN. Investigating the impact of missing value handling on Boosted trees and Deep learning for Tabular data: A Claim Reserving case study. Trans Mach Learn Res [Internet]. 2024 [cited 2026 Apr 1]; https://openreview.net/forum?id=aV6dCg1VFV. Accessed 1 Apr 2026

6. Elshaarawy MK. Metaheuristic-driven CatBoost model for accurate seepage loss prediction in lined canals. Multiscale Multidiscip Model Exp Des [Internet]. 2025 [cited 2026 Apr 1];8:235. https://doi.org/10.1007/s41939-025-00800-8

7. Elmasry NH, Elshaarawy MK. Hybrid metaheuristic optimized Catboost models for construction cost estimation of concrete solid slabs. Sci Rep [Internet]. 2025 [cited 2026 Apr 1];15:21612. https://doi.org/10.1038/s41598-025-06380-4

8. Hussain J, Fu X, Chen J, Ali N, Iqbal SM, Hussain W, et al. Estimation of rock strength parameters from petrological contents using tree-based machine learning techniques. AI Civ Eng [Internet]. 2025 [cited 2026 Apr 1];4:4. https://doi.org/10.1007/s43503-024-00047-1

9. Ruan B, Hu Y, Liu C, Zhou Z. Prediction of compression modulus of Nantong fine-grained soil based on an interpretability-driven GSA-optimized CatBoost model. Adv Eng Inform [Internet]. 2026 [cited 2026 Apr 1];69:104034. https://doi.org/10.1016/j.aei.2025.104034

10. Eini N, Janizadeh S, Bateni SM, Jun C, Heggy E. Scour depth estimation using standalone metaheuristic algorithms and their combinations with CatBoost. Sci Rep [Internet]. 2025 [cited 2026 Apr 1];15:37239. https://doi.org/10.1038/s41598-025-21120-4

11. Johnson JM, Khoshgoftaar TM. Deep Learning and Data Sampling with Imbalanced Big Data. 2019 IEEE 20th Int Conf Inf Reuse Integr Data Sci IRI [Internet]. 2019 [cited 2025 Mar 7]. p. 175–83. https://doi.org/10.1109/IRI.2019.00038

12. Zhang C, Feng F, Zhang C, Li S, Xie J. Optimizing Intermittent Pumping Duration with a Physics–Data Dual-Driven CatBoost Model Enhanced by Bayesian and Attention Mechanisms. Processes [Internet]. 2025 [cited 2026 Apr 1];13:4012. https://doi.org/10.3390/pr13124012

13. Hussein BM, Shareef SM. An Empirical Study on the Correlation between Early Stopping Patience and Epochs in Deep Learning. Kattan A, editor. ITM Web Conf [Internet]. 2024 [cited 2026 Apr 1];64:01003. https://doi.org/10.1051/itmconf/20246401003

14. A Ilemobayo J, Durodola O, Alade O, J Awotunde O, T Olanrewaju A, Falana O, et al. Hyperparameter Tuning in Machine Learning: A Comprehensive Review. J Eng Res Rep [Internet]. 2024 [cited 2026 Apr 1];26:388–95. https://doi.org/10.9734/jerr/2024/v26i61188

15. Sujon KM, Hassan R, Choi K, Samad MA. Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. J Big Data [Internet]. 2025 [cited 2026 Mar 22];12:268. https://doi.org/10.1186/s40537-025-01313-4

16. Schober P, Boer C, Schwarte LA. Correlation Coefficients: Appropriate Use and Interpretation. Anesth Analg [Internet]. 2018 [cited 2026 Apr 1];126:1763–8. https://doi.org/10.1213/ANE.0000000000002864

17. Ding W, Sun L, Qi Z, Li S, Filipović V, Wu X, et al. Conservation tillage enhances both organic and inorganic carbon in dryland: Insights from a 20-year field experiment and meta-analysis. Agric Ecosyst Environ [Internet]. 2025 [cited 2026 Apr 1];393:109845. https://doi.org/10.1016/j.agee.2025.109845

18. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat [Internet]. 2001 [cited 2026 Apr 1];29. https://doi.org/10.1214/aos/1013203451

19. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support [Internet]. arXiv; 2018 [cited 2026 Apr 1]. https://doi.org/10.48550/ARXIV.1810.11363

20. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features [Internet]. arXiv; 2019 [cited 2026 Jan 13]. https://doi.org/10.48550/arXiv.1706.09516

21. Fakhouri HN, Hudaib A, Fakhouri S, Hamad F, Alkhalaileh M. Aurora Intelligent Metaheuristic: A Novel Space-Inspired Optimizer. SN Comput Sci [Internet]. 2025 [cited 2026 Jan 13];6:971. https://doi.org/10.1007/s42979-025-04512-1

22. Xiao Y, Cui H, Khurma RA, Castillo PA. Artificial lemming algorithm: a novel bionic meta-heuristic technique for solving real-world engineering optimization problems. Artif Intell Rev [Internet]. 2025 [cited 2026 Jan 13];58:84. https://doi.org/10.1007/s10462-024-11023-7

23. Benmamoun Z, Khlie K, Bektemyssova G, Dehghani M, Gherabi Y. Bobcat Optimization Algorithm: an effective bio-inspired metaheuristic algorithm for solving supply chain optimization problems. Sci Rep [Internet]. 2024 [cited 2026 Jan 13];14:20099. https://doi.org/10.1038/s41598-024-70497-1

24. Pan Q, Tang J, Zhan J, Li H. Bacteria phototaxis optimizer. Neural Comput Appl [Internet]. 2023 [cited 2026 Jan 13];35:13433–64. https://doi.org/10.1007/s00521-023-08391-6

25. Meng X, Liu Y, Gao X, Zhang H. A New Bio-inspired Algorithm: Chicken Swarm Optimization. In: Tan Y, Shi Y, Coello CAC, editors. Adv Swarm Intell [Internet]. Cham: Springer International Publishing; 2014 [cited 2026 Jan 13]. p. 86–94. https://doi.org/10.1007/978-3-319-11857-4_10

26. Rodan A, Al-Tamimi A-K, Al-Alnemer L, Mirjalili S, Tiňo P. Enzyme action optimizer: a novel bio-inspired optimization algorithm. J Supercomput [Internet]. 2025 [cited 2026 Jan 13];81:686. https://doi.org/10.1007/s11227-025-07052-w

27. Dehghani M, Aly M, Rodriguez J, Sheybani E, Javidi G. A Novel Nature-Inspired Optimization Algorithm: Grizzly Bear Fat Increase Optimizer. Biomimetics [Internet]. 2025 [cited 2026 Jan 13];10:379. https://doi.org/10.3390/biomimetics10060379

28. Agushaka JO, Ezugwu AE, Abualigah L. Gazelle optimization algorithm: a novel nature-inspired metaheuristic optimizer. Neural Comput Appl [Internet]. 2023 [cited 2026 Feb 28];35:4099–131. https://doi.org/10.1007/s00521-022-07854-6

29. Mirjalili S, Mirjalili SM, Lewis A. Grey Wolf Optimizer. Adv Eng Softw [Internet]. 2014 [cited 2026 Jan 13];69:46–61. https://doi.org/10.1016/j.advengsoft.2013.12.007

30. Heidari AA, Mirjalili S, Faris H, Aljarah I, Mafarja M, Chen H. Harris hawks optimization: Algorithm and applications. Future Gener Comput Syst [Internet]. 2019 [cited 2026 Jan 13];97:849–72. https://doi.org/10.1016/j.future.2019.02.028

31. Amiri MH, Mehrabi Hashjin N, Montazeri M, Mirjalili S, Khodadadi N. Hippopotamus optimization algorithm: a novel nature-inspired optimization algorithm. Sci Rep [Internet]. 2024 [cited 2026 Jan 13];14:5032. https://doi.org/10.1038/s41598-024-54910-3

32. Chou J-S, Truong D-N. A novel metaheuristic optimizer inspired by behavior of jellyfish in ocean. Appl Math Comput [Internet]. 2021 [cited 2026 Jan 13];389:125535. https://doi.org/10.1016/j.amc.2020.125535

33. Sánchez Cortez JA, Peraza Vázquez H, Peña Delgado AF. A Novel Bio-Inspired Optimization Algorithm Based on Mantis Shrimp Survival Tactics. Mathematics [Internet]. 2025 [cited 2026 Jan 13];13:1500. https://doi.org/10.3390/math13091500

34. Dehghani M, Trojovský P. Osprey optimization algorithm: A new bio-inspired metaheuristic algorithm for solving engineering optimization problems. Front Mech Eng [Internet]. 2023 [cited 2026 Jan 13];8:1126450. https://doi.org/10.3389/fmech.2022.1126450

35. Lian J, Hui G, Ma L, Zhu T, Wu X, Heidari AA, et al. Parrot optimizer: Algorithm and applications to medical problems. Comput Biol Med [Internet]. 2024 [cited 2026 Jan 13];172:108064. https://doi.org/10.1016/j.compbiomed.2024.108064

36. Al-Baik O, Alomari S, Alssayed O, Gochhait S, Leonova I, Dutta U, et al. Pufferfish Optimization Algorithm: A New Bio-Inspired Metaheuristic Algorithm for Solving Optimization Problems. Biomimetics [Internet]. 2024 [cited 2026 Jan 13];9:65. https://doi.org/10.3390/biomimetics9020065

37. Chaudhary G, Rawal BS. A novel reindeer cyclone optimization algorithm (RCOA). Sci Rep [Internet]. 2025 [cited 2026 Jan 13];15:12506. https://doi.org/10.1038/s41598-025-97069-1

38. Fu Y, Liu D, Chen J, He L. Secretary bird optimization algorithm: a new metaheuristic for solving global optimization problems. Artif Intell Rev [Internet]. 2024 [cited 2026 Jan 13];57:123. https://doi.org/10.1007/s10462-024-10729-y

39. Kareem AA, Abid AJ, Hammood DA, Yaqoob SJ, Husein A, Bereznychenko V. Swift Flight Optimizer: a novel bio-inspired optimization algorithm based on swift bird behavior. Sci Rep [Internet]. 2025 [cited 2026 Jan 13];15:41597. https://doi.org/10.1038/s41598-025-25490-7

40. Hashim FA, Hussien AG. Snake Optimizer: A novel meta-heuristic optimization algorithm. Knowl-Based Syst [Internet]. 2022 [cited 2026 Jan 13];242:108320. https://doi.org/10.1016/j.knosys.2022.108320

41. Trojovský P, Dehghani M. A new bio-inspired metaheuristic algorithm for solving optimization problems based on walruses behavior. Sci Rep [Internet]. 2023 [cited 2026 Jan 13];13:8775. https://doi.org/10.1038/s41598-023-35863-5

42. Das S, Baksi S, Ghosh P. Warthog Optimization Algorithm: A New Nature-Derived Solution for Optimization Challenges. J Transform Technol Sustain Dev [Internet]. 2025 [cited 2026 Jan 13];9:4. https://doi.org/10.1007/s41314-025-00075-4

43. Azizi M, Baghalzadeh Shishehgarkhaneh M, Basiri M, Moehler RC, Fang Y, Chan M. Wolf-Bird Optimizer (WBO): A novel metaheuristic algorithm for Building Information Modeling-based resource tradeoff. J Eng Res [Internet]. 2025 [cited 2026 Jan 13];13:763–85. https://doi.org/10.1016/j.jer.2023.11.024

44. Mirjalili S, Lewis A. The Whale Optimization Algorithm. Adv Eng Softw [Internet]. 2016 [cited 2026 Jan 13];95:51–67. https://doi.org/10.1016/j.advengsoft.2016.01.008

45. Braik M, Hammouri A, Atwan J, Al-Betar MA, Awadallah MA. White Shark Optimizer: A novel bio-inspired meta-heuristic algorithm for global optimization problems. Knowl-Based Syst [Internet]. 2022 [cited 2026 Jan 13];243:108457. https://doi.org/10.1016/j.knosys.2022.108457

46. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci [Internet]. 2021 [cited 2026 Apr 1];7:e623. https://doi.org/10.7717/peerj-cs.623

47. Jierula A, Wang S, Oh T-M, Wang P. Study on Accuracy Metrics for Evaluating the Predictions of Damage Locations in Deep Piles Using Artificial Neural Networks with Acoustic Emission Data. Appl Sci [Internet]. 2021 [cited 2026 Apr 1];11:2314. https://doi.org/10.3390/app11052314

48. Khoshvaght H, Permala RR, Razmjou A, Khiadani M. A critical review on selecting performance evaluation metrics for supervised machine learning models in wastewater quality prediction. J Environ Chem Eng [Internet]. 2025 [cited 2026 Apr 1];13:119675. https://doi.org/10.1016/j.jece.2025.119675

49. Javed MF, Khan M, Fawad M, Alabduljabbar H, Najeh T, Gamil Y. Comparative analysis of various machine learning algorithms to predict strength properties of sustainable green concrete containing waste foundry sand. Sci Rep [Internet]. 2024 [cited 2026 Jan 14];14:14617. https://doi.org/10.1038/s41598-024-65255-2

50. Asgharzadeh-Bonab A, Bijanvand S, Parsaie A, Afaridegan E. Machine learning-based estimation of discharge coefficient for semicircular labyrinth weirs. Sci Rep [Internet]. 2025 [cited 2026 Jan 14];15:33002. https://doi.org/10.1038/s41598-025-18230-4

51. Hodson TO. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci Model Dev [Internet]. 2022 [cited 2026 May 28];15:5481–7. https://doi.org/10.5194/gmd-15-5481-2022

52. Tong H. Functional linear regression with Huber loss. J Complex [Internet]. 2023 [cited 2026 Apr 5];74:101696. https://doi.org/10.1016/j.jco.2022.101696

53. Simsoba K-A, Oscar N, Mageto T. A Hybrid Minkowski-Log-Cosh Loss Function for Robust LSTM-Based Time Series Forecasting. IEEE Access [Internet]. 2025 [cited 2026 Apr 5];13:187307–19. https://doi.org/10.1109/ACCESS.2025.3626795

54. Terven J, Cordova-Esparza D-M, Romero-González J-A, Ramírez-Pedraza A, Chávez-Urbiola EA. A comprehensive survey of loss functions and metrics in deep learning. Artif Intell Rev [Internet]. 2025 [cited 2026 Apr 5];58:195. https://doi.org/10.1007/s10462-025-11198-7

55. Fieldsend JE. Regression Error Characteristic Optimisation of Non-Linear Models. In: Jin Y, editor. Multi-Object Mach Learn [Internet]. Berlin, Heidelberg: Springer; 2006 [cited 2026 Mar 31]. p. 103–23. https://doi.org/10.1007/3-540-33019-4_5

56. Zou KH, O'Malley AJ, Mauri L. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. Circulation [Internet]. American Heart Association; 2007 [cited 2026 Mar 31];115:654–7. https://doi.org/10.1161/CIRCULATIONAHA.105.594929

57. Çorbacıoğlu ŞK, Aksel G. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. Turk J Emerg Med [Internet]. 2023 [cited 2026 Mar 31];23:195–8. https://doi.org/10.4103/tjem.tjem_182_23

58. Ballester R, Clemente XA, Casacuberta C, Madadi M, Corneanu CA, Escalera S. Predicting the generalization gap in neural networks using topological data analysis. Neurocomputing [Internet]. 2024 [cited 2026 May 9];596:127787. https://doi.org/10.1016/j.neucom.2024.127787

59. Wang J, Wu Z, Lu M, Ai J. An Empirical Study on the Effect of Training Data Perturbations on Neural Network Robustness. Sensors [Internet]. 2024 [cited 2026 May 9];24:4874. https://doi.org/10.3390/s24154874

60. Padró-Ferragut C, Martínez-Plumed F, Ramírez-Quintana MJ. Robustness under noise: assessing the impact of perturbed key attributes on machine learning models. Int J Data Sci Anal [Internet]. 2026 [cited 2026 May 9];22:71. https://doi.org/10.1007/s41060-025-01006-4

61. Wadoux AMJ-C, Walvoort DJJ, Brus DJ. An integrated approach for the evaluation of quantitative soil maps through Taylor and solar diagrams. Geoderma [Internet]. 2022 [cited 2026 Apr 1];405:115332. https://doi.org/10.1016/j.geoderma.2021.115332

62. Taylor KE. Summarizing multiple aspects of model performance in a single diagram. J Geophys Res Atmos [Internet]. 2001 [cited 2026 Apr 1];106:7183–92. https://doi.org/10.1029/2000JD900719

63. Moj M, Czarnecki S. Comparative analysis of selected machine learning techniques for predicting the pull-off strength of the surface layer of eco-friendly concrete. Adv Eng Softw [Internet]. 2024 [cited 2026 Jan 14];195:103710. https://doi.org/10.1016/j.advengsoft.2024.103710

64. Fouda T, Hegazy R, Alhamshary K. A simulation model to predict agricultural tractor-semi-trailer combination traction performance under different operating conditions. Sci Rep [Internet]. 2026 [cited 2026 Oct 2];16:13000. https://doi.org/10.1038/s41598-026-47522-6

65. Ally H, Wang X, Wu T, Liu T, Ge J. Improving Agricultural Tire Traction Performance Through Finite Element Analysis and Semi-Empirical Modeling. Eng [Internet]. 2025 [cited 2026 Oct 2];6:63. https://doi.org/10.3390/eng6040063

66. Yadav R, Raheman H. Machine learning-based estimation of agricultural tyre sinkage: A streamlit web application. J Terramechanics [Internet]. 2025 [cited 2026 Jan 12];119:101055. https://doi.org/10.1016/j.jterra.2025.101055

67. Jasoliya D, Untaroiu A, Untaroiu C. A review of soil modeling for numerical simulations of soil-tire/agricultural tools interaction. J Terramechanics [Internet]. 2024 [cited 2026 June 6];111:41–64. https://doi.org/10.1016/j.jterra.2023.09.003

68. Abrahám R, Majdan R, Drlička R. Possibilities of improving the wheel tractor drive force transmission to soil. Res Agric Eng [Internet]. 2015 [cited 2026 Jan 12];61:S37–42. https://doi.org/10.17221/26/2015-RAE

69. Battiato A, Diserens E. Tractor traction performance simulation on differently textured soils and validation: A basic study to make traction and energy requirements accessible to the practice. Soil Tillage Res [Internet]. 2017 [cited 2026 Jan 12];166:18–32. https://doi.org/10.1016/j.still.2016.09.005

70. Yadav R, Raheman H. Machine Learning-Based Estimation of Agricultural Tyre Sinkage Considering Wheel Slip, Drawbar Pull, Soil Strength, Normal Load, and Inflation Pressure: A Streamlit Web Application. In: Goswami S, Saha S, Basu K, Beed RS, editors. Data Manag Anal Innov [Internet]. Singapore: Springer Nature Singapore; 2026 [cited 2026 Jan 12]. p. 409–24. https://doi.org/10.1007/978-981-96-6534-1_25