Field Crops Research 287 (2022) 108640

Contents lists available at ScienceDirect

Field Crops Research

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

Interpretable machine learning methods to explain on-farm yield variability of high productivity wheat in Northwest India

Hari Sankar Nayak [a], João Vasco Silva [b], Chiter Mal Parihar [a,*], Timothy J. Krupnik [c], Dipaka Ranjan Sena [a], Suresh K. Kakraliya [d], Hanuman Sahay Jat [e], Harminder Singh Sidhu [f], Parbodh C. Sharma [f], Mangi Lal Jat [d,g], Tek B. Sapkota [h,*]

[a] ICAR-Indian Agricultural Research Institute (IARI), New Delhi, India
[b] International Maize and Wheat Improvement Center (CIMMYT), Harare, Zimbabwe
[c] International Maize and Wheat Improvement Center (CIMMYT), Dhaka, Bangladesh
[d] International Maize and Wheat Improvement Center (CIMMYT), New Delhi, India
[e] ICAR-Central Soil Salinity Research Institute (CSSRI), Karnal, India
[f] Borlaug Institute for South Asia (BISA), Ludhiana, India
[g] Resilient Farm and Food Systems Program, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, 502324, Hyderabad, India
[h] International Maize and Wheat Improvement Center (CIMMYT), El-Batan, Mexico

ARTICLE INFO

Keywords:
Random forest
Variable importance
Interaction strength
Accumulated local effect plot
Partial dependency plot
Quantile regression
Big data

ABSTRACT

The increasing availability of complex, geo-referenced on-farm data demands analytical frameworks that can guide crop management recommendations. Recent developments in interpretable machine learning techniques offer opportunities to use these methods in agronomic studies. Our objectives were two-fold: (1) to assess the performance of different machine learning methods to explain on-farm wheat yield variability in the Northwestern Indo-Gangetic Plains of India, and (2) to identify the most important drivers and interactions explaining wheat yield variability. A suite of fine-tuned machine learning models (ridge and lasso regression, classification and regression trees, k-nearest neighbor, support vector machines, gradient boosting, extreme gradient boosting, and random forest) were statistically compared using the R^2, root mean square error (RMSE), and mean absolute error (MAE). The best performing model was again fine-tuned using a grid search approach for the bias-variance trade-off. Three *post-hoc* model agnostic techniques were used to interpret the best performing model: variable importance (a variable was considered “important” if shuffling its values increased or decreased the model error considerably), interaction strength (based on Friedman’s H-statistic), and two-way interaction (i.e., how much of the total variability in wheat yield was explained by a particular two-way interaction). Model outputs were compared against empirical data to contextualize results and provide a blueprint for future analysis in other production systems. Tree-based and decision boundary-based methods outperformed regression-based methods in explaining wheat yield variability. Random forest was the best performing method in terms of goodness-of-fit and model precision and accuracy with RMSE, MAE, and R^2 ranging between 367 and 470 kg ha^{-1}, 276–345 kg ha^{-1}, and 0.44–0.63, respectively. Random forest was then used for selection of important variables and interactions. The most important management variables explaining wheat yield variability were nitrogen application rate and crop residue management, whereas the average of monthly cumulative solar radiation during February and March (coinciding with reproductive phase of wheat) was the most important biophysical variable. The effect size of these variables on wheat yield ranged between 227 kg ha^{-1} for nitrogen application rate to 372 kg ha^{-1} for cumulative solar radiation during February and March. The effect of important interactions on wheat yield was detected in the data namely the interaction between crop residue management and disease management and, nitrogen application rate and seeding rate. For instance, farmers’ fields with moderate disease incidence yielded 750 kg ha^{-1} less when crop residues were removed than when crop residues were retained. Similarly, wheat yield response to residue retention was higher under low seed and N application rates. As an inductive research approach, the appropriate application of interpretable machine learning methods can be used to extract agronomically actionable information from large-scale farmer field data.

[*] Corresponding authors.
E-mail addresses: pariharcm@gmail.com (C.M. Parihar), T.Sapkota@cgiar.org (T.B. Sapkota).

https://doi.org/10.1016/j.fcr.2022.108640
Received 3 February 2022; Received in revised form 12 July 2022; Accepted 29 July 2022
Available online 20 August 2022

1. Introduction

Interest in the application of machine learning in agronomic science is increasing in tandem with the growing availability of geo-referenced farmer field data and spatially explicit environmental data in a diversity of cropping systems (Jaenisch et al., 2021; Tseng et al., 2021). Such large datasets provide new opportunities to apply inductive research approaches examining the relationships between crop management and environmental conditions with crop yield, and may be an appropriate and cost-effective way to conduct agronomic research across large scales (Silva et al., 2020; Rattalino Edreira et al., 2017). Importantly, many of these approaches differ from those associated with hypothesis-driven, manipulative experimentation, and deductive research (de Mauro et al., 2016). Despite the wealth of studies identifying important variables governing crop yield variability (e.g., Correndo et al., 2021; Park et al., 2018), studies that employ large-scale observations of farmers' management practices to assess how these affect crop yield directly or in interaction with other practices are less common (e.g., Tseng et al., 2021; Devkota and Yigezu, 2020; Di Mauro et al., 2018).

Machine learning lies at the intersection of computer science and statistics, and has been used to unravel patterns in large datasets that are challenging to analyze with more conventional statistical approaches (Tolle et al., 2011). In agriculture, machine learning is being increasingly used to predict crop yield, although the research community remains divided on which methods are most appropriate given different data types and contexts (Ransom et al., 2019; Khaki and Wang, 2019; Van Klompenburg et al., 2020; Shook et al., 2021). Four different types of machine learning methods can be identified based on their interpretability and level of complexity: (i) regression-based methods (e.g., linear, ridge, and lasso regression), (ii) single tree- or multiple tree-based boosted methods (e.g., classification and regression trees, gradient boosting, extreme gradient boosting, and random forest), (iii) proximity-based K-nearest neighborhood (KNN), and (iv) decision boundary-based support vector regression (James et al., 2013). A method is deemed interpretable if the estimated model coefficients represent the effect size of the independent variables on the dependent variable (such as in regression-based methods; James et al., 2013). Furthermore, a method is deemed complex if it has many hyper-parameters that need to be tuned while fitting the model as compared to simpler models for which coefficients can be estimated using simple loss functions and regularization and the coefficients can not be interpretd directly for these models (James et al., 2013). This study provides an overview of the different machine learning methods available to explain on-farm yield variability and illustrates the usefulness of interpretable machine learning techniques for agronomic studies using farmer field data for irrigated wheat in the Northwestern Indo-Gangetic Plains (IGP) of India.

The rice-wheat cropping system is the predominant cropping system in the Northwestern IGP and is the key supplier of calories for food security in India, where the spring wheat is cultivated after harvest of a rainy season rice crop (Bhatt et al., 2021; Nayak et al., 2022). Spring wheat is generally cultivated during the winter season, between mid-November and mid-April. Wheat yield in this cropping system is affected by a combination of climatic conditions during the growing season, soil properties resulting from previous management, and crop management practices adopted by farmers (Singh et al., 2014; Kumar et al., 2019). For instance, in rice-wheat cropping systems in the region, rice fields are wet-tilled in a process termed puddling, and kept flooded to ensure crop growth and suppress weeds. However, intensive tillage and puddling can also lead to poor soil structure, sub-optimal permeability in subsurface soil layers, poor soil aeration, and soil compaction, which in turn can adversely affect the growth and yield of the subsequent wheat crop (Chauhan et al., 2012; Singh et al., 2014; Gathala et al., 2011). Herbicide resistant strains of *Phalaris minor*, deficiencies of micronutrients, and low use efficiency of nitrogen fertilizers are also important reducing and limiting factors for wheat productivity in the region (Bhatt et al., 2021).

Machine learning has been used to describe crop yield variability (Shendryk et al., 2021; Cao et al., 2021), but the model interpretation has been limited to the identification of important variables (Correndo et al., 2021; Krupnik et al., 2015). There are also few studies comparing the performance of different statistical methods (e.g., Mourtzinis et al., 2018). Classification and regression trees have been the preferred method to derive agronomic recommendations given its intuitive outcomes (Di Mauro et al., 2018; Krupnik et al., 2015). The advent of model-agnostic interpretable techniques makes it possible to identify the local and global effect of important variables, as well as their interactions on crop yield using different machine learning methods. The objectives of this study were: (1) to assess the performance of machine learning methods in explaining crop yield variability, (2) to identify the important variables and interactions driving wheat yield variability in the Northwestern IGP of India, and (3) to visualize the effect of important variables on wheat yield to generate evidence-based agronomic recommendations. This study consequently aims to illustrate how machine learning can be used as an inductive method in agronomy to unravel the key drivers of crop yield variability using a wealth of biophysical and management data from observations made across thousands of farmers' fields, as an alternative and/or complement to more traditional, inductive-based manipulative experimentation.

2. Material and methods

2.1. On-farm data for wheat production in the Northwestern IGP of India

A field survey was conducted during two consecutive wheat growing seasons, 2019–2020 and 2020–2021, in the three districts of Haryana (Ambala, Karnal, and Kurukshetra) and four districts of Punjab (Fatehgarh Sahib, Ludhiana, Kapurthala, and Patiala). Further details about the climatic and soil conditions across the states of Haryana and Punjab are provided in Nayak et al. (2022). The blocks and villages within the districts were purposefully selected to represent varying levels of extension outreach and technology adoption; further the farmers with in each village were randomly selected for the field survey. The surveyed fields were geo-referenced so that associated climatic and soil data could be retrieved from secondary sources. Farmers were surveyed during wheat harvest using a structured questionnaire coded on the Open-Data Kit (ODK) platform and responses were monitored real-time using a visualization dashboard. The information requested per field included: (i) variety grown, (ii) tillage practices and crop establishment method, (iii) water and nutrient management practices, (iv) pest, disease, and weed severity and control, and (v) source of inputs, labor requirements, crop production and area, and other socioeconomic characteristics (Supplementary information 1). The criteria to select biophysical and management variables was as follows: (a) variables must be agronomically meaningful to explain crop yield variability, (b) the distribution of the data for each variable must be suitable for analysis, and (c) variables can be interpreted to the farmers and extension agents. The farmer reported crop production data were verified with measured yield data obtained through crop-cut in a 2 × 2 m^2 quadrant in about 25% of the surveyed fields. A linear regression was fitted between the self-reported yield and the crop cut yield (Supplementary Fig. 1), which was then used to correct the self-reported yield for the fields where crop cuts were not conducted. The univariate outlier detection was done based on the analysis of boxplot and for bivariate outliers screening robust Mahalanobis distance complemented with expert knowledge was used, as described in Nayak et al. (2022). The total sample size across the two growing seasons was 6181 field-year combinations, from which 42 field-year combinations were identified as outliers and thus excluded in further analyses.

The weather and soil data were obtained from secondary sources using the GPS coordinates of each surveyed field. Weather data were retrieved from the ERA5 hourly re-analyzed database (Sabater, 2019)

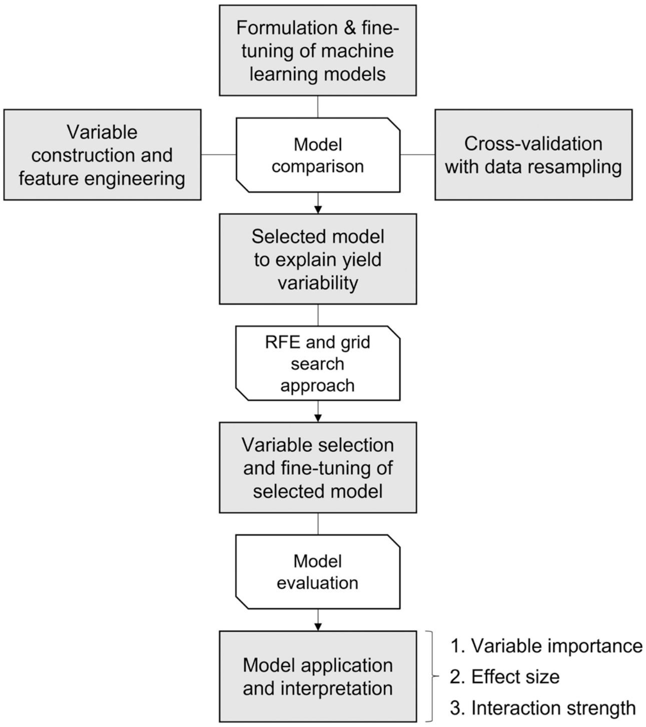
Formulation & fine-tuning of machine learning models
Variable construction and feature engineering
Model comparison
Cross-validation with data resampling
Selected model to explain yield variability
RFE and grid search approach
Variable selection and fine-tuning of selected model
Model evaluation
Model application and interpretation
1. Variable importance
2. Effect size
3. Interaction strength

Fig. 1. Analytical framework for using interpretable machine learning models in agronomic studies. The reader is referred to the main text for further explanation of each step. RFE = recursive feature elimination.

and included solar radiation (kW m^{-2}) and minimum and maximum temperatures (°C). The growing season was split into two periods, December to January and February to March, as they approximately coincide with vegetative and reproductive phase of wheat, respectively (Fischer et al., 2022). Soil textural data were retrieved from the International Soil Reference and Information Centre (ISRIC) database at a resolution of 250 m (Hengl et al., 2017). Soil texture classes were derived from the particle size distribution data from ISRIC using the USDA textural triangle classification.

2.2. Analytical approach to explain crop yield variability

The analytical framework used in this study encompassed four steps (Fig. 1). First, machine learning models were fine-tuned using random search approach and fitted to the data comprising relevant agronomic and biophysical variables using a 10-fold cross-validation scheme with data resampling. Second, the performance of the fitted models was assessed using statistical indices, which were used to identify the model explaining yield variability best. Third, the best performing model was fine-tuned using a grid search approach and recursive feature elimination to check for the bias-variance trade-off and to refine the final variables to be used in the further analysis. Lastly, the refined model was used to explain yield variability based on variable importance, quantification of effect sizes, and analysis of the interaction strength between the variables. The analytical steps are explained in greater detail in following sub-sections.

2.2.1. Formulation of machine learning models

Machine learning methods were chosen based on their working principle, level of complexity, and ease of interpretability (Table 1). Regression-based methods are parametric, and their coefficients are obtained through ordinary least-squares. Conversely, other methods like classification and regression trees, k-nearest neighbors, and support vector machines are non-parametric (James et al., 2013). Tree-based methods such as classification and regression tree, random forest, and gradient boosting rely on decision trees, and a series of if-then rules to arrive at a particular prediction or classification (Breiman, 2001a). Distance-based methods, like K-nearest mean, find the K-nearest neighbors in the feature space and provide predictions based on those K data point's outcomes (e.g., Cover and Hart, 1967). Decision boundary-based methods create a decision boundary with data projected in a higher dimension (suppose there are three variables, then their values will be projected in a 3-dimensional space) to derive a particular prediction. Regression-based methods are less complex and have easier and direct interpretation compared to tree-based or distance and decision boundary-based methods (Hastie et al., 2009).

Two regression-based methods were considered in the analysis to capture a linear relationship between the dependent variable and the independent variables: (1) ridge regression (*ridge*) and (2) lasso regression (*lasso*). Ridge and lasso regressions are a modification of linear regression. In ridge regression, the estimated coefficients shrink towards zero for the least important variables, thus reducing model complexity and multi-collinearity (Hoerl and Kennard, 1970). Lasso regression is an alternative to ridge regression in which regularization combined with a minimizing cost function eliminate variables with very small effect sizes, for which the coefficient is set to zero, hence reducing model overfitting (Tibshirani, 1996).

Four tree-based methods were included in the analysis namely classification and regression tree (*rpart*), gradient boosting (*gbm*), extreme gradient boosting (*xgbTree*), and random forest (*rf*). Classification and regression tree is a single tree-based model with simple interpretation but with unstable and poor predictive performance (Breiman et al., 2017). Random forest (Breiman, 2001), gradient boosting, and extreme gradient boosting are more complex than classification and regression trees and work based on bootstrap aggregation (bagging) technique. Gradient boosting is a technique where weak learners (data points where the prediction errors are greatest) are converted into strong learners, i.e., a higher weightage is given to data points where the prediction errors are greatest (Friedman, 2001). The loss function is a measure of how the predicted values differ from the observed values.

Table 1
Characterization of the machine learning methods used in this study based on their background algorithm and levels of complexity and interpretability. Modified by authors from James et al. (2013).

Model	Abbreviation	Type of the model	Level of complexity†	Level of interpretability‡
Classification and regression tree	rpart	Single tree	Moderate	High
Ridge regression	ridge	Regularized linear regression	Low	High
Lasso regression	lasso	Regularized linear regression	Low	High
Generalized linear model	glm	Linear regression with probability distribution and link function	Moderate	High
Gradient boosting	gbm	Sequentially constructed tree	High	Low
Extreme gradient boosting	xgbTree	Regularized gbm	High	Low
Support vector machine	svmRadial	Decision boundary	High	Low
K-nearest neighbor	kNN	Distance	Moderate	Moderate
Random forest	rf	Multiple tree	High	Low

† A model is deemed complex if it has many hyper-parameters that require fine-tuning during model fitting.
‡ A model is deemed interpretable if the estimated parameters represent the effect size of independent variables on the dependent variable.

Extreme gradient boosting works similarly to gradient boosting, except that it uses an advanced regularization to improve model generalization (i.e., to get a good performance in unseen test data; Chen and Guestrin, 2016). Random forest is also based on decision trees, but it differs from gradient boosting as it combines the results of individual decision trees using the mean or the mode to arrive at a particular prediction. Moreover, individual trees are built independently in random forest, whereas trees are built sequentially, in a forward stage manner improving shortcomings of existing weak learners, in gradient boosting (Breiman, 2001; Probst et al., 2019).

The other considered method was the K-nearest neighbor, which works based on the similarity between events in the feature space (Cover and Hart, 1967). The last method considered was support vector machine (*svmRadial*) which uses radial kernel functions to project nonlinear data into a higher dimension space to capture complex relationships between variables (Cortes and Vapnik, 1995).

2.2.2. *Variable construction and feature engineering*

Data from both growing seasons were pooled for the common variables because mean wheat yield and their distribution during both seasons was similar, resulting in a total of 6139 field-year combinations with complete data for the selected variables (Table 2). The variables were selected based on their agronomic relevance for crop production, their distribution within the surveyed fields, and their relevance for farmers and extension workers. Fourteen continuous variables were used in the analysis: six climatic variables and eight variables capturing management practices in farmers' fields. Similarly, ten categorical variables were used to describe wheat yield variability (Table 2). Variables composed of mostly unique values (i.e., more than 95% of the observations reporting the same value) were not considered in the analysis (Kuhn and Johnson, 2019). Examples of such variables included previous crop (mostly rice), use of farmyard manure (mostly no), and application rates of potassium (K) and micronutrients.

Feature engineering entails the harmonization of categorical variables to ensure a nearly similar number of observations for the different levels of each categorical variable. This was done for the following variables: tillage intensity, number of irrigations, residue level, and lodging category (Table 2). For example, there were only 158 fields where more than five irrigations were applied (the frequency of fields with more than five irrigations was small as compared to the frequency of fields with other irrigation numbers), and these were grouped together with fields reporting four irrigations. These categorical variables were one hot encoded with either zero or one using the *dummyVars ()* function in R.

Table 2
Descriptive statistics of the variables used for model comparison and further for variable screening. Categorical variables were binary coded in the analysis. Fifteen different wheat varieties not reported in this table were used in the analysis. DAS = Days after sowing.

Continuous variables	Level /Unit	Mean	Standard deviation
Wheat grain yield	kg ha^{-1}	4910	616
Sowing date	Julian days	312	6
Seeding rate	kg ha^{-1}	112.5	10.7
Total N applied	kg N ha^{-1}	161	20
Total P_2O_5 fertilizer applied	kg P_2O_5 ha^{-1}	64	11
Date of 1st urea top dress	DAS	26	3
Date of 2nd urea top dress	DAS	39	4
Days between rice harvest and wheat sowing	Days	17.9	8.4
Average maximum temperature during December and January (Vegetative stage)	°C	18.8	0.9
Average minimum temperature during December and January (Vegetative stage)	°C	7.5	0.5
Average maximum temperature during February and March (Reproductive stage)	°C	25.8	1.8
Average minimum temperature during February and March (Reproductive stage)	°C	12.3	0.9
Monthly average of cumulative radiation during December and January (Vegetative stage)	kW m^{-2}	61.1	3.6
Monthly average of cumulative radiation during February and March (reproductive stage)	kW m^{-2}	92.0	3.7
Crop duration	Days	161	8

Categorical variables	Types	% of data in each category
Tillage intensity in rice	Less than four	26.5
	Five	31.3
	Six	22.0
	Equal or more than seven	20.2
Tillage intensity in wheat	Intensive tillage	31.8
	Moderate tillage	29.2
	Zero or minimum tillage	39.0
Retention of level of rice residues	No retention	41.5
	Moderate retention	26.6
	Complete retention	31.9
Irrigation number in wheat	Equal or less than two	34.0
	Three	48.2
	Equal or more than four	17.8
Lodging event	No	60.2
	Yes	39.8
Weed infestation	Low	33.1
	Medium	66.9
Insect incidence	No	38.2
	Low	26.4
	Medium	35.2
Disease incidence	No	33.9
	Low	29.4
	Medium	36.7
Soil texture	Clay	11.2
	Clay loam	76.0
	Loam	12.8

2.2.3. *Cross-validation with data resampling*

Training and test datasets were created by partitioning the pooled data considering a 70:30 ratio, i.e., 70% of the field-year combinations were used as training dataset and the remaining 30% of the field-year combinations were used as independent test dataset. The partition of the data was done in such a way that the yield distribution was similar in both training and test datasets. This was accomplished with a stratified random sampling from the groups created using the default percentile values from *createDataPartition()* function of the *caret* R package. The machine learning methods (Table 1) were applied on the training dataset using a 10-fold cross-validation scheme with data resampling to assess the bias-variance trade-off, i.e., model overfitting or underfitting. Cross-validation iteratively creates a sub-sample from the training dataset for model fitting and evaluates the fitted model on the remaining observations of the training dataset. In this way each and every part of the data acts as both training and test dataset. This scheme was repeated three times using the *trainControl()* function of the *caret* R package (Probst et al., 2019). All the 51 variables (Table 2, with categorical variables expressed as dummy variables) were used for model development and comparison. The random search approach was used to fine-tune hyper-parameters (see Supplementary Table 1), before comparison of the different machine learning models. The *caretList()* function from the *caretEnsemble* R package (Deane-Mayer and Knowles, 2019) was used for comparing the models and for selecting the best model for further analyses. The models were compared using statistical indices (Section 2.2.4) on the cross-validated sample.

2.2.4. *Selected model to explain crop yield variability*

Three statistical indices were calculated for each 10-fold cross-validation run and used to assess the goodness-of-fit, accuracy, and precision of each model. Model accuracy refers to the ability of the model to correctly predict crop yield in the surveyed fields whereas model precision refers to the error associated with those predictions. Goodness-of-fit was assessed using the coefficient of determination (R^2), model accuracy was assessed using the root mean square error (RMSE), and model precision was assessed using the mean absolute error (MAE), which were calculated as follows:

$$R^2\ (\%) = \frac{\left(\sum_i^n (O_i - Omean)(P_i - Pmean)\right)^2}{\sum_i^n (O_i - Omean)^2 \times \sum_i^n (P_i - Pmean)^2}$$

$$RMSE\ \left(kg \Big/ ha\right) = \sqrt{\frac{\sum_{i=1}^n (O_i - P_i)^2}{n}}$$

$$MAE\ \left(kg \Big/ ha\right) = \frac{\sum_{i=1}^n |O_i - P_i|}{n}$$

where O_i is the observed yield in field-year *i*, P_i is the predicted yield by each individual model in field-year *i*, *Omean* is the mean of observed yields across all field-year combinations, and *Pmean* is the mean of

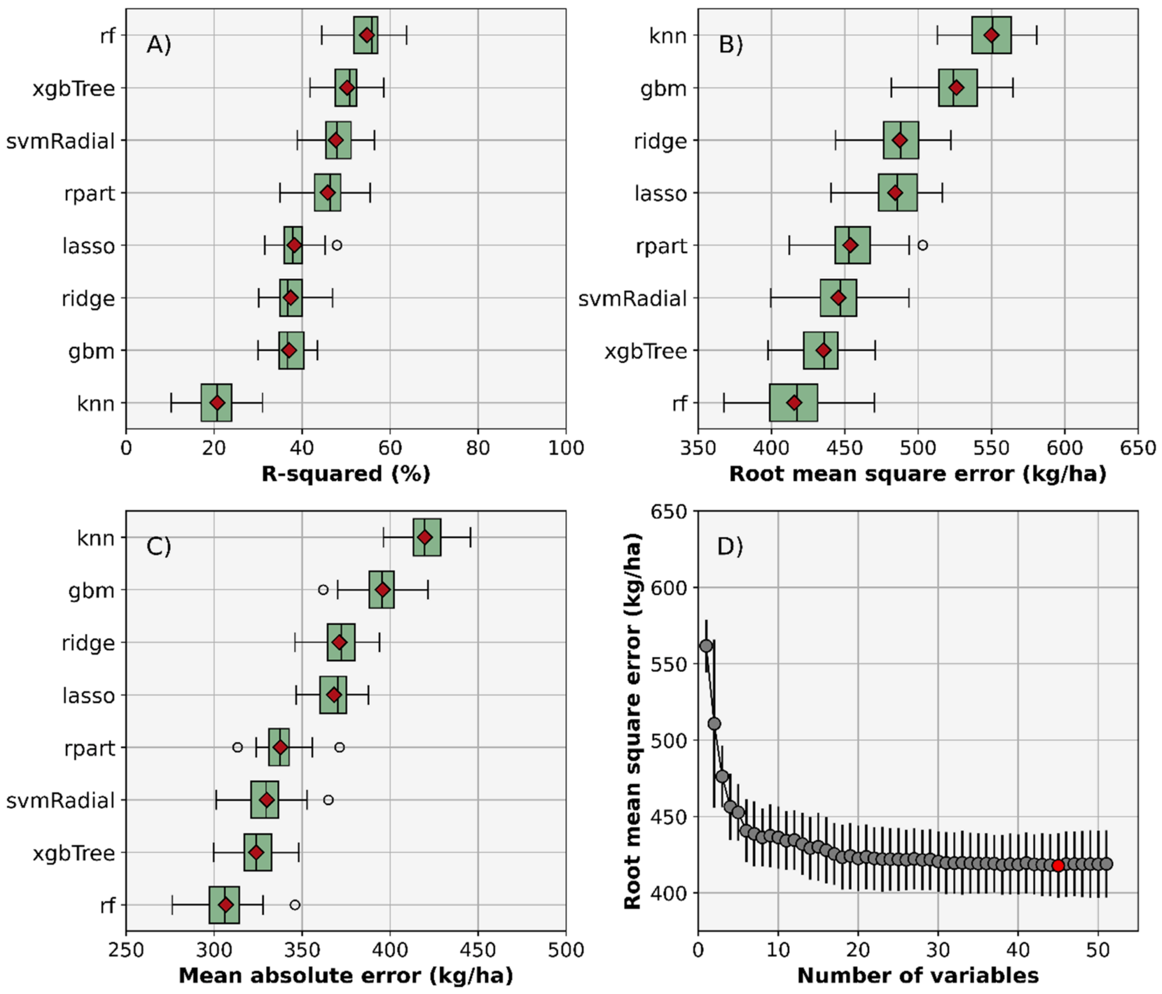

Fig. 2. Comparison of the fitted machine learning models to explain wheat yield variability in the Northwestern Indo-Gangetic Plains of India in terms of (A) coefficient of determination (R^2), (B) root mean squared error (RMSE), and (C) mean absolute error (MAE). Panel (D) shows the RMSE relative to the number of variables used for model fitting. The full name of each model, and respective abbreviation is provided in Table 2.

predicted yields by ML model across field-year combinations. R^2 values range between 0% and 100% and the greater the value the greater the variability in the data explained by the model. RMSE and MAE, both are expressed in kg ha^{-1}, are always greater than zero and the smaller the value the better the accuracy and precision of the model, respectively. The model explaining yield variability best, as indicated by the R^2, RMSE, and MAE, was selected for further analysis.

2.2.5. Variable selection and fine-tuning of selected method

The model explaining yield variability best, namely random forest (Fig. 2), was fine-tuned in two steps. The first step involved the reduction of the number of features, to avoid model overfitting (by reducing the noise) using the recursive feature elimination technique with 10-fold cross-validation. This was implemented using the *rfFuncs()* function within the *rfe* control framework of the *caret* R package (Probst et al., 2019). Recursive feature elimination implies fitting the model to the training dataset to establish the relationship between variable number and model performance based on the RMSE or other statistical indices. This was done using a backward elimination technique in which all variables were used to develop the model first, and least important variables that do not contribute in improving model accuracy were eliminated subsequently.

The second step involved fitting the model with the selected feature variables on the training dataset and evaluating it in the independent test dataset. The fitted model on reduced feature space was fine-tuned using a grid-search approach for the model specific hyper-parameters: (i) the number of variables randomly sampled as candidates at each split (*mtry* between 10 and 20), (ii) the minimum number of observations allowed in each of terminal node (*nodesize* ranging between 15 and 210 at an interval of 15) controlling the depth of the tree, and (iii) number of trees to be built *(ntree* with a value of 500, 1000 or 1500). A grid search approach was implemented for unique combinations of *nodesize* and *ntree* and for different levels of *mtry* using the *train()* and *trainControl()* functions of the *caret* R package. For each combination of *nodesize* and *ntree* values, the best *mtry* value was selected based on the greatest R^2 and lowest RMSE. After this, model performance in the training and cross-validation datasets were compared for all combinations of *nodesize* and *ntree* hyper-parameters. Hyper-parameters were chosen to reduce model overfitting which was achieved when the difference between the R^2 of training and cross-validation datasets was less than 10%.

The performance of the fine-tuned model was evaluated with Lin's concordance correlation coefficient and with a linear regression between the observed and predicted yield for the pooled data and for the training and test datasets. Data were visualized using a 1:1 plot and the Lin's concordance correlation coefficient along with R^2 was used to quantify the goodness-of-fit of the fitted models.

2.3. Model application and interpretation for wheat in the NW-IGP

Traditionally, multi-variate linear regression approaches have been used to derive insights from large and complex datasets based on the relationship between independent and dependent variables (e.g., Silva et al., 2020). Complex machine learning models can also be interpreted using model agnostic interpretation techniques related to variable importance, effect size, and interaction strength. The model fitted to the pooled dataset using the hyper-parameters from the fine-tuned model on training dataset was interpreted vis-à-vis empirical relations derived from the farmer field data for irrigated wheat in the Northwestern IGP of India.

2.3.1. Estimation of variable importance

A variable is considered "important" if shuffling its values increases or decreases the model error considerably (Fisher et al., 2019), because in this case the model relied on that variable for the prediction. The estimation of variable importance entails four steps: (1) computation of model error in the original model, (2) re-estimation of model error after reshuffling the values of a particular variable (permuted model), (3) calculation of variable importance as the ratio between the model error in the permuted model and the model error in the original model, and (4) ranking of the variables based on descending variable importance. Variable importance was estimated with the *importance()* function of the *iml* R package (Molnar et al., 2018). The two most important variables were used for detailed quantification of their effect size on wheat yield.

2.3.2. Quantification of effect sizes

Effect sizes were quantified using partial dependency plots (PDP), accumulated local effect (ALE) plots, and using quantile regressions fitted to the empirical data. PDP are commonly used to establish the relationship between important explanatory variables and the response variable (Tseng et al., 2021; Devkota and Yigezu, 2020), but they require variables to be independent and tend to over interpret the model in the range of the distribution with less or no data or if the features are correlated. Conversely, ALEs are unbiased even in the presence of a correlated feature space (Molnar et al., 2020), and hence overcome the limitations of PDP. In ALE plots, the response variable is centered at zero, i.e., the response value for each level of the independent variable in the ALE plot is the difference over to the mean outcome (here mean wheat yield), which makes model interpretation easier. PDP and ALE plots were constructed with the *iml* R package (Molnar et al., 2018). Model interpretation techniques were assessed vis-à-vis empirical relationships fitted to the farmer field data using quantile regressions for continuous variables. Non-linear quantile regressions of the form $y = a + bx + c0.99^x$ were fitted to the 90th quantile of the empirical data using the *statsmodels* library in Python.

2.3.3. Interaction strength between variables

Two measures of interaction strength were assessed namely an overall interaction strength and a two-way interaction strength. The overall interaction strength is based on Friedman's H-statistic and was used to select the most important variables interacting with other variables (Friedman and Popescu, 2008). The estimation of the overall interaction strength comprises four steps: (1) prediction of the outcome with the original model for each variable and estimation of its variance, (2) estimation of the partial dependence from the model with the variable and without the variable, (3) assessment of whether each variable is interacting with other variables by obtaining the sum of the difference of variance between the model fitted in step 1 and the partial dependence from the model with the variable and without the variable fitted in step 2, and (4) calculation of the overall interaction strength as the ratio between the differences of the variance calculated in step 3 and the variance of the model fitted in step 1. The overall interaction strength ranges between zero and one, and interactions are deemed important if the overall interaction strength is greater than 0.1-0.15.

The two-way interaction strength was further studied for the top three management variables with greatest overall interaction strength. Two-way interactions were determined with the following steps: (1) estimation of the partial dependence for two variables, *i* and *j*, say *pd(ij)*; (2) computation of the partial dependence function for each variable, i.e., *pd(i)* and *pd(j)*; (3) computation of the two-way interaction strength as the ratio of the difference in variance between the interaction (*pd(ij)*) and the individual partial dependencies (*pd(i)* and *pd(j)*) and the variance of the interaction partial dependence. The overall interaction strength and the two-way interaction strength were estimated using the *iml* R package (Molnar et al., 2018).

The three most important two-way interactions for each of the three management variables with greatest overall interaction strength were studied with descriptive scatterplots for continuous variables or faceted boxplots for categorical variables. Linear regressions were fitted to different levels of the independent variable to determine the two-way interaction empirically.

3. Results

3.1. Important machine learning models to explain wheat yield variability

The performance of eight machine learning models to explain wheat yield variability was evaluated using R^2, RMSE, and MAE values after model fine-tuning using random search approach (Fig. 2A-C). The *kNN* method performed worst, with R^2 ranging between 0.10 and 0.30 (mean equal to 0.20), RMSE ranging between 513 and 580 kg ha^{-1} (550 kg ha^{-1}), and MAE ranging between 396 and 445 kg ha^{-1} (420 kg ha^{-1}). The two regression-based models (*ridge* and *lasso*) and gradient boosting model (*gbm*) performed equally well and had slightly better performance than *kNN* with mean R^2 of 0.38, , mean RMSE of 499 kg ha^{-1}, and mean MAE of 378 kg ha^{-1}across the three models The single tree-based classification and regression tree (*rpart*) performed better than the linear regression-based models. The multi-tree-based bagged models (*xbgTree*) and *svmRadial* models performed better than linear regression, *rpart*, and *gbm* models with a mean R^2 of 0.47 and 0.50, respectively. Finally, the random forest (*rf*) model performed best considering all the statistical indices, with the R^2 ranging between 0.44 and 0.64 (mean equal to 0.55), the RMSE ranging between 367 and 470 kg ha^{-1} (424 kg ha^{-1}), and the MAE ranging between 276 and 345 kg ha^{-1}. The random forest method was thus selected for further analysis.

The random forest model was fine-tuned by reducing the number of variables not contributing to increasing the model performance (Fig. 2D) and by adjusting the hyper-parameters *mtry, ntree*, and *nodesize* to control the bias-variance trade-off (Fig. 3). The RMSE decreased with increasing number of variables used for model fitting and stabilized at approximately 45 variables, beyond which the RMSE slightly increased (Fig. 2D). The RMSE of the model with all 51 variables was 419 kg ha^{-1}, which was comparable to the RMSE of the model with 45 variables, 417 kg ha^{-1}. The R^2 of the training and cross-validation datasets was computed for each combination of the hyper-parameters *nodesize* and *ntree* against the best *mtry* value. The R^2 in the training dataset ranged between 0.81 in the model with *nodesize* 15, *ntree* 1500, and *mtry* 10, to 0.56 in the model with *nodesize* 210, *ntree* 1500, and *mtry* 17 (Fig. 3). The R^2 in the cross-validation dataset was fairly constant across different combinations of hyper-parameter values, varying between 0.52 and 0.55 (Fig. 3). The R^2 of 0.81 in the training dataset and 0.54 in cross-validation dataset clearly indicates model overfitting. The combination of hyper-parameters where the difference of R^2 between training (0.62) and cross-validation dataset (0.53) was less than 10% were *nodesize* 105, *ntree* 500, and *mtry* 13. These hyper-parameters combinations were thus used in further analyses. The list of important variables and hence the model interpretation did not change for *nodesize* values of 90, 105, and 120 (data not shown).

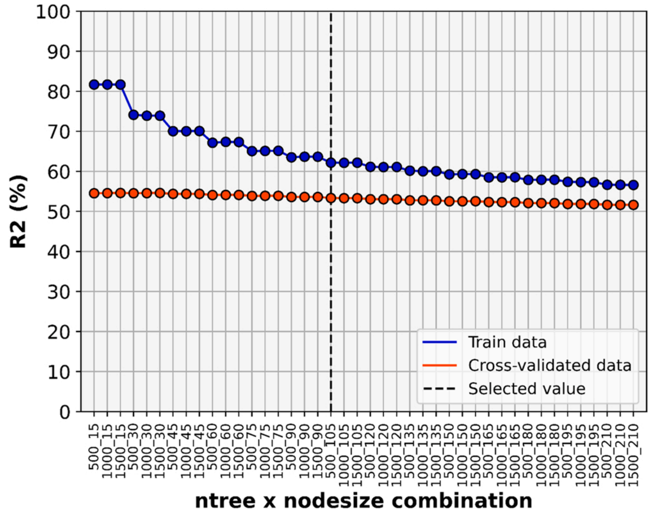

Fig. 3. Fine-tuning of the selected random forest model for the hyper-parameters' combinations of *ntrees* and *nodesize*. The hyper-parameter *ntrees* refers to the number of trees to be grown and *nodesize* refers to the number of observations in the terminal nodes. The selected values for the hyper-parameters' combination are shown by the dashed vertical line.

The final random forest model performed well in both training and test datasets and in the pooled dataset (Fig. 4). Lin's concordance correlation coefficient was 0.72 and 0.67 for the training and test dataset, respectively (Fig. 4A), and 0.74 for the pooled dataset (Fig. 4B). In the test dataset the RMSE was 414 kg ha^{-1} and the MAE was 311 kg ha^{-1}. Model performance on the pooled data was similar to that on the training dataset (R^2 of ca. 0.62) indicating the training dataset was a true sub-sample of the whole data. Model performance in the test dataset was slightly lower with a R^2 of 0.56.

3.2. Important variables affecting wheat yield and their effect sizes

Biophysical variables had a stronger impact on wheat yield variability than management variables (Fig. 5). The monthly average of cumulative solar radiation during the February and March (the period corresponding to reproductive stage of wheat) and N application rate were the most important biophysical and management variables governing the wheat yield variability, respectively (Fig. 5). Minimum and maximum temperatures during the December-January and February-March periods were also identified as important biophysical variables. Other important management variables were residue retention, seeding rate, and disease severity (whether medium or no and low). Yet, seeding rate and disease severity had a much lower contribution towards reducing model error than N application rate or residue retention (Fig. 5).

The effect size of the most important variables, i.e., cumulative solar radiation during the February and March and N application rate, on wheat yield were compared to the mean wheat yield across the surveyed fields using the Accumulated Local Effect (ALE) plot (Fig. 6). The monthly average of cumulative solar radiation during February and March varied between 83 and 98 kW m^{-2} (Fig. 6A). The ALE plot of the cumulative solar radiation during February and March indicated a maximum yield benefit of 372 kg ha^{-1} across the range observed for solar radiation. Although there was a general increase in wheat yield with increased solar radiation during February and March, two local minima were identified at cumulative solar radiation of 87 and 95 kW m^{-2} (Fig. 6A), both observed in the districts of Ambala (data not shown). The pattern described by the partial dependency plot was similar to that described by the ALE plot between solar radiation and wheat yield (Fig. 6B). The quantile regression showed a monotonous trend to solar radiation with a concave response pattern (Fig. 6B).

The ALE plot of N application rate *vs* wheat yield indicated a net yield benefit of 272 kg ha^{-1} in the range of observed N application rates (between 90 and 230 kg N ha^{-1}; Fig. 6C). The mean wheat yield of 4.9 t ha^{-1} was associated with N application rates between 140 and 170 kg N ha^{-1}, whereas increasing N application rate up to 230 kg N ha^{-1} improved wheat yield by 78 kg ha^{-1} only. The results of the ALE plot were confirmed by those of the PDP, which indicated a net yield benefit of 180 kg ha^{-1} across the N application range observed in the data (Fig. 6D). The quantile regression fitted to the 90th percentile of the data also confirmed the small increment in wheat yield to N application in the range of reported N application rates (Fig. 6C), which were likely near optimum for most fields. Yet, wheat yield response to N application was greater for N application rates below 130 kg N ha^{-1} (but very few farmers had a N application rate below 130 kg N ha^{-1}). Further, the quantile regression predicted wheat yield without N applied around 3.3 t ha^{-1} (Fig. 6C).

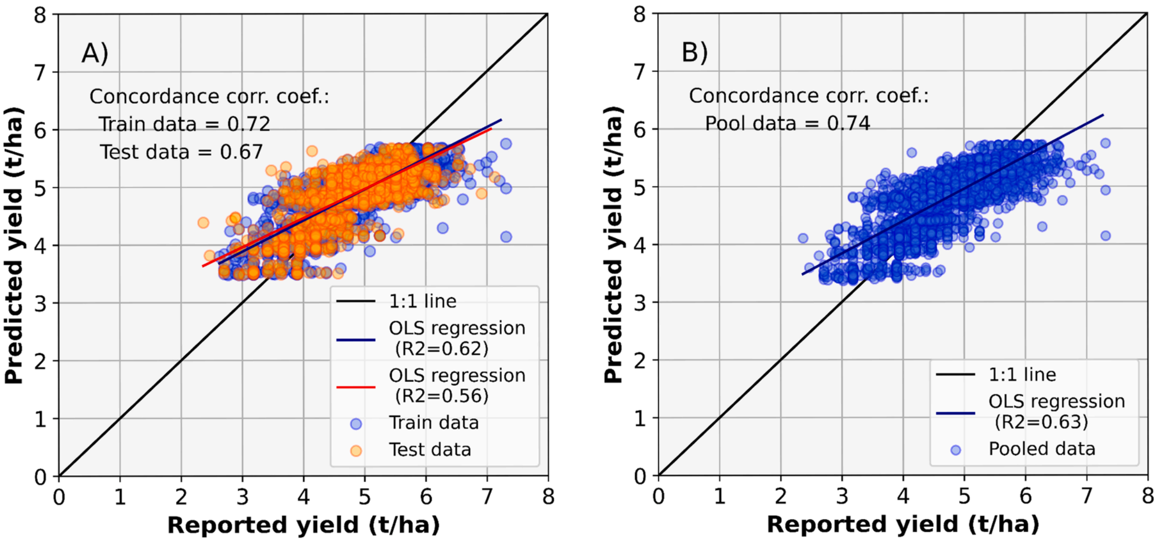

Fig. 4. Model evaluation based on (A) the relationship between observed and predicted wheat yield on the train and test datasets from the model fitted to the train dataset and predicted on test dataset, and (B) the relationship between observed and predicted wheat yield fitted on the pooled data. Solid lines show linear regressions fitted to the data, with the respective coefficient of determination (R^2) shown in the legend of each panel.

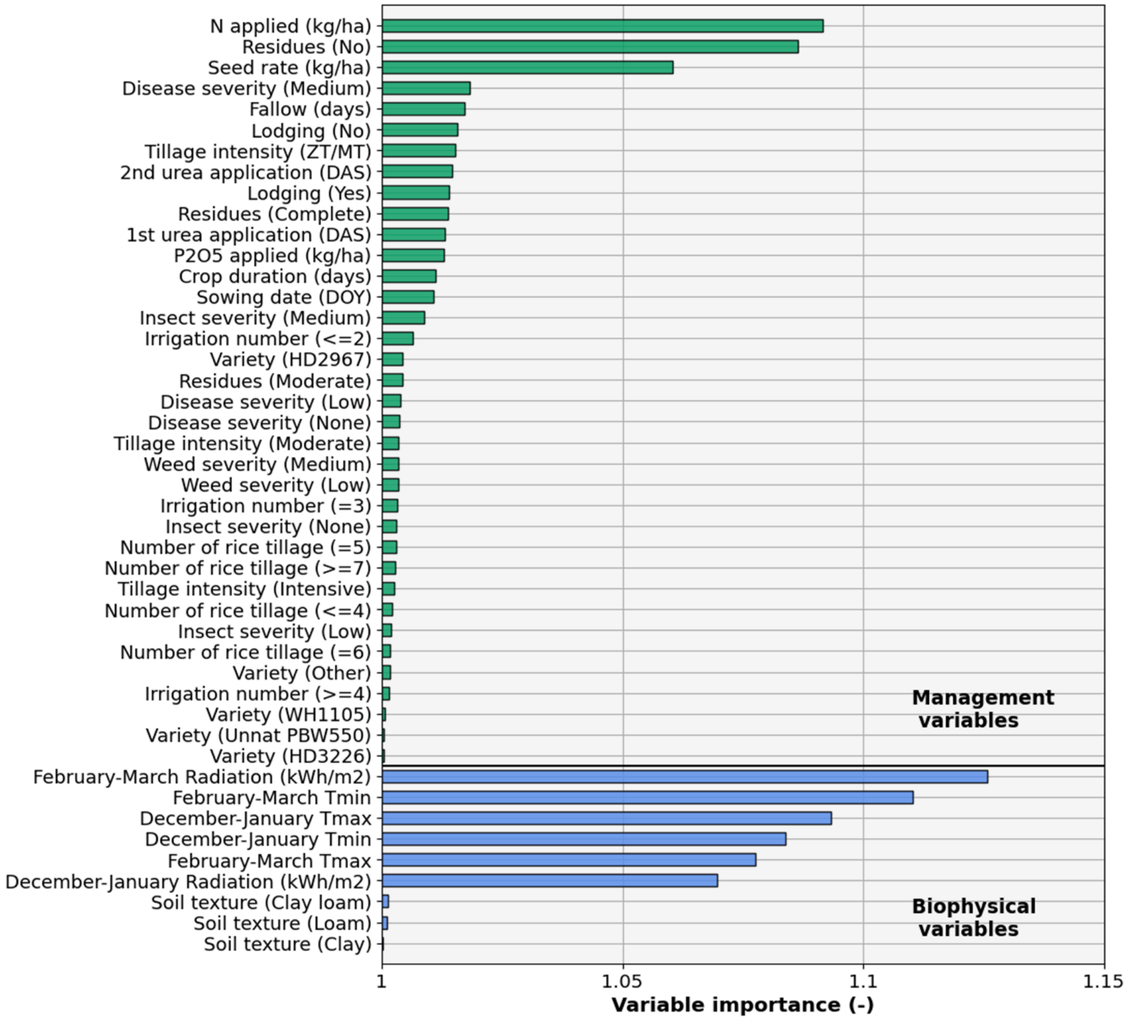

Fig. 5. Variable importance for management and biophysical variables governing wheat yield variability in the Northwestern Indo-Gangetic Plains of India. Abbreviations: DAS = days after sowing, DOY = day of the year, ZT = zero tillage, MT = minimum tillage.

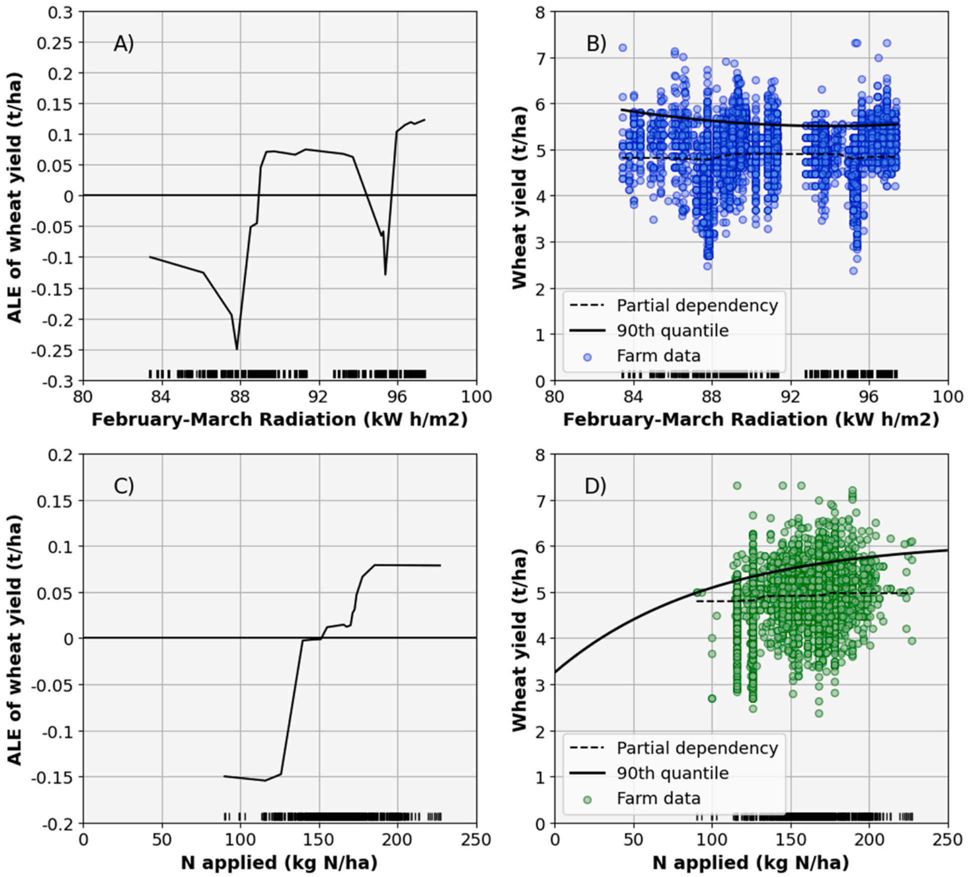

Fig. 6. Effect size of the two most important variables explaining the wheat yield variability in the Northwestern Indo-Gangetic Plains of India: (A) accumulated local effect (ALE) plot of monthly averages of cumulative solar radiation during February and March, (B) partial dependency plot (PDP) and quantile regression between monthly average of cumulative solar radiation during February and March and wheat yield, (C) ALE plot of N application rate, (D) wheat yield as a function of N application rate. Solid lines in (B) and (D) show quantile regressions fitted to the 90th quantile and dashed lines show the respective partial dependency plot.

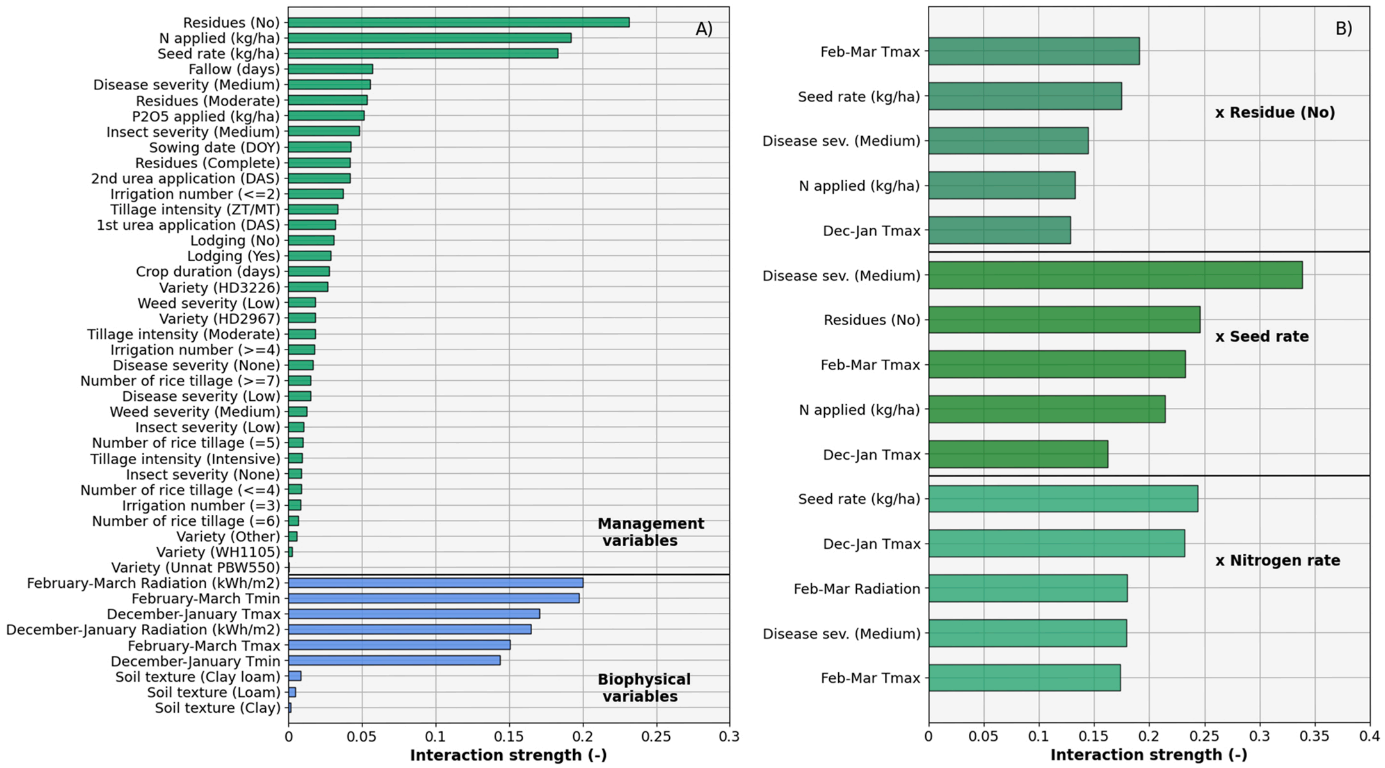

Fig. 7. Analysis of interaction terms explaining wheat yield variability in the Northwestern Indo-Gangetic Plains of India: (A) overall interaction strength for management and biophysical variables, and (B) two-way interactions for the four management variables with greatest overall interaction strength.

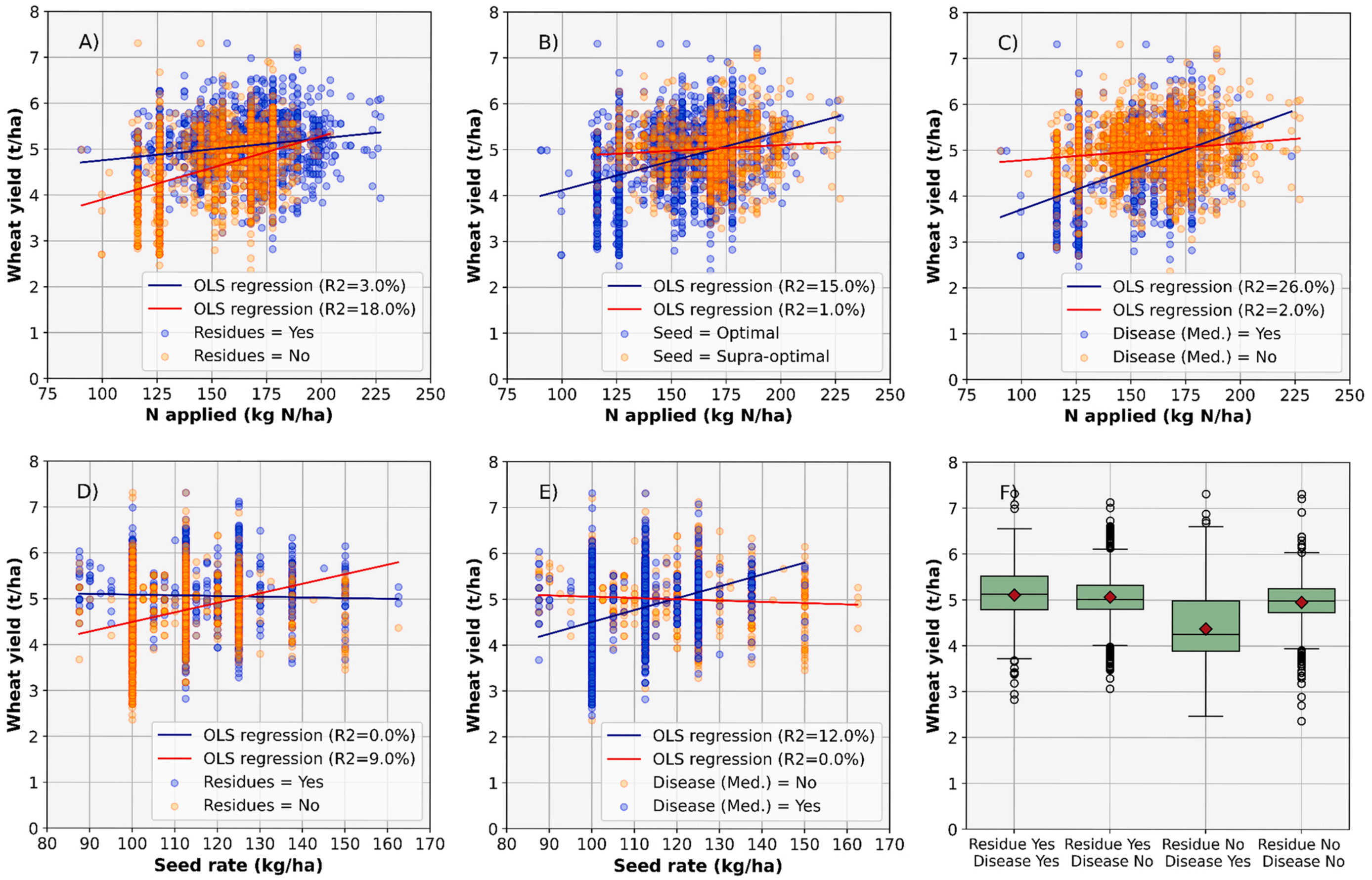

Fig. 8. Empirical analysis of important management × management interactions explaining wheat yield variability in the Northwestern Indo-Gangetic Plains of India. Solid lines in (A) – (F) show linear regressions fitted to the data and the respective coefficient of determination (R^2) is presented in the legend of each panel.

3.3. *Interactive effect of different management practices on wheat yield*

The greatest overall interaction strength among management variables was observed for residue management, N application and seeding rates, where the interaction strength was greater than 0.1 (Fig. 7A). For these variables, nearly 17–27% of the wheat yield variability was described by their interaction with other variables. Different management by management interactions were identified including: (1) residue management with seeding rate, disease incidence, and N application rate, (2) N application rate with seeding rate and disease incidence, and (3) seeding rate with disease severity (Fig. 7B). Alike the direct effect of biophysical variables, they also affected the wheat yield through interacting with other management variables. Regarding biophysical variables, the overall interaction strength was greatest for average minimum temperature and monthly cumulative solar radiation during February and March (Fig. 7A).

Wheat yield response to N applied was greater in fields where residues were removed than in fields where residues were retained (Fig. 8A). The slope of the linear regressions was equal to 4.8 and 13.8 kg yield kg^{-1} N applied for fields where residues were retained and removed, respectively. There was also a positive relationship between N applied and wheat yield when the seeding rate was close to the recommended rate of 100 kg ha^{-1} (Fig. 8B), whereas no wheat yield response to N applied was observed for fields with seeding rates beyond optimal levels (Fig. 8B). Disease severity also affected wheat yield response to N applied as fields with lower N application rates incurred greater yield losses under moderate disease incidence than fields where no or low disease infestation was reported (Fig. 8C).

Wheat yield response to seeding rate was greater when residues were removed (slope of linear regression equal to 52.3 kg yield kg^{-1} seed) than when residues were retained in the field (−3.5 kg yield kg^{-1} seed; Fig. 8D). Yet, residue retention translated into greater wheat yield than residue removal for seeding rates close to the recommended rate of 100 kg ha^{-1} and into lower wheat yield at seeding rates beyond 125 kg ha^{-1}. Wheat yield response to seeding rate was greater in fields reporting incidence of diseases (Fig. 8E) than in fields not reporting this stress. Moreover, moderate disease infestation led to larger yield losses in fields where seeding rate was close to the recommended seeding rate. The interaction effect of disease severity and residue retention revealed that, under low disease infestation, both retention and removal of residues resulted in a similar mean wheat yield, whereas residue retention resulted into 750 kg ha^{-1} higher mean wheat yield than residue removed and disease affected fields (Fig. 8F).

4. Discussion

4.1. *Comparison of machine learning models*

Very few studies evaluated the performance of a wide range of machine learning methods (Table 1) to explain crop yield variability in agronomic data (Mourtzinis et al., 2019; Paudel et al., 2021). Regression-based methods have been used traditionally in similar applications as they are easy to interpret (Silva et al., 2020; Basso and Liu, 2019), yet these models were least effective in explaining yield variability in this study (Fig. 3A–C). This was probably because of the presence of non-linear and complex relationships in the dataset, which regression-based methods cannot account for. More complex methods, such as random forest, extreme gradient boosting, or support vector machines, performed better than regression-based methods, with random forest performing best in explaining wheat yield among all models tested (Fig. 3A–C).

Tree-based methods are well-known to outperform regression-based methods in different domains (Breiman, 2001; Jeong et al., 2016) as they capture non-linear relationships and interactions between variables. Jeong et al. (2016) also observed random forest performing better than regression-based methods for predicting crop yield. Nigam et al. (2019) too concluded that random forest was the best method to predict the rice yield in India. The ALE plots and PDP (Fig. 6) clearly show the existence of non-linear relationships between wheat yield and biophysical and management variables. Moreover, the analysis of interactions also confirms the presence of complex relationships between biophysical and management factors to explain wheat yield variability (Fig. 7), which were well-captured with tree-based methods. It is thus recommended to use random forest, or variants of random forests like conditional random forest, in future agronomic studies aiming to explain crop yield variability using first order and interaction effects, as done in some other recent applications (Tseng et al., 2021; Devkota and Yigezu et al., 2020; Paudel et al., 2021; Garnaik et al., 2022).

Proper use of machine learning methods, including random forest, requires fine-tuning of model hyper-parameters (Fig. 3) to avoid the bias-variance trade-off (James et al., 2013). This is important to avoid overfitting the models on the training dataset at the expense of model performance in the test dataset. It is not recommended to interpret overfitted models as such models may capture noise and errors inherent to each dataset. This step is often ignored, or not reported, in agronomic studies which leads to unreproducible workflows and potentially biased results and conclusions.

4.2. *Determinants of wheat yield in the Northwestern IGP of India*

Wheat yield variability in the Northwestern IGP of India was mostly explained by biophysical variables than by management variables (Fig. 5), as expected in high productivity cropping system operating close to yield potential. However, the effect size of the important biophysical and management variables was generally small due to high input use for most farms (Table 2 and Fig. 6). Such small effect size of biophysical and management variables on crop yield were also observed for rice crops in the same region (Nayak et al., 2022), and for high-yielding wheat crops in other regions (Silva et al., 2020; Lollato et al., 2019). As expected, the effect size estimated with model interpretation techniques had the same sign but slightly smaller magnitude than that observed in descriptive data analysis using quantile regressions for the range of input use observed in the data (Fig. 6). This means that first-order responses observed in descriptive data analyses are modulated by non-linear relationships, errors, and interactions between different factors. The PDP and ALE plots show the repsonse for predicted yield and input use from the fitted ML model, where as the descriptive quantile regression shows the response present in the raw data which has embeded noise inside it.

The most important variables explaining wheat yield variability were monthly average of cumulative solar radiation received during February and March and N application rate (Fig. 5). The average of cumulative monthly solar radiation during February and March had a positive effect on wheat yield, which might be due to better availability of photosynthates for proper grain filling and increasing the test weight of the wheat grain during the grain filling period (Villegas et al., 2016). Alike earlier findings, we also observed important effects of minimum and maximum temperature on wheat yield (Fischer et al., 2022), but the narrow range of temperature observed in farmers' fields do not allow to derive conclusive results of their effect on wheat yield in the Northwestern IGP. Wheat yield response to N is well-document for modern wheat cultivars in the Northwestern IGP (Kaur and Ram, 2017), as those reported in this study (Table 2). Park et al. (2018) also reported a positive wheat yield response to N applied in the region using machine learning methods. Yet, interactions between N and residue retention, seeding rates, and disease incidence must be considered, along with the first-order effects of N on wheat yield (Fig. 6D). Wheat yield response to N applied was also affected by temperature during key periods of the growing season (Fig. 7B), which is in line with recent findings by Sadras et al. (2022).

Crop residue retention has been recommended for rice-wheat cropping systems in the Northwestern IGP due to the beneficial effect of residue retention on soil health and crop productivity (Jat et al., 2019; Parihar et al., 2018, 2019; Aryal et al., 2015; Sapkota et al., 2019) and potential to reduce environmental pollution by avoiding residue burning (Shyamsundar et al., 2019). Although we observed a small direct effect of residue retention on wheat yield, residue management had the largest interaction strength and interacted with seeding rate, disease infestation, and N applied (Fig. 7). Retention of crop residues improves soil structure, which in turn regulates soil moisture, and improves crop establishment (Hobbs et al., 2008), facilitating an earlier germination and initial vigorous crop growth. Bastos et al. (2020) also reported a small wheat yield response to seeding rate in a high yielding environment, which can be compared to the situation of greater wheat yield under residue retained environment. Microbial diversity increases with residue retention, which in turn improves soil fertility and crop yield (Choudhary et al., 2018). The increase in antagonistic microbial population with residue retention is also known to suppress disease infestation (Bailey, 1996). Soil health boosts early crop growth and results into greater biomass accumulation (Nayak et al., 2022a; Gathala et al., 2013), and thus help in tolerating biotic and abiotic stresses, like disease infestation. This explains why wheat crops in fields with residue retention had greater yield even under disease infestation (Fig. 8E –F).

The interaction between residue retention and N application depends on crop, soil, and longevity of residue retention (Linquist et al., 2006; Thuy et al., 2008). Often the benefits from organic forms of N are observed at low rates of inorganic N application. Alike our results (Fig. 8), Singh and Gupta (2009) observed a positive effect of residue retention on wheat yield, when the amount of N applied was about 120 kg ha^{-1}, whereas higher N application of 150 kg ha^{-1} was beneficial under residue removed or burnt conditions. Immobilization and volatilization losses of N fertilizer under residue retained conditions at high N rates might have also caused a smaller yield response to N at high N rates under residue retention conditions (Tisdale and Nelson, 1966).

Seeding rate interacted with disease incidence (Fig. 8E and F). Fields with seeding rates close to the recommended seeding rate of 100–110 kg ha^{-1} experienced larger yield losses from disease infestation than fields with seeding rates above the recommended rate, which might be linked to stand density and other factors not captured in this study. Similarly, wheat yield response to N applied was higher under optimal seeding rates only, whereas with seeding rates greater than the recommendation there might be interplant competition resulting in a smaller yield response to N applied. The role of biotic factors in reducing wheat yield response to other management and biophysical factors is another outcome of our analysis (Fig. 8) and confirms the importance of pest, disease, and weed management in high-yielding cropping systems (Nayak et al., 2022; Shah et al., 2021; Lawes et al., 2021).

4.3. *Prospects for interpretable machine learning in agronomic studies*

In this study, machine learning methods were useful to explain crop yield variability, to identify the most important variables explaining it, and to quantify the optimum level of such variables (Fig. 6). The results revealed a small yield response to management and biophysical variables with the effect sizes estimated from machine learning methods being slightly smaller than those estimated with quantile regressions (Fig. 6), which do not account for important interactions in the data and shows response from the raw data which has some noise embedded with it (Figs. 7 and 8). The small effect size of management and biophysical variables are a feature of high-yielding cropping systems, where resource use is generally high (de Wit, 1992; Silva et al., 2021). Improved crop management that tackles the twin goals of increased productivity with reduced environmental externalities in input-intensive and highly productive cropping systems will require fine-tuning management practices and better matching that with biophysical conditions during the growing season (Fig. 8; Shah et al., 2021; Silva et al., 2017).

Beyond explaining yield variability and identifying its key drivers, as done in this study, machine learning methods could also be used to predict crop yield in space and time (Van Klompenburg et al., 2020). Understanding model portability is crucial to assess the applicability of machine learning methods over multiple growing seasons and for cropping systems in different stages of intensification. The performance of machine learning models are highly data specific, i.e., it depends on different combinations of input data, total variability present in the dataset, and number of explanatory variables included. Therefore, such type of studies must be conducted across multiple environments. 'Big data' and associated data-driven analytics have been proposed as a new avenue for agronomic research in the coming decades (Vanlauwe, 2020). Such methods offer an expedient way to generate insights that can be used to guide improved crop management recommendations based on data derived from on-farm observations collected across large scales, rather than through manipulative experiments implemented in a handful of locations. However, it remains unclear how to harness these approaches and accelerate their use to aid in more robust recommendations that can inform decision making by farmers and extension services. Given the complexity of model outputs, research is therefore needed to develop systems for enhanced data interpretation and to develop guidelines for how to responsibly generate actionable recommendations from model outputs. Model stacking and ensemble techniques could also be useful for yield prediction and remote sensing-based indices can potentially help improving model performance, but interpretability of remote sensing data to make more generalizable, yet actionable, recommendations also needs further research.

The application of machine learning methods should not be limited to the identification of the key influencingy variables on yield patterns, but should also seek to quantify the effect sizes and their interactions in driving yield variability across farmers' fields. Machine learning interpretation techniques are suitable to screen many interactions and identify those with the largest effect on crop yield (Figs. 7 and 8). This is often hard to do with parametric statistical methods, which require a certain number of degrees of freedom for reliable model fitting. As a next step, the optimum range of such variables should be quantified specific to the production environment, but also to consider the influence of a range of socioeconomic variables and how they may affect crop management and in turn, yield patterns. Such analyses could be further complemented with the analyses of trade-offs between crop yield and other indicators of sustainability, namely profitability, resource-use efficiency, and environmental footprints, to explore new pathways to more productive, profitable, and environmentally sound cropping systems.

5. Conclusion

A large database characterizing wheat production in the Northwestern Indo-Gangetic Plains of India was used to explain wheat yield variability in the region using machine learning methods. Machine learning models with different levels of complexity and interpretability, including regression-, tree-, and decision boundary-based methods, were compared for their ability to explain wheat yield variability. Random forest outperformed the eight other tested models, with a mean R^2 of 52% and a RMSE of 430 kg ha^{-1} in the test dataset. Biophysical variables had a stronger effect on wheat yield variability than management variables. Cumulative radiation over the months of February and March was the most important biophysical variable explaining wheat yield variability with an effect size of 373 kg ha^{-1}. N application rate was the second most important management variable explaining wheat yield variability in the region. Interactions between biophysical and management variables explained up to 25% of wheat yield variability. Wheat yield response to the most important variables and their interaction was generally small, which is a feature of high-yielding cropping systems. These small effect sizes were confirmed by accumulated local effect plots, partial dependency plots, and analysis of the primary data with quantile regressions. Yet, increases in effect size are to be expected in cropping systems with greater variability in crop yield and input use than observed in this study, which should be explored with the application of machine learning methods to other cropping systems. Further research is also needed to better understand trade-offs between crop yield, profitability, and a range of sustainability indicators, in order to identify pathways for more productive, remunerative, and environmentally sound wheat production in Northwestern Indo-Gangetic Plains of India.

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.

Data availability

Data will be made available on request.

Acknowledgments

HSN sincerely acknowledges the Indian Council of Agricultural Research (ICAR)-Indian Agricultural Research Institute (IARI) and the International Maize and Wheat Improvement Center (CIMMYT) for support of PhD research work. The research was carried out by CIMMYT, ICAR-Central Soil Salinity Research Institute (CSSRI), ICAR-IARI, Borlaug Institute for South Asia-CIMMYT with the support of CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS; https://ccafs.cgiar.org). CCAFS' work was funded through CGIAR Fund Donors and through bilateral funding agreements (https://ccafs.cgiar.org/donors). Portions of this work were also supported by the Cereal Systems Initiative for South Asia (https://csisa.org) funded by the USAID and Bill and Melinda Gates Foundation (BMGF). We acknowledge Dr Zia Ahmed, Dr Raj Singh, Dr V K Singh, Dr Rajkumar Dhakar, Dr S L Jat, Dr D K Sharma, Dr B N Mandal, Gokul Paudel, Asif Faisal, Khaled Hossain, Saral Karki, Sanjay Pothireddy, and Noufa Cheerakkollil Konath for technical assistance and Manish Kumar, Deepak Bejarniya, Dr Kajod Mal Choudhary, Yogesh Kumar, Love Singh, Sushil Kumar, and Kailash Kalvania for help during data collection. This manuscript should not be taken as endorsement by CCAFS, USAID or the US government, or BMGF, and shall not be used for advertising purposes.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fcr.2022.108640.

References

Aryal, J.P., Sapkota, T.B., Jat, M.L., Bishnoi, D.K., 2015. On-farm economic and environmental impact of zero-tillage wheat: a case of North-West India. Exp. Agric. 51 (1), 1–16.

Bailey, K.L., 1996. Diseases under conservation tillage systems. Can. J. Plant Sci. 76 (4), 635–639.

Basso, B., Liu, L., 2019. Seasonal crop yield forecast: Methods, alications, and accuracies. Adv. Agron. 154, 201–255.

Bastos, L.M., Carciochi, W., Lollato, R.P., Jaenisch, B.R., Rezende, C.R., Schwalbert, R., Vara Prasad, P.V., Zhang, G., Fritz, A.K., Foster, C., Wright, Y., 2020. Winter wheat yield response to plant density as a function of yield environment and tillering potential: A review and field studies. Front. Plant Sci. 11, 54.

Bhatt, R., Singh, P., Hossain, A., Timsina, J., 2021. Rice–wheat system in the northwest Indo-Gangetic plains of South Asia: issues and technological interventions for increasing productivity and sustainability. Paddy Water Environ. 1–21.

Breiman, L., 2001, Random forests. Mach. Learn. 45, 5–32. 〈https://link.springer.com/content/pdf/10.1023%2FA%3A1010933404324.pdf〉.

Breiman, L., 2001a. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Stat. Sci. 16 (3), 199–231.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 2017. Classification and Regression Trees, 1st ed..,. Routledge. https://doi.org/10.1201/9781315139470.

Cao, J., Zhang, Z., Luo, Y., Zhang, L., Zhang, J., Li, Z., Tao, F., 2021. Wheat yield predictions at a county and field scale with deep learning, machine learning, and google earth engine. Eur. J. Agron. 123, 126204.

Chauhan, B.S., Mahajan, G., Sardana, V., Timsina, J., Jat, M.L., 2012. Productivity and sustainability of the rice–wheat croing system in the Indo-Gangetic Plains of the Indian subcontinent: problems, opportunities, and strategies. Adv. Agron. 117, 315–369.

Choudhary, M., Datta, A., Jat, H.S., Yadav, A.K., Gathala, M.K., Sapkota, T.B., Das, A.K., Sharma, P.C., Jat, M.L., Singh, R., Ladha, J.K., 2018. Changes in soil biology under conservation agriculture based sustainable intensification of cereal systems in Indo-Gangetic Plains. Geoderma 313, 193–204.

Chen, T., Guestrin, C., 2016, Xgboost: A scalable tree boosting system. In Proceedings of the 2nd international conference on knowledge discovery and data mining, 785–794.

Correndo, A.A., Hefley, T.J., Holzworth, D.P., Ciampitti, I.A., 2021. Revisiting linear regression to test agreement in continuous predicted-observed datasets. Agric. Syst. 192, 103194.

Cortes, C., Vapnik, V., 1995. Suort-vector networks. Mach. Learn. 20 (3), 273–297.

Cover, T., Hart, P., 1967. Nearest neighbour pattern classification. IEEE Trans. Inf. Theory 13 (1), 21–27.

de Mauro, A., Greco, M., Grimaldi, M., 2016. A formal definition of big data based on its essential features. Libr. Rev. 65, 122–135.

Di Mauro, G., Cipriotti, P.A., Gallo, S., Rotundo, J.L., 2018. Environmental and management variables explain soybean yield gap variability in Central Argentina. Eur. J. Agron. 99, 186–194.

de Wit, C.D., 1992. Resource use efficiency in agriculture. Agric. Syst. 40 (1–3), 125–151.

Deane-Mayer, Z.A., Knowles, J.E., 2019. caretEnsemble: Ensembles of Caret Models. R. Package Version 2 (1), 35.

Devkota, M., Yigezu, Y.A., 2020. Explaining yield and gross margin gaps for sustainable intensification of the wheat-based systems in a Mediterranean climate. Agric. Syst. 185, 102946.

Fischer, T., Honsdorf, N., Lilley, J., Mondal, S., Monasterio, I.O., Verhulst, N., 2022. Increase in irrigated wheat yield in north-west Mexico from 1960 to 2019: Unravelling the negative relationship to minimum temperature. Field Crops Res. 275, 108331.

Fisher, A., Rudin, C., Dominici, F., 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. J. Mach. Learn. Res. 20 (177), 1–81.

Friedman, J.H., 2001. Greedy function aroximation: a gradient boosting machine. Ann. Stat. 1189–1232.

Friedman, J.H., Popescu, B.E., 2008. Predictive learning via rule ensembles. Ann. Alied Stat. 2 (3), 916–954.

Garnaik, S., Samant, P.K., Mandal, M., Mohanty, T.R., Dwibedi, S.K., Patra, R.K., Mohapatra, K.K., Wanjari, R.H., Sethi, D., Sena, D.R., Sapkota, T.B., 2022. Untangling the effect of soil quality on rice productivity under a 16-years long-term fertilizer experiment using conditional random forest. Comput. Electron. Agric. 197, 106965.

Gathala, M.K., Ladha, J.K., Saharawat, Y.S., Kumar, V., Kumar, V., Sharma, P.K., 2011. Effect of tillage and crop establishment methods on physical properties of a medium-textured soil under a seven-year rice− wheat rotation. Soil Sci. Soc. Am. J. 75 (5), 1851–1862.

Gathala, M.K., Kumar, V., Sharma, P.C., Saharawat, Y.S., Jat, H.S., Singh, M., Kumar, A., Jat, M.L., Humphreys, E., Sharma, D.K., Sharma, S., Ladha, J.K., 2013. Optimizing intensive cereal-based cropping systems addressing current and future drivers of agricultural change in the northwestern Indo-Gangetic Plains of India. Agric., Ecosyst. Environ. 177, 85–97. https://doi.org/10.1016/j.agee.2013.06.002.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12 (2), e0169748.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

Hobbs, P.R., Sayre, K., Gupta, R., 2008. The role of conservation agriculture in sustainable agriculture. Philos. Trans. R. Soc. B: Biol. Sci. 363 (1491), 543–555.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.

Jaenisch, B.R., Munaro, L.B., Bastos, L.M., Moraes, M., Lin, X., Lollato, R.P., 2021. On-farm data-rich analysis explains yield and quantifies yield gaps of winter wheat in the US central Great Plains. Field Crops Res. 272, 108287.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, Vol. 112. Springer, New York.

Jat, S.L., Parihar, C.M., Singh, A.K., Nayak, H.S., Meena, B.R., Kumar, B., Parihar, M.D., Jat, M.L., 2019. Differential response from nitrogen sources with and without residue management under conservation agriculture on crop yields, water-use and economics in maize-based rotations. Field Crops Res. 236, 96–110.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.M., Gerber, J.S., Reddy, V.R., Kim, S.H., 2016. Random forests for global and regional crop yield predictions. PLoS One 11 (6), e0156571.

Kaur, H., Ram, H., 2017. Nitrogen management of wheat cultivars for higher productivity-A review. J. Alied Nat. Sci. 9 (1), 133–143.

Khaki, S., Wang, L., 2019. Crop yield prediction using deep neural networks. Front. Plant Sci. 10, 621. https://doi.org/10.3389/fpls.2019.00621.

Krupnik, T.J., Ahmed, Z.U., Timsina, J., Yasmin, S., Hossain, F., Al Mamun, A., Mridha, A.I., McDonald, A.J., 2015. Untangling crop management and environmental influences on wheat yield variability in Bangladesh: an application of non-parametric approaches. Agric. Syst. 139, 166–179.

Kuhn, M., Johnson, K., 2019. Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press.

Kumar, V., Gathala, M.K., Saharawat, Y.S., Parihar, C.M., Kumar, R., Kumar, R., Jat, M.L., Jat, A.S., Mahala, D.M., Kumar, L., Nayak, H.S., 2019. Impact of tillage and crop establishment methods on crop yields, profitability and soil physical properties in rice–wheat system of Indo-Gangetic Plains of India. Soil Use Manag. 35 (2), 303–313.

Lawes, R., Chen, C., Whish, J., Meier, E., Ouzman, J., Gobbett, D., Vadakattu, G., Ota, N., van Rees, H., 2021. Applying more nitrogen is not always sufficient to address dryland wheat yield gaps in Australia. Field Crops Res. 262, 108033.

Linquist, B.A., Brouder, S.M., Hill, J.E., 2006. Winter straw and water management effects on soil nitrogen dynamics in California rice systems. Agron. J. 98, 1050–1059. https://doi.org/10.2134/agronj2005.0350.

Lollato, R.P., Ruiz Diaz, D.A., DeWolf, E., Kna, M., Peterson, D.E., Fritz, A.K., 2019. Agronomic practices for reducing wheat yield gaps: a quantitative appraisal of progressive producers. Crop Sci. 59 (1), 333–350.

Molnar, C., Casalicchio, G., Bischl, B., 2018. iml: An R package for interpretable machine learning. J. Open-Source Softw. 3 (26), 786.

Molnar, C., Gruber, S., Koer, P., 2020. Limitations of interpretable machine learning methods.

Mourtzinis, S., Edreira, J.I.R., Grassini, P., Roth, A.C., Casteel, S.N., Ciampitti, I.A., Kandel, H.J., Kyveryga, P.M., Licht, M.A., Lindsey, L.E., Mueller, D.S., 2018. Sifting and winnowing: Analysis of farmer field data for soybean in the US North-Central region. Field Crops Res. 221, 130–141.

Mourtzinis, S., Specht, J.E., Conley, S.P., 2019. Defining optimal soybean sowing dates across the US. Sci. Rep. 9 (1), 1–7.

Nayak, H.S., Parihar, C.M., Mandal, B.N., Patra, K., Jat, S.L., Singh, R., Singh, V.K., Jat, M.L., Garnaik, S., Nayak, J., Abdallah, A.M., 2022a. Point placement of late vegetative stage nitrogen splits increase the productivity, N-use efficiency and profitability of tropical maize under decade long conservation agriculture. Eur. J. Agron. 133, 126417.

Nayak, H.S., Silva, J.V., Parihar, C.M., Kakraliya, S.K., Krupnik, T.J., Bijarniya, D., Jat, M.L., Sharma, P.C., Jat, H.S., Sidhu, H.S., Sapkota, T.B., 2022. Rice yield gaps and nitrogen-use efficiency in the Northwestern Indo-Gangetic Plains of India: Evidence based insights from heterogeneous farmers' practices. Field Crops Res. 275, 108328.

Nigam, A., Garg, S., Agrawal, A., Agrawal, P., 2019, Crop yield prediction using machine learning algorithms. In 2019 Fifth International Conference on Image Information Processing (ICIIP), 125–130. IEEE.

Parihar, C.M., Parihar, M.D., Sapkota, T.B., Nanwal, R.K., Singh, A.K., Jat, S.L., Nayak, H. S., Mahala, D.M., Singh, L.K., Kakraliya, S.K., Stirling, C.M., 2018. Long-term impact of conservation agriculture and diversified maize rotations on carbon pools and stocks, mineral nitrogen fractions and nitrous oxide fluxes in inceptisol of India. Sci. Total Environ. 640, 1382–1392.

Parihar, C.M., Nayak, H.S., Rai, V.K., Jat, S.L., Parihar, N., Aggarwal, P., Mishra, A.K., 2019. Soil water dynamics, water productivity and radiation use efficiency of maize under multi-year conservation agriculture during contrasting rainfall events. Field Crops Res. 241, 107570.

Park, A.G., Davis, A.S., McDonald, A.J., 2018. Priorities for wheat intensification in the Eastern Indo-Gangetic Plains. Glob. Food Secur. 17, 1–8.

Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., Athanasiadis, I.N., 2021. Machine learning for large-scale crop yield forecasting. Agric. Syst. 187, 103016.

Probst, P., Wright, M.N., Boulesteix, A.L., 2019. Hyperparameters and tuning strategies for random forest. WIREs Data Min. Knowl. Disco, e1301. https://doi.org/10.1002/widm.1301.

Ransom, C.J., Kitchen, N.R., Camberato, J.J., Carter, P.R., Ferguson, R.B., Fernández, F. G., Franzen, D.W., Laboski, C.A.M., Myers, D.B., Nafziger, E.D., Sawyer, J.E., Shanahan, J.F., 2019. Statistical and machine learning methods evaluated for incorporating soil and weather into corn nitrogen recommendations. Comput. Electron. Agric. 164, 104872 https://doi.org/10.1016/j.compag.2019.104872.

Rattalino Edreira, J.I., Mourtzinis, S., Conley, S.P., Roth, A.C., Ciampitti, I.A., Licht, M. A., Kandel, H., Kyveryga, P.M., Lindsey, L.E., Mueller, D.S., Naeve, S.L., Nafziger, E., Specht, J.E., Stanley, J., Staton, M.J., Grassini, P., 2017. Assessing causes of yield gaps in agricultural areas with diversity in climate and soils. Agric. . Meteorol. 247, 170–180. https://doi.org/10.1016/j.agrformet.2017.07.010.

Sabater, M.J. 2019, ERA5-Land hourly data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS).

Sapkota, T.B., Vetter, S.H., Jat, M.L., Sirohi, S., Shirsath, P.B., Singh, R., Jat, H.S., Smith, P., Hillier, J., Stirling, C.M., 2019. Cost-effective oortunities for climate change mitigation in Indian agriculture. Sci. Total Environ. 655, 1342–1354.

Sadras, V.O., Giordano, N., Correndo, A., Cossani, C.M., Ferreyra, J.M., Caviglia, O.P., Coulter, J.A., Ciampitti, I.A., Lollato, R.P., 2022. Temperature-Driven Developmental Modulation of Yield Response to Nitrogen in Wheat and Maize. Front. Agron. 4, 903340 https://doi.org/10.3389/fagro.

Shah, D.A., Butts, T.R., Mourtzinis, S., Rattalino Edreira, J.I., Grassini, P., Conley, S.P., Esker, P.D., 2021. A machine learning interpretation of the contribution of foliar fungicides to soybean yield in the north-central United States. . Sci. Rep. 11 (1), 1–11.

Shendryk, Y., Davy, R., Thorburn, P., 2021. Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning. Field Crops Res. 260, 107984.

Shook, J., Gangopadhyay, T., Wu, L., Ganapathy subramanian, B., Sarkar, S., Singh, A.K., 2021. Crop yield prediction integrating genotype and weather variables using deep learning. Plos One 16 (6), e0252402.

Shyamsundar, P., Springer, N.P., Tallis, H., Polasky, S., Jat, M.L., Sidhu, H.S., Krishnapriya, P.P., Skiba, N., Ginn, W., Ahuja, V., Cummins, J., 2019. Fields on fire: Alternatives to crop residue burning in India. Science 365 (6453), 536–538.

Silva, J.V., Reidsma, P., van Ittersum, M.K., 2017. Yield gaps in Dutch arable farming systems: Analysis at crop and crop rotation level. Agric. Syst. 158, 78–92.

Silva, J.V., Tenreiro, T.R., Spätjens, L., Anten, N.P., van Ittersum, M.K., Reidsma, P., 2020. Can big data explain yield variability and water productivity in intensive croing systems? Field Crops Res. 255, 107828.

Silva, J.V., Reidsma, P., Baudron, F., Laborte, A.G., Giller, K.E., van Ittersum, M.K., 2021. How sustainable is sustainable intensification? Assessing yield gaps at field and farm level across the globe. Glob. Food Secur. 30, 100552.

Singh, Y., Gupta, R.K., Singh, Gurpreet, Singh, Jagmohan, Sidhu, H.S., Singh, Bijay, 2009. Nitrogen and residue management effects on agronomic productivity and nitrogen use efficiency in rice–wheat system in Indian Punjab. Nutr. Cycl. Agroecosyst. 84 (2), 141–154.

Singh, Y., Kukal, S.S., Jat, M.L., Sidhu, H.S., 2014. Improving water productivity of wheat-based croing systems in South Asia for sustained productivity. Adv. Agron. 127, 157–258.

Tisdale, S.L., Nelson, W.L., 1966. Soil fertility and fertilizers. Soil Sci. 101 (4), 346.

Thuy, N.H., Shan, Y., Wang, K., Cai, Z., Buresh, R.J., 2008. Nitrogen supply in rice-based cropping systems as affected by crop residue management. Soil Sci. Soc. Am. J. 72 (2), 514–523.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc.: Ser. B (Methodol.) 58 (1), 267–288.

Tolle, K.M., Tansley, S., Hey, T., 2011. The fourth paradigm: data-intensive scientific discovery. Proc. IEEE 99, 1334–1337. https://doi.org/10.1109/JPROC.2011.2155130.

Tseng, M.C., Roel, Á., Macedo, I., Marella, M., Terra, J., Zorrilla, G., Pittelkow, C.M., 2021. Field-level factors for closing yield gaps in high-yielding rice systems of Uruguay. Field Crops Res. 264, 108097.

Van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agric. 177, 105709.

Vanlauwe, B., Dobermann, A., 2020. Sustainable intensification of agriculture in sub-Saharan Africa: first things first. Front. Agric. Sci. Eng. 7 (4), 376–382.

Villegas, D., Alfaro, C., Ammar, K., Cátedra, M.M., Crossa, J., García del Moral, L.F., Royo, C., 2016. Daylength, temperature and solar radiation effects on the phenology and yield formation of spring durum wheat. J. Agron. Crop Sci. 202 (3), 203–216.