Computers and Electronics in Agriculture 238 (2025) 110831

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

eXplainable expert systems for potato tuber yield prediction in the Maritime provinces of Canada

Mehdi Jamei[a,b,*], Aitazaz Ahsan Farooque[a,c,*], Mumtaz Ali[d], Masoud Karbasi[a], Hassan Afzaal[c], Saad Javed Cheema[a], Qamar Uz Zaman[e], Kok Sin Woon[f], Paul Sheridan[g]

[a] *Canadian Centre for Climate Change and Adaptation, University of Prince Edward Island, St Peters Bay, PE, Canada*
[b] *Department of Civil Engineering, Faculty of Civil Engineering and Architecture, Shahid Chamran University of Ahvaz, Ahvaz, Iran*
[c] *Faculty of Sustainable Design Engineering, University of Prince Edward Island, Charlottetown, PE, Canada*
[d] *UniSQ College, University of Southern Queensland, QLD 4305, Australia*
[e] *Engineering Department, Faculty of Agriculture, Dalhousie University, Truro, NS, Canada*
[f] *Carbon Neutrality and Climate Change Thrust, Society Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, China*
[g] *School of Mathematical and Computational Sciences, University of Prince Edward Island, Charlottetown, PE, Canada*

ARTICLE INFO

Keywords:
Potato tuber yield
Explainable machine learning
MOORA, Boruta-SHAP
LightGBM
CatBoost

ABSTRACT

The potato crop is vital to the economy of Canada's Maritime provinces. *Prince Edward Island* (PEI) and *New Brunswick* (NB) contribute significantly to Canadian potato production and gross domestic product (GDP). Estimating potato tuber yields helps farmers to make informed decisions for sustainable and profitable farming. This study investigated fluctuations in tuber yield based on 30 soil properties gathered over four seasons through experimental trials. An emerging eXplainable high-dimensional feature vote-based ensemble framework explained with SHAP (SHapley Additive exPlanations) tool was employed to estimate potato tuber yield accurately. In order to develop the model, the most influential feature was first filtered using the Boruta-SHAP feature selection. Afterwards, the most deserved combinations (four scenarios) were ascertained using the best subset regression (BSR) integrated with two Multi-Criteria Decision-Making (MCDM), namely Weighted Aggregated Sum Product Assessment (WASPAS) and Multi-Objective Optimization methods were adopted based on the Ratio Analysis (MOORA). To estimate potato tuber yield, we adopted a novel explainable ensemble machine learning model, called VOTE-LGCB, that combines voted Categorical Boosting and the Light Gradient-Boosting Machine framework. We evaluated our approach against Least Absolute Shrinkage and Selection Operator (LASSO) regression, elastic net regression, Extra Tree, classical LightGBM, and classical CatBoost baselines. Six metric performances such as correlation coefficient (R), root mean square error (RMSE), and reliability were implemented to validate the multi-process ML models. All the metrics were singularized using WASPAS and MOORA to determine the best input combination related to each model separately. We found that the VOTE-LGCB-Combo 3 outperformed baseline methods (R = 0.8958, RMSE = 5088.5087, Reliability = 93.7500, WASPAS = 0.00023, and MOORA = 0.3788). Moisture content was identified as the most significant feature, followed by the Normalized Difference Vegetation Index (NDVI). The modeling framework we advance can be used as a reliable simulation system for various aspects of agricultural production systems that involve high-dimensional features.

1. Introduction

Potatoes rank as the fifth most significant primary agricultural product in Canada. In 2022, they contributed to the Canadian economy with an impressive1.7 billion Canadian Dollars (CAD) in farm cash receipts. The export of potatoes and potato products bolstered the economy with a substantial 3.4 billion CAD in the fiscal year 2022–2023 (Canada, 2023). *Prince Edward Island* and *New Brunswick* are major potato (Solanum tuberosum) producing provinces in Canada, contributing over 38 % of the country's total potato production (Agriculture and Agri-Food Canada (AAFC), 2018). Potato production in both provinces contributes significantly to the provincial gross domestic product (GDP) and provides a livelihood for many residents. Optimizing the growth and yield of potatoes is crucial for maintaining food stability amidst the rise of the world's population (Liu et al., 2023). Prompt and efficient

[*] Corresponding authors at: Canadian Centre for Climate Change and Adaptation, University of Prince Edward Island, St Peters Bay, PE, Canada.
E-mail addresses: jmehdi@upei.ca (M. Jamei), afarooque@upei.ca (A.A. Farooque), koksin.woon@xmu.edu.my (K.S. Woon), pasheridan@upei.ca (P. Sheridan).

https://doi.org/10.1016/j.compag.2025.110831
Received 19 July 2024; Received in revised form 29 July 2025; Accepted 30 July 2025

Nomenclature

List of abbreviations and symbols used.

(PC)	Amemiyas's prediction criterion index
(BSR)	Best-subset regression
(Ca)	Calcium
(CatBoost)	Categorical boosting
(CatBoost)	Categorical boosting
(CEC)	Cation exchange capacity
Ad-R^2	Adjusted R^2
(CV)	Coefficient of variance
(ENET)	Elastic Net
(ELNET)	elastic net regression
(GIS)	Geographic information system
(GS +)	Geostatistics for environmental sciences
(GPS)	Global positioning systems
(GDP)	Gross domestic product
(HCP)	Horizontal coplanar geometry
(IDW)	Inverse distance weighting iron
(KGE)	Kling–gupta efficiency
(LASSO)	Lasso regression
(LightGBM)	Light gradient-boosting
(LI)	Lime Index
(LOI)	Loss on ignition
(ML)	Machine learning
(C_p)	Mallow's factor
(MZs)	Management zones
(MZSA)	Maximum Z-score
(MSE)	mean square error
(θ)	Moisture content
(MCDM)	Multi-criteria decision-making
(MOORA)	Multi-objective optimization method based on ratio analysis
(rNIR)	Near-infrared
(NDVI)	Normalized difference vegetation index
(N/S)	Nugget-to-sill ratio
(P)	Phosphorous
(K)	Potassium
(PA)	Precision agriculture
(PEI)	*Prince Edward Island*
(RTK-GPS)	Real-time kinematic global positioning system
(RLR)	Red-light reflectance
(RSS)	Residual sum of squares
(RMSE)	Root mean square error
(SHAP)	Shapley additive explanation
(SMSMP)	Slightly modified shoemaker-McLean-pratt
SOMC	Soil organic matter content
SD	Standard deviation
(TS)	Target Statistics

Time domain reflectometry

VOTE-LGCB	Voting-based super ensemble model,
(WASPAS)	Weighted aggregated sum product assessment

surveillance of potatoes is essential for enhancing fertilizer utilization and forecasting agricultural yield (Liu et al., 2024). Conventionally, growers rely on their experiences, past weather, and crop yield data to make crucial decisions to increase long-term sustainability and short-term profitability (Arbuckle and Rosman, 2014). The recommended practices of applying crop and soil inputs such as macronutrients, micronutrients, organic matter, and water vary spatially with crop-specific requirements. Furthermore, the interaction of various soil properties adds complexity to this problem. Under these circumstances, there is a dire need to study individual factors influencing crop yield and their co-dependencies to predict potato crop yield accurately.

Various techniques have been used to understand the relationship between crop inputs, soil properties, and crop yield. Multiple linear regression was evaluated by Kravchenko & Bullock (2000) and Kbakural et al. (1999). However, the results are unsatisfactory due to collinearities among some predictor variables. Machine learning (ML) has emerged as a powerful tool for crop yield prediction, providing improved crop yield predictions by learning valuable patterns and relationships from input data. ML algorithms have been applied to predict crop yields with more encouraging results (Klompenburg et al., 2020). A study by Kuradusenge et al. (2023) applied ML techniques to predict crop yield based on weather data for Musanze District, Rwanda. The study found that a random forest model was the best model for early yield prediction. Another study by Das et al. (2023) introduced a novel hybrid approach, combining ML algorithms with feature selection, for efficient modeling and forecasting of crop yield. The proposed ML hybrid models outperformed the individual models.

Recent developments in explainable artificial intelligence (XAI) have demonstrated encouraging applications in processing agricultural data. A thorough review by (Ahmed et al., 2025) demonstrated that XAI approaches, including SHAP and LIME, are increasingly utilized for spectroscopic agricultural quality evaluation, enhancing the transparency and reliability of predictive models. Likewise, (Liu et al., 2025) illustrated that UAV-based hyperspectral remote sensing, in conjunction with agronomic characteristics, can proficiently monitor potato growth and precisely estimate yield, highlighting the significance of amalgamating spectral and soil-related data. Furthermore, (Paudel et al., 2023) underscored the importance of interpretability in deep learning models for crop yield forecasting, proposing that transparent and elucidative models improve decision-making in precision agriculture. These studies collectively highlight the necessity for precise and interpretable models, prompting our emphasis on the SHAP-driven Vote-LGCB framework for predicting potato yield in the Maritime provinces of Canada. Multiple soil, environmental, and agronomic factors all affect potato tuber yield. Climatic variables, including temperature (Ezekeil, 1997), humidity (Pereira et al., 2009), and precipitation (Wurr et al., 2001), delineate the climatic conditions for potato cultivation, directly influencing tuber size, quality, and overall production. In addition to climate, biotic stresses such as pests and diseases (Hernandez Nopsa et al., 2014) can significantly diminish yield and crop health if inadequately handled. Soil parameters, including moisture content, pH, organic matter, and cation exchange capacity, are equally vital, since they affect nutrient availability, water retention, and plant growth during the growing season. Although significant, the influence of soil parameters on yield prediction has been insufficiently highlighted in current predictive modeling research, which frequently depends predominantly on satellite or meteorological data. This work systematically analyzes soil variables alongside explainable machine learning algorithms to enhance the interpretability and accuracy of yield estimates.

However, in this study we aim to model potato tuber yield in relation to soil properties and assess how variations in soil composition, such as nutrient levels, pH, and texture, interact with climatic conditions to influence overall yield performance. Several studies on potato yield modeling suggested the importance of soil fertility and soil moisture (Abbas et al., 2020). Ensign (1935) suggested that soil moisture and temperature influence potato yield. However, the effect of individual factors and their co-dependence needs to be investigated for specific regions and crops. Yield modeling is a complex procedure with multiple variables influencing the response, causing challenges in studying these responses. Both belowground and above-ground factors influence crop yields. Notable research by Liu et al., 2024, and Liu et al. 2022 investigated the aboveground biomass of potatoes across various growth stages, revealing a unique pattern compared to other crops such as wheat and maize. Unlike these crops, potato biomass initially increases during growth but subsequently decreases. This distinct trend underscores the critical role of the vegetative growth stage in developing yield prediction models for potatoes. In the current study, below and aboveground data were integrated, with readily accessible soil test reports and proximal sensor measurements being leveraged to provide detailed yield estimates over the growing season.

Feature selection is an accurate and precise method to estimate crop yield within a given dataset by filtering out the most relevant data in large databases. It helps reduce redundancy, remove irrelevant data, increase learning accuracy, and improve result comprehensibility (Anukrishna and Paul, 2017). Different feature selection methods have been developed and applied in various fields (Parmar and Bhatt, 2022). These methods have evolved to cope with the challenges posed by the advent of big data and the increasing dimensionality of datasets (Ray et al., 2021). High-dimensional feature filtering using a multi-level ensemble approach is an advanced ML and data analysis method. This approach is particularly effective in dealing with high-dimensional data, where traditional methods may struggle due to the curse of dimensionality (Ben Brahim and Limam, 2018). The multi-level ensemble approach combines independent feature subsets to better approximate the optimal subset of features. It aims to provide unique and stable feature selection without compromising predictive accuracy (Kumar and Minz, 2016). Recently, several studies used Shapley Additive Explanation (SHAP) (Parsa et al., 2020; Wieland et al., 2021) to understand and interpret ML models. Shapley (1953) introduced SHAP to provide insights into individual and combined feature contributions in model predictions. Lundberg & Lee (2017) developed a visualisation-based Python package implementation of SHAP to aid human intuition in model interpretation and predictions. For a particular prediction, a specific value is assigned to each feature by SHAP (Lundberg and Lee, 2017). The SHAP values identify a new class of feature importance measures and determines a theoretical unique solution with desirable feature properties.

This research first conducted a comprehensive field investigation to measure fluctuations in tuber yield based on 30 soil physical properties (such as soil moisture, electrical conductivity, and slope) and chemical properties (including micronutrients, pH level, and organic matter). Potato fields were divided into 36 to 40 spatial grids in Prince Edward Island (PEI) and New Brunswick (NB), Canadian provinces. To model potato yield, based on ratio analysis (MOORA), we employed a novel eXplainable high-dimensional feature vote-based ensemble framework coupled with two multi-criteria Decision-Making (MCDM) methods, namely weighted aggregated sum product assessment (WASPAS) and multi-objective optimisation methods. To do this, Boruta-SHAP feature selection, best subset regression (BSR), and WASPAS & MOORA schemes were incorporated to indicate the best input combination among 30 features. Here, a novel explainable ensemble model entailing voted CatBoost and LightGBM, called (VOTE-LGCB), has been developed to estimate potato tuber yield accurately. The ELNET, LASSO, Extra Tree, LightGBM, and CatBoost were considered comparative models to validate the robustness of the main model. Also, the SHAP tool was adopted to explain the influence of each feature during the training process. We used statistical indices, graphical investigations, and diagnostic analyses to validate and assess the accuracy of the various procedures.

This research presents a novel approach for predicting potato tuber yields by employing a comprehensive field investigation of physical and chemical soil properties across extensive spatial grids with an eXplainable high-dimensional feature-based ensemble framework and an explainable ensemble model. The potato tuber yield models will provide a valuable tool for optimising crop management and enhancing food production efficiency.

Table 1
Details about research sites, years, datasets, ML algorithm training and testing data, and potato fields utilized for data collecting.

Province	Year	Dataset name	Training points	Testing points	Fields location
Prince Edward Island	2019	PE-2019	80	40	Field 1 Field 2 Field 3
	2020	PE-2020	80	40	Field 1 Field 2 Field 3
New Brunswick	2019	NB-2019	80	40	Field 1 Field 2 Field 3
	2020	NB-2020	80	40	Field 1 Field 2 Field 3

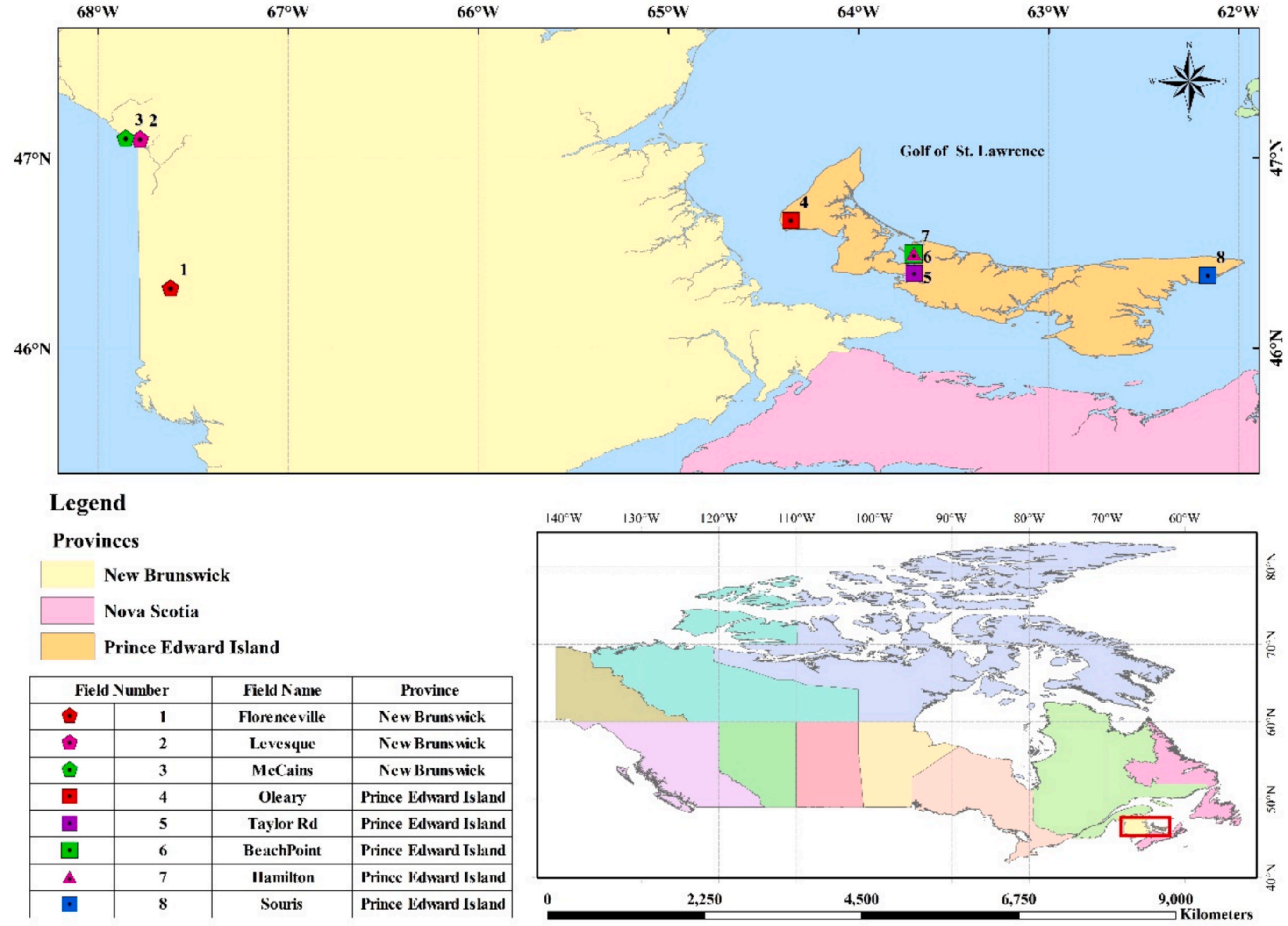
2. Material and methods

2.1. Study area and filed investigation description

Physicochemical property data were gathered from three fields in PEI and three in NB over the growing seasons of 2019 and 2020, respectively (Table 1). Soil samples were collected from each field following a grid plan. Approximately 36 to 40 30-meter-by-30-meter grids were generated utilizing a Topcon Positioning System Inc. Differential Global Positioning System (Topcon Positioning Systems, Inc., Livermore, USA). During the summer seasons of 2019 and 2020, four data collection events occurred: the initial sampling occurred in late May when seeds were planted; the second sampling occurred in mid-late June when plants emerged; the third sampling occurred in late July, over the eighty-day period, and the fourth sampling occurring in late August. The various samplings were performed during the growing season to get insight into the seasonal variations of chosen variables. In each year, one dataset was compiled from each province's three fields (Fig. 1) to represent the variation across many fields inside a single dataset. Every field was planted with the potato variety of Russet Burbank. The cut seeds were seeded during the 2019 and 2020 growing seasons, and the selected fields were harvested in early October, respectively. The research fields were comprised of sandy loam soil (Orthic Humo-Ferric Podzol). Over the previous decade, traditional agronomic procedures were maintained in all fields for various crop cycles, including the potato as a primary rotation crop (Farooque et al., 2019). Inter-row spacing was 0.9 m, whilst the gap between plants was 0.3 m.

2.2. Proximal sensors data

Physiochemical characteristics of the fields were assessed at each sampling date for both years using sensors to quantify the following parameters: slope, normalized difference vegetation index (NDVI), and volumetric moisture content; soil electrical conductivity parameters (specifically HCP and PRP arrays (Taylor, n.d.). A DualEM-2 sensor (DualEM Inc., Milton, Canada) was manually positioned alongside potato furrows to obtain HCP and PRP measurements, ensuring no metallic items encountered the instrument. Five readings were recorded at each grid within a two-meter radius. Near the locations where HCP/PRP measurements were recorded, five random volumetric moisture content values at 15 cm depths were obtained using a FieldScout TDR 350 (Spectrum Technologies, Aurora, USA). Using a portable slope meter (Mastercraft Torpedo Level, Vonore, USA), the field slope was measured three times in a parallel direction to the plant furrows at each position. At 0.5 m from the potato plants, the FieldScout CM 1000 NDVI Meter (Spectrum Technologies, Aurora, USA) was used to determine the NDVI, representing plant growth. Certain NDVI measurements were excluded during the planting phase in the absence of vegetation. A representative



Field Number		Field Name	Province
⬟	1	Florenceville	New Brunswick
⬟	2	Levesque	New Brunswick
⬟	3	McCains	New Brunswick
🟥	4	Oleary	Prince Edward Island
🟪	5	Taylor Rd	Prince Edward Island
🟩	6	BeachPoint	Prince Edward Island
▲	7	Hamilton	Prince Edward Island
🟦	8	Souris	Prince Edward Island

Fig. 1. The field investigation geographically to assess the tuber potato yield in Atlantic Canada.

measurement for all sensor data was calculated by averaging five readings at each site.

2.3. Soil sampling

In each growing season, soil samples were collected at each grid point within selected fields during the first and third samplings. Using a soil auger, three soil samples were extracted from a depth of 15 cm at each sampling position. Soil cores were mixed to create a representative sample from each sampled grid point. Protocol-compliant procedures were used to analyze the soil samples at the PEI Analytical Laboratory (Charlottetown, Canada). The degradation of organic matter (SOM), cation exchange capacity (CEC), and soil pH were accomplished using established methodologies, such as titration with a PC titration instrument (ManSci Inc., Orlando, USA) (Taylor, n.d.), the loss-on-ignition technique (Condie, 1993) employing a Combustion Analyzer model CN628 (LECO Corporation, St. Joseph, USA), and the Sodium Acetate Method (Carter and Gregorich, 2007). Soil macro and micronutrients were analyzed using standard methods at the analytical laboratory.

2.4. Tuber yield data collection

During the harvesting season, tuber yield samples from each grid were collected manually to explain the role of soil variables in fluctuating crop yield. The yield samples were collected in early October each year. A designated area of 2.7 m^2 was marked out within each grid to collect tuber yield samples. This area was used to dig potato furrows and collect tuber yield samples. The collected samples were placed in separate plastic buckets, and the weight of the samples was recorded in kilograms (kg) using a computerized field weighing balance. After weighing and recording the weights, the dug potato tubers were put back in the soil for harvesting.

2.5. Data Description

Table 2 lists descriptive statistics for tilled potato soil properties according to several statistical indices. The maximum kurtosis and skewness are related to slope4 and slope1 (1.99 and 1.14), and minimum kurtosis and skewness are related to the NDVI2 and BS1 (−1.53 and 0.01), respectively.

- HCP (horizontal co-planar) represents the configuration in dual-EM instruments, where the coils are aligned horizontally on the same plane, capturing shallow subsurface electrical conductivity. PRP (Perpendicular Co-Planar) denotes the perpendicular alignment of the coils, measuring electrical conductivity at greater depths and providing complementary insights into soil properties. MC (Moisture Content) denotes the volumetric water content of the soil. Slope (Terrain Slope) indicates the gradient of the land surface recorded from a handheld slope sensor. NDVI (Normalized Difference Vegetation Index) is derived from a handheld sensor to assess vegetation health. OM (Organic Matter) represents the percentage of decomposed plant and animal residues in the soil. PH (Soil pH) measures soil acidity or alkalinity. The phosphorus-to-aluminium (PAL) ratio

Table 2
Statistical properties of all the field features gathered from the PEI and NB potato fields.

Index	Minimum	Maximum	Mean	Std Dev	Cov	Skewness	Kurtosis
HCP1	1.66	13.10	5.91	1.98	33.56 %	0.84	0.90
PRP1	0.90	11.18	4.30	1.75	40.80 %	0.78	0.83
MC1	3.20	41.62	13.60	6.68	49.11 %	0.70	−0.24
Slope1	0.10	8.74	2.29	1.35	59.08 %	1.14	1.99
NDVI1	0.00	0.43	0.07	0.08	111.9 %	0.72	0.10
OM1	0.50	7.00	2.98	1.03	34.51 %	0.87	0.80
PH1	4.80	7.30	5.92	0.38	6.341 %	0.35	0.91
PAl1	1.51	32.67	12.37	4.29	34.65 %	0.84	2.29
CEC1	3.00	17.00	8.67	3.02	34.86 %	0.55	0.05
BS1	17.30	99.60	66.89	19.52	29.19 %	0.01	−0.60
HCP2	2.12	12.80	6.69	2.07	30.88 %	0.25	−0.31
PRP2	1.05	13.00	5.07	1.98	39.01 %	0.73	0.41
MC2	3.80	38.66	14.22	5.74	40.39 %	0.59	0.31
Slope2	0.10	7.90	2.41	1.33	55.25 %	0.80	0.79
NDVI2	0.10	0.89	0.53	0.28	51.53 %	−0.49	−1.53
HCP3	1.72	15.28	6.73	2.29	34.06 %	0.41	0.41
PRP3	0.83	12.18	5.24	2.05	39.18 %	0.32	−0.28
MC3	3.11	36.42	13.98	5.94	42.50 %	0.87	1.04
Slope3	0.10	8.54	2.43	1.39	57.24 %	0.91	0.98
NDVI3	0.41	0.99	0.82	0.09	11.33 %	−1.05	2.24
OM3	0.80	6.80	3.02	0.95	31.39 %	0.79	0.68
PH3	4.25	7.10	5.37	0.46	8.489 %	0.56	0.65
PAl3	0.05	36.90	15.06	5.70	37.82 %	0.72	0.93
CEC3	5.00	21.00	12.01	3.10	25.84 %	0.27	−0.56
BS3	19.00	110.00	55.17	18.21	33.01 %	0.67	−0.17
HCP4	1.51	12.80	5.67	1.92	33.86 %	0.50	0.16
PRP4	0.56	9.90	4.20	1.67	39.69 %	0.57	0.64
MC4	0.65	26.33	9.59	5.34	55.71 %	0.09	−0.33
Slope4	0.10	10.48	2.35	1.38	58.54 %	1.10	2.36
NDVI4	0.62	1.00	0.86	0.07	8.187 %	−0.81	−0.01
Yield	6.10	23.20	13.08	3.08	23.55 %	0.43	0.16

evaluates the relationship between available phosphorus (P) in the soil and extractable aluminium (AL), which can influence phosphorus availability in plants. CEC (Cation Exchange Capacity, like calcium, magnesium, and potassium) reflects the soil's ability to retain and exchange positively charged ions. BS (Base Saturation) is the portion of the cation exchange capacity occupied by basic cations such as calcium, magnesium, potassium, and sodium. It should be mentioned that the unit of yield is Kg/ha.

- Data sampling seasonal stages: 1: May; 2: July; 3: August; 4: September.

They reveal that the leptokurtic distribution is between (−3 and 3) due to kurtosis. At the same time, the remaining datasets, due to falling into the allowable range of kurtosis (−3 and 3) (Jamei et al., 2021c), are taken into account as the mesokurtic (near 3) and platykurtic (less than 3) distributions. Fig. 2 (above) demonstrates the normalized values of all the datasets used for predicting potato tuber yield utilising a novel, explainable, intelligent expert system. Furthermore, Fig. 2 (bottom) illustrates the relationship between 30 input features and the goal variable (yield). An accurate evaluation of the Pearson correlation coefficient indicates that the highest values among the inputs are 0.62 (MC3) and 0.6 (MC2), highlighting the low linearity and complexity of the datasets. A straightforward analysis indicates the need for a robust high non-linear feature identifier to eliminate redundant features.

3. Soft computing methods and data analysis techniques description

The current research will use feature selection techniques and machine learning methods to predict potato tuber yield in Atlantic Canada. Two different multi-objective optimisation techniques were used in feature selection. In Section 3, these techniques, methods, the SHAP explainer tool, and evaluation metrics will be introduced and discussed briefly.

3.1. Feature selection techniques

3.1.1. Boruta SHAP feature selection

The present study employed the Boruta algorithm to identify significant variables for predicting potato yield. The Boruta method is a feature selection approach that uses the random forest (RF) algorithm, which is known for its robustness (Kursa et al., 2010). There are five steps in the Boruta algorithm technique (Ba-Alawi et al., 2023; Gholami et al., 2021). Initially, all features in the dataset are duplicated. Second, to remove the correlations with the target factor, create shadow features by rearranging the values of the duplicated features. Third, Z-scores are calculated by running the RF algorithm over the expanded dataset. Fourth, the maximum Z-score (MZSA) among the shadow attributes is found, and the significance of the characteristic is contrasted with it. If the feature's relevance is less than MZSA, it is permanently eliminated from the dataset. If not, the characteristic is retained in the dataset. After the fifth phase of eliminating shadow characteristics, all preceding procedures are carried out again until all unnecessary variables are eliminated from the dataset.

3.1.2. Best subset regression

The Best-subset Regression (BSR) model, sometimes referred to as the "all possible models" technique and "all possible regression," is a significant strategy, particularly when selecting more prevalent variables (Jamei et al., 2021c; Singh et al., 2022). This method of model selection involves examining each conceivable combination of predictor variables. The optimal model is then selected using a predefined statistical criterion, such as Akaike's information criterion (AIC), adjusted coefficient of determination (Ad-R^2), mean square error (MSE), Mallows' C_p, or Amemiya's Prediction Criterion index (PC) (Jamei et al., 2021c). The explanations for the formulations of the criteria are as follows:

$$C_p = \frac{RSS_k}{MSE_J} + 2M - N, J > M \quad (1)$$

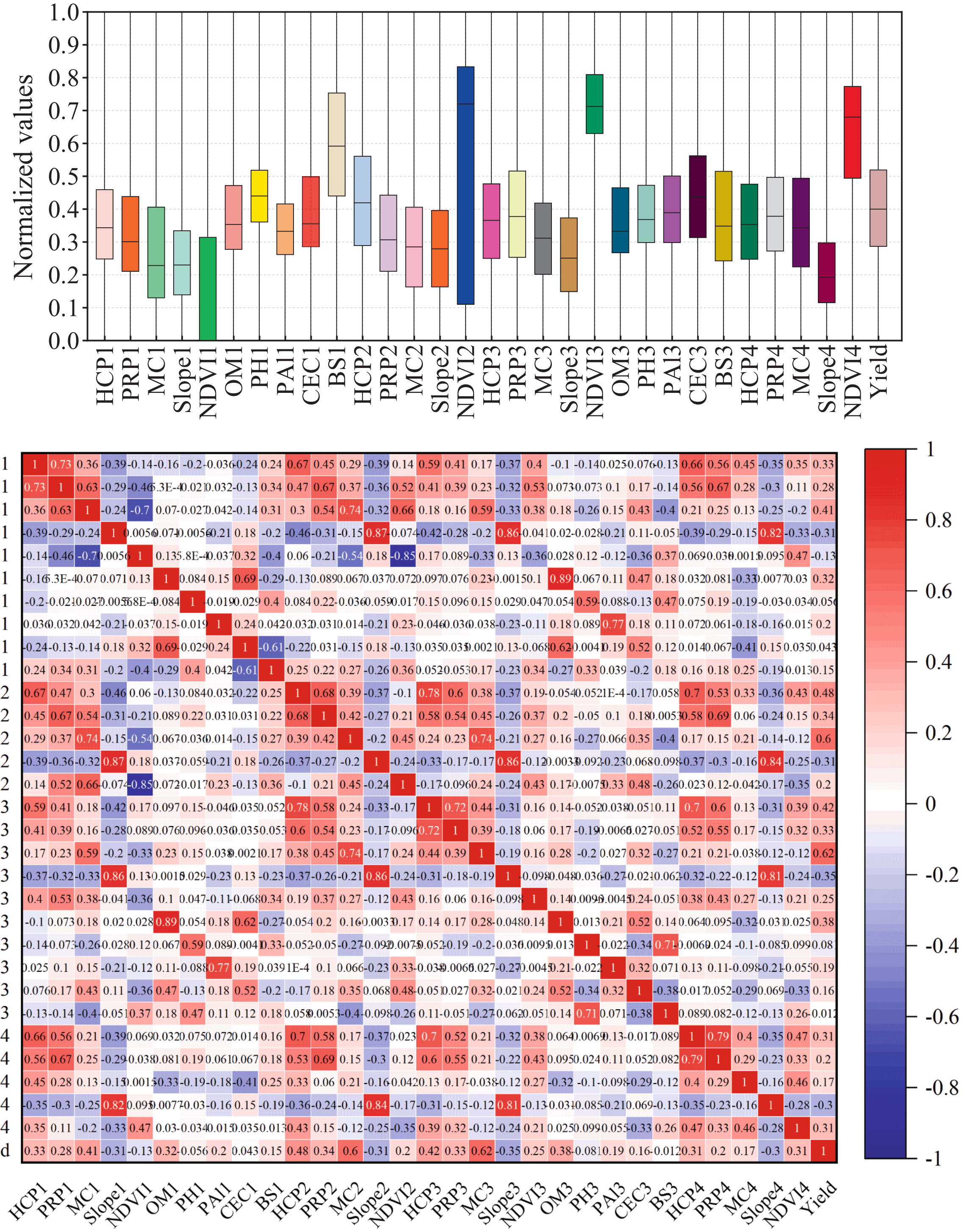


Fig. 2. Statistically exploring the trend of normalized data gathered from the fields to simulate the yield values (Above); Correlogram associate with all excisting features and yiled (bottom).

$$AIC = 2k + N\ln\left(\frac{1}{N}\sum_{i=1}^{N}\hat{e}_i^2\right) \tag{2}$$

$$PC = \frac{1}{(N-k)}\sum_{i=1}^{N}\hat{e}_i^2\left(1+\frac{k}{N}\right) \tag{3}$$

In the above equations, M represents the total number of variables, N signifies the total number of samples, MSE_J represents the mean square error, RSS_k denotes the residual sum of squares utilized in the regression models, and $\hat{e}$ represents the ith residual value. The minimum values of MSE, C_p, PC, and AIC are preferred (Wang et al., 2001).

3.2. *Multi-objective optimizing methods*

3.2.1. *Multi-Objective Optimization method on the basis of ratio analysis (MOORA)*

Multi-objective optimisation simultaneously optimises two or more

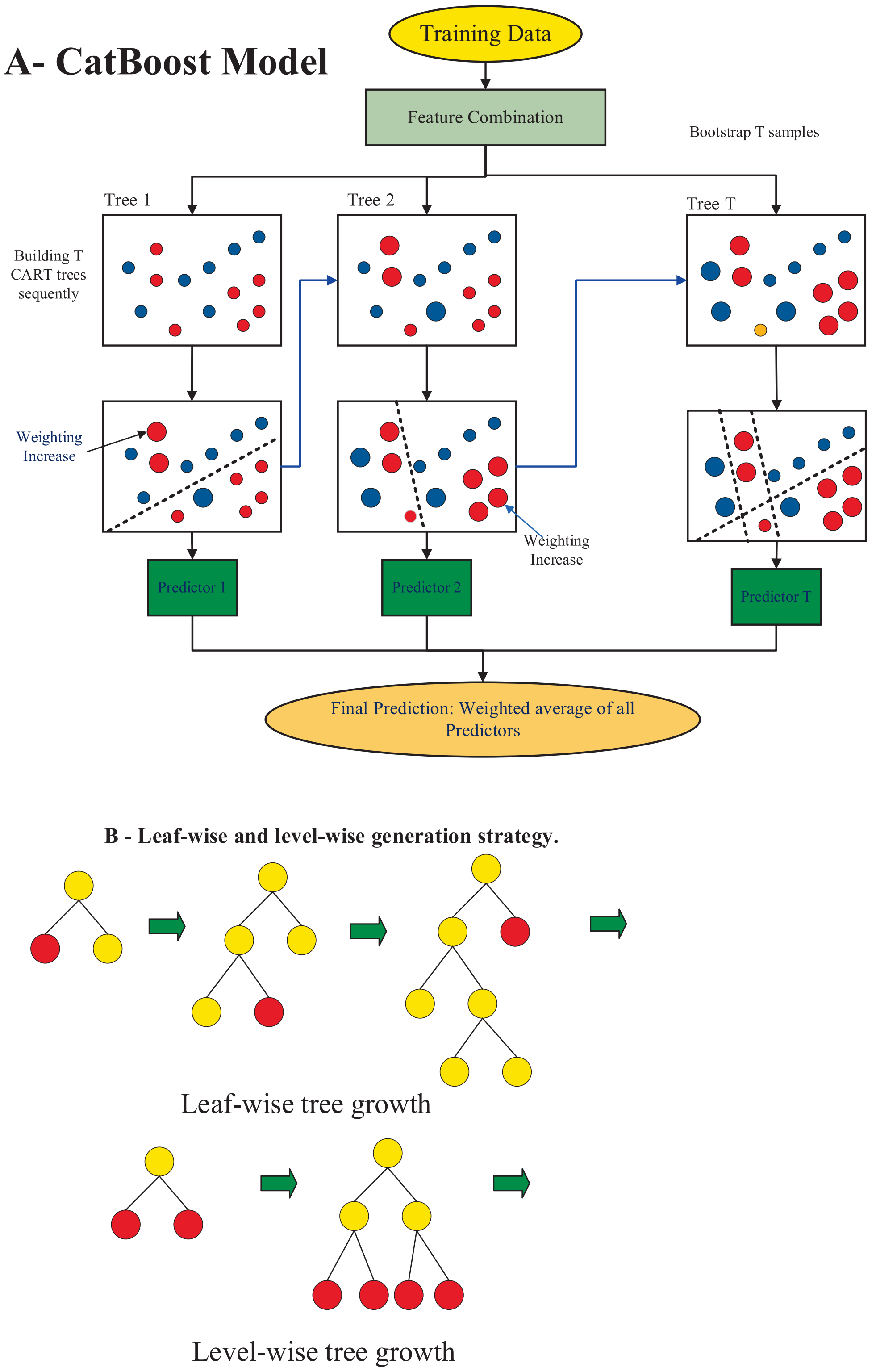


Fig. 3. A) CatBoost Structure – B) Leaf-wise and level-wise generation strategy.

competing criteria (objectives) while adhering to specific constraints. The multi-objective optimisation based on ratio analysis (MOORA) method ranks or selects one or more alternatives from a set of available options, taking into account both advantageous and disadvantageous objectives (criteria) (Brauers et al., 2010; Chakraborty, 2011). This approach initiates with a decision matrix that illustrates the performance of various alternatives for various criteria. Subsequently, the decision matrix is normalized to eliminate all dimensions and ensure that every element is comparable. The normalisation procedure is a ratio system where the performance of one alternative concerning a particular criterion is compared to a denominator that serves as a representative value for all other options. The following equation has been proposed for normalisation (Karande and Chakraborty, 2012):

$$x^*_{ij} = x_{ij} \Big/ \left[\sum_{i=1}^{m} x_{ij}\right] (j = 1, 2, \cdots, n) \tag{4}$$

where x_{ij} represents the i th alternative's performance measure on the j th criterion. The number of criteria is n, and the number of other options is m. The normalized performances for beneficial criteria are added to the MOORA method, while those for non-beneficial criteria are subtracted in the manner specified in the expression below:

$$y_i = \sum_{j=1}^{g} x_{ij}^{*} - \sum_{j=g+1}^{n} x_{ij}^{*} \tag{5}$$

where y_i is the assessment value of the ith alternative about all criteria, g is the number of criteria to be maximized, and $n-g$ is the number of criteria to be minimized. The option with the highest assessment value is best when arranged in descending order.

3.2.2. *WASPAS MCDM (weighted aggregated sum product assessment)*

The "WASPAS technique" is a novel utility theory-based model proposed by Zavadskas et al. (2012). This approach has been widely implemented for a variety of objectives. This method has the following steps (Debnath et al., 2023):

In the first stage, the alternative (A_i) and criteria (C_j) are chosen for evaluation. In the given set $i = 1, \cdots\cdots m$ and $j = 1, \cdots\cdots n$.

In the second step, one of the MCDM methodologies is utilized to compute the weights of the criteria. SWARA was employed to quantify the weights of the criteria in this investigation.

Using Eqs. (6) and (7), the decision matrix is normalized in Step 3. In order to maximize the benefit (beneficiary),

$$\overline{X}_{ij} = X_{ij}/\max X_{ij} \tag{6}$$

For minimum optimum value (non-beneficiary)

$$\overline{X}_{ij} = \min X_{ij}/X_{ij} \tag{7}$$

In the fourth stage, the "Weighted Sum Model" is implemented to calculate the initial total relative significance value $(\mathrm{Q}_{\mathrm{i}}^{(1)})$ using Eq. (8).

$$Q_i^{(1)} = \sum_{j=1}^{n} \overline{X}_{ij} W_j \tag{8}$$

Step 5: Using Equation (9), the "Weighted Product Model (WPM)" is implemented to calculate the second total relative significance value $(\mathrm{Q}_{\mathrm{i}}^{(2)})$.

$$Q_i^{(2)} = \prod_{j=1}^{n} (\overline{X}_{ij})^{W_j} \tag{9}$$

Eq. (10) is used in Step 6 to calculate the aggregate total relative significance value (Q_{i}), where λ denotes the coefficient value of Q_i:

$$Q_i = \lambda Q_i^{(1)} + (1-\lambda) Q_i^{(2)} \tag{10}$$

3.3. *Machine learning techniques*

3.3.1. *Categorical boosting (CatBoost)*

Prokhorenkova et al. (2018) have suggested the novel gradient boosting technique Categorical Boosting (CatBoost). The technique handles categorical features with minimal loss. Fig. 3-A shows a flowchart for the CatBoost model. In the flowchart, the initial N samples and M features are explicitly designated. Subsequently, a sequential construction of T regression trees (CART) are undertaken by integrating their respective characteristics. Ultimately, a prediction is determined through the computation of the weighted average aggregate of all predictors (Huang et al., 2019). Moreover, utilising the K-Fold cross-validation procedure during network training is a preventive measure against overfitting.

Consider a dataset of observations, denoted as $D = \{X_i, Y_i\} i = 1, \cdots, n$. If we have a permutation $\theta = (\sigma_1, \sigma_2, \cdots, \sigma_n)_n^T$, it can be modified using the method proposed by Prokhorenkova et al. (2018):

$$x_{\sigma_{p,k}} = \frac{\sum_{j=1}^{p-1} \left[x_{\sigma j,k} = x_{\sigma_{p,k}}\right] \times Y_{\sigma_j} + \beta \times P}{\sum_{j=1}^{p-1} \left[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}\right] + \beta} \tag{11}$$

In this context, the symbol β represents the weight assigned to the prior, whereas P denotes the prior value. Within the dataset, the prior refers to the mean value of the labels, to mitigate the presence of noise associated with categories that occur infrequently.

3.3.2. *Light gradient-boosting (LightGBM)*

LightGBM is an ML framework designed to implement gradient-boosting algorithms (Ke et al., 2017). This approach's use of a histogram technique and leaf-wise growth strategy results in reduced memory consumption and enhanced data separation (Fan et al., 2019). The LightGBM methodology uses a leaf-wise strategy to identify and split the leaf with the maximum scattering gain among all the current leaves, as depicted in Fig. 3-B. LightGBM employs a leaf-wise growth strategy combined with a maximum depth constraint to optimize computational efficiency and mitigate the risk of overfitting. The level-wise tree development technique involves the cultivation of trees in a hierarchical manner, with each level representing a distinct stage of growth. This strategy consists of dividing information by each node, focusing on the nodes closest to the root of the tree (Shakeel et al., 2023).

3.3.3. *Lasso regression (LASSO)*

Robert Tibshirani introduced the acronym LASSO (Shrinkage, 2016). The resilient method completes two fundamental objectives: feature selection and regularization. Particularly in models with high-dimensional predictors, including a penalty item in linear regression can significantly minimize the variance of the model by effectively reducing the size of the estimated coefficients (Zhang et al., 2021). The following describes the optimized goal function of LASSO Regression (LASSO-Reg):

$$\sum_{i=1}^{n} \left(y_i - \lambda_0 - \sum_{j=1}^{p} \lambda_j x_{ij} \right)^2 + \tau \sum_{j=1}^{p} |\lambda_j| \tag{12}$$

The symbol λ_0 represents the LASSO-Reg shift, whereas λ_j represents the x_{ij} coefficients. Within this context, the parameter is denoted as τ function, as a regulator.

3.3.4. *Elastic net (ENET)*

Elastic-net regression (ENET) originated as a reaction to criticisms leveled against the LASSO regression model, which was criticized for its potentially unstable variable selection that was overly dependent on the data. In fact, the objective is to minimize the subsequent loss function (Hastie et al., 2009).

$$L^{ENET}(\beta) = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \zeta \left(\frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right) \tag{13}$$

The mixing parameter between the Ridge ($\alpha = 0$) and LASSO ($\alpha = 1$) is denoted by its symbol α. Currently, two parameters, ζ and α, require tuning. When there are several correlated features, the ENET is a very beneficial regularised regression approach since it linearly integrates the penalties L1 and L2 of the LASSO and Ridge regression methods.

3.3.5. *Extra trees*

The ML technique, referred to as "Extremely randomised trees" or "extra trees" (Geurts et al., 2006), is a recently developed ensemble algorithm. The algorithm was created to continue the random forest algorithm to decrease the probability of overfitting a dataset (Geurts et al., 2006). The additional trees technique, similar to random forest, employs a random subset of features to train each base estimator (Singh et al., 2022). However, it sets itself apart by randomly choosing the most optimal feature and its corresponding value to divide each node (John et al., 2016).

3.4. SHAP explainer tool

The essential concept that forms the foundation of SHAP is the allocation of a numerical value to each characteristic inside a specific data point, which measures the extent to which that characteristic influenced the prediction made by the model. The scores provided are derived from applying Shapley values, a concept rooted in cooperative game theory. This theoretical framework aims to allocate the value generated by a collaborative endeavour to each person involved. The precise computation of the Shapley values for a given feature is as follows (Lundberg and Lee, 2017):

$$EM = \varphi_0 + \sum_{i=1}^{n} \varphi_i t_i \quad (14)$$

$$\varphi_i(ML, x) = \sum_{t \subseteq x} \frac{|t|!(n - |t| - 1)!}{n!}[ML(t) - ML(t\backslash i)] \quad (15)$$

t_i represents the simplification of the input variable numbers, where n represents the input variable numbers. The variable's contribution to the ML model is denoted by i, $\varphi_i \in \mathbb{R}$, and the differences notation for set operations is represented by \.

3.5. Evaluation indicators

Evaluation of the performance of ML techniques is essential in prediction tasks. In the present study, six different statistical metrics (Correlation Coefficient (R), Root Mean Square Error (RMSE), Uncertainty coefficient (U95%), Reliability coefficient, Mean Absolute Percentage Error (MAPE), and Kling–Gupta Efficiency (KGE) (Gupta et al., 2009)) were used to evaluate the accuracy of the techniques used. The following are formulas for the mentioned metrics (Jamei et al., 2022, 2020):

$$R = \sum_{i=1}^{N}(Yield_{o,i} - \frac{\overline{Yield_o})\ (Yield_{p,i} - \overline{Yield_p})}{\sqrt{\sum_{i=1}^{N}\left(Yield_{o,i} - \overline{Yield_o}\right)^2 \sum_{i=1}^{N}\left(Yield_{p,i} - \overline{Yield_p}\right)^2}} \quad (16)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Yield_{o,i} - Yield_{p,i}\right)^2} \quad (17)$$

$$U_{95\%} = 1.96\sqrt{SD_e^2 + RMSE^2} \quad (18)$$

$$Reliability = \frac{\sum_{i=1}^{N} K_i}{N} \times 100\% \quad (19)$$

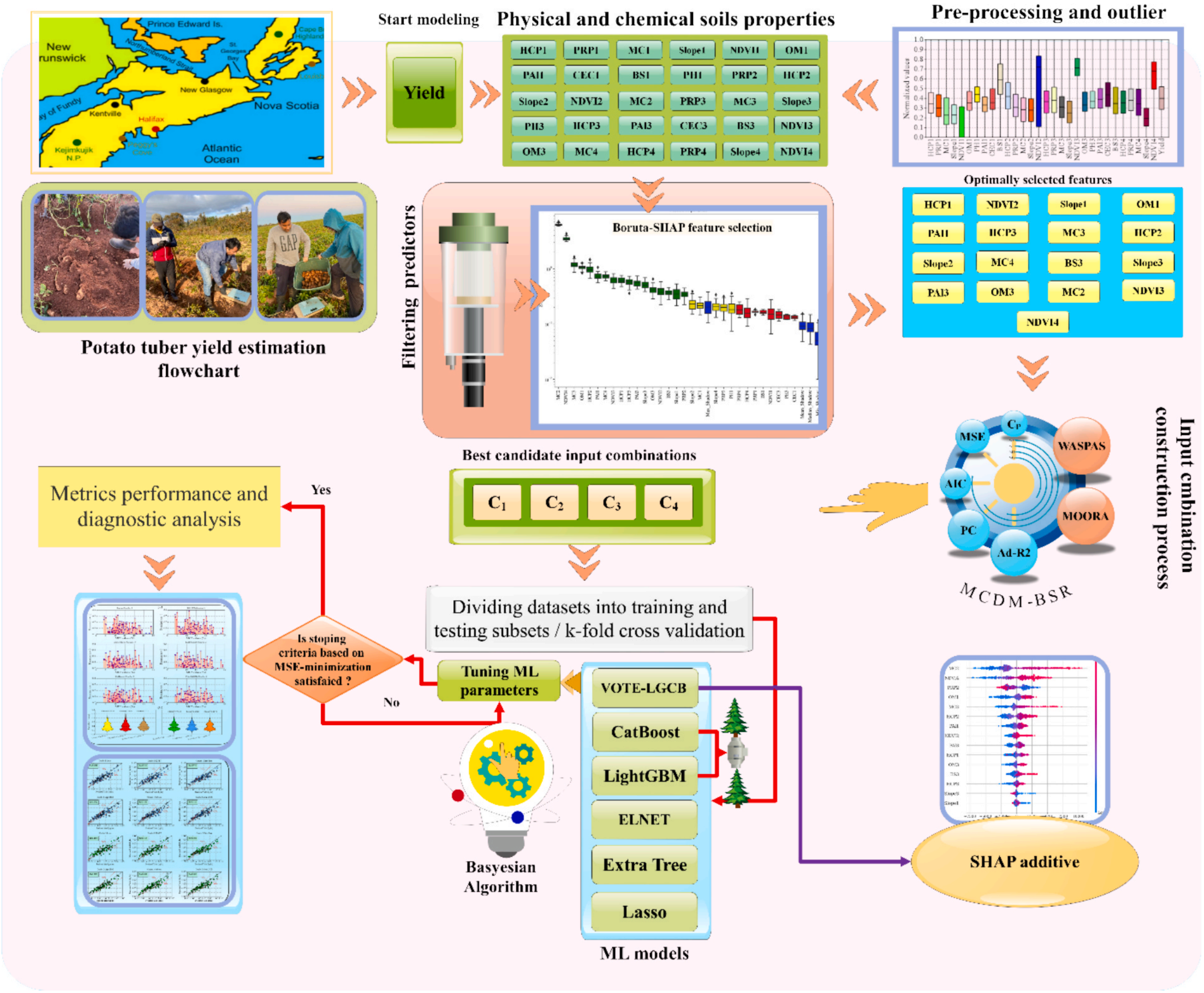


Fig. 4. Workflow of a multi-level preprocessing-based super ensemble ML framework to accurately estimate the crop potato in Atlantic Canada.

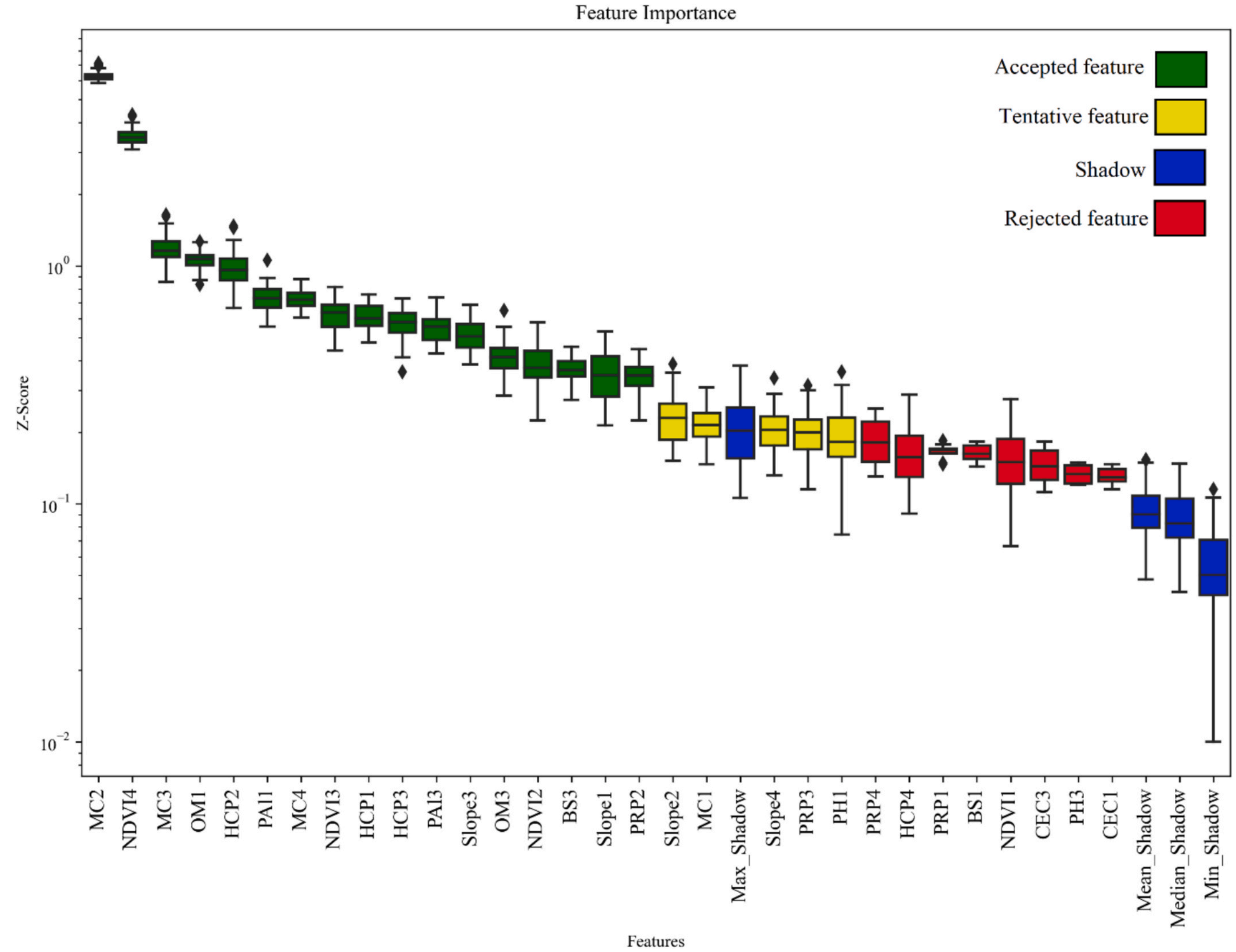


Fig. 5. Boruta-SHAP feature selection results based on Z-score values to recognize the most influential field features for yield potato monitoring in the Atlantic province of Canada.

$$K_i = \begin{cases} 1, if(RAE_i \leq \delta) \\ 0else \end{cases} \tag{20}$$

$$RAE_i = \frac{|Yield_{o,i} - Yield_{p,i}|}{Yield_{o,i}} \times 100\%, RAE \geq 0 \tag{21}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Yield_{o,i} - Yield_{p,i}}{Yield_{o,i}}\right| \times 100 \tag{22}$$

$$KGE = 1 - \sqrt{(R-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{23}$$

here $Yield_{o,i}$ is observed yield, $Yield_{p,i}$ is predicted yield, N is the total number of data, SD_e is the standard deviation of errors ($Yield_{o,i} - Yield_{p,i}$), β represents the ratio of the average predicted yield value to the average observed yield value, while α represents the relative standard deviation of the predicted and observed yield values. Correlation coefficient (R), measures the strength and direction of the linear relationship between observed and predicted values, with a range of -1 to 1. If R = 1, it has a perfect linear relationship and if R = 0 than there is no linear relationship; if the value is -1, it has a perfect negative linear relationship. High positive values (close to 1) indicate a strong positive correlation (Malik et al., 2022a). High negative values (close to -1) indicate a strong negative correlation. Values close to 0 suggest no correlation. Root Mean Square Error (RMSE) measures the average magnitude of the prediction error. It squares the errors before averaging, thus penalizing more significant errors than smaller ones. It has range from 0 to ∞. RMSE = 0 is a perfect fit, while higher RMSE values result in worse model performance. Uncertainty Coefficient (U95%) provides a confidence interval around the predictions, indicating the range within which the actual values are expected to lie with 95 % confidence (Jamei et al., 2021a). It ranges from 0 to ∞, while lower U95% values have less uncertainty and higher U95% values have more significant uncertainty. Reliability Coefficient typically measures the proportion of the total variation in observed values that is explained by the predicted values. It has range from 0 to 1. Coefficient = 1 has Perfect reliability, and coefficient = 0 has no reliability. Mean Absolute Percentage Error (MAPE), measures the accuracy of predictions as a percentage, showing the average absolute error as a percentage of the observed values. It has range from 0 to ∞. MAPE = 0 has perfect accuracy, and higher MAPE values result in worse model performance (Jamei et al., 2022). Kling–Gupta Efficiency (KGE) is a composite metric that combines correlation, variability, and bias components to provide a holistic measure of model performance. It has a range from $-\infty$ to 1. KGE = 1 perfectly matches observed and predicted values, while KGE = 0 model performance is as good as the mean of observed values. The KGE < 0 model performs worse than simply using the mean of the observed values. Understanding these indicators and their implications helps evaluate and compare predictive

Table 3
MCDM-based candidate input combination identifying using hybridized best subset regression with MOORA and WASPAS.

# Var	Variables	MSE	Ad-R^2	Cp	AIC	PC	WASPAS	MOORA
10	PAl1 / HCP2 / PRP2 / MC2 / NDVI2 / HCP3 / MC3 / OM3 / BS3 / NDVI4	2.878	0.697	25.577	598.673	0.309	5.61E-06	0
11	Slope1 / PAl1 / HCP2 / PRP2 / MC2 / NDVI2 / MC3 / Slope3 / OM3 / BS3 / NDVI4	2.847	0.700	20.469	593.565	0.306	4.41E-06	0.4794
12	HCP1 / Slope1 / PAl1 / HCP2 / PRP2 / MC2 / HCP3 / MC3 / Slope3 / OM3 / BS3 / NDVI4	2.823	0.702	16.924	589.957	0.304	3.59E-06	0.8120
13	HCP1 / Slope1 / PAl1 / HCP2 / PRP2 / MC2 / NDVI2 / HCP3 / MC3 / Slope3 / OM3 / BS3 / NDVI4	2.812	0.704	15.666	588.635	0.303	3.31E-06	0.9300
14	HCP1 / Slope1 / OM1 / PAl1 / HCP2 / PRP2 / MC2 / NDVI2 / HCP3 / MC3 / Slope3 / OM3 / BS3 / NDVI4	2.803	0.704	14.928	587.828	0.303	3.15E-06	0.9960
15	HCP1 / Slope1 / OM1 / PAl1 / HCP2 / PRP2 / MC2 / NDVI2 / HCP3 / MC3 / Slope3 / OM3 / PAl3 / BS3 / NDVI4	2.800	0.705	15.289	588.140	0.303	3.23E-06	0.9662
16	HCP1 / Slope1 / OM1 / PAl1 / HCP2 / PRP2 / MC2 / NDVI2 / HCP3 / MC3 / Slope3 / OM3 / PAl3 / BS3 / MC4 / NDVI4	2.799	0.705	16.222	589.039	0.304	3.44E-06	0.8787
17	HCP1 / Slope1 / OM1 / PAl1 / HCP2 / PRP2 / MC2 / NDVI2 / HCP3 / MC3 / Slope3 / NDVI3 / OM3 / PAl3 / BS3 / MC4 / NDVI4	2.803	0.704	18.000	590.810	0.305	3.84E-06	0.7121

models' performance effectively. Lower RMSE, MAPE, and U95% values generally indicate better model performance, while higher R, reliability coefficient, and KGE values indicate stronger and more reliable predictive capabilities (Jamei et al., 2021b; Malik et al., 2022b).

4. Model development and adjustment

This section describes the workflow for monitoring potato tuber yield in Canada's Atlantic provinces (*Prince Edward Island* and *New Brunswick*) using a high-dimensional feature MCDM-multi-preprocessing explainable VOTE-LGCB scheme. In this regard, 556 data points (241 in *New Brunswick* and 315 in PEI) including the 30 field features, namely HCP1, PRP1, MC1, Slope1, NDVI1, OM1, PH1, PAl1, CEC1, BS1, HCP2, PRP2, MC2, Slope2, NDVI2, HCP3, PRP3, MC3, Slope3, NDVI3, OM3, PH3, PAl3, CEC3, BS3, HCP4, PRP4, MC4, Slope4, NDVI4, were used to construct the predictive models. This is the first time such a number of features has been used to estimate the yield values of every agricultural product. Comparative ML approaches (VOTE-LGCB, CatBoost, LightGBM, Extra Tree, ELNET, and LASSO) have also been adopted. The main hybridized model is comprised of the Botura-

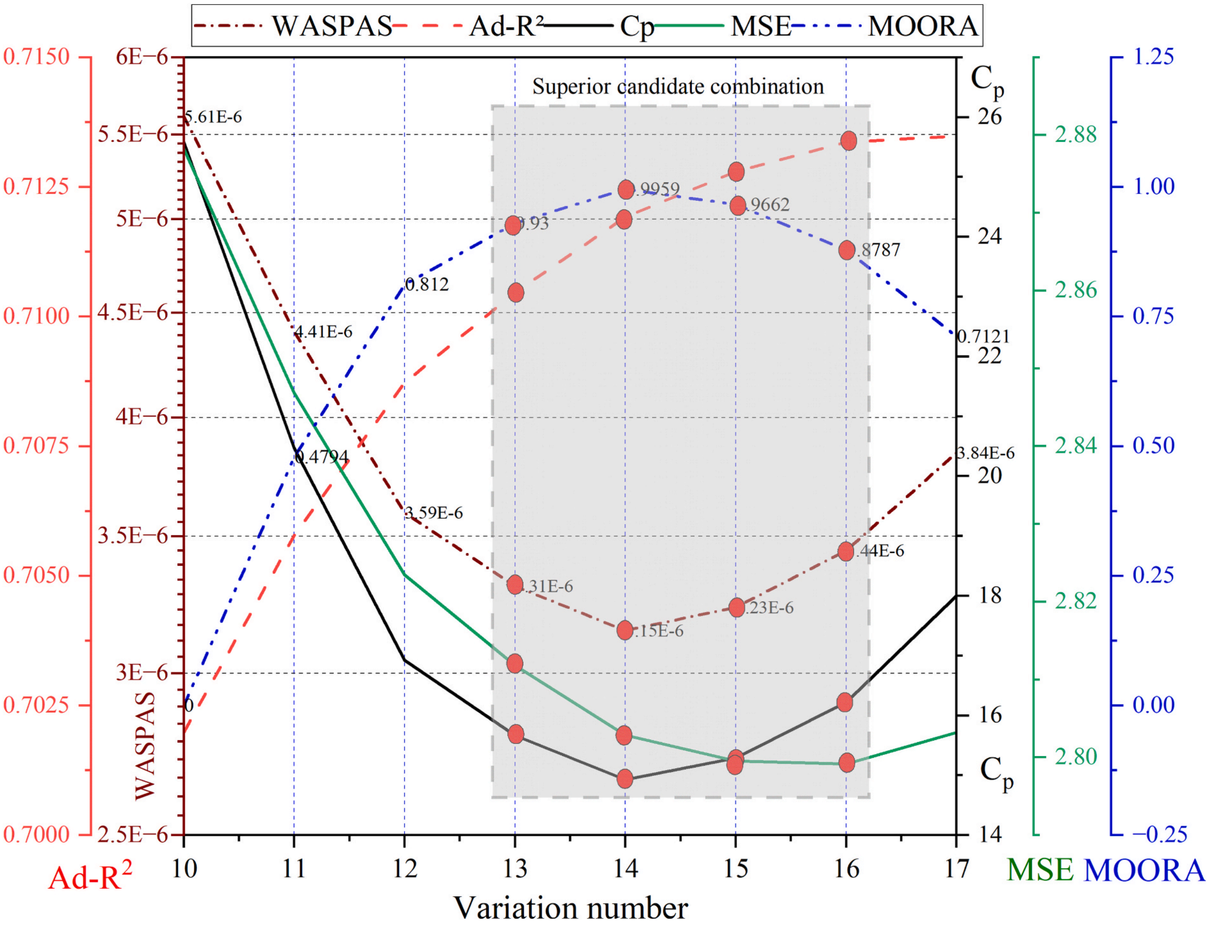


Fig. 6. MCDM-based BSR outcomes to discover the best input combination for feeding ML schemes aim to the prediction of yield values in Atlantic zones of Canada.

Table 4
Bayesian-based optimum hyperparameters of all the ML models constructing the yield predictive models.

Model	Hyperparameters			
Combos	C1	C2	C3	C4
LASSO	τ: 1.5	τ: 1.5	τ: 1.5	τ: 1.5
ELNET	ζ = 0.05, α: 0.8	ζ = 0.1, α: 0.8	ζ = 0.05, α: 0.8	ζ = 0.05, α: 0.8
Extra tree	max_depth:10, min_samples_split: 2, N_estimators: 200; criterion=SE	max_depth:10, min_samples_split: 2, N_estimators: 300; criterion=SE	max_depth:10, min_samples_split: 2, N_estimators: 200; criterion=SE	max_depth:10, min_samples_split: 2, N_estimators: 100; criterion=SE
LightGBM	learning_rate: 0.2, max_depth: 8, N_estimators: 100	learning_rate: 0.1, max_depth: 6, N_estimators: 200	learning_rate: 0.1, max_depth: 8, N_estimators: 100	learning_rate: 0.1, max_depth: 8, N_estimators: 300
CatBoost	learning_rate: 0.1, max_depth: 8, N_estimators: 100	learning_rate: 0.05, max_depth: 5, N_estimators: 200	learning_rate: 0.1, max_depth: 8, N_estimators: 100	learning_rate: 0.15, max_depth: 6, N_estimators: 100
Vote-LGCB	Meta learners: LGBMRegressor and CatBoostRegressor; verbose = 0			

*SE=Square Error.

SHAP feature selection integrated with two MCDM schemes (WASPAS and MOORA) and the best subset regression (BSR), as a multi-level preprocessing scheme, and a voting-based super ensemble model, VOTE-LGCB. In pursuit of this objective, all the ML models (approaches: CatBoost, LightGBM, Extra Tree, ELNET, and LASSO) are structured on the open-source Scikit-learn performing on the Python platform. Besides, Boruta-SHAP and SHAP explainers were executed based on the Boruta and SHAP open-source libraries, whereas WASPAS and MOORA were performed with NumPy and Pandas. Notably, every calculation is performed on a personal laptop equipped with an 8.0 GB RAM

Table 5
Goodness-of-fit indices are used to assess the performance and robustness of the provided hybrid expert systems and monitor the yield values in the Atlantic province of Canada.

Model	Combo	Phase	R	RMSE	MAPE	KGE	U95%	Reliability
LASSO	Combo 1	TRN	0.8346	5772.4868	10.7191	0.7655	16009.5442	87.3874
		TST	0.8669	5576.5244	10.0283	0.7821	15470.7317	91.0714
	Combo 2	TRN	0.8356	5757.5656	10.6341	0.7668	15968.1614	87.8378
		TST	0.8675	5562.6600	10.0273	0.7811	15440.4160	91.0714
	Combo 3	TRN	0.8361	5748.6457	10.6134	0.7676	15943.4228	87.3874
		TST	0.8677	5555.5712	9.9449	0.7830	15421.2866	91.0714
	Combo 4	TRN	0.8364	5743.4654	10.5878	0.7680	15929.0555	87.8378
		TST	0.8680	5546.7496	9.9759	0.7840	15399.5834	91.0714
ELNET	Combo 1	TRN	0.8141	6087.6948	11.3206	0.7266	16883.7491	86.9369
		TST	0.8383	6093.2330	10.9620	0.7585	16892.4917	84.8214
	Combo 2	TRN	0.8102	6144.5366	11.3125	0.7238	17041.3955	86.0360
		TST	0.8319	6189.4433	11.1125	0.7540	17172.8473	85.7143
	Combo 3	TRN	0.8160	6059.6421	11.1743	0.7300	16805.9472	86.9369
		TST	0.8380	6087.0044	10.8528	0.7595	16889.8307	85.7143
	Combo 4	TRN	0.8215	6028.7504	11.0176	0.7428	16719.3969	86.9369
		TST	0.8406	6050.9010	10.6221	0.7604	16779.2983	85.7143
Extra tree	Combo 1	TRN	0.9939	1313.2269	2.3625	0.9348	3642.1330	98.8229
		TST	0.8778	5672.3515	10.6139	0.6907	15690.2353	90.1786
	Combo 2	TRN	0.5846	8823.3574	15.0968	0.5430	24470.8970	77.0270
		TST	0.8806	5611.5876	10.5752	0.6917	15544.2459	91.0714
	Combo 3	TRN	0.9941	1286.5623	2.3145	0.9364	3568.1807	99.7748
		TST	0.8910	5402.3815	10.0673	0.7060	14961.1986	91.0714
	Combo 4	TRN	0.9946	1234.8211	2.2298	0.9387	3424.6804	99.7748
		TST	0.8889	5434.6035	10.1449	0.7038	15061.8066	91.0714
LightGBM	Combo 1	TRN	0.9868	1764.8306	3.1585	0.9400	4894.6206	99.5495
		TST	0.8732	5479.2504	9.7550	0.7672	15215.7693	91.9643
	Combo 2	TRN	0.9862	1801.2839	3.2870	0.9380	4995.7211	99.7748
		TST	0.8860	5279.4948	9.5391	0.7586	14659.3198	91.0714
	Combo 3	TRN	0.9876	1707.3528	3.1424	0.9422	4735.2104	99.5495
		TST	0.8756	5420.9857	9.8876	0.7751	15057.0567	89.2857
	Combo 4	TRN	0.9893	1586.8958	2.8623	0.9473	4401.1325	98.5600
		TST	0.8830	5300.7073	9.6300	0.7700	14723.0394	90.1786
CatBoost	Combo 1	TRN	0.9931	1295.1831	2.4384	0.9532	3592.0889	98.7556
		TST	0.8860	5393.2011	9.4958	0.7273	14946.0631	91.0714
	Combo 2	TRN	0.9822	2017.2630	3.6965	0.9367	5594.6854	99.7748
		TST	0.8806	5366.5627	9.5436	0.7645	14888.0089	91.0714
	Combo 3	TRN	0.9937	1239.8431	2.3369	0.9560	3438.6056	96.1250
		TST	0.8875	5415.2456	9.6639	0.7137	15013.0142	91.9643
	Combo 4	TRN	0.9937	1241.4561	2.3469	0.9551	3443.0799	97.2501
		TST	0.8864	5453.4631	9.6251	0.7135	15089.3683	88.3929
Vote-LGCB	Combo 1	TRN	0.9851	1885.1531	3.5105	0.9314	5228.3259	99.7748
		TST	0.8810	5357.9742	9.4376	0.7634	14870.5359	89.2857
	Combo 2	TRN	0.9857	1844.4627	3.4433	0.9326	5115.4709	99.7748
		TST	0.8855	5235.6550	9.4110	0.7610	14527.8179	95.5357
	Combo 3	TRN	0.9860	1824.9885	3.4402	0.9342	5061.4640	99.7748
		TST	0.8958	5088.5087	8.9913	0.7652	14127.6637	93.7500
	Combo 4	TRN	0.9850	1896.6947	3.5036	0.9289	5260.3356	99.7748
		TST	0.8934	5149.6633	9.4761	0.7569	14305.8586	91.9643

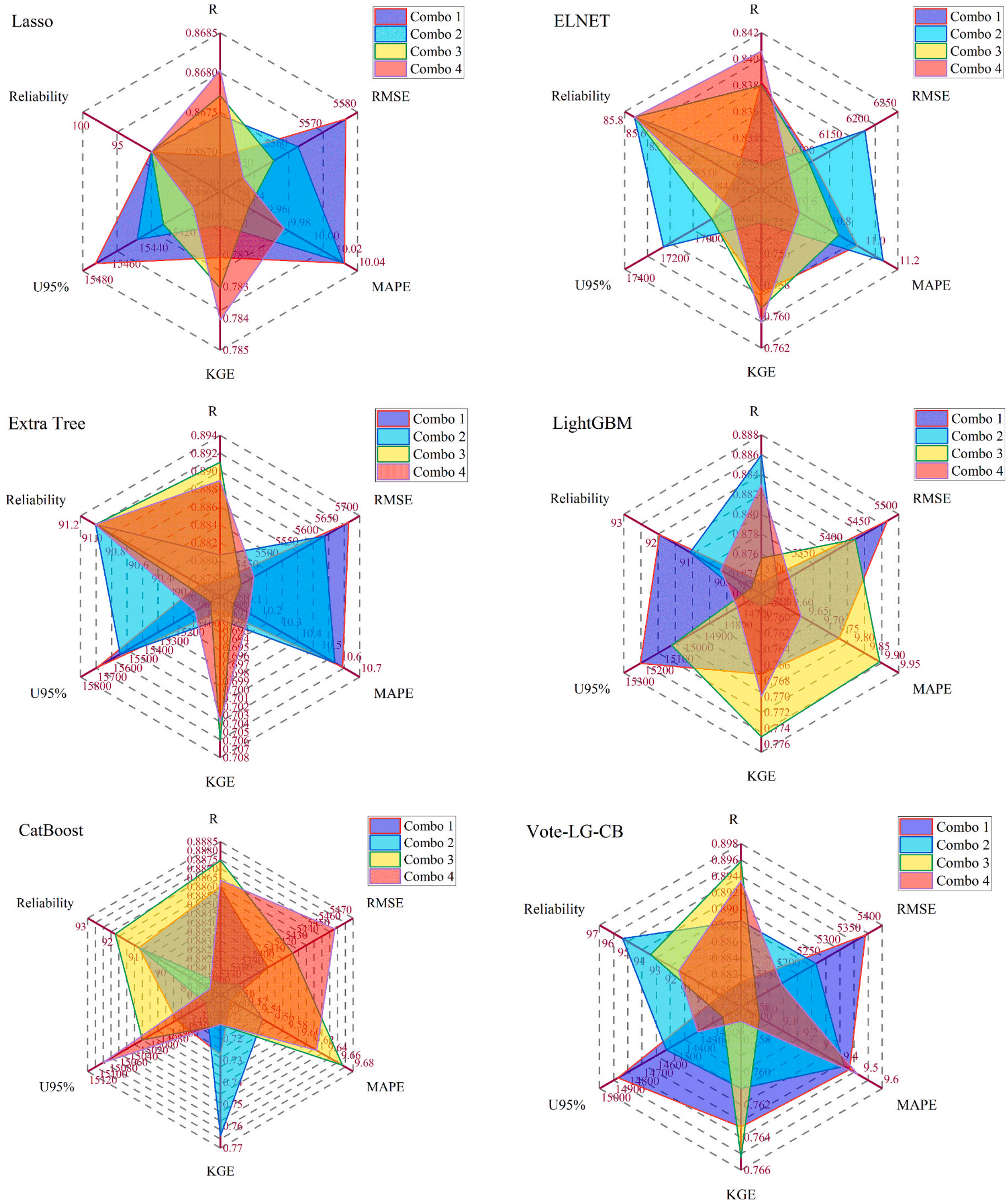


Fig. 7. Spider plot measuring whole metric performance (i.e., R, RMSE, MAPE, KGE, $U_{95\%}$, and Reliability) in testing phase simulation of the yield value using super ensemble and five comparative ML models.

configuration and an Intel (R) Core i7 processor operating at 3.0–3.20 GHz. Fig. 4 depicts each phase of modelling potato tuber yield value with a unique high-dimensional filtering super ensemble model. Given the lack of predictive power exhibited by the accumulated agricultural datasets concerning yield, it becomes necessary to identify the most influential predictors by implementing a robust methodology that accounts for the nonlinear correlation between predictors and targets. The Boruta-SHAP FS approach was employed to achieve this objective, which uses an important factor derived using a Z-score value to identify the most significant predictors among numerous input features for yield estimation. To this end, it is critical to use an appropriate benchmark criterion, the so-called Max-Shadow, to remove the redundant predictors in order to gain the most efficient performance. Boxplots in Fig. 5 illustrate the outcomes of the Boruta-SHAP method, indicating the significant input predictors based on Z-score values to construct the predictive model of potato tuber yield. As shown in Fig. 5, green boxes reveal the 17 accepted features, including HCP1, Slope1, OM1, PAl1, HCP2, PRP2, MC2, NDVI2, HCP, MC3, Slope3, NDVI3, OM3, PAl3, BS3, MC4, and NDVI4. In contrast, features corresponding red boxes show the rejected ones based on the Max-Shadow benchmark values. In the next pre-processing stage, the optimal candidate input combinations with 10–17 components were computed using the BSR scheme. According to the literature, the lowest values of MSE, Cp, AIC, and PC and the highest value of Ad-R^2 are vital in indicating the best possible combinations.

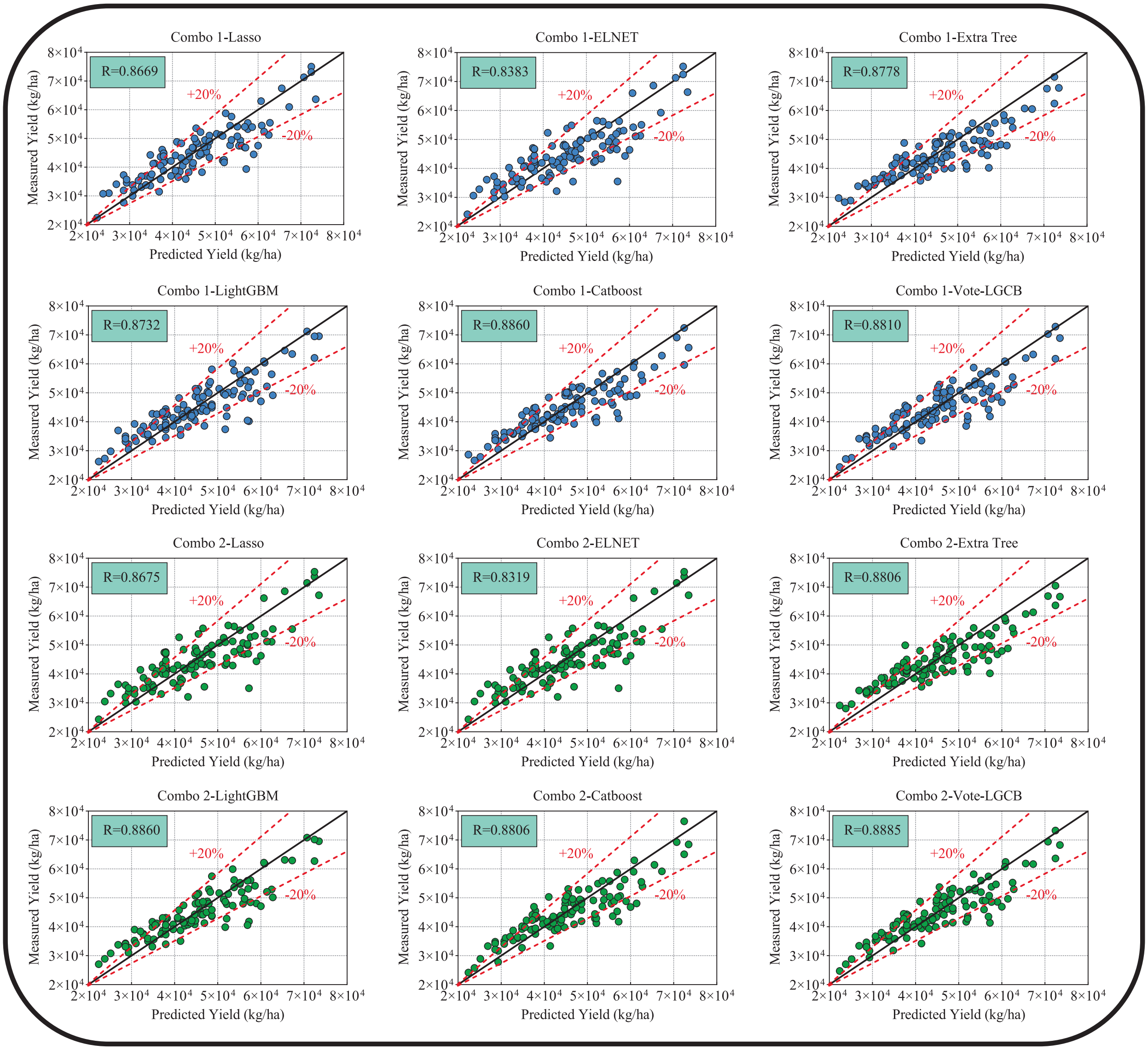


Fig. 8. Scatter plots were associated with simulated and measured yield values using the super ensemble and five comparative ML models, using four input combinations selected in the testing phase.

However, no specific law determines which ones are more decisive. Aiming to address this drawback, two MCDM schemes, namely WASPAS and MOORA, were coupled with the BSR technique to specify the optimal combinations accurately. Table 3 reports all the outcomes of the second pre-processing stage based on the abovementioned criteria. Basically, the minimal values of WASPAS and the maximal values of MOORA reveal the superior candidate input combinations. In order to do this, the four best input combinations with 13–16 features were found using WASPAS and MOORA values and then evaluated using ML models. To make the ML modelling evaluation easier, the following combinations are used: Combo 1 (WASPAS = 3.31E-06 and MOORA = 0.93), Combo 2 (WASPAS = 3.15E-06 and MOORA = 0.9960), Combo 3 (WASPAS = 3.23E-06 and MOORA = 0.9962), and Combo 4 (WASPAS = 3.44E-06 and MOORA = 0.8787). To better understand how the MCDM-based BSR scheme worked, the Ad-R^2, MSE, Cp, WASPAS, and MOORA values were shown on a multi-scale plot (Fig. 6) for all the possible input combinations.

The 556 data points were randomly partitioned to create an 80/20 training and testing data split. The training subset was subjected to k-fold cross-validation (with five folds) in this investigation to prevent overfitting and ensure that each dataset had an equal opportunity throughout both the training and testing stages. Also, Before running the ML techniques, the training and testing datasets were normalized to a range of [0, 1] to improve convergence and maintain model stability.

The hyperparameters associated with a soft computing framework are critical in its preparation since they directly impact model precision (Jamei et al., 2022). In the current research, the LightGBM and CatBoost models consider the main *meta*-learner, which contains three key hyperparameters, namely learning_rate, max_depth, and N_estimators.

The Bayesian optimisation approach adjusted the Extra tree, LightGBM, and CatBoost hyperparameters. The setting of the VOTE-LGCB super ensemble model is directly dependent on two *meta*-learner functions (LGBMRegressor and CatBoostRegressor) in the Scikit-learn library. Consequently, it has no specific hyperparameter. The optimal values and ranges of the hyperparameters for all the proposed ML (LASSO, ELNET, Extra tree, LightGBM, CatBoost, and VOTE-LGCB) systems are shown in Table 4.

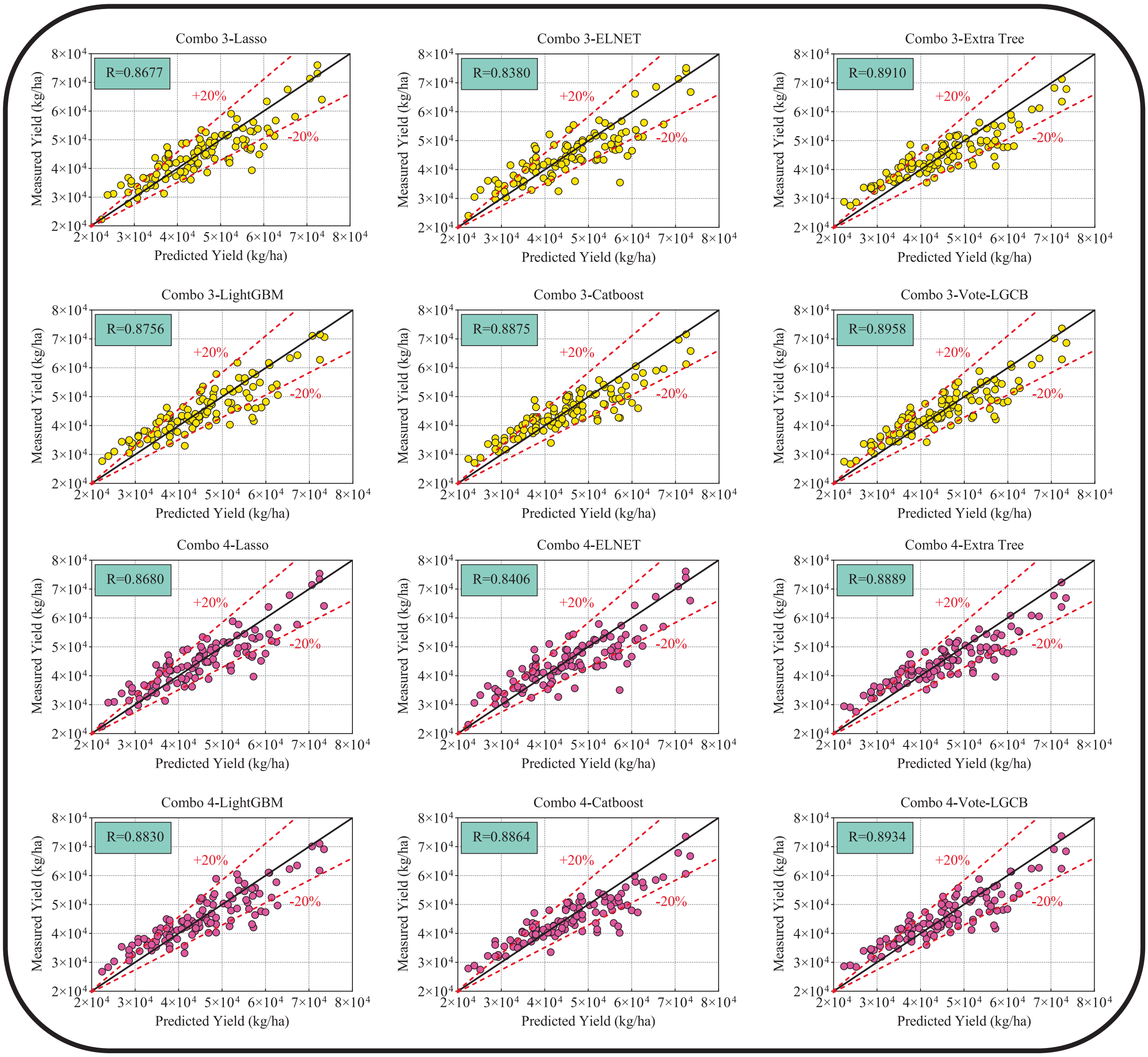


Fig. 8. (*continued*).

5. Application results and computational assessment

The prediction results generated by the LASSO, ELNET, Extra tree, LightGBM, CatBoost, and Vote-LGCB models are discussed in detail in four scenarios of input combinations Combo 1, Combo 2, Combo 3, Combo 4. Table 5 shows values for the statistical metrics R, RMSE, MAPE, KGE, U95%, and Reliability in training (TRN) and testing (TST) phases to monitor potato tuber yield in Atlantic Canada Province, Edward Island. The proposed models have also been examined with the help of distinct diagnostic plots to assess their suitability and applicability in predicting potato tuber yield.

Table 5 illustrates the performance of the LASSO, ELNET, Extra tree, LightGBM, CatBoost, and Vote-LGCB models using the input combinations Combo 1, Combo 2, Combo 3, and Combo 4. By analyzing, the LASSO model shows slightly better precision in terms of (R = 0.8364, RMSE = 5743.4654, MAPE = 10.5878, KGE = 0.7680, U95% = 15929.0555, Reliability = 87.8378)-TRN and (R = 0.8680, RMSE = 5546.7496, MAPE = 9.9759, KGE = 0.7840, U95% = 15399.5834, Reliability = 91.0714)-TST phases by incorporating the Combo 4 input subset followed by Combo 3, Combo 2, and Combo 1 to predict potato tuber yield. Similarly, the ELNET model produces better prediction accuracy for Combo 4 than Combo 1, Combo 2, and Combo 3 to predict potato tuber yield.

The Extra tree, CatBoost, and Vote-LGCB models appeared more precise with Combo 3-based statistical metrics in training and testing phases against Combo 1, Combo 2, and Combo 4 sets of input combinations to predict potato tuber yield. Meanwhile, the LightGBM acquired better accuracy in terms of Combo 1 concerning other input combinations. Overall, the Vote-LGCB model shows highest prediction accuracy [R = 0.9860, RMSE = 1824.9885, MAPE = 3.4402, KGE = 0.9342, U95% = 5061.4640, Reliability = 99.7748]-TRN and (R = 0.8958, RMSE = 5088.5087, MAPE = 8.9913, KGE = 0.7652, U95% = 14127.6637, Reliability = 93.7500]-TST as compared to LASSO, ELNET, Extra tree, LightGBM, CatBoost models to predict potato tuber yield. Moreover, Combo 3 appeared to be a slightly better input combination to predict potato tuber yield than Combo 1, Combo 2, and Combo 4. Therefore, the Vote-LGCB model appeared to be the most accurate model based on Table 5 compared to other models.

Fig. 7 exhibits the spider plots of LASSO, ELNET, Extra tree, LightGBM, CatBoost, and Vote-LGCB models in terms of R, RMSE, MAPE, KGE, U95%, and Reliability metrics to predict potato tuber yield for all four input combinations Combo1(purple), Combo 2 (cyan), Combo 3 (yellow) and Combo 4 (pink). It is quickly realized that the Vote-LGCB model reports better accuracy in these spider plots for all 4 input combinations but outperformed input combination C3. The comparing models (i.e., LASSO, ELNET, Extra tree, LightGBM, CatBoost)

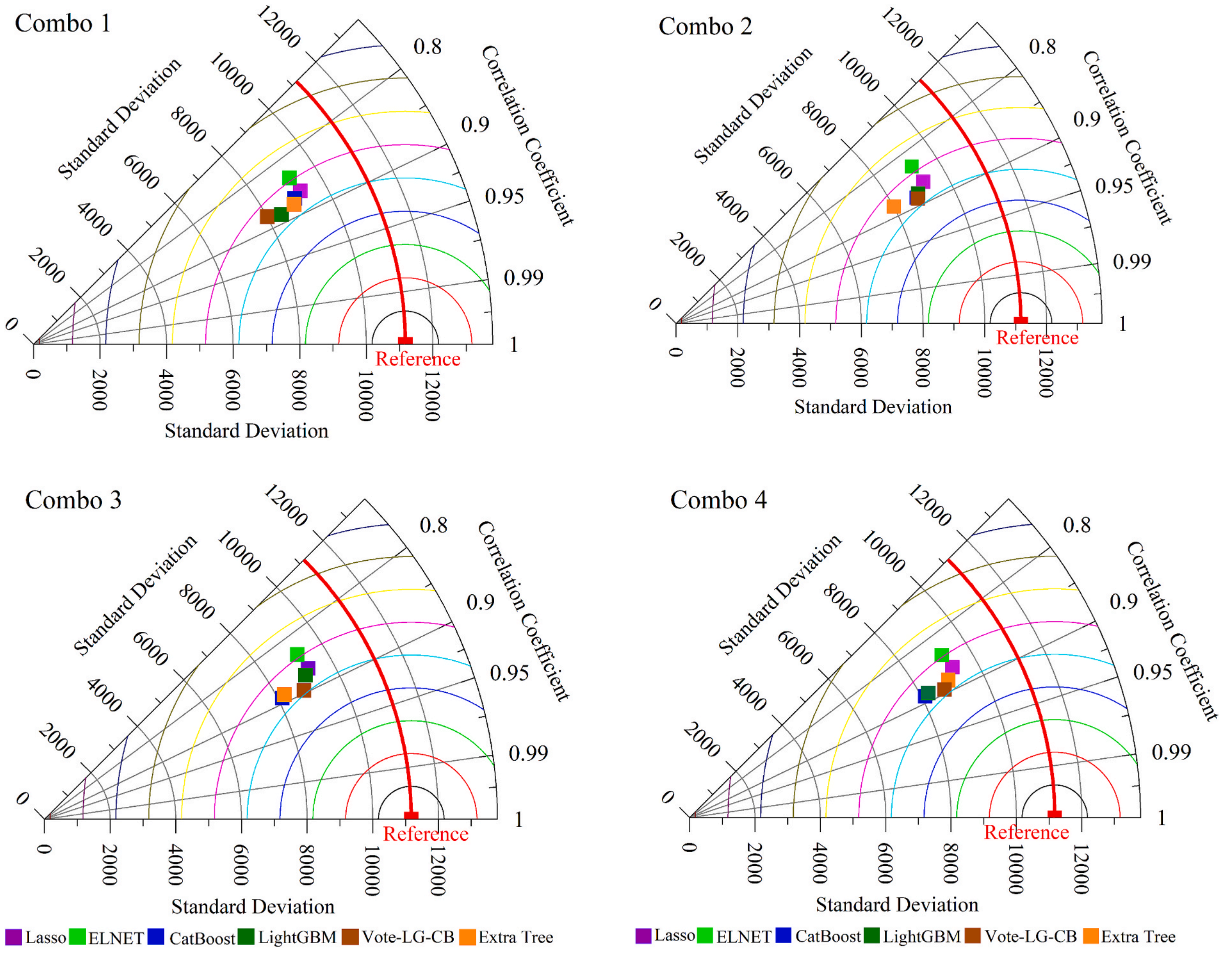


Fig. 9. Taylor diagram of simulated yield values using the super ensemble and five comparative ML models through selected four input combinations (Combo 1, Combo 2, Combo 3, Combo 4) against the reference (i.e., actual yield) point related to the measured yield values.

are reasonably good in terms of spider plots in all 4 input combination scenarios to predict potato tuber yield. However, the Vote-LGCB model surpasses all the models with Combo 3 in predicting potato tuber yield by attaining precise assessment metric scores.

The scatter plots in Fig. 8 inspect the efficiency of the LASSO, ELNET, Extra tree, LightGBM, CatBoost, and Vote-LGCB models between the predicted and measured potato tuber yields based on Combo 1, Combo 2, Combo 3, and Combo 4. Scatterplots further elaborate the models' prediction competence by including the R metric and the 20 % upper and lower bounds. The Vote-LGCB model with Combo 3 accomplished the highest precision with better prediction capacity by obtaining R = 0.8958, followed by Vote-LGCB with Combo 4 (R = 0.8934) and Extra Tree with Combo 3 (R = 0.8910) to predict potato tuber yield as compared to other models. Thus, Fig. 8 authenticates that the Vote-LGCB model with Combo 3 is a reasonably good potato tuber yield prediction model.

The Taylor diagrams in Fig. 9 discussed the LASSO, ELNET, Extra tree, LightGBM, CatBoost, and Vote-LGCB model's performance more tangibly and concretely between the referenced and predicted potato tuber yields in Combo 1, Combo 2, Combo 3, and Combo 4 scenarios. Taylor diagrams are branded a comprehensive valuation to inspect the models' comparability based on standard deviation and correlation coefficient.

For Combo 2, Combo 3, and Combo 4, the clearly Vote-LGCB model

Table 6
Finalization of six predictive models through possible input combinations based on two MCDM schemes (i.e., WASPAS and MOORA) for concentrating on superior performances of models.

Model	Combo	WASPAS	MOORA	Sup-Combo	Model	Combo	WASPAS	MOORA	Sup-Combo
LASSO	Combo 1	0.00025	0.6828	3	LightGBM	Combo 1	0.00026	0.4800	2
	Combo 2	0.00024	0.7256			Combo 2	0.00023	0.4357	
	Combo 3	0.00024	0.2834			Combo 3	0.00025	0.6075	
	Combo 4	0.00024	0.3951			Combo 4	0.00024	0.4594	
ELNET	Combo 1	0.00024	0.6110	4	CatBoost	Combo 1	0.00024	0.5611	2
	Combo 2	0.00025	0.6719			Combo 2	0.00025	0.1952	
	Combo 3	0.00024	0.4557			Combo 3	0.00025	0.6352	
	Combo 4	0.00024	0.3040			Combo 4	0.00024	0.8427	
Extra Tree	Combo 1	0.00026	0.6188	3	Vote-LG-CB	Combo 1	0.00025	0.6546	3
	Combo 2	0.00025	0.6194			Combo 2	0.00026	0.4320	
	Combo 3	0.00023	0.3812			Combo 3	0.00023	0.3788	
	Combo 4	0.00023	0.3853			Combo 4	0.00024	0.5944	

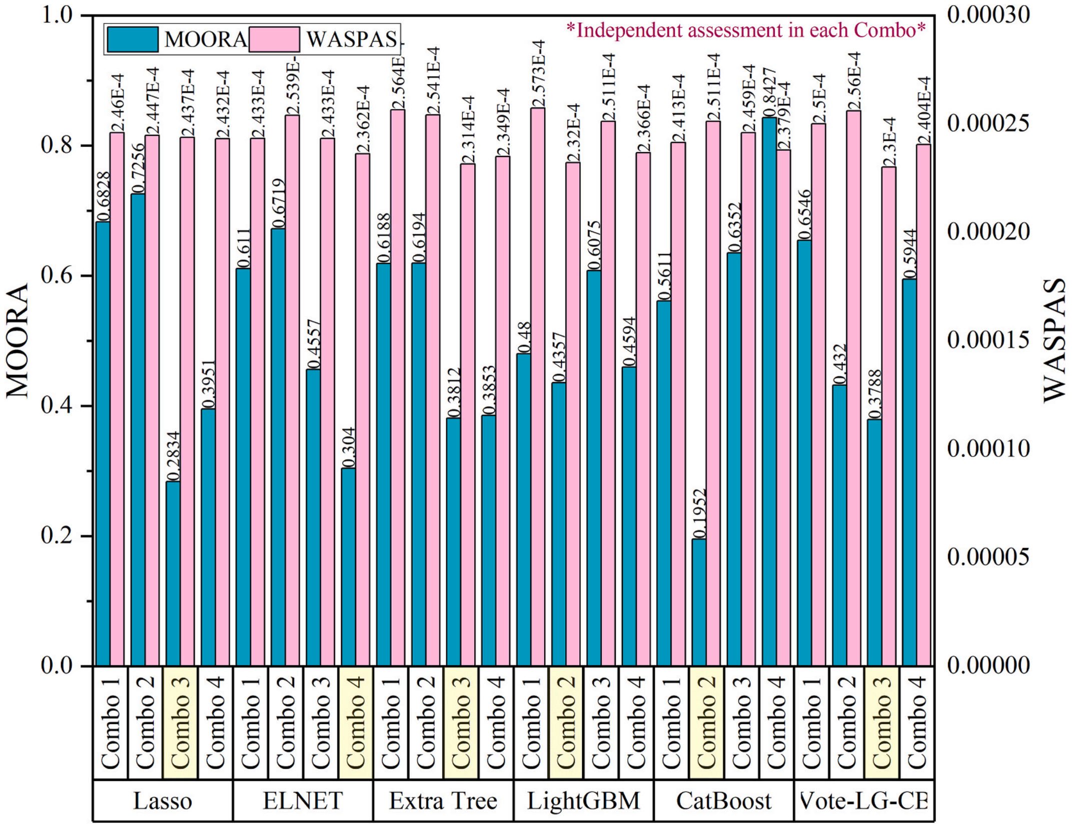


Fig. 10. Grouped Columns – Indexed data plot to represent the MOORA and WASPAS values associated with all the models through four input combinations to ascertain the best combination in each model in the testing phase.

positioned itself slightly closer to the reference potato tuber yield with a correlation coefficient between 0.90 and 0.95, except for Combo 1. The comparison of models LASSO, ELNET, Extra Tree, LightGBM, and CatBoost shows acceptable accuracy. However, it could not surpass the Vote-LGCB model. Therefore, Fig. 9 further confirmed the appropriateness of the Vote-LGCB model in monitoring potato tuber yield using Combo 1, Combo 2, Combo 3, and Combo 4.

6. Further analysis and interpretably assessment of outcomes

Table 6 identifies the most accurate predictive model using the input combination Combo 1, Combo 2, Combo 3, and Combo 3 based on the two MCDM schemes (i.e., WASPAS and MOORA). Observing Table 6, Combo 3, based on WASPAS and MOORA schemes, appeared to be the most optimum input combination for the LASSO model to predict potato tuber yield. Likewise, the WASPAS and MOORA schemes depict that Extra tree performs best on Combo 3. Similarly, the Vote-LGCB model reports better accuracy in combination with Combo 3 using the WASPAS and MOORA schemes. The WASPAS and MOORA schemes confirmed that Combo 2 is better for LightGBM and CatBoost. In contrast, Combo 4 is the optimum set of input for the ELNET model to predict potato tuber yield accurately. However, the WASPAS and MOORA schemes support and validate the highest performance of the Vote-LGCB model over other counterpart models by achieving the lowest scores based on Combo 3 (Table 6).

The bar graphs in Fig. 10 represent the WASPAS and MOORA scores attained by the LASSO, ELNET, Extra tree, LightGBM, CatBoost, and Vote-LGCB models in all four combinations (i.e., Combo 1, Combo 2, Combo 3, and Combo 4) to predict potato tuber yield. The Vote-LGCB model obtained the smaller values of WASPAS and MOORA scores to predict potato tuber yield using Combo 1, Combo 2, Combo 3, and especially Combo 4, where the magnitudes of these MCA schemes are much less than others. The LASSO, ELNET, Extra tree, LightGBM, and CatBoost models displayed relatively lower accuracy based on WASPAS and MOORA scores using all four input combinations to predict potato tuber yields. Thus, Fig. 10 established that overall, the Vote-LGCB model displays higher accuracy using WASPAS and MOORA scores.

Fig. 11 offers a more in-depth analysis using the normalized error distribution and MC2 values (%) in terms of stem plots for the most significant input feature. Additionally, the violin plot distribution of the normalized error was also inserted to assess the robustness of the LASSO, ELNET, Extra tree, LightGBM, CatBoost, and Vote-LGCB models in their corresponding optimal input combinations (i.e., Combo 2, Combo 3, and Combo 4). The Vote-LGCB model with Combo 4 exhibited lower normalized error distribution and MC2 values (%) as compared to LASSO (with Combo 3), ELNET (with Combo 3), Extra tree (with Combo 3), and LightGBM (with Combo 2), CatBoost (with Combo 2) to predict potato tuber yield. Moreover, the violin plot distribution generated by the Vote-LGCB model with Combo 4 is also consistent with the minor normalized error along with Mean = 0.0675 as compared to all other models. Hence, the Vote-LGCB model with Combo 4 accomplishes better prediction accuracy for potato tuber yield monitoring.

The expected time-series trend plots in Fig. 12 compare the measured and predicted potato tuber yield (pink) and their residuals (blue) using six predictive models in the best possible input combination. The Vote-LGCB model with Combo 4 accomplished better accuracy in terms of parallel and consistent trends against the measured potato tuber yield as compared to the LASSO (with Combo 3), ELNET (with Combo 3), Extra tree (with Combo 3), and LightGBM (with Combo 2), and CatBoost (with Combo 2) models. The Vote-LGCB model with Combo 3 also generated smaller residuals, which further confirmed the prediction accuracy was higher during potato tuber yield prediction. Thus, the Vote-LGCB model with the Combo 3 model outperforms other models in predicting potato tuber yield accurately.

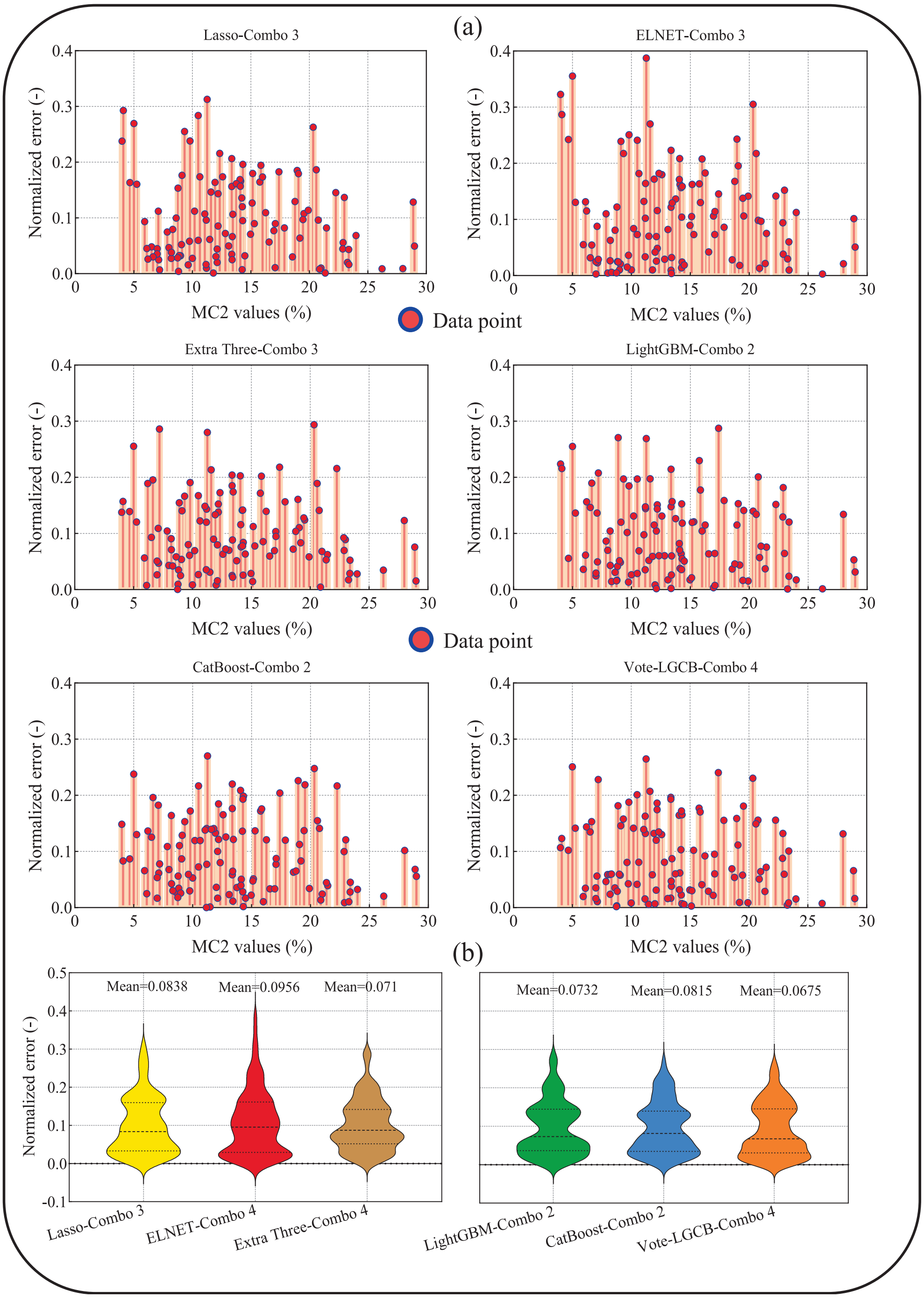


Fig. 11. Normalized error distribution versus the MC2, as the most significant input feature (a) and violin plot of normalized error values (b) to assess the robustness of the super ensemble approach compared to the predictive models (according to the optimal input combination) aim to the yield modeling.

The absolute forecast error |FE| are given in Fig. 13 using empirical cumulative distribution function (ECDF) and a 95 % confidence interval to compare the Vote-LGCB model with Combo 3 against other benchmarking models to portray a more tangible view. The ECDF of the Vote-LGCB model with Combo 3 exhibited a very close profile corresponding to 60 % of the cumulative probability, and the minimal |FE| value is around 10 within the 95 % confidence interval. The comparing models LASSO (with Combo 3), ELNET (with Combo 3), Extra tree (with Combo 3), LightGBM (with Combo 2), and CatBoost (with Combo 2) exhibit slightly higher |FE| values. Hence, these figures further confirm the suitability of the Vote-LGCB model in predicting potato tuber yield.

Fig. 14 reports explainability and interpretability during the potato tuber yield model prediction, which consists of a Force plot (top), summary plot (middle), and correspondence plot (below). The forced and summary plots are exhibited for the pre-defined point of dividing training and testing subsets. Fig. 14, extracted by the SHAP explainer, appraises the impact and influence of every significant input predictor on the Vote-LGCB model's prediction. The SHAP values resulted in the LightGBM and CatBoost average as the voting method's *meta*-leaners (i.e., Vote-LGCB), representing the super ensemble SHAP outcomes. The Force plot indicates that the input predictors MC2, NDVI4, HCP2, Slope3, NDVI2, and PAI1 in red have contributed significantly to the model's output prediction with score value MC2 = 17.81 appeared to be the most positively contributed predictor, followed by PAI1, and HCP2.

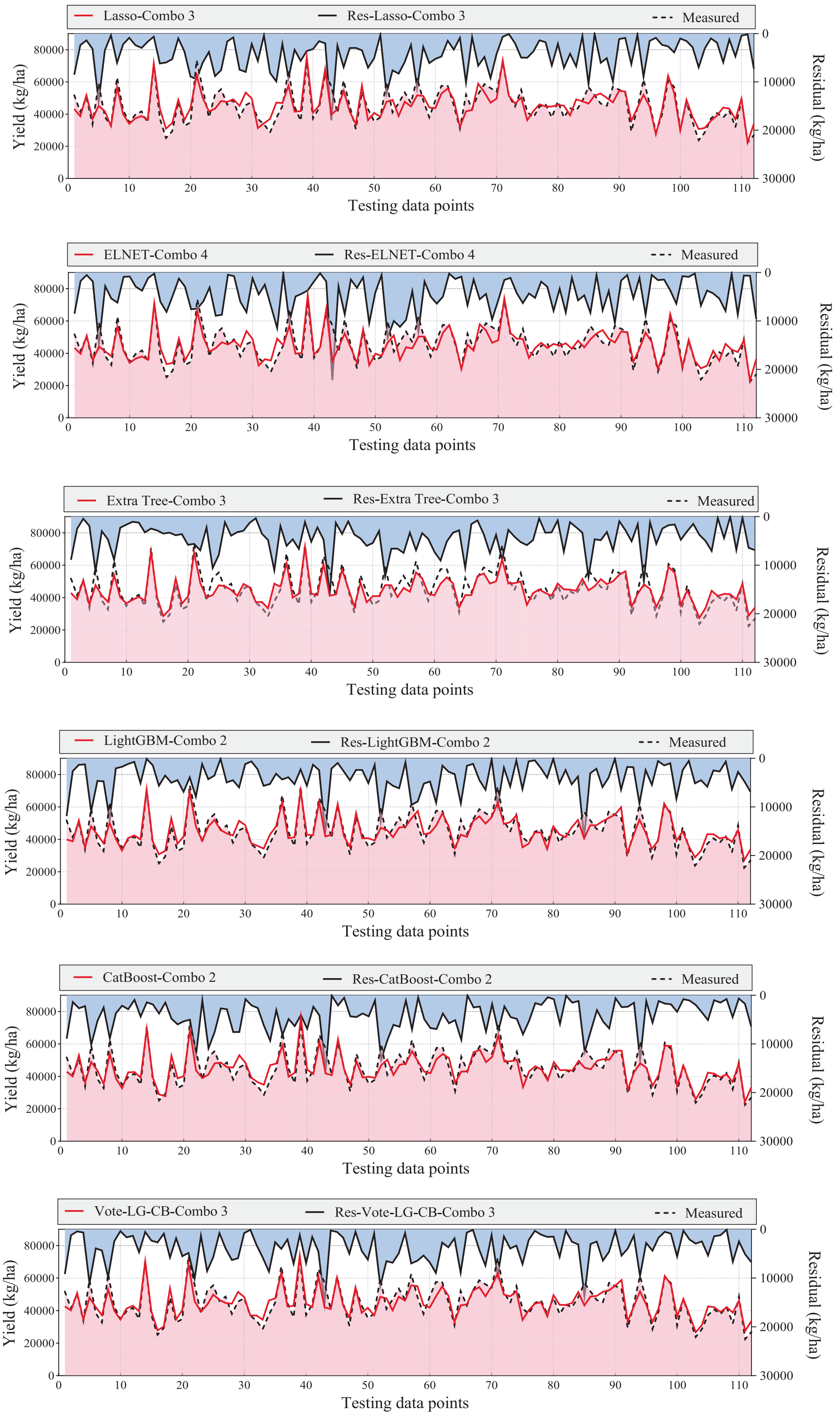


Fig. 12. Physical expected trends related to yield values were measured using six predictive models in the best possible input combination, and yield values were calculated and predicted in the testing phase (pink colour) and residual values (blue colour). Here, the term Res refers to the residuals.

The predictors PRP2 and BS3 in blue colour show low feature values in the model's prediction.

The red dots in the summary and correspondence plots display that the corresponding predictors (i.e., MC2, NDVI4, OM1, HC3, HCP2, PAI1, and NDVI2) have a higher impact and influence on the model's output prediction. In contrast, the variable in the blue dot describes lower and poor effects. Based on Fig. 14, the numerical values in the waterfall plots are the model's score. The input predictor MC2, with a score of 17.81, contributed significantly to potato tuber yield prediction and pushed the Vote-LGCB model to attain the highest score value. Accordingly, the

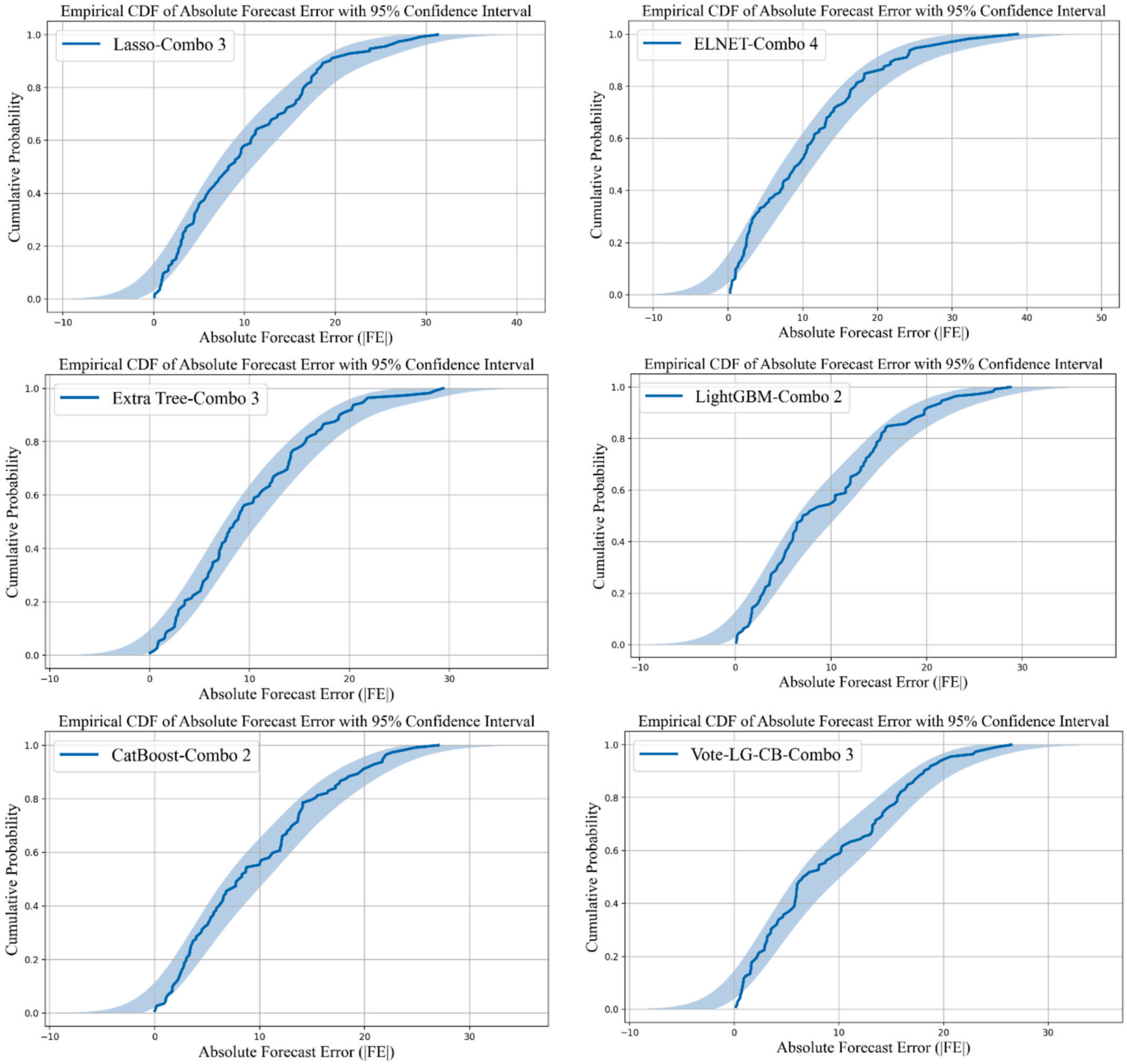


Fig. 13. Empirical cumulative distribution function (ECDF) versus absolute forecast error |FE| and 95% confidence interval for six predictive models (superior input combination) in the testing datasets of yield modeling.

input feature MC2, followed by NDVI4, OM1, HC3, HCP2, PAI1, and NDVI2, are the significant features for the model's accuracy in prediction.

7. Further discussion

The manuscript presents a novel approach for accurately estimating potato tuber yields in Atlantic Canadian provinces by leveraging a comprehensive soil property data collected across multiple fields and growing seasons. An advanced explainable ML framework employing techniques like Boruta-SHAP feature selection, best subset regression, and multi-criteria decision-making methods (WASPAS and MOORA) has been constructed to identify the most influential soil variables affecting potato tuber yield. A unique ensemble model called VOTE-LGCB, which combines CatBoost and LightGBM through a voting scheme, has been developed and demonstrated to outperform other comparing models (i.e., LASSO, ELNET, Extra tree, LightGBM, CatBoost) with high correlation and low error rates, highlighting the advantages of the ensemble approach. The voting ability caused the ensemble robustness of both classical CatBoost and LightGBM schemes to capture the nonlinearities of the understudy target efficiently. However, the proposed framework involves advanced techniques like ensemble modelling, multi-criteria decision-making, and feature selection, which can be computationally intensive, especially for larger datasets or real-time applications. As the current study is the first ground-based big data modeling endeavour using 30 feature inputs, developing a robust model that captures the nonlinearities between yield and available features is deeply challenging. Thus, comparing the present research achievements with previous literature is restricted to satellite image-based modeling conducted by (Gómez et al., 2019) in Spain and (Salvador et al., 2020) in

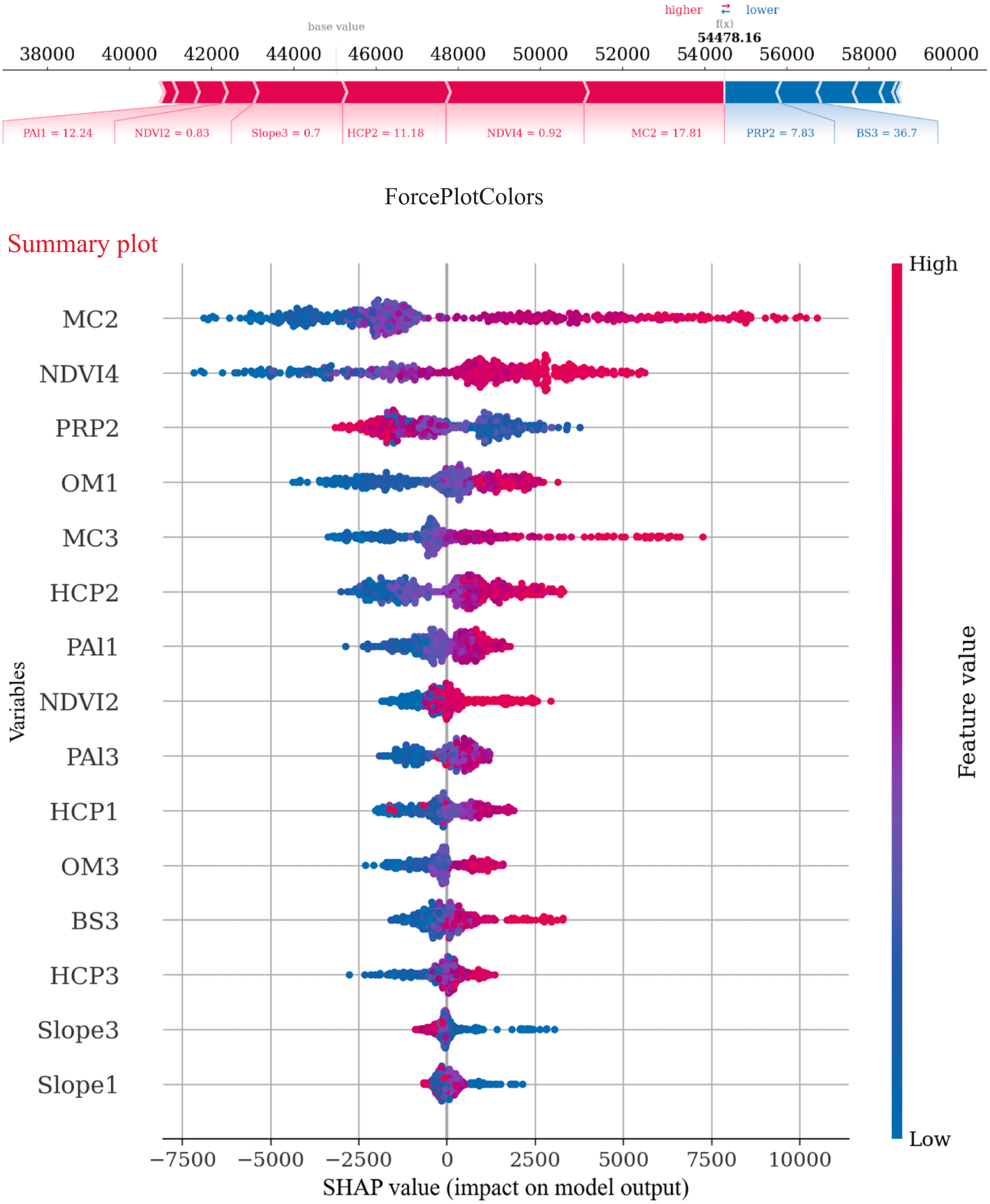


Fig. 14. Interpretable results during the training phase of modeling. Forced plot, summary plot (at the pre-defined point of dividing training and testing sub-sets), and correspondence plot.

Mexico. In Spain, potato yield was estimated using images from the twin Sentinel 2 satellites (European Space Agency Copernicus Programme) based on the Support Vector Machine Radial (SVR Radial) (Gómez et al., 2019). The SVR Radial accuracy resulted in R = 0.964 and RMSE = 11.7, compared to R = 0.8875 in the current study. In Mexico, the superior ML model based on ERA5′s meteorological data and satellite imagery from TERRA, Support Vector Machine Polynomial (SVMP), led to R = 0.926 and RMSE = 14.9 (Salvador et al., 2020). In the abovementioned research, modelling was performed using the classification regarding the fewer input features. Consequently, their reliability and efficiency are fewer than those of the VOTE-LGCB platform in modeling potato tuber yield based on numerous input data.

This study demonstrated that SHAP analysis and the VOTE-LGCB model consistently identified soil moisture content (MC), soil pH, Cation Exchange Capacity (CEC), and the Normalized Difference Vegetation Index (NDVI) as the primary variables influencing potato tuber

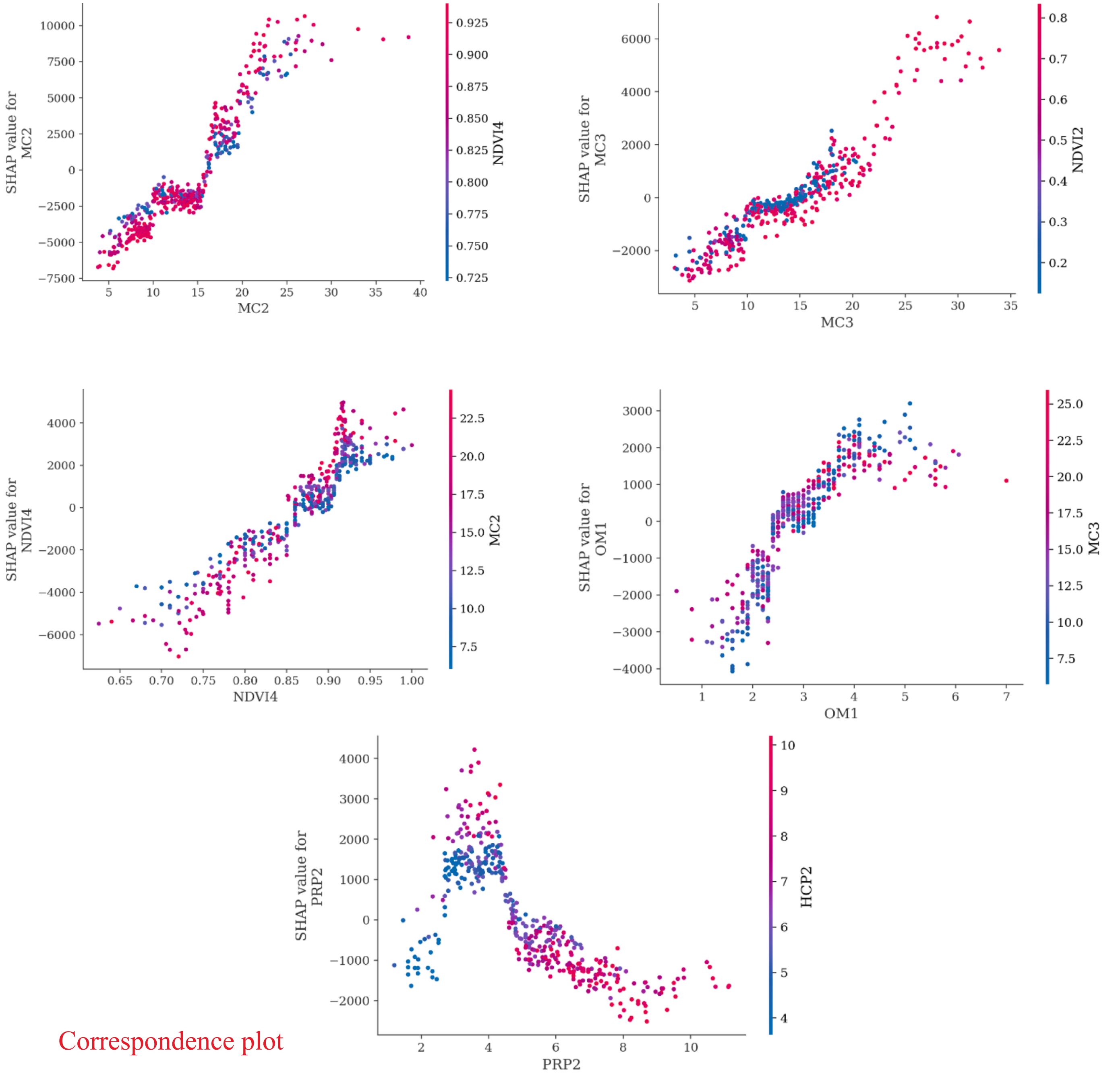


Fig. 14. (*continued*).

yield, corroborating established agronomic principles. Soil moisture is a crucial determinant of tuber initiation and bulking; insufficient water availability during these phases reduces tuber size and uneven development. All the Analyses indicate that yield fluctuations strongly correlate with variations in moisture content throughout the season, emphasizing the necessity of precise irrigation management in the Maritime regions. Similarly, soil pH affects nutrient availability, especially phosphate and micronutrients essential for tuber development; the current model results show a pronounced decrease in expected yield when pH levels deviate significantly from the optimal range (5.5–6.5) recommended for potatoes. Cation Exchange Capacity (CEC), indicative of the soil's ability to retain and exchange vital cations like $Ca2^+$, $Mg2^+$, and K^+, is directly associated with soil fertility and the plant's capacity to maintain consistent tuber development. Our analysis indicates that lower CEC values (<8 cmol/kg) correlate with significant yield losses, suggesting that soils with inadequate cation exchange characteristics may require targeted fertilization or organic matter improvements.

Furthermore, NDVI, derived from proximal sensing, serves as a reliable metric for canopy vitality and photosynthetic performance, establishing a direct connection between soil conditions and aboveground plant health. By integrating these variables, the provided methodology achieves high forecast accuracy and provides interpretable insights into the underlying mechanisms influencing potato production variability. The findings underscore that the proposed model functions as both a predictive tool and a decision-support system, enabling site-specific soil and crop management strategies in the region. The existing approach exhibits commendable predictive performance, although other avenues could augment its robustness. Integrating physical models with the suggested framework may elucidate the fundamental agronomic processes, enhancing interpretability and reliability.

The VOTE-LGCB model exhibited robust predictive accuracy for potato yield estimation in the examined fields of PEI and NB; however, its applicability to other regions with varying soil types, climatic conditions, or management practices requires external validation. Future endeavours will entail evaluating the framework on autonomous datasets and varied agro-climatic circumstances to determine its transferability and resilience across different production contexts. This study examined 30 soil parameters; however, future research should investigate additional aspects, including climatic variations, pest and disease prevalence, and management strategies. These improvements will augment the model's practical utility across many agricultural contexts. On top of that, to augment the applicability of the current methodology for real-world scenarios, it has discerned a minimized subset of significantly impactful soil characteristics (e.g., moisture content, pH, and CEC) by Boruta-SHAP and Best Subset Regression, facilitating yield calculation independent of the complete dataset. Furthermore, we advocate for integrating this diminished feature set with proximal sensing instruments, such as NDVI and soil moisture sensors, which can offer dynamic crop condition indicators with fewer field measurements. This method reduces expenses and labor requirements while ensuring strong yield forecast effectiveness and can be considered for future purposes.

8. Conclusion

This research presents an advanced, eXplainable high-dimensional feature filtering and Bayesian super ensemble framework (Vote-LGCB) for predicting potato tuber production through a combination of soil physical and chemical parameters. Integrating multi-criteria decision-making methodologies (WASPAS and MOORA), we discovered four best input combinations (Combo 1–4) for model training and evaluation. Combo 3 consistently surpassed the other input sets, exhibiting enhanced predictive capability. The Vote-LGCB model, augmented with SHAP-based interpretability, attained superior accuracy (R = 0.9860, RMSE = 1824.99, MAPE = 3.44, KGE = 0.93, U95% = 5061.46, Reliability = 99.77) in comparison to the LASSO, ELNET, Extra Tree, LightGBM, and CatBoost models, hybridised with preprocessing procedure. This research's principal innovation involves converting the opaque Vote-LGCB model into an interpretable framework through SHAP analysis, which identified soil moisture content, NDVI, organic matter (SOM), ground conductivity, and phosphorus-to-aluminium ratio (PAL) as the most significant predictors of tuber yield. This improved comprehension of soil-yield dynamics can inform precision agricultural methodologies by pinpointing the aspects most significantly influence yield variability. In addition to predicting potato yields, the suggested approach has potential applications in agriculture, hydrology, environmental management, and renewable resource optimisation, facilitating data-driven decision-making to enhance resource allocation and sustainability. Future research should integrate meteorological factors (temperature, humidity, and precipitation), disease prevalence (e.g., late blight, early blight, common scab), and pesticide usage (type, frequency, and efficacy) to enhance model precision. Extending this methodology to forecast yields of additional crops, including blueberries and soybeans, would enhance its relevance and influence.

CRediT authorship contribution statement

Mehdi Jamei: Writing – review & editing, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Aitazaz Ahsan Farooque:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Mumtaz Ali:** Validation, Resources. **Masoud Karbasi:** Resources, Methodology. **Hassan Afzaal:** Methodology, Investigation. **Saad Javed Cheema:** Methodology. **Qamar Uz Zaman:** Writing – review & editing, Writing – original draft. **Kok Sin Woon:** Writing – review & editing. **Paul Sheridan:** Writing – review & editing, Supervision.

Funding

Natural Sciences and Engineering Research Council of Canada.

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.

Acknowledgements

The authors thank the Natural Sciences and Engineering Research Council of Canada (Alliance Grant) for providing financial support in conducting this research. A special thanks to the Sustainable Agriculture Research Team at the University of Prince Edward Island for their assistance during data collection.

Ethics approval: Not applicable

Data availability

The data that has been used is confidential.

References

Abbas, F., Afzaal, H., Farooque, A.A., Tang, S., 2020. Crop yield prediction through proximal sensing and machine learning algorithms. Agronomy 10, 1046. https://doi.org/10.3390/agronomy10071046.

Ahmed, M.T., Ahmed, M.W., Kamruzzaman, M., 2025. A systematic review of explainable artificial intelligence for spectroscopic agricultural quality assessment. Comput. Electron. Agric. 235, 110354.

Anukrishna, P.R., Paul, V., 2017. A review on feature selection for high dimensional data. In: Proc. Int. Conf. Inven. Syst. Control. ICISC 2017. https://doi.org/10.1109/ICISC.2017.8068746.

Ba-Alawi, A.H., Heo, S., Aamer, H., Chang, R., Woo, T., Kim, M., Yoo, C., 2023. Development of transparent high-frequency soft sensor of total nitrogen and total phosphorus concentrations in rivers using stacked convolutional auto-encoder and explainable AI. J. Water Process Eng. 53, 103661.

Ben Brahim, A., Limam, M., 2018. Ensemble feature selection for high dimensional data: a new method and a comparative study. ADAC 12, 937–952. https://doi.org/10.1007/S11634-017-0285-Y/TABLES/7.

Brauers, W.K.M., Ginevičius, R., Podvezko, V., 2010. Regional development in Lithuania considering multiple objectives by the MOORA method. Technol. Econ. Dev. Econ. 16, 613–640.

Chakraborty, S., 2011. Applications of the MOORA method for decision making in manufacturing environment. Int. J. Adv. Manuf. Technol. 54, 1155–1166.

Das, P., Jha, G.K., Lama, A., Parsad, R., 2023. Crop yield prediction using hybrid machine learning approach: a case study of lentil (Lens culinaris Medik.). Agric 13, 596. https://doi.org/10.3390/AGRICULTURE13030596/S1.

Debnath, B., Bari, A.B.M.M., Haq, M.M., de Jesus Pacheco, D.A., Khan, M.A., 2023. 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, 100001.

Ensign, M.R., 1935. Factors Influencing the Growth and Yield of Potatoes in Florida. JSTOR 10.

Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., Zeng, W., 2019. Light Gradient Boosting Machine: an efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. Agri. Water Manag. 225, 105758.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42.

Gholami, H., Mohammadifar, A., Golzari, S., Kaskaoutis, D.G., Collins, A.L., 2021. Using the Boruta algorithm and deep learning models for mapping land susceptibility to atmospheric dust emissions in Iran. Aeolian Res. 50, 100682.

Gómez, D., Salvador, P., Sanz, J., Casanova, J.L., 2019. Potato yield prediction using machine learning techniques and sentinel 2 data. Remote Sens. 11, 1745.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377, 80–91.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J., Yu, X., Zeng, W., Zhou, H., 2019. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. J. Hydrol. 574, 1029–1041.

Jamei, M., Ahmadianfar, I., Chu, X., Yaseen, Z.M., 2021a. Estimation of triangular side orifice discharge coefficient under a free flow condition using data-driven models. Flow Meas. Instrum. 77, 101878. https://doi.org/10.1016/j.flowmeasinst.2020.101878.

Jamei, M., Ahmadianfar, I., Chu, X., Yaseen, Z.M., 2021b. Estimation of triangular side orifice discharge coefficient under a free flow condition using data-driven models. Flow Meas. Instrum. 77, 101878.

Jamei, M., Ahmadianfar, I., Chu, X., Yaseen, Z.M., 2020. Estimation of triangular side orifice discharge coefficient under a free flow condition using data-driven models. Flow Meas. Instrum., 101878

Jamei, M., Karbasi, M., Adewale Olumegbon, I., Moshraf-Dehkordi, M., Ahmadianfar, I., Asadi, A., 2021c. Specific heat capacity of molten salt-based nanofluids in solar thermal applications: a paradigm of two modern ensemble machine learning methods. J. Mol. Liq. 335, 116434. https://doi.org/10.1016/j.molliq.2021.116434.

Jamei, M., Karbasi, M., Alawi, O.A., Kamar, H.M., Khedher, K.M., Abba, S.I., Yaseen, Z. M., 2022. Earth skin temperature long-term prediction using novel extended Kalman filter integrated with artificial intelligence models and information gain feature selection. Sustain. Comput. Inform. Syst. 35, 100721.

John, V., Liu, Z., Guo, C., Mita, S., Kidono, K., 2016. Real-time lane estimation using deep features and extra trees regression. Springer, pp. 721–733.

Karande, P., Chakraborty, S., 2012. Application of multi-objective optimization on the basis of ratio analysis (MOORA) method for materials selection. Mater. Des. 37, 317–324.

Kbakural, B.R., Robert, P.C., Huggins, D.R., 1999. Variability of Corn/Soybean Yield and Soil/Landscape Properties across A Southwestern Minnesota Landscape 573–579. DOI: 10.2134/1999.precisionagproc4.c51.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. LightGBM: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Proces. Syst. 3149–3157.

Kravchenko, A.N., Bullock, D.G., 2000. Correlation of corn and soybean grain yield with topography and soil properties. Agron. J. 92, 75–83. https://doi.org/10.2134/agronj2000.92175x.

Kumar, V., Minz, S., 2016. Multi-view ensemble learning: an optimal feature set partitioning for high-dimensional data classification. Knowl. Inf. Syst. 49, 1–59. https://doi.org/10.1007/S10115-015-0875-Y/TABLES/9.

Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, A., Uwitonze, C., Ngabonziza, J., Uwamahoro, A., 2023. Crop yield prediction using machine learning models: case of irish potato and maize. Agric 13, 225. https://doi.org/10.3390/AGRICULTURE13010225.

Kursa, M.B., Jankowski, A., Rudnicki, W.R., 2010. Boruta – a system for feature selection. Fundam. Informaticae 101, 271–285. https://doi.org/10.3233/FI-2010-288.

Liu, Y., Feng, H., Fan, Y., Yue, J., Chen, R., Ma, Y., Bian, M., Yang, G., 2024. Improving potato above ground biomass estimation combining hyperspectral data and harmonic decomposition techniques. Comput. Electron. Agric. 218, 108699.

Liu, Y., Feng, H., Fan, Y., Yue, J., Yang, F., Fan, J., Ma, Y., Chen, R., Bian, M., Yang, G., 2025. Utilizing UAV-based hyperspectral remote sensing combined with various agronomic traits to monitor potato growth and estimate yield. Comput. Electron. Agric. 231, 109984.

Liu, Y., Feng, H., Yue, J., Li, Z., Yang, G., Song, X., Yang, X., Zhao, Y., 2022. Remote-sensing estimation of potato above-ground biomass based on spectral and spatial features extracted from high-definition digital camera images. Comput. Electron. Agric. 198, 107089. https://doi.org/10.1016/j.compag.2022.107089.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process Syst. 30.

Malik, A., Jamei, M., Ali, M., Prasad, R., Karbasi, M., Yaseen, Z.M., 2022. Multi-step daily forecasting of reference evapotranspiration for different climates of India: a modern multivariate complementary technique reinforced with ridge regression feature selection. Agric Water Manag 272, 107812. https://doi.org/10.1016/j.agwat.2022.107812.

Parmar, K.P., Bhatt, T., 2022. Crop yield prediction based on feature selection and machine learners: a review. In: Proc. 2nd Int. Conf. Artif Intell. Smart Energy, ICAIS, pp. 354–358. https://doi.org/10.1109/ICAIS53314.2022.9742891.

Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S., Mohammadian, A.(Kouros), 2020. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accid. Anal. Prev. 136, 105405. https://doi.org/10.1016/j.aap.2019.105405.

Paudel, D., De Wit, A., Boogaard, H., Marcos, D., Osinga, S., Athanasiadis, I.N., 2023. Interpretability of deep learning models for crop yield forecasting. Comput. Electron. Agric. 206, 107663.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. Catboost: unbiased boosting with categorical features. Adv. Neural Inf. Proces. Syst. 6637–6647.

Ray, P., Reddy, S.S., Banerjee, T., 2021. Various dimension reduction techniques for high dimensional data analysis: a review. Artif. Intell. Rev. 54, 3473–3515. https://doi.org/10.1007/S10462-020-09928-0/METRICS.

Salvador, P., Gómez, D., Sanz, J., Casanova, J.L., 2020. Estimation of potato yield using satellite data at a municipal level: a machine learning approach. ISPRS Int. J. Geo-Information 9, 343.

Shakeel, A., Chong, D., Wang, J., 2023. District heating load forecasting with a hybrid model based on LightGBM and FB-prophet. J. Clean. Prod. 409, 137130.

Shrinkage, R., 2016. Regression Shrinkage and Selection via the Lasso Author (s): Robert Tibshirani Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 58 , No . 1 (1996), Published by : Wiley for the Royal Statistical Society Stable URL 58, 267–288.

Singh, U.K., Jamei, M., Karbasi, M., Malik, A., Pandey, M., 2022. Application of a modern multi-level ensemble approach for the estimation of critical shear stress in cohesive sediment mixture. J. Hydrol., 127549

Wieland, R., Lakes, T., Nendel, C., 2021. Using Shapley additive explanations to interpret extreme gradient boosting predictions of grassland degradation in Xilingol. China. Geosci. Model Dev. 14, 1493–1510. https://doi.org/10.5194/gmd-14-1493-2021.

Zavadskas, E.K., Turskis, Z., Antucheviciene, J., Zakarevicius, A., 2012. Optimization of weighted aggregated sum product assessment. Elektron. Ir Elektrotechnika 122, 3–6.

Zhang, S., Wu, J., Jia, Y., Wang, Y.G., Zhang, Y., Duan, Q., 2021. A temporal LASSO regression model for the emergency forecasting of the suspended sediment concentrations in coastal oceans: Accuracy and interpretability. Eng. Appl. Artif. Intel. 100, 104206. https://doi.org/10.1016/j.engappai.2021.104206.