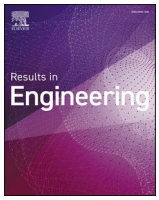
Results in Engineering 24 (2024) 103297

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering



# Monitoring of greenhouse gas emission drivers in Atlantic Canadian Potato production: A robust explainable intelligent glass-box

Mehdi Jamei [a,c,i,*], Muhammad Hassan [b], Aitazaz A. Farooque [a,b,*], Mumtaz Ali [d], Masoud Karbasi [a,e], Gurjit S. Randhawa [f], Zaher Mundher Yaseen [g,h], Ross Dwyer [a]

[a] Canadian Centre for Climate Change and Adaptation, University of Prince Edward Island, St Peters Bay, PE, Canada
[b] Faculty of Sustainable Design Engineering, University of Prince Edward Island, Charlottetown, PE, C1A 4P3, Canada
[c] Faculty of Civil Engineering and Architecture, Shahid Chamran University of Ahvaz, Ahvaz, Iran
[d] UniSQ College, University of Southern Queensland, QLD, 4350, Australia
[e] Water Engineering Department, Faculty of Agriculture, University of Zanjan, Zanjan, Iran
[f] School of Computer Science, University of Guelph, Guelph, ON, N1G 2W1, Canada
[g] Civil and Environmental Engineering Department, King Fahd University of Petroleum & Minerals, Dhahran, 31261, Saudi Arabia
[h] Interdisciplinary Research Center for Membranes and Water Security, King Fahd University of Petroleum & Minerals, Dhahran, 31261, Saudi Arabia
[i] New Era and Development in Civil Engineering Research Group, Scientific Research Center, Al-Ayen University, Thi-Qar, Nasiriyah, 64001, Iraq

ARTICLE INFO

*Keywords:*
Greenhouse gas emission
Precision agriculture
Runge-Kutta optimizer
Gradient-boosted decision tree
BSLR-WASPAS

ABSTRACT

In this research, a novel explainable multi-level ensemble learning framework has been developed to accurately monitor the greenhouse gas (GHG) emission drivers of the Atlantic Canada's potato crop system i.e., Carbon dioxide ($CO_2$), nitrous oxide ($N_2O$), and water vapour ($H_2O$). For this purpose, alongside the GHG emission drivers, the hydro-meteorological and soil properties information was collected from three Canadian sites, two in Prince Edward Island (PEI) and one in New Brunswick. This advanced framework includes a transparent multi-level pre-processing module and a Runge-Kutta optimizer (RUN), integrated with an eXplainable gradient-boosted decision Tree (GBDT) machine learning (ML) technique. The preprocessing scheme meticulously selects the most effective input combinations from the hydro-meteorological and soil properties datasets using hybridization of Boruta-GBDT for feature selection, Best Subset Lasso Regression (BSLR), and Weighted Aggregated Sum Product Assessment (WASPAS). The optimal combinations were then analyzed using the GBDT-RUN and compared against two algorithms: LightGBM coupled with RUN optimizer (LightGBM-RUN) and classical GBDT. The explainability of the primary model was enhanced using SHapley Additive exPlanations (SHAP). Model validation employed various metrics, such as the correlation coefficient (R), squared deviation (SquD), and a range of sophisticated statistical graphics. Results demonstrated that the GBDT-RUN model exhibited superior performance in monitoring GHG emissions ($CO_2|R = 0.8431$, SquD=17.1759, WASPAS=1.88E-07; $N_2O|R = 0.8431$, SquD=17.1759, WASPAS=1.88E-07; $H_2O|R = 0.8431$, SquD=17.1759, WASPAS=1.88E-07), outperforming both LightGBM-RUN and classical GBDT. Furthermore, the explainability analysis identified dew point and soil temperature as the most influential factors in the $CO_2$, $N_2O$, and $H_2O$ emissions scenarios.

## 1. Introduction

Anthropogenic greenhouse gas (GHG) emissions have increased significantly over the last several years as a result of changing land use, deforestation, and increased usage of fossil fuels. Around the world, there has been an increase in the temperature of the atmosphere. Global climate variability has severely impacted all spheres of life, making it necessary to study and comprehend future climate variability at a site-specific and regional scale to minimize and adapt to these changes [1]. The IPCC has highlighted the effects of greenhouse gas emissions, specifically those of carbon dioxide ($CO_2$), nitrous oxide ($N_2O$), troposphere ozone (O3), methane (CH4), and chlorofluorocarbons (CFCs) on climate change [2,3]. Over the past 50 years, average global GHG concentrations have increased more than in any other year [4]. Due to

* Corresponding authors.
*E-mail addresses:* jmehdi@upei.ca (M. Jamei), mhassan11153@upei.ca (M. Hassan), afarooque@upei.ca (A.A. Farooque), mumtaz.ali@unisq.edu.au (M. Ali), mkarbasi@upei.ca (M. Karbasi), randhawg@uoguelph.ca (G.S. Randhawa), zaheryaseen88@gmail.com (Z.M. Yaseen).

https://doi.org/10.1016/j.rineng.2024.103297
Received 2 September 2024; Received in revised form 25 October 2024; Accepted 2 November 2024
Available online 4 November 2024

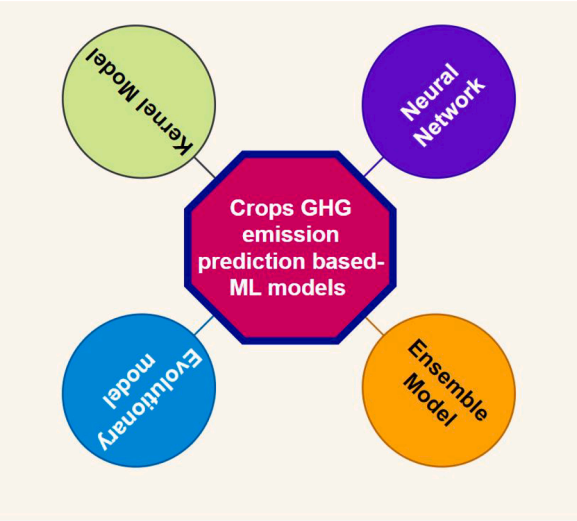



**Fig. 1.** The primary ML-based model's versions applied for crops GHG emissions; (i) neural network: e.g., Artificial neural network [50,51] and Deep Learning [52], (ii) ensemble models e.g., random forest [53,54], (iii) kernel model: e.g., Support vector machine [16], (iv) evolutionary model: e.g.: gene expression programming [55].

intensification and changes in land use, anthropogenic activities in the forestry and agricultural sectors have produced GHG emissions that have upset the ecosystem's equilibrium [5]. This is particularly evident on the Canadian continent, where agriculture is one of the primary sources of greenhouse gas emissions [6,7].

All over the world, potatoes are grown under a wide range of altitude, latitude, and climatic conditions; studies have suggested that no other food crop can match the potato's yield of food energy and food value per unit area [8]. For instance, potatoes, when compared to other vegetable crops, have one of the highest input requirements for fertilizer. Furthermore, the proportion of nitrogen (N), phosphorous (P), and potassium (K) required for potato cultivation in comparison to tomato or pepper production are, respectively, 100 %, 100 %, and 33 % higher [9]. Every province in Canada grows potatoes as the main crop, and the nation is in the top ten exporters of both fresh potatoes and potato seeds [10]. In terms of overall potato production by area, Atlantic Canada produces 39.9 % of all potatoes produced in Canada, followed by Western Canada (37.9 %) and Central Canada (22.2 %) [11]. Approximately 34,803 hectares of potatoes are farmed in the PEI, accounting for 25 % of all potatoes grown in Canada [11]. Additionally, with over 1.3 billion in revenue generated annually for the province, potato cultivation is the main driver of the PEI economy [10].

Determining greenhouse gas emissions from soil is a well-known and urgent environmental issue because soil degassing is responsible for significant $CO_2$, $N_2O$, and $N_2O$ emissions [12]. The frequent soil disturbances due to tillage techniques, excessive nutrient inputs, cover crops (CC) or crop type in rotation systems, and irrigation account for approximately 14 % of all GHG emissions [13]. In addition to agricultural intensification and management due to the world's growing need for food, soil and climate factors have been the main regulators of greenhouse gas emissions. Temperature, humidity, air pressure, precipitation, and other climatic factors change soil's biological, chemical, and physical characteristics, which in turn affects the hydrological and biogeochemical cycles and, ultimately, the greenhouse gas emissions from agricultural fields [14].

Greenhouse gas emissions and sinks can both occur in agricultural soils [15]. Reducing greenhouse gas emissions is facilitated by climate-smart agronomic practices, which include appropriate use of irrigation, tillage, drainage, and fertilization; bulk density, microbial activity, and organic matter in the soil; and environmental factors like soil temperatures and precipitation. A more creative approach to lowering greenhouse gas emissions is needed in agriculture, given food security, population increase, and climate change concerns. Flux towers and confined chambers are typically used to measure soil emissions [16]. Several biophysical models, including the Denitrification-Decomposition model (DNDC), the Root Zone Water Quality Model (RZWQM2), the daily version of the CENTURY model, and the Decision Support System for Agrotechnology Transfer, have been effectively designed to simulate greenhouse gas emissions [17]. Despite their effectiveness and widespread use, these biophysical models have certain limitations, such as (i) the need for skilled and knowledgeable users with agro-environmental expertise, knowledge, and skills; (ii) pre-procedures and protocols for model calibration and validation; and (iii) the availability of various site-specific input parameters (such as forests, savannahs, rangelands, and agricultural fields).

Due to recent advancements in computer-assisted models, statistical techniques are often employed to examine and/or summarize the results of large-scale simulations involving complicated cropping system models [18,19]. Finding patterns in huge datasets is a strong fit for machine learning (ML)-based algorithms. Four kinds of ML-based algorithms—neural networks, kernel models, ensemble models, and evolutionary models have been identified throughout the literature review for the topic of greenhouse gas emissions (Fig. 1). ML-based models have mainly been used to analyze data with several associated incident variables, diverse data types, data interactions, and small or unbalanced datasets. This variability is used by regression and classification trees to group classes of predictors that are significant for a desired result; many such trees can be computed through random subsampling from a population. The literature evidence several attempts on the utility of ML models for GHG emissions of plants [20–24]. In particular for potato GHG emission, there are limited number of researches established recently confirmed the potential ML models including hybrid random forest (RF) model [25], standalone RF model [26]. Predicting greenhouse gas emissions from potato farms based on climate variability is the ultimate objective of the current research. Since numerous factors in the crop system play crucial roles, it is currently challenging to develop mitigation and adaptation techniques to maximize crop productivity and GHG emissions [27]. In light of the exhibited literature, the development of computer aid models is remarkably interested in predicting GHG emissions for different crops.

Addressing the limitations of existing ML models for crop greenhouse gas (GHG) emission prediction, such as learning process restrictions, difficulty in explaining predictor significance, and feature selection challenges, this research introduces an innovative, optimized ML model. By hybridizing Gradient Boosted Decision Tree (GBDT) and LightGBM models with a Runge-Kutta optimizer (RUN), we developed a robust prediction model for GHG emissions in Atlantic Canadian Potato production.

Our study's main objectives are:

- To pioneer the application of an advanced ML model combining GBDT and LightGBM with a RUN algorithm for GHG emission prediction in Atlantic Canadian Potato production.
- To conduct a comprehensive feature selection investigation using various algorithms, including Boruta, Best Subset Lasso Regression, and Weighted Aggregated Sum Product Assessment, to identify the most effective input combinations.
- To incorporate SHapley Additive exPlanations (SHAP) as an explainer to address the critical need for interpreting the physical significance of predictors in relation to output parameters, a crucial step in ML modeling.

This intelligent framework effectively filters the most significant features from extensive datasets, reducing computational time costs while improving prediction accuracy and providing valuable insights into GHG emission drivers in potato production.

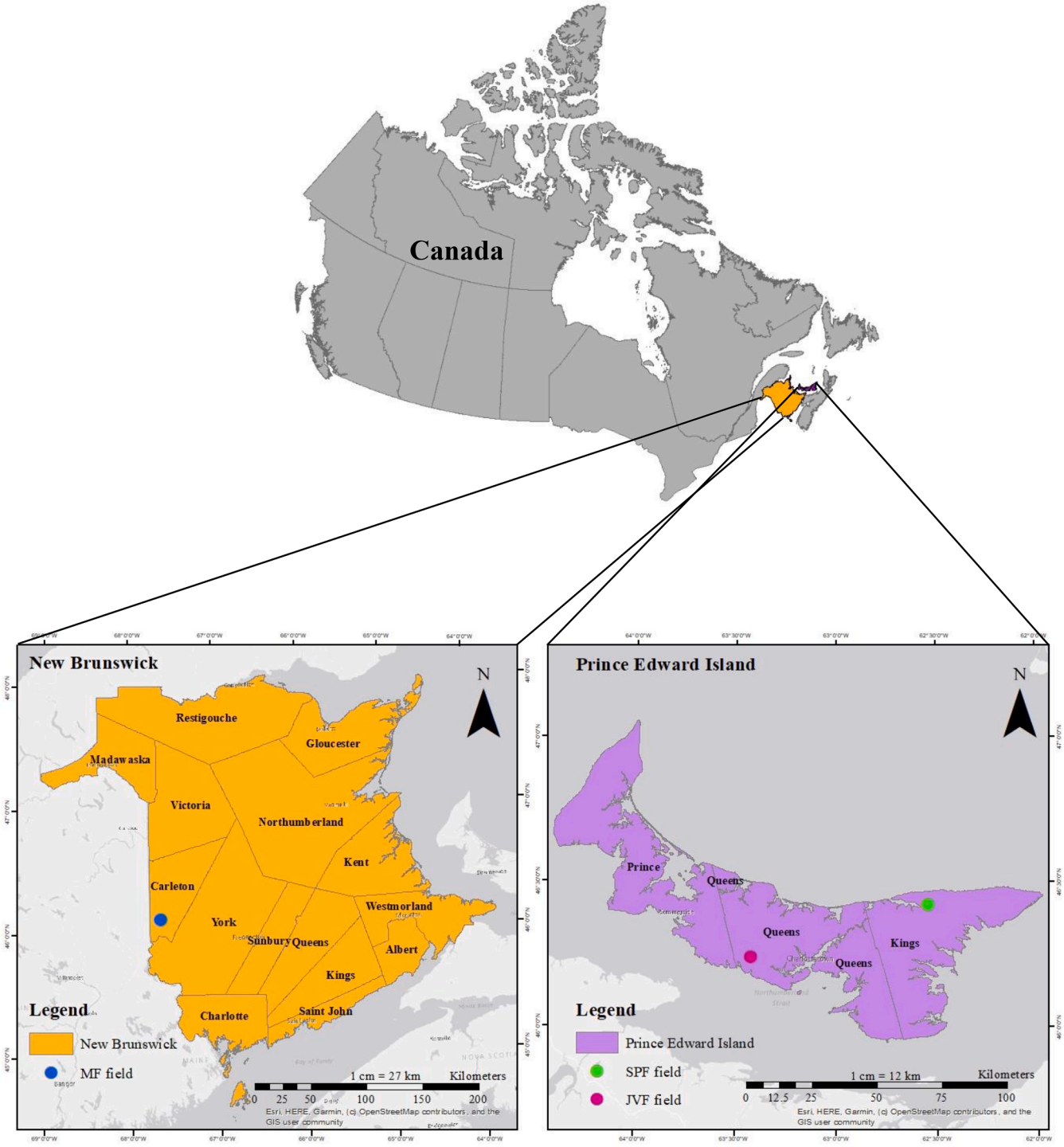



**Fig. 2.** Study area map and geographical field coordinates of Prince Edward Island (right) and New Bruswick (left).

## 2. Materials and methods

### 2.1. Study area

The research was conducted across two distinct provinces in Atlantic Canada, Prince Edward Island (PEI) and New Brunswick (NB), leveraging their unique climatic and soil conditions conducive to potato cultivation. PEI is characterized by its surrounding Atlantic Ocean, contributing to a predominantly humid climate, and experiencing an extended winter season, contrasted with relatively short summers. The winters are marked by frequent snow and blizzard events driven by weather systems from the Atlantic Ocean and the Gulf of Mexico. Spring brings moderate coolness, aiding in winter's snow and ice dissipation. Summer months are characterized by warmth, with temperatures often soaring above 30 ˚C during the day. The autumn season is notably rainy. The soil is primarily sandy loam and is derived from forest soils, making it suitable for potato cultivation. Two fields were selected within PEI for the study: Jordan Visser Farm (JVF) and St. Peter's Farm (SPF), spanning 33 ha and 6 ha and located at coordinates 46.2219° N, 63.5150° W and 46.4210° N, 62.5793° W, respectively.

Meanwhile, the NB, positioned along Canada's eastern coast, experiences a humid continental climate heavily influenced by the Atlantic

**Table 1**
Detailed statistics analysis of collected datasets.

| Parameter | Symbol | Minimum | Median | Maximum | Mean | Std. D | COV | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| Soil Moisture | SM [$cm^3/cm^3$] | 0.009 | 0.059 | 0.196 | 0.068 | 0.035 | 52.04 % | 1.294 | 1.595 |
| Soil Temperature | ST [°C] | 15.570 | 22.840 | 28.520 | 22.660 | 2.920 | 12.88 % | −0.276 | −0.767 |
| Soil EC | SEC [S/m] | 0.001 | 0.003 | 0.015 | 0.003 | 0.002 | 65.12 % | 2.327 | 6.561 |
| Air Temperature | AT [°C] | 12.300 | 16.900 | 24.440 | 18.240 | 3.361 | 18.43 % | 0.228 | −0.907 |
| Precipitation | P [mm] | 0.000 | 0.000 | 7.600 | 0.567 | 1.656 | 292.1 % | 3.653 | 12.620 |
| Relative Humidity | RH [%] | 69.500 | 86.000 | 99.000 | 84.390 | 7.681 | 9.101 % | −0.118 | −0.590 |
| Wind Speed | WS [m/s] | 0.400 | 1.100 | 17.000 | 3.336 | 4.738 | 142.0 % | 1.887 | 2.261 |
| Dew Point | DP [°C] | 7.200 | 16.000 | 21.000 | 15.280 | 3.973 | 25.99 % | −0.427 | −1.079 |
| $CO_2$ | $CO_2$ [$\mu mol/m^2/s$] | 0.757 | 3.468 | 9.726 | 3.602 | 1.614 | 44.81 % | 0.681 | 0.401 |
| $N_2O$ | $N_2O$ [$nmol/m^2/s$] | 0.014 | 0.313 | 10.280 | 0.895 | 1.472 | 164.6 % | 3.358 | 13.290 |
| $H_2O$ | $H_2O$ [mmol/mol] | 13.200 | 23.460 | 33.240 | 23.420 | 4.689 | 20.02 % | −0.022 | −1.060 |

Ocean and varied topography. NB stands as one of the warmer areas within Canada, maintaining an average daily high temperature of approximately 11 °C like Central European climates. It experiences a predominantly cold and wet climate, with a brief period of warmer summer months. The Mclean Farm (MF) field in Woodstock, NB, covering 11 hectares and situated at 46.1106° N, 67.6574° W, features silty loam soil, differing from the sandy loam observed in PEI. Traditional farm management practices were employed for potato cultivation in each of the selected fields. The soil parameters, $CO_2$ and $N_2O$ flux rates, $H_2O$ concentrations and climatic data were meticulously gathered throughout the growing season of 2023, observing different potato growth stages from planting to harvesting. Fig. 2 exhibits the study areas and the coordinates of the filed investigation in PEI and NB.

### 2.2. Data collection and exploration

Data collection commenced with the analysis of soil samples from each field before potato planting. These samples were analyzed from the PEI Analytical Lab (PEIAL) for texture analysis, facilitating a detailed understanding of the soil composition within the study areas. To monitor the soil's GHG emissions effectively, we used a LI-COR survey-based soil gas flux system that includes a smart chamber connected to a laser-based Trace Gas Analyzer, which precisely facilitated the measurement of GHG fluxes. This system employs Optical Feedback-Cavity Enhanced Absorption Spectroscopy (OF-CEAS) for accurate GHG emissions from the soil, as outlined in the studies of [28,29]. Twenty-four PVC soil collars were strategically placed between the plants and on the ridges of the potato crops [30]. Following the recommendations by LI-COR Biosciences (2022), these collars were inserted at a depth of 9 cm with an offset height of 2 cm above the soil surface, considering the field's topography to encompass top hill, mid-slope, and depression areas. Additional parameters such as soil moisture (SM), soil temperature (ST), and electrical conductivity (SEC) were measured using Steven's hydra probe, which was connected to the smart chamber.

The LI-COR Trace Gas Analyzer followed a comprehensive calibration procedure, utilizing zero or baseline calibration gas and air-balanced gases of known volumes for $CO_2$ and $N_2O$ span calibration. This process, detailed in the LI-COR TGA Manual (2022), was essential for accurate GHG emissions measurements from the soil. Furthermore, a HOBO RX3000 weather station [31] was installed in each field to record critical weather parameters, including air temperature (AT), precipitation (P), relative humidity (RH), wind speed (WS), and dew point (DP). This data was instrumental in conducting a statistical analysis of the collected datasets (Table 1), providing insights into the environmental conditions impacting potato growth in PEI and NB.

Table 1 presents a detailed statistical analysis of soil and weather parameters measured from PEI and NB fields influencing GHG emissions. The standard deviation (Std. D) reflects the amount of variability from the mean. It is 1.614 for $CO_2$, suggesting the fluctuations of $CO_2$ emissions due to varying agricultural practices, soil conditions and environmental factors, 1.472 for $N_2O$, highlighting the potential for periodic emissions events influenced by soil management practices, SM and ST, and 4.689 for $H_2O$ reflecting considerable fluctuations in water vapor concentrations due to changing weather patterns, SM and plant transpiration rates. Kurtosis describes the shape of a distribution's tails concerning its overall shape by comparing the peakedness or flatness of the data with normal distribution. For $CO_2$, it is 0.401, indicating that $CO_2$ emissions data does not have tails and is relatively balanced around the mean. It is 13.290 for $N_2O$, which is slightly higher than the normal distribution and indicates the presence of outliers and extreme emission events, which are crucial to understanding the overall impact of agricultural practices. $H_2O$ is calculated as −1.060, representing a flatter distribution to the normal distribution with lighter tails that suggest variations in $H_2O$ vapor concentrations. These insights are vital for understanding the dynamics of GHG emissions and water vapor in agricultural soils, informing strategies for mitigation and management.

## 3. Computational and mathematical aspects

### 3.1. Gradient-boosted decision trees (GBDT)

The gradient-boosting decision tree (GBDT) is a repetitive tree approach that integrates a sequential set of weak prediction models, commonly called classification and regression trees (CART). During the iterative procedure, the CART is trained using the residuals obtained from the preceding tree [32]. The output is determined by aggregating the regression outcomes of all trees [33]. The GBDT algorithm's fundamental principle is to approximate the basic model's loss value to the loss function's negative gradient value to construct the basic model for the subsequent round [34]. The primary objective of GBDT is to compile the outcomes of all trees into a single final result. GBDT will continually reduce the residual by fitting an additional regression tree in the general direction of the gradient of the previous residual reduction in each iteration [35].

In the context of the GBDT approach, $x$ was designated to represent a collection of predictor variables. $F(x)$ is a function that denotes the anticipated values. The objective of the model is to minimize the loss function $L(y,F(x)) = [y - F(x)]^2$. This is achieved by utilizing the $F_M(x)$ function and M decision trees, which are based on the training set $T = \{(x_1,y_1),\cdots,(x_N,y_N)\}$. The function $F_M(x)$ is defined as follows:

$$F_M(x) = \sum_{m=1}^{M} f_m(x) = \sum_{m=1}^{M} \theta_m t(x;\mu_m) \tag{1}$$

The symbol $\mu_m$ represents the average number of split locations and terminal nodes in a single tree $t(x;\mu_m)$. The value of $\theta_m$ was established by minimizing the loss function. The optimization approach started with the consideration of the following function:

$$f_0(x) = \gamma = \text{argmin}_\gamma \sum_{i=1}^{N} L(y_i,\gamma). \tag{2}$$

The value of the iteration count m was adjusted within the range of 1

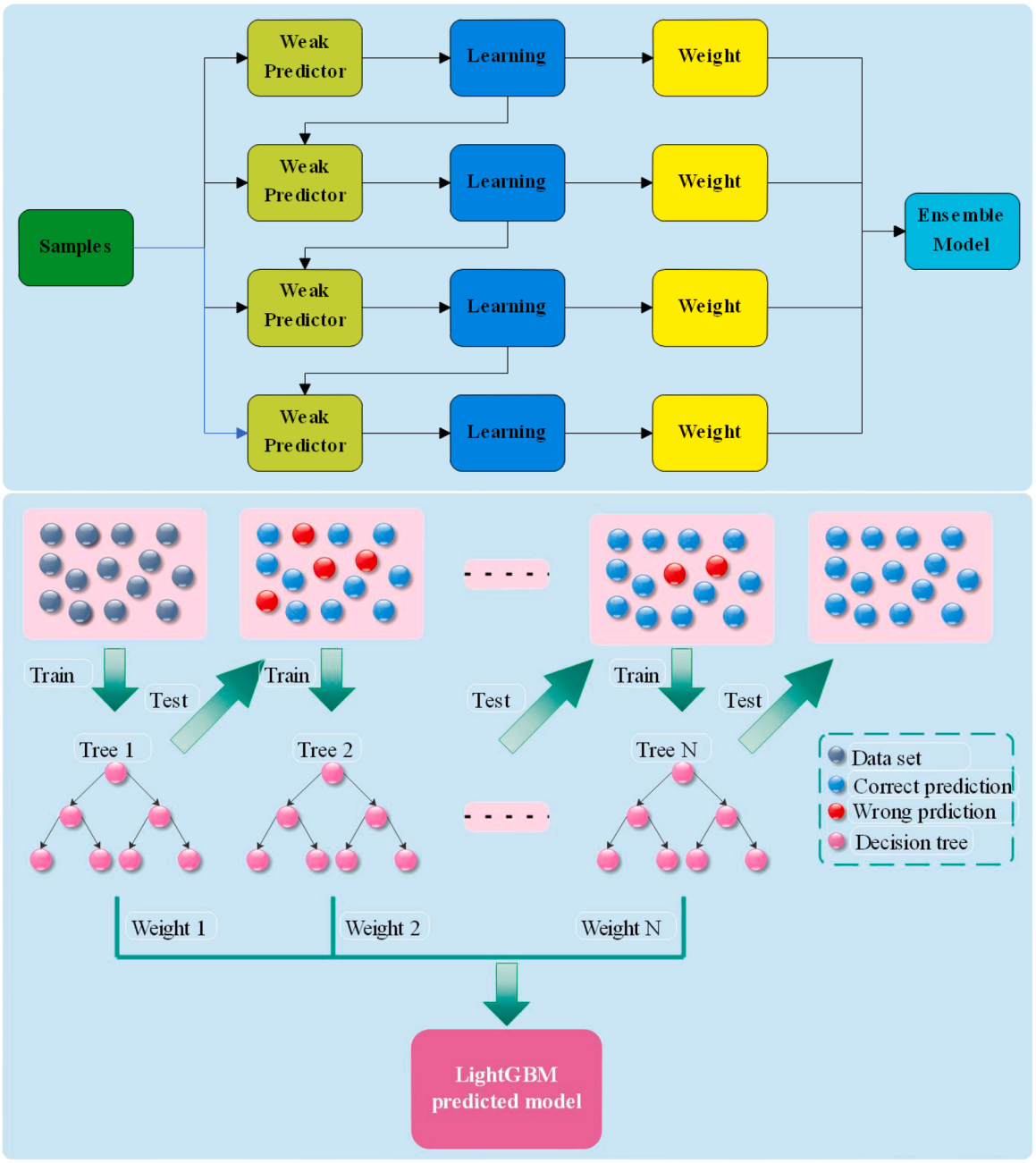



**Fig. 3.** Structure of gradient boosting decision tree (GBDT) **(above)** LightGBM (**bottom**) models.

to *N*, and the direction of the boosting steps was dictated by the negative gradient $g_{im}$ for each individual data sample *i*.

$$-g_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \tag{3}$$

In order to minimize the squared error, the regression tree $t(x;\mu_m)$ was set up to predict the negative gradient $g_{im}$. The step length $\theta_m$ was then computed according to the following formula:

$$\theta_m = \text{argmin}_\theta \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \theta t(x;\mu_m)) \tag{4}$$

Consequently, the optimization method may be altered in accordance with the following Equation:

$$F(x) = F_{m-1}(x) + \theta_m t(x;\mu_m). \tag{5}$$

Fig. 3 (above) shows the structure of the GBDT model. When utilizing GBDT, it is crucial to consider three primary hyperparameters: the learning rate, the maximum depth of the individual regression estimators (max_depth), and the number of iterative learning cycles (n_estimators). Since gradient boosting is not susceptible to overfitting, many iterative learning cycles frequently enhance performance. The maximum depth restricts the number of nodes in the tree. The optimization of efficacy can be achieved through the adjustment of this parameter. The interaction of the input variables determines the optimal value. The rate of learning diminishes the contribution of each tree. In general, smaller values are preferred as they enhance the model's ability to generalize by increasing its resistance to the individual properties of the tree [36].

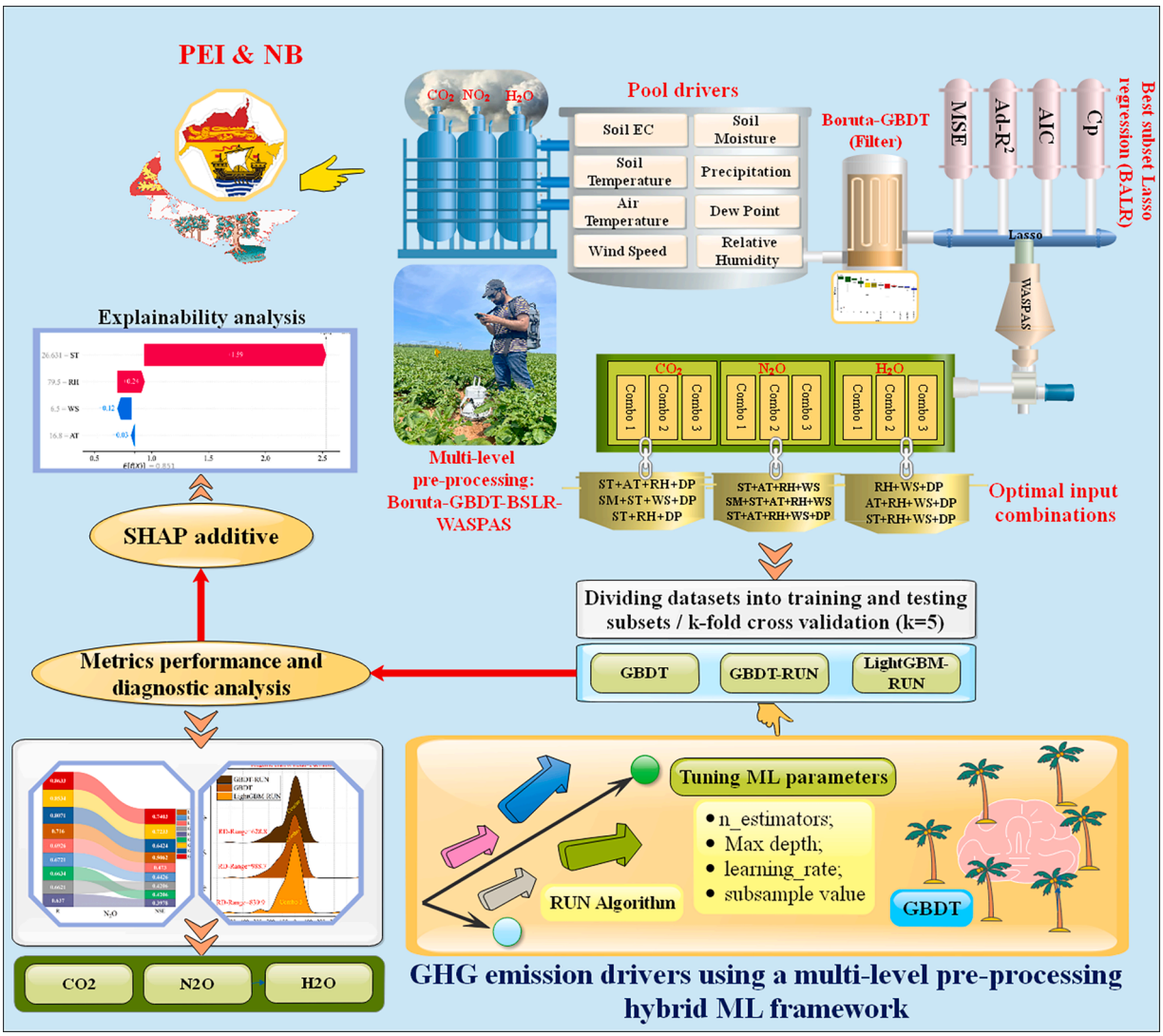



**Fig. 4.** Schematic work path of the multi-level preprocessing-based inspired intelligent framework to monitor the GHG emission drivers.

### 3.2. Light gradient-boosting machine (LightGBM)

LightGBM is one of the newest and most widely used ML techniques based on the gradient-boosting decision tree model [37]. Microsoft developed it in 2017 and it stands out as a highly efficient ML algorithm for addressing classification, regression, and sorting problems [38]. LightGBM mainly uses an advanced histogram algorithm, making it a powerful tool for predictive modelling in various domains [39]. Its ability to reduce memory usage for its computational speed without sacrificing the accuracy in handling large datasets outperforms other gradient-boosting decision tree models like extreme gradient boosting (XGBoost) and stochastic gradient boosting (SGB) [37]. LightGBM optimizes decision tree construction through leaf-wise splitting and data binning, efficiently managing computational resources by organizing feature data into histograms that capture gradient information and instance counts [40,41]. The key parameters that enhance the performance of LightGBM include learning_rate, number of trees (num_iterations), number of leaves per tree (num_leaves), max_depth, bagging_fraction, and a subset of features on each iteration (feature_fraction). Fig. 3 (bottom) illustrates the creation of a LightGBM prediction model by training successive decision trees, each informed by the error gradients of prior ones, and combining them with respective accuracy-based weights to form an ensemble model with enhanced predictive capability.

### 3.3. RUN optimization

The Runge-Kutta method served as the basis for the development of the RUN optimization algorithm [42]. The RUN employs a logical and promising seeking mechanism for global optimization, utilizing the logic of slope changes computed by the Runge-Kutta technique. The Runge Kutta technique is a widely used procedure for solving ordinary differential equations. This method can be implemented to develop a numerical technique with high precision by employing functions that do not necessitate their high-order derivatives [43,44]. The two operators comprising the algorithm are the Runge-Kutta search mechanism ($S_M$) and the enhanced solution quality operator. The primary stages of the RUN algorithm are detailed in the subsequent subsections.

#### 3.3.1. Update solutions

The RUN algorithm uses the search method $S_M$ to update the solutions at each generation and generate a new solution ($x_{new1}$) using Eq. (6).

$$x_{new1} = \begin{cases} (x_{n1} + \mu.A_F.b.x_{n1}) + A_F.S_M + \rho.randn.(x_{n2} - x_{n1}) \ \textit{if rand} < 0.5 \\ (x_{n2} + \mu.A_F.b.x_{n2}) + A_F.S_M + \rho.\ randn.(x_{c1} - x_{c2}) \ \textit{otherwise} \end{cases} \tag{6}$$

The variable μ represents an integer with a value of either 1 or −1. The variable $b$ represents a random number between 0 and 2. The

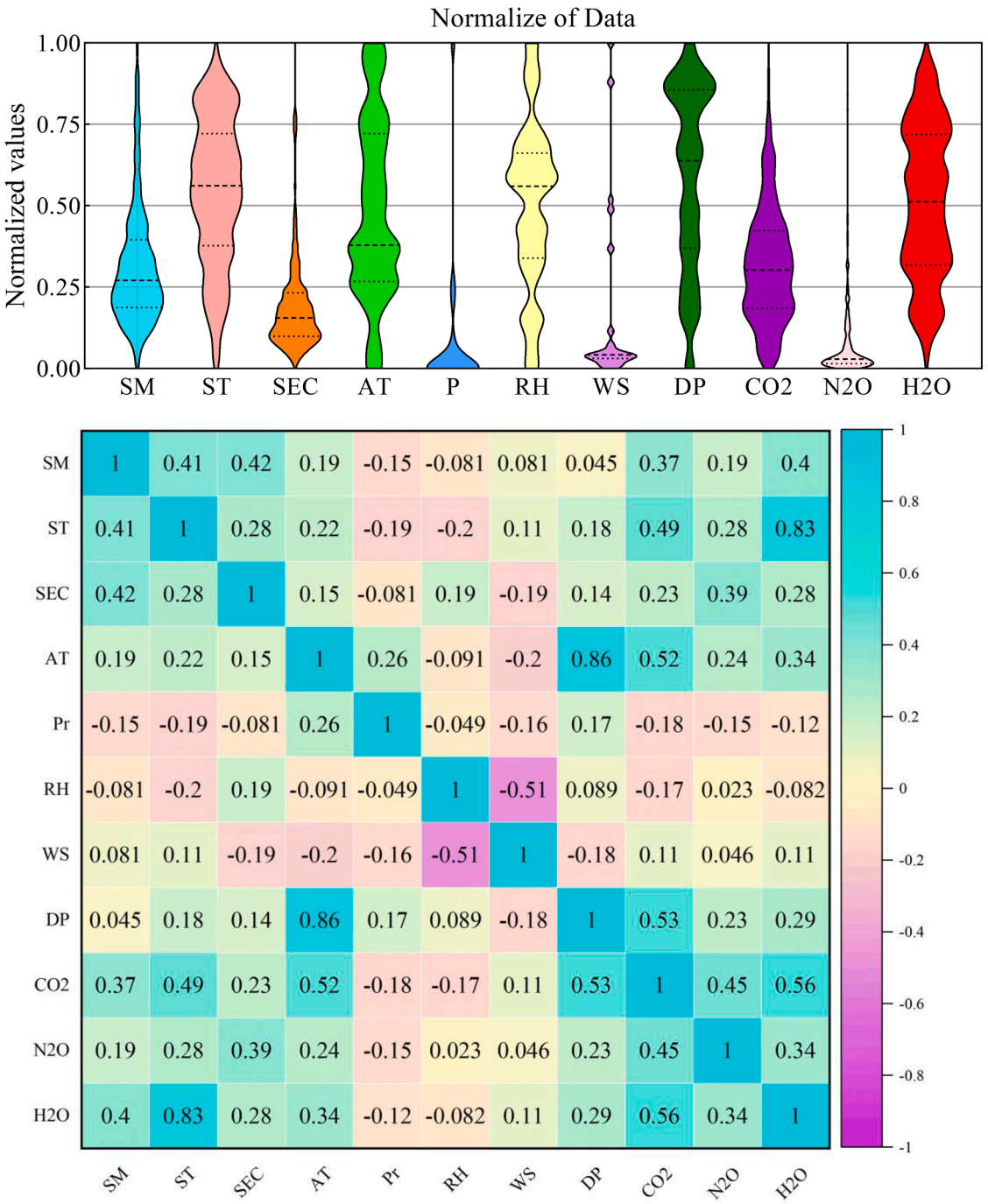



**Fig. 5.** The normalized distribution of datasets (above) correlogram (bottom) of all the datasets aims to model the $CO_2$, $N_2O$, and $H_2O$ drivers.

variables $x_{c1}$ and $x_{c2}$ are two randomly selected positions within the solution space are chosen from the interval between 1 and *Np*. *Np* represents the size of the population. $A_F$ represents an adaptive coefficient, and $\rho$ represents a random number. The formulation of $S_M$ is as follows:

$$S_M = \frac{1}{6}(V_{RK})\Delta x \tag{7}$$

$$V_{RK} = u_1 + 2 \times u_2 + 2 \times u_3 + u_4 \tag{8}$$

$$u_1 = \frac{1}{2\Delta x}(rand.x_{\mathrm{wt}} - \sigma.x_{\mathrm{bst}}) \tag{9}$$

$$\sigma = round(1 + rand).(1 - rand) \tag{10}$$

$$u_2 = \frac{1}{2\Delta x}(rand.(x_{\mathrm{wt}} + a_1.u_1.\Delta x) - (\sigma.x_{\mathrm{bt}} + a_2.u_1.\Delta x)) \tag{11}$$

$$u_3 = \frac{1}{2\Delta x}\left(rand.\left(x_{\mathrm{wt}} + a_1.\left(\frac{1}{2}u_2\right).\Delta x\right) - (\sigma.x_{\mathrm{bt}} + a_2.\left(\frac{1}{2}u_2\right).\Delta x)\right) \tag{12}$$

$$u_4 = \frac{1}{2\Delta x}(r.(x_{\mathrm{wt}} + a_1.u_3.\Delta x) - (\sigma.x_{\mathrm{bt}} + a_2.u_3.\Delta x)) \tag{13}$$

The variables $x_{\mathrm{wt}}$ and $x_{\mathrm{bt}}$ represent the worst and best locations determined during each generation, respectively. The variables $a_1$ and $a_2$ represent two random integers chosen from the range of values between 0 and 1. The calculation of $\Delta x$ is determined using the following equations:

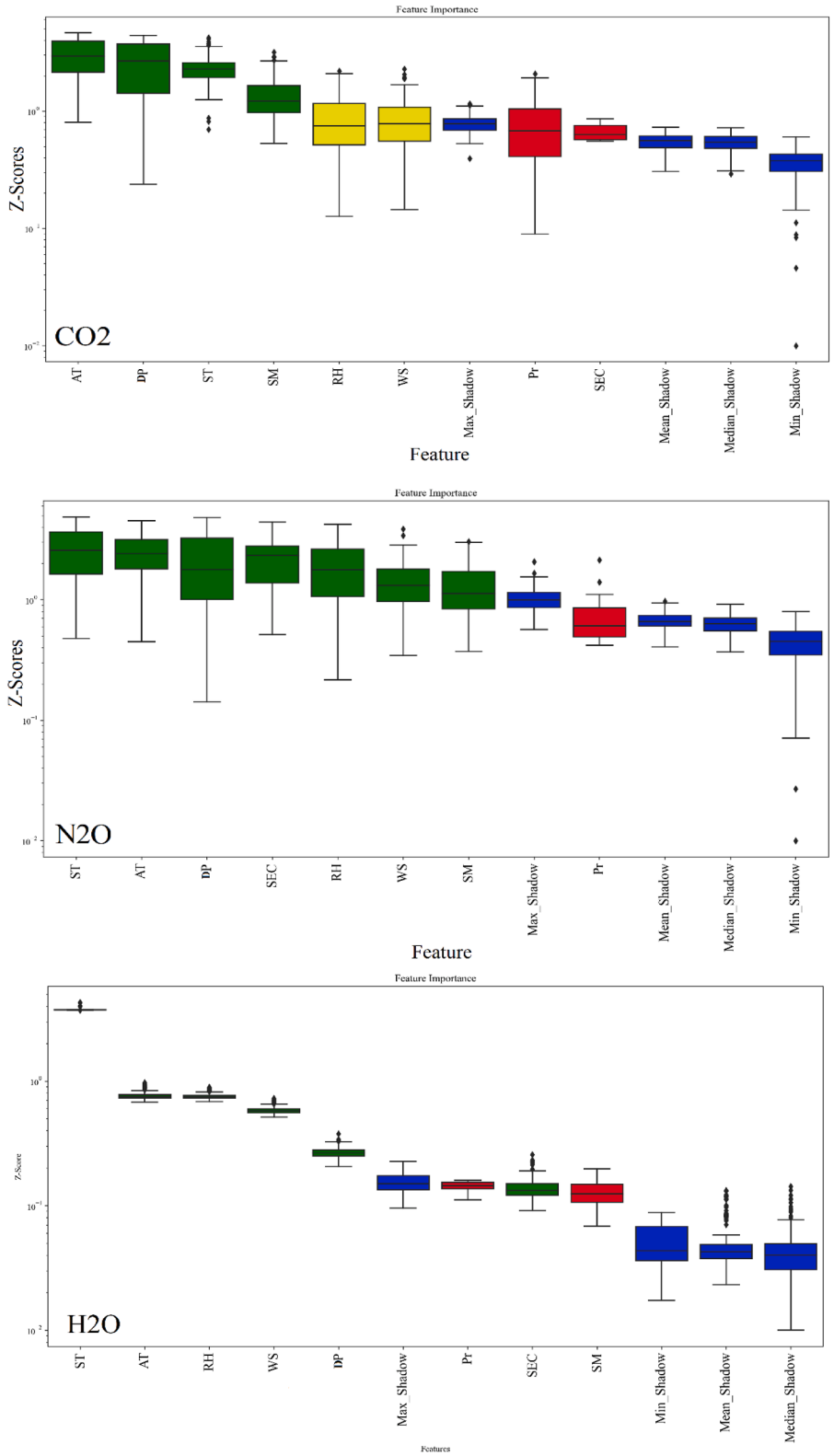



**Fig. 6.** Outcomes of Boruta-GBDT feature selection based on the Z-score values aim to the first stage filtering the most influential features for each GHG emission drivers.

**Table 2**
The second stage of the pre-processing outcomes uses the BSLR-WASPAS schemes to exhibit the $CO_2$, $N_2O$, and $H_2O$ drivers.

| Tar | Num | Input combination | CP | Ad-R2 | MSE | AIC | WASPAS |
|---|---|---|---|---|---|---|---|
| $CO_2$ | 1.0000 | ST+RH+WS+DP | 1.4055 | 0.4636 | 1.3837 | 172.9650 | 0.6090 |
| | 2.0000 | SM+ST+RH+WS+DP | 1.4109 | 0.4625 | 1.3837 | 174.9650 | 0.4965 |
| | 3.0000 | Combo 1: ST+AT+RH+DP | 1.4281 | 0.4549 | 1.4060 | 181.0950 | 0.3414 |
| | 4.0000 | ST+AT+RH+WS+DP | 1.3993 | 0.4669 | 1.3723 | 170.7700 | 0.6263 |
| | 5.0000 | SM+ST+AT+WS+DP | 1.4020 | 0.4659 | 1.3749 | 171.7450 | 0.6119 |
| | 6.0000 | ST+WS+DP | 1.4116 | 0.4602 | 1.3952 | 175.1680 | 0.5389 |
| | 7.0000 | Combo 2: SM+ST+WS+DP | 1.4171 | 0.4591 | 1.3952 | 177.1680 | 0.4175 |
| | 8.0000 | Combo 3: ST+RH+DP | 1.4251 | 0.4551 | 1.4085 | 179.9840 | 0.3755 |
| | 9.0000 | ST+AT+WS+DP | 1.3966 | 0.4670 | 1.3749 | 169.7450 | 0.6592 |
| | 10.0000 | SM+ST+AT+RH+WS+DP | 1.4047 | 0.4659 | 1.3723 | 172.7700 | 0.5602 |
| $N_2O$ | 1.0000 | Combo 1: ST+AT+RH+WS | 1.9014 | 0.1278 | 1.8719 | 326.4910 | 0.1984 |
| | 2.0000 | Combo 2: SM+ST+AT+RH+WS | 1.9087 | 0.1261 | 1.8719 | 328.4910 | 0.2012 |
| | 3.0000 | Combo 3: ST+AT+RH+WS+DP | 1.9087 | 0.1261 | 1.8719 | 328.4910 | 0.2014 |
| | 4.0000 | SM+ST+AT+RH+WS+DP | 1.9161 | 0.1243 | 1.8719 | 330.4910 | 0.2625 |
| | 5.0000 | ST+SEC+AT+RH+WS | 1.9087 | 0.1261 | 1.8719 | 328.4910 | 0.2012 |
| | 6.0000 | SM+ST+SEC+AT+RH+WS+DP | 1.9235 | 0.1226 | 1.8719 | 332.4910 | 0.3587 |
| | 7.0000 | ST+SEC+AT+RH+WS+DP | 1.9161 | 0.1243 | 1.8719 | 330.4910 | 0.2625 |
| | 8.0000 | ST+AT+RH | 1.9284 | 0.1137 | 1.9059 | 333.6300 | 0.8179 |
| | 9.0000 | ST+RH+WS+DP | 1.9313 | 0.1141 | 1.9014 | 334.4350 | 0.7962 |
| | 10.0000 | SM+ST+SEC+AT+RH+WS | 1.9161 | 0.1243 | 1.8719 | 330.4910 | 0.2625 |
| $H_2O$ | 1.0000 | ST+SEC+AT+RH | 6.013 | 0.7281 | 5.92 | 911.39584 | 0.499542 |
| | 2.0000 | ST+AT+RH+WS+DP | 5.667 | 0.7442 | 5.558 | 881.33076 | 0.519603 |
| | 3.0000 | ST+AT+RH | 5.99 | 0.7286 | 5.92 | 909.39584 | 0.519891 |
| | 4.0000 | Combo 1: ST+RH+WS+DP | 6.0777 | 0.7251 | 5.9835 | 916.8149 | 0.4939 |
| | 5.0000 | ST+AT+RH+DP | 6 | 0.7286 | 5.907 | 910.31522 | 0.501138 |
| | 6.0000 | Combo 2: ST+SEC+RH+WS+DP | 6.1013 | 0.7246 | 5.9835 | 918.8149 | 0.4804 |
| | 7.0000 | ST+SEC+AT+RH+WS+DP | 5.689 | 0.7437 | 5.558 | 883.33076 | 0.506871 |
| | 8.0000 | Combo 3: ST+SEC+AT+RH+DP | 6.0237 | 0.7281 | 5.9074 | 912.3152 | 0.4817 |
| | 9.0000 | ST+AT+RH+WS | 5.703 | 0.7421 | 5.615 | 884.49718 | 0.533646 |
| | 10.0000 | ST+SEC+AT+RH+WS | 5.725 | 0.7416 | 5.615 | 886.49718 | 0.518341 |

$$\Delta x = 2.rand.|\gamma| \tag{14}$$

$$\gamma = rand.((x_{bt} - rand.x_m) + \tau) \tag{15}$$

$$\tau = rand.(x_k - rand.(U - L)).\exp\left(-4.\frac{g}{Maxg}\right) \tag{16}$$

The symbol $\gamma$ represents the step size used in each generation. $L$ and $U$ represent the lower and upper bounds of the issue. $g$ represents the current generation number. $Maxg$ represents the maximum number of generations. $x_m$ represents the average location at each generation. The values of $x_{\mathrm{wt}}$ and $x_{\mathrm{bs}}$ are calculated using the following Equation:

$$\begin{aligned}
&\textbf{if } f(x_k) < f(x_{bs,k})\\
&\quad x_{bt} = x_k\\
&\quad x_{wt} = x_{bs}\\
&\quad \textbf{else}\\
&\quad x_{bt} = x_{bs,k}\\
&\quad x_{wt} = x_k\\
&\quad \textbf{end}
\end{aligned} \tag{17}$$

$x_{bs,k}$ represents the optimal solution obtained from three randomly selected places ($x_{c1}$, $x_{c2}$, and $x_{c3}$). $A_F$ is expressed in the following manner:

$$A_F = 2.(0.5 - rand) \times \omega \tag{18}$$

$$\omega = 10 \times exp\left(-12.rand.\left(\frac{g}{Maxg}\right)\right) \tag{19}$$

The values of $x_{n1}$ and $x_{n2}$ are determined using the following formulae:

$$x_{n1} = \beta \times x_k + (1 - \beta) \times x_{c1} \tag{20}$$

$$x_{n2} = \beta \times x_{bst} + (1 - \beta) \times x_{cbst} \tag{21}$$

$\beta$ represents a random value between 0 and 1. $x_{bst}$ represents the best position found so far, and $x_{cbst}$ represents the best position attained in the current generation.

### 3.3.2. Enhanced solution quality

The RUN algorithm enhances the quality of solutions and avoids becoming stuck in local situations. The suggested approach utilizes the enhanced solution quality (ESQ) to generate a new solution ($V_{ESQ}$), which is formulated as follows:

$$\begin{aligned}
&\textbf{if } rand < 0.5\\
&\quad \textbf{if } \psi < 1\\
&\quad V_{ESQ} = x_{new2} + \chi.\eta.|(x_{new2} - x_m) + randn|\\
&\quad \textbf{else}\\
&\quad V_{ESQ} = (x_{new2} - x_m) + \chi.\eta.|(2.rand.x_{new2} - x_m) + randn|\\
&\quad \textbf{end}\\
&\textbf{end}
\end{aligned} \tag{22}$$

$$\eta = rand(0,\ 2).exp\left(-a.\left(\frac{g}{Maxg}\right)\right) \tag{23}$$

$$x_m = \frac{x_{c1} + x_{c2} + x_{c3}}{3} \tag{24}$$

$$x_{new2} = \theta \times x_m + (1 - \theta) \times x_{bst} \tag{25}$$

$\theta$ represents a random number between 0 and 1. $a$ represents a random number that is equal to 5 times a randomly generated number. $\chi$ represents an integer value that can be either 1, 0, or −1.

In order to explore the possibility of finding a better location, a new solution called $x_{new3}$ is developed, given that the fitness function of the solution $V_{ESQ}$ may not be superior to the present solution $x_k$ (i.e., $f(V_{ESQ}) > f(x_k)$).

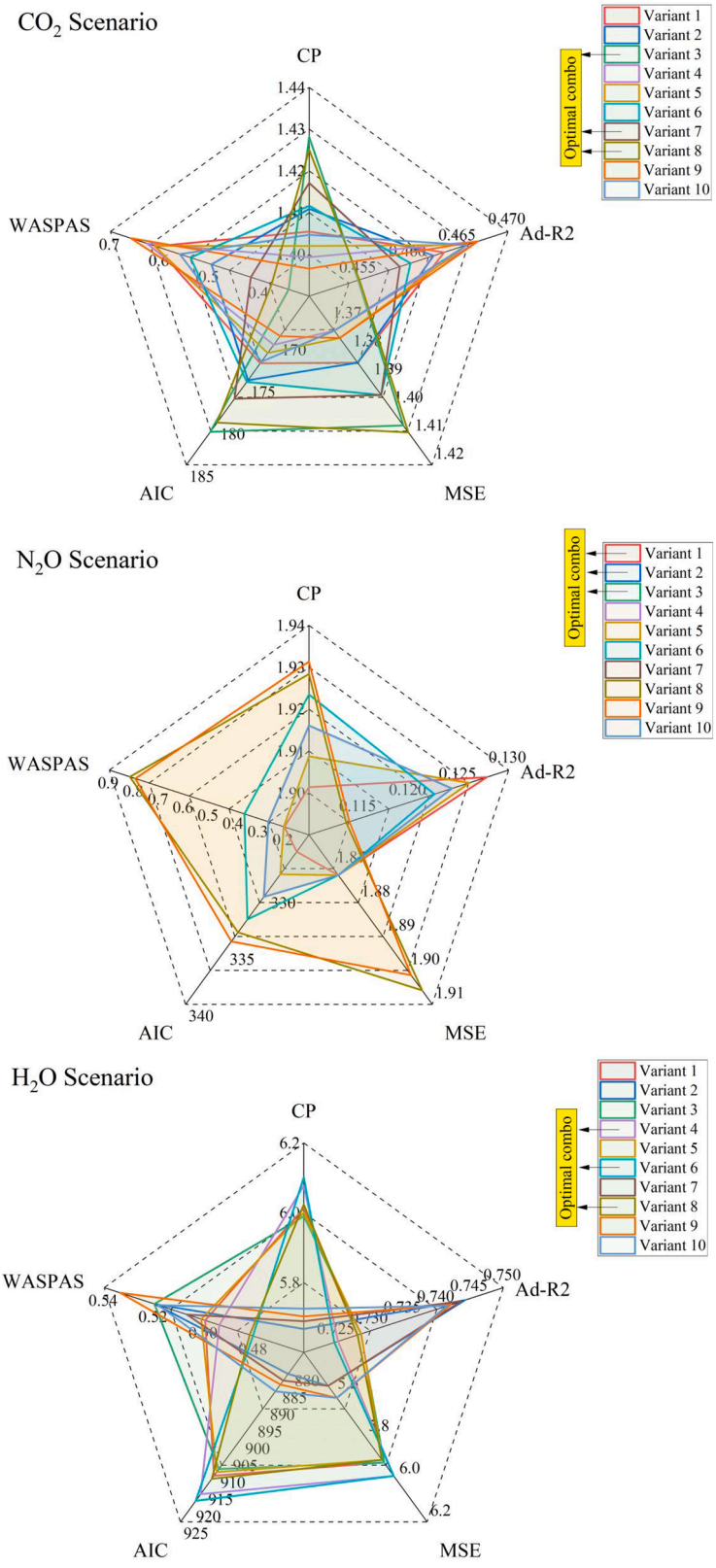



**Fig. 7.** The spider plots of second-stage preprocessing using the BSLR-WASPAS scheme among the top-ten possible input combinations of the GHG emission drivers aim to obtain the three best optimal candidate ones.

**Table 3**
Superior hyperparameters of all the predictive inspired RUN-based models and GBDT to simulate the GHG emission drivers.

| Tar | Model/ Combo | Combo 1 | Combo 2 | Combo 3 | RUN setting |
|---|---|---|---|---|---|
| $CO_2$ | GBDT-RUN | n_estimators=100; Max depth=4; learning_rate=0.024; subsample value=0.90 | n_estimators=100; Max depth=4; learning_rate=0.024; subsample value=0.90 | n_estimators=115; Max depth=4; learning_rate=0.02; subsample value=0.90 | Population size=15; Epoch=15; pr=0.50; beta_min=0.2; beta_max=1.2 |
| | GBDT | n_estimators=100; Max depth=3; learning_rate=0.1; subsample value=0.90 | n_estimators=100; Max depth=3; learning_rate=0.1; subsample value=0.90 | n_estimators=100; Max depth=3; learning_rate=0.1; subsample value=0.90 | – |
| | LightGBM-RUN | n_estimators=100; learning_rate=0.47; num_leaves=2; Max depth =18 | n_estimators=40; learning_rate=0.12; num_leaves=25; Max depth =8 | n_estimators=32; learning_rate=0.50; num_leaves=30; Max depth =3 | Population size=15; Epoch=15; pr=0.50; beta_min=0.20; beta_max=1.2 |
| $N_2O$ | GBDT-RUN | n_estimators=435; Max depth=6; learning_rate=0.045; subsample value=0.10 | n_estimators=1000; Max depth=14; learning_rate=0.01; subsample value=0.10 | n_estimators=1000; Max depth=15; learning_rate=0.024; subsample value=0.10 | Population size=15; Epoch=15; pr=0.50; beta_min=0.20; beta_max=1.2 |
| | GBDT | n_estimators=100; Max depth=3; learning_rate=0.10; subsample value=0.9 | n_estimators=100; Max depth=3; learning_rate=0.10; subsample value=0.9 | n_estimators=100; Max depth=3; learning_rate=0.10; subsample value=0.9 | – |
| | LightGBM-RUN | n_estimators=100; learning_rate=0.44; num_leaves=4; Max depth =11 | n_estimators=98; learning_rate=0.44; num_leaves=10; Max depth =3 | n_estimators=98; learning_rate=0.48; num_leaves=4; Max depth =9 | Population size=15; Epoch=15; pr=0.50; beta_min=0.20; beta_max=1.2 |
| $H_2O$ | GBDT-RUN | n_estimators=888; Max depth=14; learning_rate=0.0; subsample value=0.10 | n_estimators=940; Max depth=4; learning_rate=0.01; subsample value=0.90 | n_estimators=328; Max depth=3; learning_rate=0.06; subsample value=0.90 | Population size=15; Epoch=15; pr=0.50; beta_min=0.20; beta_max=1.2 |
| | GBDT | n_estimators=100; Max depth=3; learning_rate=0.10; subsample value=0.90 | n_estimators=100; Max depth=3; learning_rate=0.10; subsample value=0.90 | n_estimators=100; Max depth=3; learning_rate=0.10; subsample value=0.90 | – |
| | LightGBM-RUN | n_estimators=10; learning_rate=0.49; num_leaves=9; Max depth =13 | n_estimators=10; learning_rate=0.48; num_leaves=29; Max depth =8 | n_estimators=10; learning_rate=0.50; num_leaves=27; Max depth =20 | Population size=15; Epoch=15; pr=0.50; beta_min=0.20; beta_max=1.2 |

$$\textbf{if } \boldsymbol{rand} < \psi$$
$$x_{new3} = (V_{ESQ} - rand.V_{ESQ}) + A_F.(rand.V_{RK} + (2.rand.x_{bst} - V_{ESQ})) \quad (26)$$
$$\textbf{End}$$

### 3.4. Boruta-GBDT

Feature selection (FS) is a critical component of ML-based prediction models, as it is necessary to identify the most relevant features or variables that contribute to prediction accuracy. Boruta has garnered attention among the numerous feature selection algorithms due to its ability to identify statistically significant features and manage large data sets [45]. The Boruta algorithm, which serves as a robust feature set wrapper, was initially derived from the RF, which derives its name from a mythical Slavic god associated with forests [46]. In order to assess the significance of the features involved in classification and regression tasks, the fundamental concept behind this algorithm is to incorporate additional randomization into the original system and combine them [47]. In recent times, the development of open-source frameworks for ensemble ML has enabled the formulation of this algorithm utilizing novel robust ensemble tree-based techniques, including XGBoost, Catboost, and decision tree. Instead of RF, the GBDT algorithm was utilized in this study to prioritize the significant features in order to input the ML models by eliminating the insignificant features iteratively.

The following is a comprehensive protocol for Boruta FS [48]:

1- The feature set may be enhanced by introducing randomization by creating scrambled duplicates, known as shadow features, of all the features. These shadow features can then be combined with the original features to produce an expanded feature set.
2- Develop a GBDT model using the expanded feature set and assess the significance of each feature by calculating the average decreased accuracy ($Z$ value). The shadow feature's biggest $Z$ value, written as $Z_{max}$, corresponds to a larger $Z$ value, indicating its greater importance.
3- In every iteration, if the $Z$ value of the feature exceeds $Z_{max}$, the feature is deemed significant and retained. Alternatively, if the feature is considered to be of minimal significance, it will be eliminated from the feature set.
4- The aforementioned procedure concludes when either all features have been verified or rejected or when the Binary Feature Set (Boruta FS) reaches its maximum number of repetitions.

### 3.5. Best subset lasso regression (BSLR)

The best subset regression (BSR) is a method for selecting models in which, following the evaluation of every potential combination of predictor variables in the initial preprocessing phase, the optimal model is determined using statistical criteria [49]. Here, the classical linear regression has been replaced with the least absolute shrinkage and selection operator (Lasso) regression 36 to efficiently capture the non-linearity between inputs and targets. As this scheme is constructed in the Python platform,

The optimal subset selection for $k$ independent variables can be briefly described as follows:

1. Evaluate any potential model comprising one, two, or more variables, up to a maximum of $k$ variables.
2. Thereafter, the optimal model of size $k$ is chosen, followed by the optimal model of sizes one and two. In the end, the model with the highest overall quality is chosen from the finalists. In the end, the optimal subset selection chooses one model from a set of $2^k$ alternative models. For optimal subset selection, statistical criteria, including Mallow's coefficient ($C_p$) [50], mean square error (MSE), adjusted $R^2_{adj}$, Akaike's Information Criterion (AIC) [51], and determination coefficient ($R^2$) are applied. The conventional BSR core regression is simple linear regression but in the current research, lasso regression [52], as a robust regression technique, is used as a core regressor. The aforementioned criteria may be computed utilizing the subsequent equations [53]:

$$C_p = \frac{RSS_k}{MSE_m} + 2k - N,\ m > k \quad (27)$$

**Table 4**
Goodness-of-fit metrics related to understudy GHG emission components using three ensemble ML schemes to validate the robustness of the main hybrid model (GBDT-RUN).

| Target | Model | Combo | Phase | R | RMSE | MAE | MAPE | NSE | U95 % | SquD |
|---|---|---|---|---|---|---|---|---|---|---|
| $CO_2$ | GBDT-RUN | Combo 1 | Training | 0.8741 | 0.8097 | 0.5974 | 19.8317 | 0.7440 | 2.2460 | 30.9405 |
| | | | Testing | 0.8431 | 0.9337 | 0.7006 | 25.2936 | 0.6758 | 2.5847 | 18.1759 |
| | | Combo 2 | Training | 0.9353 | 0.6262 | 0.4790 | 16.7307 | 0.8469 | 1.7368 | 20.1629 |
| | | | Testing | 0.8201 | 0.9518 | 0.7216 | 26.1152 | 0.6631 | 2.6400 | 19.0841 |
| | | Combo 3 | Training | 0.8699 | 0.8277 | 0.6171 | 20.6314 | 0.7325 | 2.2959 | 32.2480 |
| | | | Testing | 0.8240 | 0.9525 | 0.7209 | 26.7216 | 0.6626 | 2.6373 | 19.3806 |
| | GBDT | Combo 1 | Training | 0.9067 | 0.6851 | 0.5041 | 16.6430 | 0.8167 | 1.9005 | 22.7022 |
| | | | Testing | 0.7942 | 1.0047 | 0.7150 | 25.0560 | 0.6246 | 2.7791 | 20.2269 |
| | | Combo 2 | Training | 0.9262 | 0.6220 | 0.4628 | 15.4295 | 0.8490 | 1.7252 | 19.0727 |
| | | | Testing | 0.8107 | 0.9627 | 0.7044 | 25.3297 | 0.6553 | 2.6715 | 19.2386 |
| | | Combo 3 | Training | 0.9018 | 0.7034 | 0.5225 | 17.4432 | 0.8069 | 1.9510 | 24.2502 |
| | | | Testing | 0.8117 | 0.9658 | 0.7408 | 26.4020 | 0.6531 | 2.6708 | 19.4553 |
| | LightGBM-RUN | Combo 1 | Training | 0.8091 | 0.9417 | 0.6875 | 22.1463 | 0.6538 | 2.6122 | 40.1141 |
| | | | Testing | 0.8105 | 0.9708 | 0.7247 | 24.9808 | 0.6495 | 2.6810 | 18.9495 |
| | | Combo 2 | Training | 0.8843 | 0.7544 | 0.5542 | 17.6424 | 0.7778 | 2.0926 | 25.3662 |
| | | | Testing | 0.8049 | 0.9737 | 0.7180 | 25.6911 | 0.6474 | 2.7027 | 19.9187 |
| | | Combo 3 | Training | 0.8305 | 0.8929 | 0.6464 | 20.7984 | 0.6888 | 2.4768 | 36.0346 |
| | | | Testing | 0.8121 | 0.9669 | 0.7157 | 25.1301 | 0.6523 | 2.6707 | 18.8564 |
| $N_2O$ | GBDT-RUN | Combo 1 | Training | 0.9153 | 0.5569 | 0.2992 | 75.7307 | 0.8338 | 1.5447 | 35.8082 |
| | | | Testing | 0.8633 | 0.8585 | 0.4339 | 84.3531 | 0.7403 | 0.7600 | 27.6479 |
| | | Combo 2 | Training | 0.9378 | 0.4864 | 0.2462 | 66.5599 | 0.8732 | 0.9622 | 24.7464 |
| | | | Testing | 0.8071 | 1.0074 | 0.5308 | 104.2082 | 0.6424 | 0.8693 | 32.2075 |
| | | Combo 3 | Training | 0.9357 | 0.4908 | 0.2513 | 70.1303 | 0.8709 | 0.9618 | 26.2942 |
| | | | Testing | 0.8534 | 0.8861 | 0.4493 | 93.0211 | 0.7233 | 0.9075 | 27.1144 |
| | GBDT | Combo 1 | Training | 0.9631 | 0.3807 | 0.2247 | 67.9726 | 0.9223 | 0.9779 | 19.9047 |
| | | | Testing | 0.6634 | 1.2823 | 0.5630 | 122.4798 | 0.4206 | 0.7908 | 39.7093 |
| | | Combo 2 | Training | 0.9781 | 0.2973 | 0.1916 | 69.2825 | 0.9526 | 0.9870 | 17.1551 |
| | | | Testing | 0.6370 | 1.3073 | 0.5754 | 130.5415 | 0.3978 | 0.7504 | 42.3224 |
| | | Combo 3 | Training | 0.9614 | 0.3856 | 0.2247 | 67.3268 | 0.9203 | 0.9775 | 20.3497 |
| | | | Testing | 0.6621 | 1.2822 | 0.5685 | 126.7158 | 0.4206 | 0.7891 | 39.9863 |
| | LightGBM-RUN | Combo 1 | Training | 0.8818 | 0.6478 | 0.3468 | 98.1628 | 0.7751 | 0.9292 | 46.3923 |
| | | | Testing | 0.6926 | 1.2229 | 0.5244 | 82.5732 | 0.4730 | 0.7740 | 36.1169 |
| | | Combo 2 | Training | 0.9036 | 0.5921 | 0.3305 | 115.7939 | 0.8121 | 0.9418 | 52.3001 |
| | | | Testing | 0.6721 | 1.2576 | 0.5699 | 103.5138 | 0.4426 | 0.7640 | 41.4100 |
| | | Combo 3 | Training | 0.8630 | 0.6931 | 0.3630 | 100.7754 | 0.7426 | 0.9168 | 45.4727 |
| | | | Testing | 0.7160 | 1.1837 | 0.5185 | 81.2189 | 0.5062 | 0.7949 | 34.0221 |
| $H_2O$ | GBDT-RUN | Combo 1 | Training | 0.9908 | 0.6214 | 0.4518 | 2.0262 | 0.9816 | 1.7237 | 3.1020 |
| | | | Testing | 0.9763 | 1.0665 | 0.7864 | 3.4579 | 0.9526 | 2.9604 | 3.8300 |
| | | Combo 2 | Training | 0.9952 | 0.4551 | 0.3514 | 1.5755 | 0.9901 | 1.2624 | 1.6352 |
| | | | Testing | 0.9691 | 1.2087 | 0.8434 | 3.7421 | 0.9391 | 3.3550 | 4.8535 |
| | | Combo 3 | Training | 0.9955 | 0.4383 | 0.3464 | 1.5447 | 0.9908 | 1.2158 | 1.5151 |
| | | | Testing | 0.9735 | 1.1251 | 0.8106 | 3.5815 | 0.9473 | 3.1223 | 4.2232 |
| | GBDT | Combo 1 | Training | 0.9908 | 0.6214 | 0.4518 | 2.0262 | 0.9816 | 1.7237 | 3.1020 |
| | | | Testing | 0.9688 | 1.2143 | 0.8671 | 3.8503 | 0.9386 | 3.3714 | 4.9380 |
| | | Combo 2 | Training | 0.9867 | 0.7537 | 0.5637 | 2.5527 | 0.9729 | 2.0905 | 4.6361 |
| | | | Testing | 0.9680 | 1.2310 | 0.8945 | 3.9636 | 0.9369 | 3.4160 | 5.0505 |
| | | Combo 3 | Training | 0.9879 | 0.7207 | 0.5523 | 2.4748 | 0.9752 | 1.9990 | 4.1012 |
| | | | Testing | 0.9684 | 1.2227 | 0.8754 | 3.8996 | 0.9377 | 3.3942 | 5.0086 |
| | LightGBM-RUN | Combo 1 | Training | 0.9747 | 1.0281 | 0.7170 | 3.2317 | 0.9496 | 2.8517 | 8.4892 |
| | | | Testing | 0.9720 | 1.1529 | 0.8398 | 3.7271 | 0.9446 | 3.2008 | 4.4654 |
| | | Combo 2 | Training | 0.9807 | 0.8995 | 0.6572 | 2.9053 | 0.9614 | 2.4951 | 6.2328 |
| | | | Testing | 0.9713 | 1.1699 | 0.8610 | 3.8288 | 0.9430 | 3.2459 | 4.6545 |
| | | Combo 3 | Training | 0.9800 | 0.9128 | 0.6563 | 2.9100 | 0.9603 | 2.5319 | 6.5052 |
| | | | Testing | 0.9707 | 1.1893 | 0.8679 | 3.8032 | 0.9411 | 3.2975 | 4.7106 |

$$AIC = 2p + N.ln\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{e_i}^{\,2}\right) \tag{28}$$

The variables denoted as $N$, $p$, and k represent the number of data, model parameters, and variables, respectively, the residual ($\widehat{e_i}$) and the residual sum of squares ($RSS_k$). *AIC*, $C_p$, and *MSE* values with the lowest values are optimal [54].

### 3.6. WASPAS scheme

The "weighted aggregated sum product assessment" (WASPAS) method, which was introduced by Zavadskas et al., has proven to be a functional method for Multi-Criteria Decision Making (MCDM) in a variety of domains [55]. This method is a combination of the Weighted Product Model (WPM) and the Weighted Sum Model (WSM). The WASPAS approach is more precise than WPM and WSM [55,56]. This approach has been implemented in numerous decision-making scenarios and circumstances [57–59]. This method's widespread adoption and swift growth can be attributed to its easy and straightforward computation, which yields reasonably precise outcomes when evaluating and selecting specific options in opposition to conflicting criteria. The subsequent procedures constitute the mathematical description of the WASPAS method [60]:

1- During the initial phase, the criteria $(C_j)$ and alternative $(A_i)$ are selected for assessment. Given the set $i = 1,\ldots\ldots m$ and $j = 1,\ldots\ldots n$.

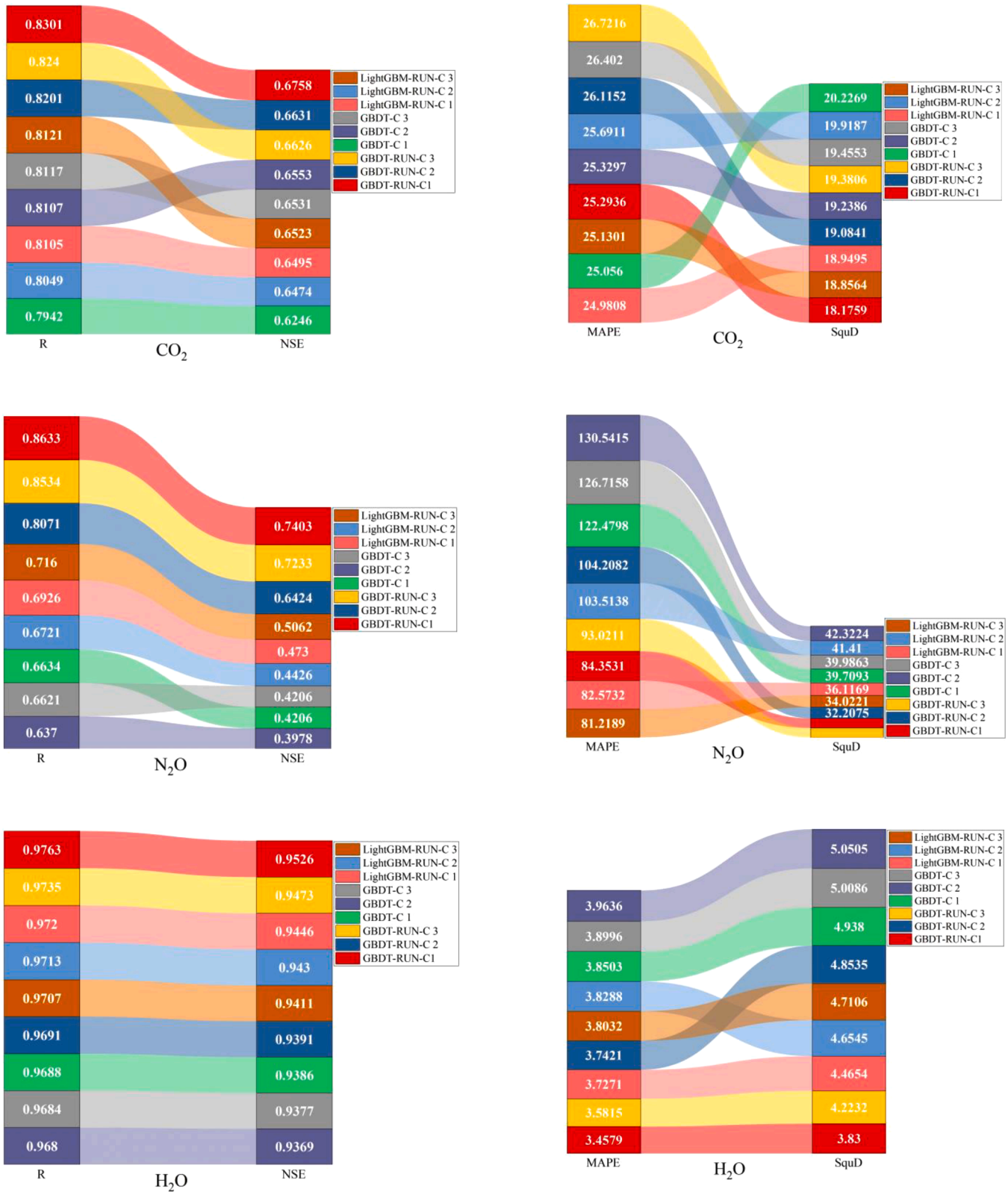



**Fig. 8.** Ribbon plot of metric performances for each ML model's variant input combinations to predict the GHG emission's $CO_2$, $N_2O$, and $H_2O$ components. C denotes the Combo.

2- One of the MCDM methodologies is employed to calculate the weights of the criteria in the second phase. In this investigation, the weights of the criteria were quantified utilizing SWARA.
3- In Step 3, the decision matrix is normalized by employing Eqs. (29) and (30). To optimize the benefit for the beneficiary,

$$\overline{X}_{ij} = X_{ij} / \max X_{ij} \tag{29}$$

For minimum optimum value (non-beneficiary)

**Table 5**
Multi-criteria decision-making (WASPAS)-based assessment of model performance to predict the GHG emission components.

| | Combo | $CO_2$ | | $N_2O$ | | $H_2O$ | |
|---|---|---|---|---|---|---|---|
| Model | | Waspas | Rank | Waspas | Rank | Waspas | Rank |
| GBDT-RUN | Combo 1 | 1.88E-07 | 1.0000 | 1.29E-07 | 1.0000 | 1.23E-07 | 1.0000 |
| | Combo 2 | 2.12E-07 | 5.0000 | 2.48E-07 | 8.0000 | 2.27E-07 | 6.0000 |
| | Combo 3 | 2.21E-07 | 8.0000 | 1.72E-07 | 4.0000 | 1.60E-07 | 2.0000 |
| GBDT | Combo 1 | 2.16E-07 | 6.0000 | 2.37E-07 | 6.0000 | 2.47E-07 | 7.0000 |
| | Combo 2 | 2.02E-07 | 3.0000 | 2.42E-07 | 7.0000 | 2.75E-07 | 9.0000 |
| | Combo 3 | 2.24E-07 | 9.0000 | 2.48E-07 | 9.0000 | 2.59E-07 | 8.0000 |
| LightGBM-RUN | Combo 1 | 2.02E-07 | 4.0000 | 1.48E-07 | 2.0000 | 1.90E-07 | 3.0000 |
| | Combo 2 | 2.17E-07 | 7.0000 | 2.14E-07 | 5.0000 | 2.15E-07 | 4.0000 |
| | Combo 3 | 2.00E-07 | 2.0000 | 1.50E-07 | 3.0000 | 2.24E-07 | 5.0000 |

Grey and cream colour hatches denote the 1st and 2nd ranks, respectively.

$$\overline{X}_{ij} = \min X_{ij} / X_{ij} \tag{30}$$

4- The "Weighted Sum Model" is employed in the fourth stage to compute the initial total relative significance value $\left(Q_i^{(1)}\right)$ by utilizing Eq. (31).

$$Q_i^{(1)} = \sum_{j=1}^{n} \overline{X}_{ij} W_j \tag{31}$$

5- In Step 5, the "Weighted Product Model (WPM)" is executed to compute the second total relative significance value $\left(Q_i^{(2)}\right)$ utilizing Eq. (32).

$$Q_i^{(2)} = \prod_{j=1}^{n} \left(\overline{X}_{ij}\right)^{W_j} \tag{32}$$

6- In Step 6, the aggregate total relative significance value $(Q_i)$ is calculated using Eq. (33), where $\lambda$ represents the coefficient value of $Q_i$.

$$Q_i = \lambda Q_i^{(1)} + (1-\lambda) Q_i^{(2)} \tag{33}$$

The multi-level pre-processing technique to filter out the best possible input combination of all the scenarios has been comprised of the BSLR and WASPAS (BSLR-WASPAS). More details are comprehensively presented in the following sections. More details on application of WASPAS in feature selection are comprehensively presented in 3.9 section.

### 3.7. SHapley Additive exPlanations (SHAP) explainer

SHAP is a model-agnostic tool that Lundberg & Lee (2017) developed, representing a significant advancement in ML interpretability. It draws upon the concept introduced by Shapley in 1953, utilizing game theory principles to assess the impact of individual features on model predictions. The essence of SHAP values lies in their ability to quantify how observing a particular feature shifts the model's output, thus providing a nuanced understanding of feature importance and its effects across the dataset. This method distinguishes itself by examining each input parameter's contribution to the ML model's output through an explanation model (EM) selected based on the problem [61]. SHAP's versatility and model-agnostic nature have facilitated its successful application across various domains, underscoring its utility in enhancing the performance and interpretability of several ML algorithms [62,63]. Mathematically, SHAP is defined as:

$$M = \varphi_0 + \sum_{i=1}^{N} \varphi_i t_i \tag{34}$$

Where $\varphi_i$ is the attribute of the feature $i$, $t_i$ represents the coalition vector, i.e., if the feature is present ($t_i$=1) or absent ($t_i$=0), and N denotes the number of input features. SHAP values are the Shapley values of a conditional exception function $f_x$, and can be calculated as follows:

$$\phi_i = \sum_{S \subseteq Z \setminus \{i\}} \frac{|S|!(N-|S|-1)!}{N!} \left[f_x(S \cup \{i\}) - f_x(S)\right] \tag{35}$$

Where $S$ denotes a subset of features, $Z$ represents the set of all input features, and N indicates the number of input features. Furthermore, the Python SHAP library offers tools for visualizing the importance of features in tree-based models from Scikit-learn, facilitating the interpretation of ML model outcomes [64].

### 3.8. Goodness-of-fit indicators

To examine the robustness of provided ML-based frameworks, seven goodness of fit statistics were utilized including the coefficient of correlations (R), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), Nash–Sutcliffe efficiency (NSE), uncertainty coefficient with 95 % confidence level ($U_{95\,\%}$), and squared deviation (SquD). The mathematical definitions of the mentioned indices are defined as Equations [65–67] (36 to 42):

$$R = \frac{\sum_{i=1}^{N} \left(X_{\mathrm{m},i} - \overline{X_{\mathrm{m}}}\right)\left(X_{\mathrm{p},i} - \overline{X_{\mathrm{p}}}\right)}{\sqrt{\sum_{i=1}^{N} \left(X_{\mathrm{m},i} - \overline{X_{\mathrm{m}}}\right)^2 \sum_{i=1}^{N} \left(X_{\mathrm{p},i} - \overline{X_{\mathrm{p}}}\right)^2}} \tag{36}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(X_{m,i} - X_{p,i}\right)^2} \tag{37}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left|X_{m,i} - X_{p,i}\right| \tag{38}$$

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left|\frac{X_{\mathrm{m},i} - X_{\mathrm{p},i}}{X_{\mathrm{m},i}}\right| \tag{39}$$

$$NSE = 1 - \frac{\sum_{i=1}^{N} \left(X_{P,i} - X_{m,i}\right)^2}{\sum_{i=1}^{N} \left(X_{m,i} - \overline{X}\right)^2} \tag{40}$$

$$U_{95\%} = 1.96 \sqrt{StD_e{}^2 + RMSE^2} \tag{41}$$

$$SquD = \sum_{i=1}^{N} \frac{\left(X_{m,i} - X_{p,i}\right)^2}{X_{m,i} + X_{p,i}} \tag{42}$$

where $N$ is the number of data points, $X_m$ and $X_p$ are the measured and predicted values of the model outcomes, respectively. $\overline{X_m}$ and $\overline{X_p}$ are the mean values of measured and predicted outcomes of the model,

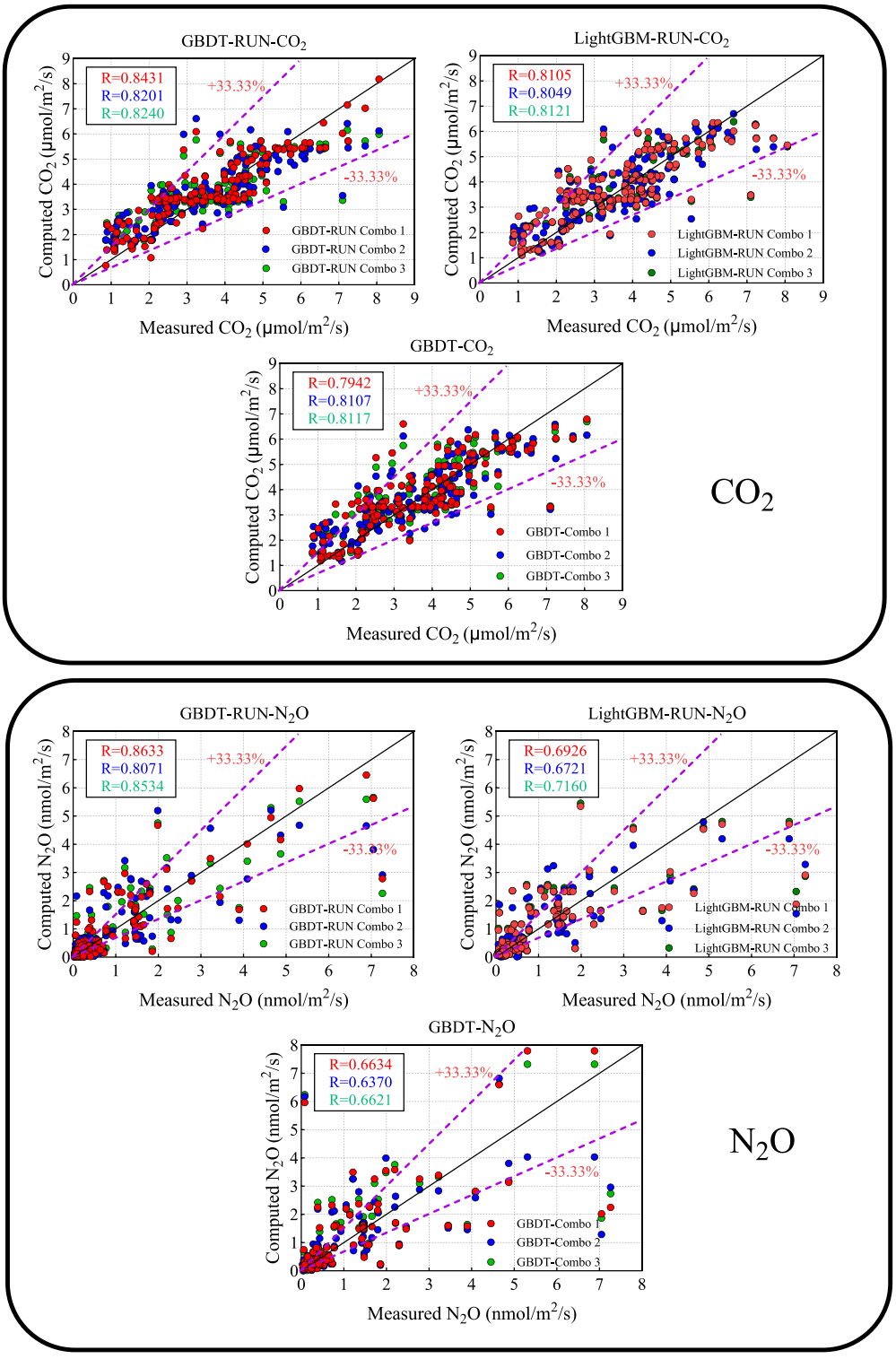



**Fig. 9.** Scatter plot of $CO_2$, $N_2O$, and $H_2O$ variables in the testing phase to compare the compatibility of computed and measured values of targets using all the understudy models considering the candidate input combinations extracted with the multilevel pre-processing.

respectively, and *StDe* is the standard deviation error.

### 3.9. Computational configuration and scenarios

In this research, along with the experimental investigation on the GHG emission drivers of potatoes in PEI and New Brunswick, Canada, a novel explainable multi-level intelligent framework comprised of the new pre-processing technology integrated with the inspired ML scheme has been designed. For this aim, the Boruta-GBDT feature selection incorporated with the BSLR and WASPAS-MCDM to accurately ascertain the optimal input combination of three scenarios of the $CO_2$, $N_2O$, and $H_2O$ GHG emission drivers. All the gathered input data are introduced as the SM, ST, SEC, AT, P, RH, WS, and DP. The computational modelling performed in the Python environment and the implemented libraries

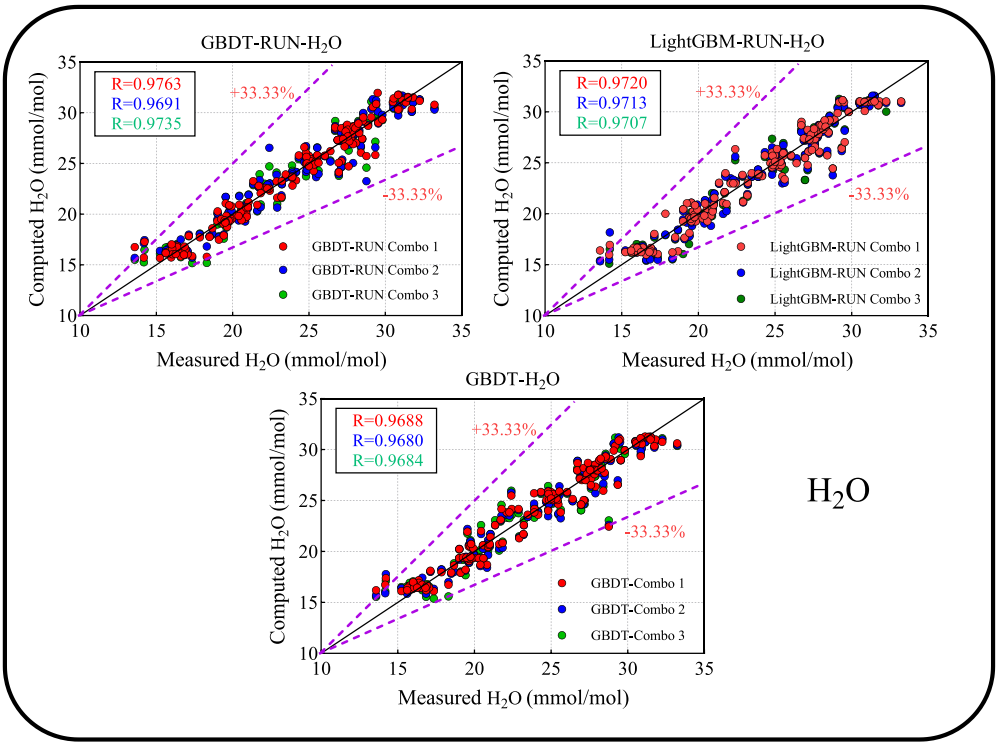



**Fig. 9.** (*continued*).

were Scikit-learning, Boruta, SHAP, and MEALPY for the ML models, feature selection, and optimization algorithm, respectively. Here, a PC system supported by the Intel(R) Core (TM) i9–9900 K CPU @ 3.60 GHz and 16.00 GB RAM has been used to arrange the computational efforts. This corporation was developed for the first time in this research to solve highly non-linear environmental problems. Fig. 4 shows the schematic work path of a multi-level preprocessing-based inspired intelligent framework to monitor the GHG emission drivers.

In the first stages, the normalized distribution of all features in form of a viloin plot shows in Fgure 5 (up). Then, all the linear correlation interaction between inputs and targets has been assessed using the Pearson correlation coefficient-based correlogram depicted in Fig. 5 (bottom). According to the linear analysis, in the $CO_2$, $N_2O$, and $H_2O$ scenarios, (DP and AT), (SEC and ST), (ST and AT) have the highest linear interaction, respectively. To ascertain the optimal input combinations for each scenario, the Boruta-GBDT feature selection based on the Z-score and Max-Shadow criteria was implemented, and its outcomes are shown in Fig. 6. The green, yellow, and red colors demonstrate the accepted, tentative, and rejected features, while the dark blue color shows the Max-Shadow criteria. Here, tentative features were considered as the selected feature aims for more assessment in the second pre-processing stage. According to (upper panel)-Fig. 6, the $CO_2$ scenario reveals that the AT, DP, ST, SM, RH, and WS are identified as the most influential parameters, whereas the ST, AT, DP, SEC, RH, WS, and SM considering (middle panel)-Fig. 6 and ST, AT, RH, WS, DP, and SEC considering (bottom panel)-Fig. 6 likewise are designated for $N_2O$ and $H_2O$ scenarios, respectively.

In the subsequent stage of pre-processing, the BSLR-WASPAS method was utilized to compute the most favorable input combinations given the pre-selected feature in every scenario. Previous research suggests that specific criteria, such as the lowest values of MSE, Cp, AIC, and PC, as well as the highest values (Adj-$R^2$), play a crucial role in identifying the best combinations. However, there is no definitive rule for determining which criteria are more influential. To address this gap, the BSLR scheme, which has a high potential for capturing non-linearity between the datasets, was combined with an MCDM scheme, WASPAS. Based on the criteria above, the second pre-processing stage's outcomes are reported in Table 2. The role of WASPAS is to introduce three superior candidate input combinations by singularization of the metrics mentioned above. In the WASPAS setting, all weight values were set to have a unit mean equality across all the BSLR metrics effects.

Essentially, the minimum values of WASPAS indicate the superior candidate input combinations. To facilitate the evaluation of ML models, the three best input combinations with 4–7 features were identified using the WASPAS values in the $CO_2$, $N_2O$, and $H_2O$ scenarios. The outcomes computationally reveal that for the $CO_2$ scenario, [ST+AT+RH+DP], [SM+ST+WS+DP], and [ST+RH+DP] given the lower values of WASPAS (0.3414, 0.4175, and 0.3755, respectively) are considered to aim at more ML-based assessment. In contrast, in the $N_2O$ scenario, [ST+AT+RH+WS|0.1984], [SM+ST+AT+RH+WS|0.2012], and [ST+AT+RH+WS+DP|0.2014] and in the $H_2O$ scenario [ST+RH+WS+DP|0.4939], [ST+SEC+RH+WS+DP|0.4804], and [ST+SEC+AT+RH+DP|0.4817] have been ascertained to feed the novel intelligent frameworks. Fig. 7 shows the spider plot of the BSLR-WASPAS pre-processing ascertaining the optimal candidate input combinations of $CO_2$, $N_2O$, and $H_2O$ GHG emissions drivers by the yellow color box indicator.

Tuning hyperparameters is one of the most challenging stages in ML-based schemes for solving engineering linear problems. Negligence of ML hyperparameters might lead to decreased accuracy even in advanced paradigms. In this research, a state-of-the-art optimization algorithm, well-known RUN, has been employed to compute GBDT and LightGBM hyperparameters to overcome this defect. The objective hyperparameters of GBDT-RUN and LightGBM are n_estimators, learning_rate, subsample value, and n_estimators, learning_rate, num_leaves, and Max depth, respectively. Also, the setting parameters of RUN are population size, Epoch, pr, beta_min, and beta_max. The default values have been utilized for the classical GBDT hyperparameters. Table 3 lists all the setting hyperparameters of all the provided intelligent frameworks to predict the $CO_2$, $N_2O$, and $H_2O$ scenarios.

To construct the models, the whole dataset was divided into training and testing parts using an 80 % to 20 % ratio. Then, a k-fold cross-

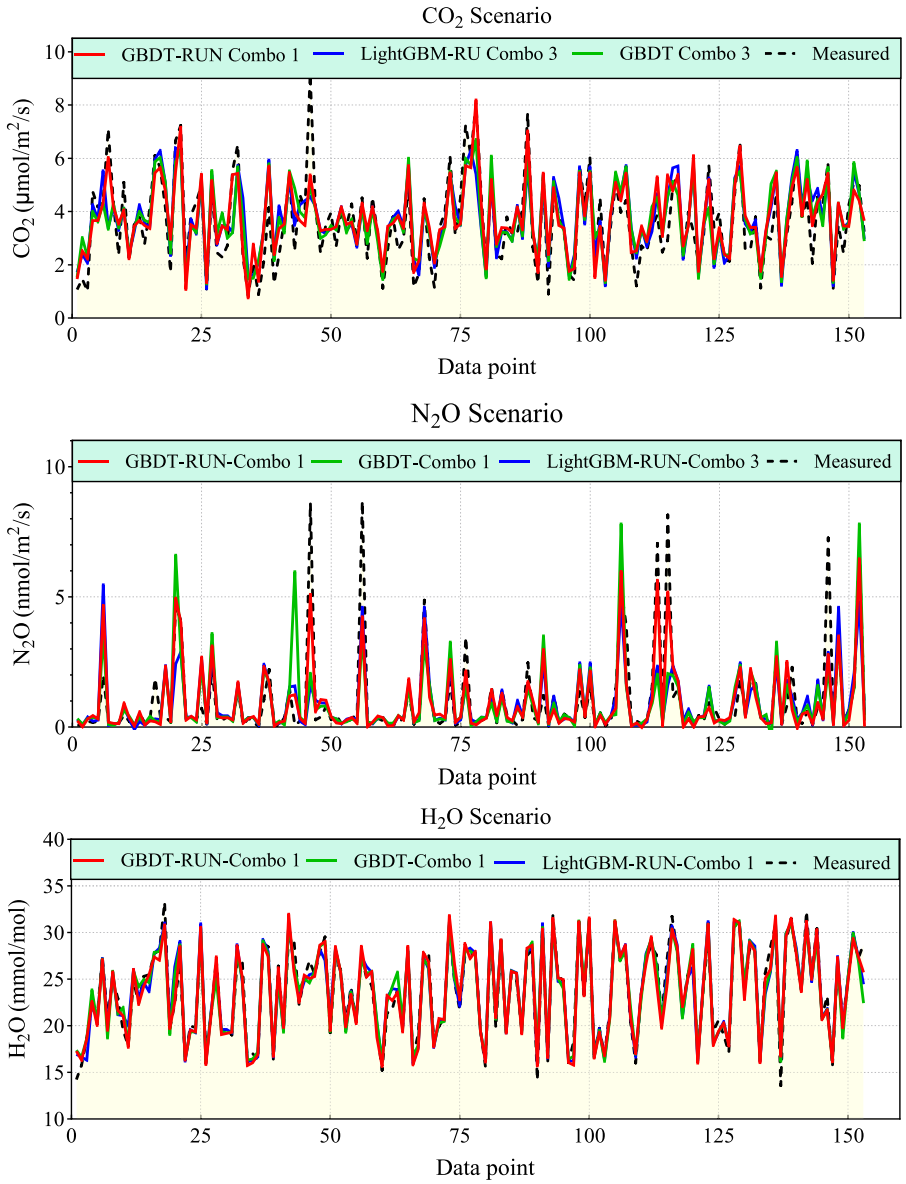



**Fig. 10.** Physical trend expectations related to the $CO_2$, $N_2O$, and $H_2O$ variables in the best input combination potential to assess the robustness of the main model (GBDT-RUN) and other comparative models in the testing phase.

validation has been adopted, avoiding overfitting during the training procedure by ($k = 5$). In addition, before feeding the ML model, all the inputs and output variables were scaled into a range of zero to one to handle different units and distributions, improve convergence, and void feature domination.

## 4. Application results and accuracy assessment

The experimental results and accuracy assessment have been thoroughly discussed in three scenarios of input combination Combo 1, Combo 2, and Combo 3. The primary hybrid GBDT-RUN model was compared against LightGBM-RUN and GBDT models to predict $CO_2$, $N_2O$ and $H_2O$ using R, RMSE, MAE, MAPE, NSE, U95 %, and SquD assessment metrics in both training and testing periods for Atlantic Canada Provinces, PEI and NB.

Table 4 provides the performance accuracy of the GBDT-RUN, LightGBM-RUN and GBDT models for Combo 1, Combo 2, and Combo 3 to predict $CO_2$. By examining, the hybrid GBDT-RUN model shows higher efficiency in terms of ($R = 0.8741$, RMSE = 0.8097, MAE = 0.5974, MAPE = 19.8317, NSE = 0.7440, U95 % = 2.2460, SquD = 30.9405) and ($R = 0.8431$, RMSE = 0.9337, MAE = 0.7006, MAPE = 25.2936, NSE = 0.6758, U95 % = 2.5847, SquD = 18.1759) in training and testing phases respectively based on the Combo 1 (best input combination) followed by Combo 3, and Combo 2. For LightGBM-RUN and GBDT, Combo 3 appeared to be the optimum choice for input combination to predict $CO_2$. However, the GBDT-RUN model achieved the highest performance with Combo 1 (in all three combinations of inputs) compared to LightGBM-RUN and GBDT models in predicting $CO_2$.

To predict $N_2O$, again, the GBDT-RUN model turns out to be the best and most accurate choice with Combo 1, followed by Combo 3 and Combo 1 using R, RMSE, MAE, MAPE, NSE, U95 %, and SquD. When compared, the GBDT-RUN model provides better prediction in both training and testing phases against the LightGBM-RUN and GBDT models to predict $N_2O$. Analyzing the performance individually, the LightGBM-RUN model is better with Combo 3 against Combo 1 and 2. In contrast, the GBDT model appeared good with Combo1 to predict $N_2O$ but could not surpass the GBDT-RUN model

Considering $H_2O$ prediction, the GBDT-RUN model displays higher assessment metrics values such as $R = 0.9908$, 0.9763; RMSE = 0.6214,

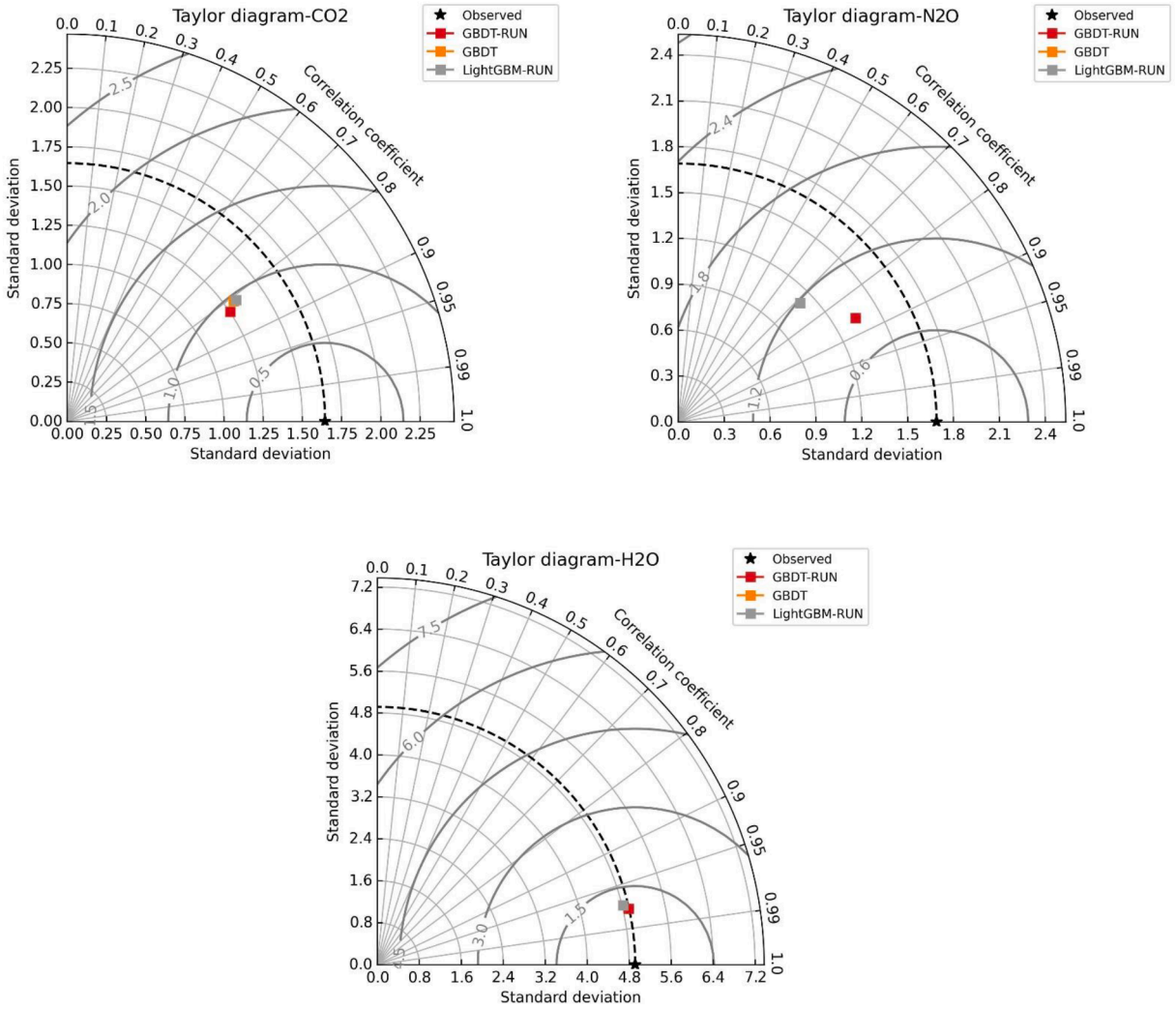



**Fig. 11.** Taylor diagram to reveal the robustness of three models in corresponding optimal performance aims to estimate the GHG emission components, $CO_2$, $N_2O$, and $H_2O$, in the testing phase.

1.0665; MAE = 0.4518, 0.7864; MAPE = 2.0262, 3.4579; NSE = 0.9816, 0.9526; U95 % = 1.7237, 2.9604; SquD = 3.1020, 3.8300 when using Combo 1, followed by Combo 3 and Combo 2. The LightGBM-RUN and GBDT models are reasonably good in predicting $H_2O$ but could not exceed the GBDT-RUN model. Thus, overall, the GBDT-RUN model is better at predicting $CO_2$, $N_2O$, and $H_2O$ with Combo 1 than the LightGBM-RUN and GBDT models.

It is easily understood that the GBDT-RUN model reports higher precision for all Combo 1, Combo 2, and Combo 3 input combinations but outperformed Combo 1, achieving maximum $R$ = 0.8301, NSE = 0.6758 values and MAPE = 25.2936 and SquD = 18.1759. The GBDT-RUN is again positioned in 2nd place with Combo 2, followed by the Combo 3 input combination to predict $CO_2$. The comparison of model GBDT with Combo 2 and LightGBM-RUN with Combo 3 is reasonably good in all input combinations for predicting $CO_2$ of GHG emission. However, the GBDT-RUN model exceeds all the models with Combo 1 compared to other models in predicting $CO_2$. Similarly, the GBDT-RUN with Combo 1 model acquired the top performance by obtaining the highest R, NSE and lowest MAPE and Squd magnitudes values to predict $N_2O$ and $H_2O$ of GHG emission against comparing models and input combinations Combo 2 and Combo 3.

Fig. 8 represents the ribbon plot of compatibility (R and NSE) and diagnostic performance (MAPE and SquD) of all the models in each scenario. Substantially, the GBDT-RUN with Combo 1 in compatibility assessment has been recognized as the best model in all the scenarios. However, in the diagnostic analysis of $CO_2$, $N_2O$, and $H_2O$ scenarios, the LightGBM-RUN in Combo 1, LightGBM-RUN in Combo 3, and GBDT-RUN in Combo 1 have superior performance.

In order to recognize the best predictive models in all the scenarios based on all the metric indicators considering equity weight, Table 5 classifies these models in ranking to determine the most precise model based on Combo 1, Combo 2, and Combo 3 using the WASPAS-MCDM method. Looking at Table 5, the GBDT-RUN model ranked 1st in predicting $CO_2$, $N_2O$, and $H_2O$ components of GHG emission using the input combination Combo 1 compared to LightGBM-RUN and GBDT models. Further, Combo 1, based on the WASPAS method, appeared to be an optimal combination of inputs to predict all three components of GHG emission. The GBDT-RUN with Combo 3 attained 2nd position for $H_2O$ prediction, whereas the LightGBM-RUN model placed 2nd ranking for the prediction of $CO_2$ (with Combo 3) and $N_2O$ (with Combo 1) components of GHG emission. Thus, the WASPAS scheme established that Combo 1 is an optimal choice of input combinations with GBDT-RUN for accurately predicting $CO_2$, $N_2O$, and $H_2O$ components of the GHG emission.

Fig. 9 examines the competence of the GBDT-RUN, LightGBM-RUN, and GBDT model between the predicted and measured $CO_2$, $N_2O$, and $H_2O$ components of the GHG emission in terms of scatter plots using Combo 1, Combo 2, and Combo 3. The scatter plots further intricately

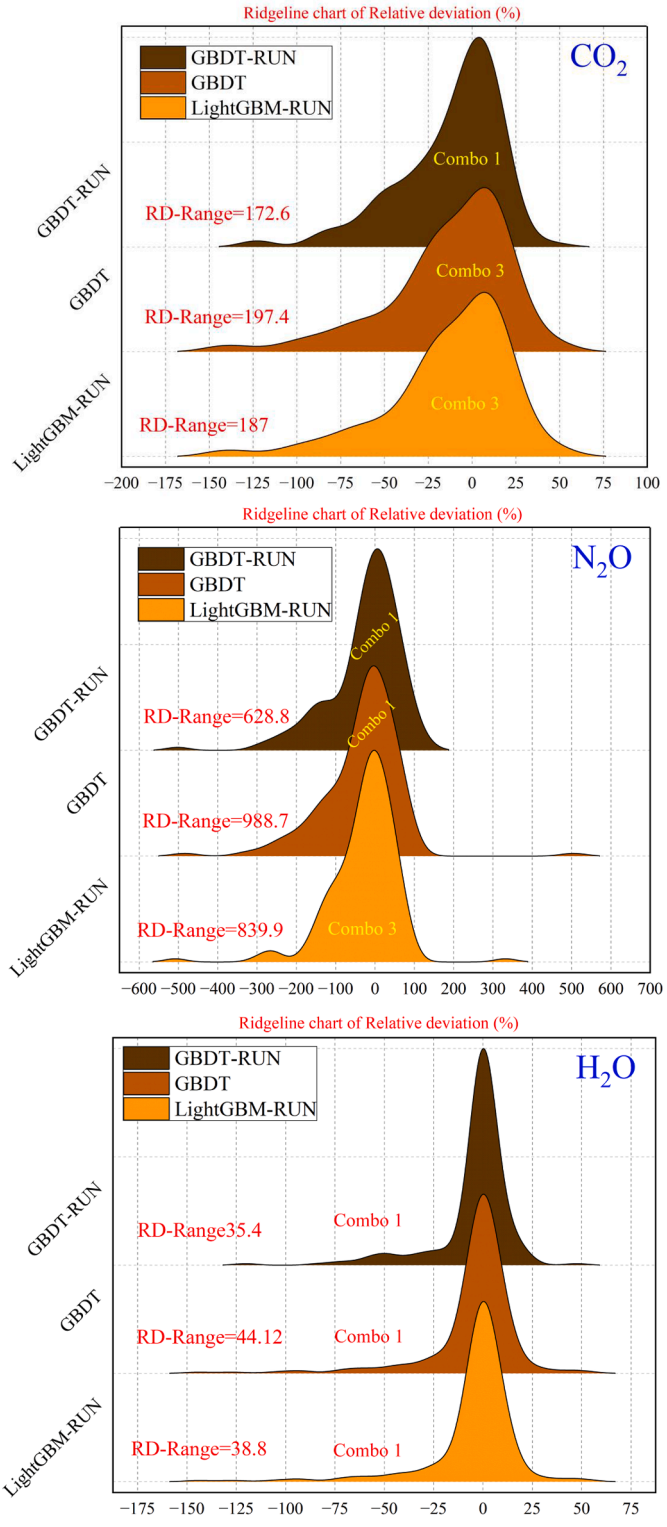



**Fig. 12.** Ridgeline chart relative deviation percentage associated with the superior operation of each model to predict the GHG emission components for the testing phase.

measure the models' prediction capability by counting the R metric and the 33 % upper and lower bounds limits. The GBDT-RUN model utilizing Combo 1 portrayed the highest precision with $R = 0.8431$, followed by GBDT-RUN with Combo 3 ($R = 0.8240$) and Combo 2 ($R = 0.8201$) as compared to LightGBM-RUN, and GBDT model to predict $CO_2$ component. Likewise, the GBDT-RUN model with Combo 1 again appeared superior in predicting the GHG emission's $N_2O$ and $H_2O$ components. Thus, Fig. 9 affirmed that the GBDT-RUN based on Combo 1 is a competitive predictive model for GHG emissions' $CO_2$, $N_2O$, and $H_2O$ components.

Fig. 10 exhibits the Physical trend expectations between the measured and predicted $CO_2$, $N_2O$, and $H_2O$ components of GHG emission in the best input combination to assess the robustness of the main GBDT-RUN and other comparative models. The GBDT-RUN model with Combo 1 is proficient in better accuracy by showing parallel and consistent trends against the measured $CO_2$, $N_2O$, and $H_2O$ components compared to the LightGBM-RUN and GBDT models Combo 1 and Combo 3. The physical trend expectation plots further prove the higher prediction accuracy of the GBDT-RUN model during $CO_2$, $N_2O$, and $H_2O$ prediction than other models.

### 4.1. *Further discussion and explanation of models*

The Taylor diagrams in Fig. 11 elaborate the GBDT-RUN, LightGBM-RUN and GBDT model's performance more palpably and concretely between the observed and predicted $CO_2$, $N_2O$, and $H_2O$ in Combo 1, Combo 2, and Combo 3 scenarios. Taylor diagrams are depicted as a complete inclusive assessment to examine the models' comparison depending on standard deviation and correlation coefficient. The GBDT-RUN model with Combo 1 is slightly closer to the measured $CO_2$, $N_2O$, and $H_2O$, with a correlation coefficient between 0.80 and 0.95. Comparing LightGBM-RUN and GBDT models are reasonably okay but could not surpass the GBDT-RUN model, and this established the suitability of the GBDT-RUN model to predict the $CO_2$, $N_2O$, and $H_2O$ components of GHG emission.

Fig. 12 compares the Ridgeline chart of the relative deviation percentage of GBDT-RUN, LightGBM-RUN and GBDT models with the optimum input combinations to predict the GHG emission components. Ridgeline charts accurately compare the models' prediction capacity and relative deviation (RD range) values for each model in Combo 1, Combo 2, and Combo 3 scenarios. From Fig. 12, it is noticeable that the GBDT-RUN model offered higher precision with RD-range = 172.6 ($CO_2$), 628.8 ($N_2O$), and 35.4 ($H_2O$) using the optimal Combo 1 as compared to other models. Thus, based on Ridgeline plots, GBDT-RUN models achieve accurate $CO_2$, $N_2O$, and $H_2O$ prediction of GHG emission components.

Fig. 13 describes the explainability and interpretability of the GBDT-RUN model prediction using the SHAP waterfall plot at the splitting point (a) and SHAP summary plot (b) to predict $CO_2$, $N_2O$, and $H_2O$ components of the GHG emission. The SHAP explainer in Fig. 13 evaluates the influence and effect of the input predictor on the GBDT-RUN models' prediction. The SHAP waterfall plot at the splitting point specifies that the input predictors DP in blue have significantly contributed to the model's output prediction with magnitude = −0.48, which appeared to be the most negatively contributed factor whereas the predictor ST = +0.36 is contributing positively to the model's prediction. The predictors AT, and RH in blue demonstrate lower feature values, negatively impacting the model's prediction. For $N_2O$, ST is again the main contributing predictor, followed by RH with +1.59 and +0.24 scores, respectively. Similarly, ST, WS, and RH are the main components of the model's prediction of $H_2O$.

The red dots in summary plots (b) exhibit that the corresponding predictors (i.e., ST, DP, AT, and RH) have a high and positive impact on the model's output prediction. In contrast, the blue dots label lower and adverse effects during the prediction. For $CO_2$, the input predictor DP shows a higher impact corresponding to the feature value bar, followed by ST than other predictors. Similarly, the ST input depicts a higher positive contribution to predicting $N_2O$ and $H_2O$ components of the GHG emission.

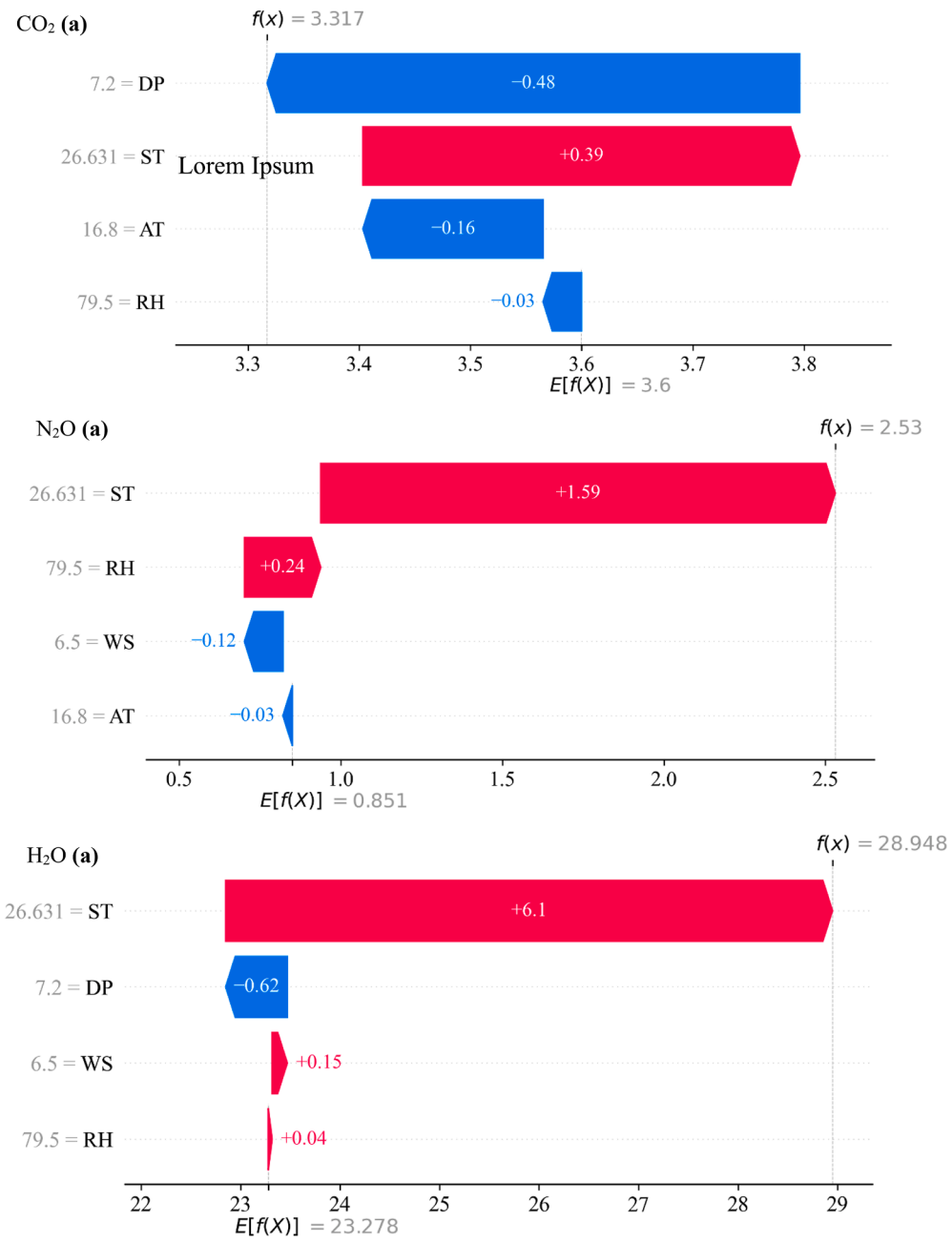



**Fig. 13.** SHAP waterfall plot at the splitting point **(a)** and SHAP summary plot **(b)** related to the best model (GBDT-RUN) to predict the GHG emission components for the training phase.

### *4.2. Applications and future work*

The developed GHG prediction model in potato farming will be a valuable tool for improving farm management and promoting sustainable agriculture. The model will provide farmers with actionable insights to optimize irrigation and soil management, reducing emissions while maintaining productivity by analyzing key emission drivers such as soil moisture and temperature. This approach supports climate-smart farming by promoting efficient use of resources like fertilizers and water, ultimately minimizing the environmental impact of potato cultivation. Additionally, the model will serve as a decision-support tool for agricultural advisors, facilitating data-driven recommendations, and its insights can guide regional planning and policymaking to promote sustainable practices across the potato industry, reducing the sector's carbon footprint.

Future research should focus on testing this model in diverse geographic locations and climates, particularly in other major potato-producing provinces in Canada and globally, to enhance its general applicability. These regions should encompass a variety of soil types, different temperature regimes, moisture levels, and diverse climates, such as arid, temperate, tropical and cold regions, to assess the robustness and adaptability of the GHG prediction model. Expanding the model to include additional GHGs and environmental factors for various crops and utilizing advanced data fusion techniques and satellite data will further strengthen its predictive capacity. Furthermore, integration with process-based models like DNDC (Denitrification-Decomposition) could also support the development of best management practices to reduce GHG emissions. Long-term studies will be crucial to track

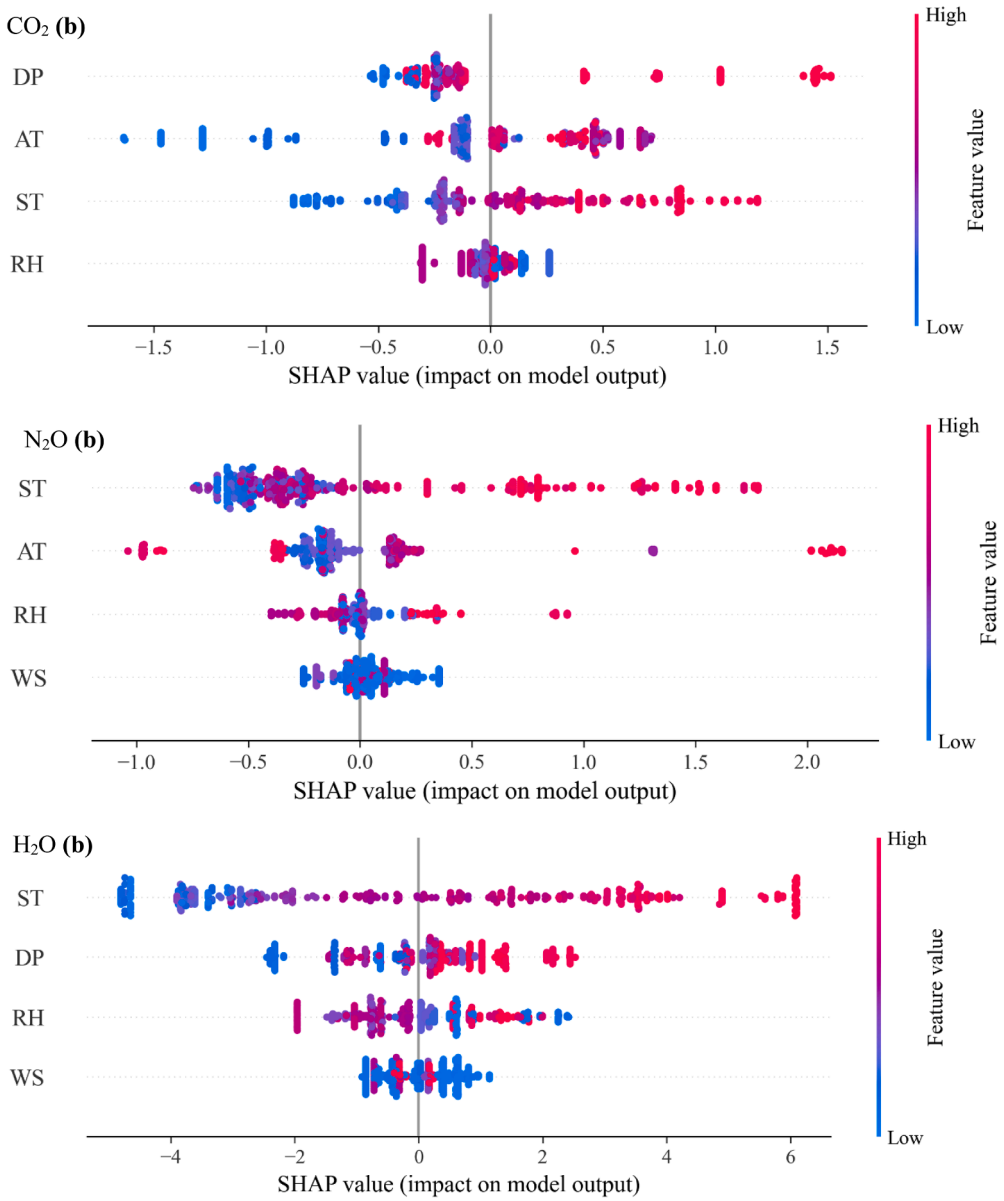



**Fig. 13.** (*continued*).

emissions trends and refine the model over time. Additionally, the creation of a user-friendly app or web-based platform will enable farmers to input field data and access real-time GHG emissions predictions, facilitating sustainable practices. Future work should also focus on improving machine learning algorithms for enhanced accuracy, transparency, and practicality, with interdisciplinary collaboration being key to addressing agricultural challenges in the context of climate change.

## 5. Conclusion

This paper proposed a meticulous multi-level explainable inspired-ensemble algorithm for experimental and computational monitoring of greenhouse gas emission components (i.e., to predict $CO_2$, $N_2O$, and $H_2O$) in Atlantic Canada during potato production. The modelling strategy adopted Boruta-GBDT feature selection based on the Z-score values in the first stage to filter the most influential features. The BSLR and WASPAS schemes are implemented to decide the best optimal combination of inputs among the top-ten possible subsets of the inputs to obtain the three best optimal candidate sets, namely Combo 1, Combo 2, and Combo 3. Next, these input combinations were used in the GBDT-RUN, LightGBM-RUN and GBDT models to predict $CO_2$, $N_2O$, and $H_2O$ components. Here, the RUN algorithm played an important role in the optimization technique. Finally, the SHAP method was implemented to explain the GBDT-RUN model's predictions. Using several goodness-of-fit metrics, the GBDT-RUN with Combo 1 discloses the highest performance in predicting $CO_2$, $N_2O$, and $H_2O$ against LightGBM-RUN and GBDT models.

The GBDT-RUN model displays maximum accuracy [$R$ = 0.8431, RMSE = 0.9337, MAE = 0.7006, MAPE = 25.2936, NSE = 0.6758, U95 % = 2.5847, SquD = 18.1759] and [$R$ = 0.9763, RMSE = 1.0665, MAE = 0.7864, MAPE = 3.4579, NSE = 0.9526, U95 % = 2.9604, SquD = 3.8300] with Combo 1 as compared to LightGBM-RUN and GBDT models to predict $CO_2$, and $H_2O$. Similarly, the GBDT-RUN model with Combo 1 surpasses the comparing models in predicting $N_2O$ components of GHG emission. This modelling strategy can be applied in other sectors, namely agronomy, water monitoring, environment, and renewables, to extend the scope and help authorities make intelligent, on-time decisions.

## Materials and correspondence

Please contact the corresponding author for data requests.

## CRediT authorship contribution statement

**Mehdi Jamei:** Visualization, Validation, Software, Methodology, Data curation, Conceptualization. **Muhammad Hassan:** Investigation, Data curation, Conceptualization, Methodology, Writing – review & editing. **Aitazaz A. Farooque:** Writing – review & editing, Supervision, Funding acquisition, Formal analysis. **Mumtaz Ali:** Writing – review & editing, Validation, Methodology. **Masoud Karbasi:** Validation, Methodology. **Gurjit S. Randhawa:** Writing – review & editing, Supervision, Funding acquisition, Project administration. **Zaher Mundher Yaseen:** Writing – original draft, Methodology. **Ross Dwyer:** Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement and funding

The authors would like to thank the Natural Sciences and Engineering Council of Canada, the Department of Environment, Energy and Climate Action, the Government of Prince Edward Island, and the Atlantic Canada Opportunities Agency for supporting the project. Special thanks go to the Research Group at the Centre of Excellence in Food Security and Sustainability, University of Prince Edward Island, for their lab and field experimentation assistance. Sincere gratitude to the participating growers in this research study.

## Data availability

The data that has been used is confidential.

## References

[1] J. Maqsood, A.A. Farooque, X. Wang, F. Abbas, B. Acharya, H. Afzaal, Contribution of climate extremes to variation in potato tuber yield in Prince Edward Island, Sustainability 12 (2020) 4937.
[2] J. Bogner, R. Pipatti, S. Hashimoto, C. Diaz, K. Mareckova, L. Diaz, P. Kjeldsen, S. Monni, A. Faaij, G. Qingxian, Z. Tianzhu, A.A. Mohammed, R.T. M. Sutamihardja, R. Gregory, Mitigation of global greenhouse gas emissions from waste: conclusions and strategies from the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report. Working Group III (Mitigation), Waste Manag. Res. (2008), https://doi.org/10.1177/0734242X07088433.
[3] M. Cellura, M.A. Cusenza, S. Longo, Energy-related GHG emissions balances: IPCC versus LCA, Sci. Total Environ. (2018), https://doi.org/10.1016/j.scitotenv.2018.02.145.
[4] V. Masson-Delmotte, P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, Global warming of 1.5 C, An IPCC Spec. Rep. Impacts Glob. Warm. 1 (2019) 93–174.
[5] G. Althor, J.E.M. Watson, R.A. Fuller, Global mismatch between greenhouse gas emissions and the burden of climate change, Sci. Rep. 6 (2016) 20281.
[6] A. Haider, The determinants of greenhouse gas emissions: empirical evidence from Canadian Provinces, Sustainability 16 (2024) 2498.
[7] U.A. Bhatti, M.A. Bhatti, H. Tang, M.S. Syam, E.M. Awwad, M. Sharaf, Y.Y. Ghadi, Global production patterns: understanding the relationship between greenhouse gas emissions, agriculture greening and climate variability, Environ. Res. 245 (2024) 118049.
[8] R. Raymundo, S. Asseng, R. Robertson, A. Petsakos, G. Hoogenboom, R. Quiroz, G. Hareau, J. Wolf, Climate change impact on global potato production, Eur. J. Agron. 100 (2018) 87–98.
[9] A. Mohammadi, A. Tabatabaeefar, S. Shahin, S. Rafiee, A. Keyhani, Energy use and economical analysis of potato production in Iran a case study: Ardabil province, Energy Convers. Manag. (2008), https://doi.org/10.1016/j.enconman.2008.07.003.
[10] J. Maqsood, Machine Learning Based Climate Projections for Sustainable Potato Production in Prince Edward Island, 2021.
[11] Agriculture and Agri-Food Canada, Potato market information review 2022-2023, (2023) 1–40.
[12] C. Oertel, J. Matschullat, K. Zurba, F. Zimmermann, S. Erasmi, Greenhouse gas emissions from soils—a review, Chemie Der Erde (2016), https://doi.org/10.1016/j.chemer.2016.04.002.
[13] A. Mehmandoost Kotlar, J. Singh, S. Kumar, Prediction of greenhouse gas emissions from agricultural fields with and without cover crops, Soil Sci. Soc. Am. J. (2022), https://doi.org/10.1002/saj2.20429.
[14] S.M. Ogle, C. Alsaker, J. Baldock, M. Bernoux, F.J. Breidt, B. McConkey, K. Regina, G.G. Vazquez-Amabile, Climate and soil characteristics determine where no-till management can store carbon in soils and mitigate greenhouse gas emissions, Sci. Rep. (2019), https://doi.org/10.1038/s41598-019-47861-7.
[15] S.R. Ehsani Amrei, L. Babu-Saheer, C. Luca, ML-Based prediction of carbon emissions for potato farms in Iran, in: IFIP Adv. Inf. Commun. Technol., 2023, https://doi.org/10.1007/978-3-031-34107-6_28.
[16] K. Lloyd, C.A. Madramootoo, K.P. Edwards, A. Grant, Greenhouse gas emissions from selected horticultural production systems in a cold temperate climate, Geoderma (2019), https://doi.org/10.1016/j.geoderma.2019.04.030.
[17] A. Hamrani, A. Akbarzadeh, C.A. Madramootoo, Machine learning for predicting greenhouse gas emissions from agricultural soils, Sci. Total Environ. (2020), https://doi.org/10.1016/j.scitotenv.2020.140338.
[18] M. Taki, S. Abdanan Mehdizadeh, A. Rohani, M. Rahnama, M. Rahmati-Joneidabad, Applied machine learning in greenhouse simulation; new application and analysis, Inf. Process. Agric. (2018), https://doi.org/10.1016/j.inpa.2018.01.003.
[19] G.A. Taiwo, T.O. Akinwole, O.B. Ogundepo, Statistical analysis of stakeholders perception on adoption of AI/ML in sustainable agricultural practices in rural development, in: Int. Congr. Inf. Commun. Technol., Springer, 2024, pp. 123–131.
[20] H. Dehghanisanij, B. Yargholi, S. Emami, H. Emami, H. Fujimaki, A hybrid extreme learning machine approach for modeling the effectiveness of irrigation methods on greenhouse gas emissions, Environ. Dev. Sustain. (2024), https://doi.org/10.1007/s10668-024-04644-z.
[21] S. Sharafi, A. Kazemi, Z. Amiri, Estimating energy consumption and GHG emissions in crop production: a machine learning approach, J. Clean. Prod. (2023), https://doi.org/10.1016/j.jclepro.2023.137242.
[22] C. Wang, X. Xu, Y. Zhang, Z. Cao, I. Ullah, Z. Zhang, M. Miao, A stacking ensemble learning model combining a crop simulation model with machine learning to improve the dry matter yield estimation of greenhouse Pakchoi, Agronomy 14 (2024) 1789, 2024.
[23] G. Kaur, Rajni, J.S. Sivia, Integrating data envelopment analysis and machine learning approaches for energy optimization, decreased carbon footprints, and wheat yield prediction across north-western India, J. Soil Sci. Plant Nutr. 24 (2024) 1424–1447.
[24] E. Harsányi, M. Mirzaei, S. Arshad, F. Alsilibe, A. Vad, A. Nagy, T. Ratonyi, M. Gorji, M. Al-Dalahme, S. Mohammed, Assessment of advanced machine and deep learning approaches for predicting CO2 emissions from agricultural lands: insights across diverse agroclimatic zones, Earth Syst. Environ. (2024) 1–17.
[25] M. Hassan, K. Khosravi, A.A. Farooque, T.J. Esau, A. Boluwade, R. Sadiq, Prediction of carbon dioxide emissions from Atlantic Canadian potato fields using advanced hybridized machine learning algorithms–nexus of field data and modelling, Smart Agric. Technol. 9 (2024) 100559.
[26] T.K. Abdelkader, H.A.A. Sayed, M. Refai, M.M. Ali, Y. Zhang, Q. Wan, I. Khalifa, Q. Fan, Y. Wang, M.A. Abdelhamid, Machine learning, mathematical modeling and 4E (energy, exergy, environmental, and economic) analysis of an indirect solar dryer for drying sweet potato, Renew. Energy. 227 (2024) 120535.
[27] K.G. Liakos, P. Busato, D. Moshou, S. Pearson, D. Bochtis, Machine learning in agriculture: a review, Sensors 18 (2018) 2674.
[28] S. Petrakis, A. Seyfferth, J. Kan, S. Inamdar, R. Vargas, Influence of experimental extreme water pulses on greenhouse gas emissions from soils, Biogeochemistry 133 (2017) 147–164, https://doi.org/10.1007/s10533-017-0320-2.
[29] S. Petrakis, J. Barba, B. Bond-Lamberty, R. Vargas, Using greenhouse gas fluxes to define soil functional types, Plant Soil 423 (2018) 285–294, https://doi.org/10.1007/s11104-017-3506-4.
[30] Z. He, R. Larkin, W. Honeycutt, Sustainable potato production: global case studies, Sustain. Potato Prod. Glob. Case Stud 9789400741 (2012) 1–539, https://doi.org/10.1007/978-94-007-4104-1.
[31] Onset, HOBO® RX3000 remote monitoring station manual, (2023).
[32] H. Tao, S.M. Awadh, S.Q. Salih, S.S. Shafik, Z.M. Yaseen, Integration of extreme gradient boosting feature selection approach with machine learning models: application of weather relative humidity prediction, Neural Comput. Appl. (2021), https://doi.org/10.1007/s00521-021-06362-3.
[33] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.
[34] H. Zhang, W. Wu, H. Wu, TOC prediction using a gradient boosting decision tree method: a case study of shale reservoirs in Qinshui Basin, Geoenergy Sci. Eng. 221 (2023) 111271.
[35] H. Qian, B. Wang, M. Yuan, S. Gao, Y. Song, Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree, Expert Syst. Appl. 190 (2022) 116202.
[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[37] A. Shehadeh, O. Alshboul, R.E. Al Mamlook, O. Hamedat, Machine learning models for predicting the residual value of heavy construction equipment: an evaluation of modified decision tree, LightGBM, and XGBoost regression, Autom. Constr. 129 (2021) 103827, https://doi.org/10.1016/j.autcon.2021.103827.
[38] M.R. Machado, S. Karray, I.T. De Sousa, LightGBM: an effective decision tree gradient boosting method to predict customer loyalty in the finance industry, in:

14th Int. Conf. Comput. Sci. Educ. ICCSE 2019, 2019, pp. 1111–1116, https://doi.org/10.1109/ICCSE.2019.8845529.
[39] Z.M. Yaseen, O.A. Alawi, A.M. Alshammari, A. Alsuwaiyan, M.O. Oyedeji, A. Y. Oudah, Development of advanced data-intelligence models for radial gate discharge coefficient prediction: modeling different flow scenarios, Water Resour. Manag. (2023), https://doi.org/10.1007/s11269-023-03624-8.
[40] J. Guo, S. Yun, Y. Meng, N. He, D. Ye, Z. Zhao, L. Jia, L. Yang, Prediction of heating and cooling loads based on light gradient boosting machine algorithms, Build. Environ. 236 (2023) 110252, https://doi.org/10.1016/j.buildenv.2023.110252.
[41] A.M. Al-Areeq, S.I. Abba, B. Halder, I. Ahmadianfar, S. Heddam, V. Demir, H. C. Kilinc, A.A. Farooque, M.L. Tan, Z.M. Yaseen, Flood subsidence susceptibility mapping using elastic-net classifier: new approach, Water Resour. Manag. (2023) 1–22.
[42] I. Ahmadianfar, A.A. Heidari, A.H. Gandomi, X. Chu, H. Chen, RUN beyond the metaphor: an efficient optimization algorithm based on Runge Kutta method, Expert Syst. Appl. 181 (2021) 115079, https://doi.org/10.1016/j.eswa.2021.115079.
[43] C. Runge, Über die numerische Auflösung von Differentialgleichungen, Math. Ann. 46 (1895) 167–178.
[44] W. Kutta, Beitrag Zur Näherungsweisen Integration totaler Differentialgleichungen, Teubner, 1901.
[45] G. Manikandan, B. Pragadeesh, V. Manojkumar, A.L. Karthikeyan, R. Manikandan, A.H. Gandomi, Classification models combined with Boruta feature selection for heart disease prediction, Informatics Med. Unlocked. 44 (2024) 101442.
[46] M.B. Kursa, A. Jankowski, W.R. Rudnicki, Boruta – A System for Feature Selection, Fundam. Informaticae 101 (2010) 271–285, https://doi.org/10.3233/FI-2010-288.
[47] M.B. Kursa, W.R. Rudnicki, Feature selection with the Boruta package, J. Stat. Softw. 36 (2010), https://doi.org/10.18637/jss.v036.i11.
[48] Q. Qiao, A. Yunusa-Kaltungo, R.E. Edwards, Developing a machine learning based building energy consumption prediction approach using limited data: boruta feature selection and empirical mode decomposition, Energy Rep. 9 (2023) 3643–3660.
[49] A. Jović, K. Brkić, N. Bogunović, A review of feature selection methods with applications, in: 2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron., IEEE, 2015, pp. 1200–1205.
[50] M. Kobayashi, S. Sakata, Mallows' Cp criterion and unbiasedness of model selection, J. Econom. 45 (1990) 385–395.
[51] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Contr. 19 (1974) 716–723.
[52] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B. 58 (1996) 267–288.
[53] M. Jamei, F. Karimi, M. Ali, B. Karimi, M. Karbasi, Y. Aminpour, Experimental and computational assessment of wetting pattern for two-layered soil profiles in pulse drip irrigation: designing a novel optimized Bidirectional deep learning paradigm, J. Hydrol. (2022) 128496.
[54] U.K. Singh, M. Jamei, M. Karbasi, A. Malik, M. Pandey, Application of a modern multi-level ensemble approach for the estimation of critical shear stress in cohesive sediment mixture, J. Hydrol. (2022) 127549.
[55] E.K. Zavadskas, Z. Turskis, J. Antucheviciene, A. Zakarevicius, Optimization of weighted aggregated sum product assessment, Elektron. Ir Elektrotechnika 122 (2012) 3–6.
[56] A. Mardani, M. Nilashi, N. Zakuan, N. Loganathan, S. Soheilirad, M.Z.M. Saman, O. Ibrahim, A systematic review and meta-Analysis of SWARA and WASPAS methods: theory and applications with recent fuzzy developments, Appl. Soft Comput. 57 (2017) 265–292.
[57] S. Chakraborty, E.K. Zavadskas, Applications of WASPAS method in manufacturing decision making, Informatica 25 (2014) 1–20.
[58] M. Vafaeipour, S.H. Zolfani, M.H.M. Varzandeh, A. Derakhti, M.K. Eshkalag, Assessment of regions priority for implementation of solar projects in Iran: new application of a hybrid multi-criteria decision making approach, Energy Convers. Manag. 86 (2014) 653–663.
[59] O. Bozorg-Haddad, A. Azarnivand, S.-M. Hosseini-Moghari, H.A. Loáiciga, Development of a comparative multiple criteria framework for ranking pareto optimal solutions of a multiobjective reservoir operation problem, J. Irrig. Drain. Eng. 142 (2016) 4016019.
[60] B. Debnath, A.B.M.M. Bari, M.M. Haq, D.A. de Jesus Pacheco, M.A. Khan, An integrated stepwise weight assessment ratio analysis and weighted aggregated sum product assessment framework for sustainable supplier selection in the healthcare supply chains, Supply Chain Anal 1 (2023) 100001.
[61] K. Topuz, A. Bajaj, I. Abdulrashid, Interpretable machine learning, in: Proc. Annu. Hawaii Int. Conf. Syst. Sci., 2023, pp. 1236–1237, https://doi.org/10.1201/9780367816377-16, 2023-Janua.
[62] M. Jamei, M. Ali, M. Karbasi, B. Karimi, N. Jahannemaei, A.A. Farooque, Z. M. Yaseen, Monthly sodium adsorption ratio forecasting in rivers using a dual interpretable glass-box complementary intelligent system: hybridization of ensemble TVF-EMD-VMD, Boruta-SHAP, and eXplainable GPR, Expert Syst. Appl. 237 (2024) 121512.
[63] A. El Bilali, T. Abdeslam, N. Ayoub, H. Lamane, M.A. Ezzaouini, A. Elbeltagi, An interpretable machine learning approach based on DNN, SVR, Extra Tree, and XGBoost models for predicting daily pan evaporation, J. Environ. Manage. 327 (2023) 116890, https://doi.org/10.1016/j.jenvman.2022.116890.
[64] M. Kuzlu, U. Cali, V. Sharma, Ö. Güler, Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools, IEEE Access 8 (2020) 187814–187823.
[65] M. Jamei, I. Ahmadianfar, X. Chu, Z.M. Yaseen, Estimation of triangular side orifice discharge coefficient under a free flow condition using data-driven models, Flow Meas. Instrum. (2020) 101878.
[66] M. Jamei, M. Karbasi, O.A. Alawi, H.M. Kamar, K.M. Khedher, S.I. Abba, Z. M. Yaseen, Earth skin temperature long-term prediction using novel extended Kalman filter integrated with Artificial Intelligence models and information gain feature selection, Sustain. Comput. Informatics Syst. 35 (2022) 100721.
[67] I. Ahmadianfar, M. Jamei, X. Chu, Prediction of local scour around circular piles under waves using a novel artificial intelligence approach, Mar. Georesources Geotechnol. 0 (2019) 1–12, https://doi.org/10.1080/1064119X.2019.1676335.