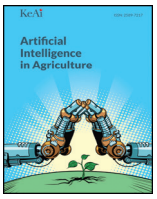
Artificial Intelligence in Agriculture 6 (2022) 257–265

Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/



Explainable artificial intelligence and interpretable machine learning for agricultural data analysis

Masahiro Ryo [a,b,*]

[a] Leibniz Centre for Agricultural Landscape Research (ZALF), Eberswalder Str. 84, 15374 Müncheberg, Germany
[b] Brandenburg University of Technology Cottbus–Senftenberg, Platz der Deutschen Einheit 1, 03046 Cottbus, Germany

ARTICLE INFO

Article history:
Received 22 September 2022
Received in revised form 14 November 2022
Accepted 15 November 2022
Available online 17 November 2022

Keywords:
Interpretable machine learning
Explainable artificial intelligence
Agriculture
Crop yield
No-tillage
XAI

ABSTRACT

Artificial intelligence and machine learning have been increasingly applied for prediction in agricultural science. However, many models are typically black boxes, meaning we cannot explain what the models learned from the data and the reasons behind predictions. To address this issue, I introduce an emerging subdomain of artificial intelligence, explainable artificial intelligence (XAI), and associated toolkits, interpretable machine learning. This study demonstrates the usefulness of several methods by applying them to an openly available dataset. The dataset includes the no-tillage effect on crop yield relative to conventional tillage and soil, climate, and management variables. Data analysis discovered that no-tillage management can increase maize crop yield where yield in conventional tillage is <5000 kg/ha and the maximum temperature is higher than 32°. These methods are useful to answer (i) which variables are important for prediction in regression/classification, (ii) which variable interactions are important for prediction, (iii) how important variables and their interactions are associated with the response variable, (iv) what are the reasons underlying a predicted value for a certain instance, and (v) whether different machine learning algorithms offer the same answer to these questions. I argue that the goodness of model fit is overly evaluated with model performance measures in the current practice, while these questions are unanswered. XAI and interpretable machine learning can enhance trust and explainability in AI.

1. Introduction

Artificial intelligence (AI) and machine learning are increasingly used for prediction in agriculture (Benos et al., 2021; Liakos et al., 2018). They often outperform conventional statistical parametric models like generalized linear models in predictive performance (Breiman, 2001a). A general linear regression, for example, needs the variables to follow normality and linearity; therefore, data transformation is often needed. Meanwhile, random forests and artificial neural networks do not need such transformation procedures. In addition, machine learning algorithms can automatically discover nonlinearity and variable interactions (Ryo and Rillig, 2017). These tools are now easy to learn because various online courses are nowadays available, lowering the hurdle for students and researchers to start using machine learning in their projects.

AI and machine learning make statistical modeling more predictive, but it comes at a cost. It sacrifices interpretability. Machine learning algorithms that achieve a higher predictive performance tend to be more complex, like random forests, gradient boosting, and artificial neural networks (Breiman, 2001a). Increasing model complexity (with regularization) is key to enhancing predictability. However, the most accurate model is often too complex for human beings to interpret the logic behind a prediction, the so-called black box. We cannot explain what the model learned from the data, why it predicts a certain value for a given instance, and when it tends to make a mistake. In general, there is a trade-off between the accuracy and interpretability of statistical models (Breiman, 2001a).

Achieving both high accuracy and interpretability is challenging (Breiman, 2001a), but most researchers would agree that the employed model should be both accurate and easy to interpret. Providing interpretable predictions is more important than providing accurate predictions with a black-box model for decision-making (Rudin, 2019; Rudin et al., 2022). For instance, an AI model suggests a farmer change the current field management from conventional tillage to no-tillage so that yield can increase by 10%. Surely, the farmer wants to know why the model predicted so. The model developer should also know if the model learned agriculturally meaningful patterns from the data and if the reasons behind each prediction make sense. What if the model

* Corresponding author at: Leibniz Centre for Agricultural Landscape Research (ZALF), Eberswalder Str. 84, 15374 Müncheberg, Germany.
E-mail address: Masahiro.Ryo@zalf.de (M. Ryo).

https://doi.org/10.1016/j.aiia.2022.11.003

discovers a strange but interesting pattern? One can investigate it further to evaluate if the discovery is important or not. For these purposes, AI and machine learning need to be interpretable and explainable (Meske and Bunde, 2020; Ribeiro et al., 2016).

To fulfill this demand, we can make use of the emerging subfield of the AI domain, explainable AI (XAI), especially a set of tools, interpretable machine learning (Adadi and Berrada, 2018; Doshi-Velez and Kim, 2017; Molnar, 2019; Murdoch et al., 2019; Rudin et al., 2022). XAI aims to develop tools for enhancing the interpretability of complex algorithms without sacrificing predictability (Carvalho et al., 2019). The XAI domain has been gaining much attention in the past decade, and its potential has been disseminated to several natural science fields, such as biodiversity research (Ryo et al., 2021), geoscience (Mamalakis et al., 2022), and hydrological/climatic science (Başağaoğlu et al., 2022). In the agricultural domain, several previous studies have started applying the techniques since 2020 (Fig. 1): Crop yield estimate (Sihi et al., 2022; Wolanin et al., 2020); crop type and trait classification using satellite (Newman and Furbank, 2021; Orynbaikyzy et al., 2020); soil texture classification (Zhou et al., 2022); leaf disease classification (Wei et al., 2022); water flux and quality assessment (Garrido et al., 2022; Zhang et al., 2022); IoT based smart agriculture system (Sabrina et al., 2022); biomethane production (De Clercq et al., 2020); agricultural land identification (Viana et al., 2021). However, these studies use only a few particular methods. Moreover, potentially several articles are using XAI methods without emphasizing the usage. Nevertheless, I argue that the XAI concept and several useful techniques remain largely unintroduced to the agricultural domain.

This article aims to demonstrate the potential of XAI, especially interpretable machine learning techniques, for analyzing agricultural datasets. After a brief introduction to the concept of interpretable machine learning, I show how interpretable machine learning methods can be used for discovering novel patterns from a tabular dataset. As a case study, I use the global dataset for crop production under conventional tillage and no-tillage systems openly available from Su et al. (2021). The analysis gives a novel insight into under which conditions no-tillage management can improve Maize crop yield compared to conventional tillage management (see section 2.2 for the detailed description of the dataset). I made the analysis fully reproducible with the data and R script available on GitHub, hoping that to facilitate readers' hands-on learning (https://github.com/masahiroryo/2022_IML_Agriculture.git).

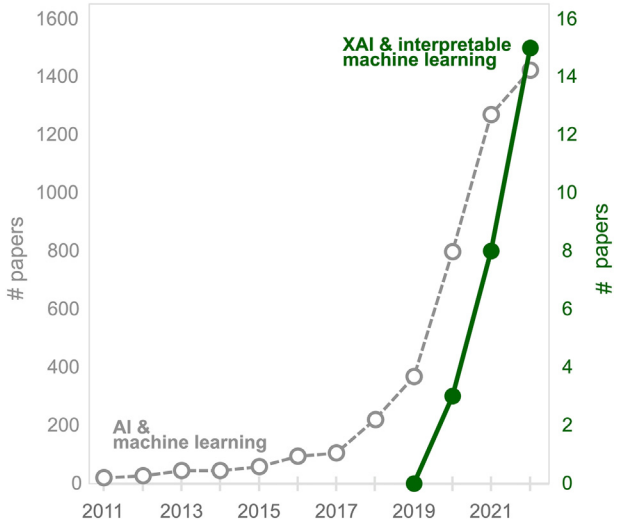


Fig. 1. Publication trend in "AI and machine learning" and "XAI and interpretable machine learning" in agricultural science according to the Web of Science Core Collection. XAI: Explainable artificial intelligence. The search queries were.

[("machine learning" OR "artificial intelligence") AND "agricul*"] (topic) and [("interpretable machine learning" OR "explainable artificial intelligence" OR "explainable machine learning" OR "XAI" OR "interpretable ML" OR "explainable AI") AND "agricul*"] (topic), respectively. The search was done on 30.08.2022, and the number in 2022 was multiplied by 1.5 so that it can be an estimate for the end of the year, compatible with the past years.

2. Methods

2.1. Interpretable machine learning: An overview

Machine learning algorithms can make accurate predictions, but understanding the rationales behind predictions is often difficult. The lack of interpretability makes scientists and stakeholders wonder how much they should trust what the models predict regardless of accuracy (Meske and Bunde, 2020; Ribeiro et al., 2016). This problem developed the idea of XAI and various tools, namely, interpretable machine learning (Murdoch et al., 2019). XAI aims to develop tools for enhancing the interpretability of complex machine learning algorithms without sacrificing accuracy (Carvalho et al., 2019). XAI has been gaining popularity rapidly in recent years, and many new interpretable machine learning methods have been proposed, reviewed, and applied in various scientific fields recently (Boehmke and Greenwell, 2020; Molnar, 2019; Murdoch et al., 2019; Ryo et al., 2021).

Most interpretable machine learning methods are categorized in model selection, method generality, and explanation scale (Adadi and Berrada, 2018; Molnar, 2019; Murdoch et al., 2019). Firstly, model selection is either model-based or post-hoc. Model-based means that a machine learning algorithm used for the study is rather simple and directly interpretable (e.g., decision tree and generalized additive model), while post-hoc means that a complex machine learning algorithm (e.g., random forests and gradient boosting) is used for the study. Then the fitted model is analyzed with some statistical methods. Secondly, method generality is either model-specific or model-agnostic. Some methods can be used only for the corresponding algorithm (e.g., Gini importance for tree-based algorithms), but many methods are developed and can be used for any algorithm, so-called model-agnostic. Thirdly, explanation scale is either global or local. Global means interpreting what the model learned from the entire variable distribution (e.g., if predictor X is positively associated with the response). Local means interpreting the rationale behind every single prediction given by the model (e.g., the model predicts this plant is sick, but why does it predict so?). Note that different terminology may also be used for method classification in other studies because the XAI domain is still at emergence and dynamic.

2.2. Dataset

I use the global dataset for crop production under conventional tillage and no-tillage systems, which are openly available from Su et al. (2021). The dataset contains paired yield observations comparing conventional tillage and no-tillage conditions for eight major staple crops in 50 countries. The dataset reports crop yield, crop growing season, management practices, soil characteristics, and key climate parameters throughout the experimental year.

Conventional tillage is a tillage system using cultivation as the primary means of seedbed preparation and weed control with emphasis on soil preparation, including a sequence of soil tillage, such as plowing and harrowing, and the removal of most of the plant residue from the previous crop (OECD, 2001). No-tillage (also zero tillage) is a minimum tillage practice where the crop is sown directly into the soil not tilled since the harvest of the previous crop, weed control is achieved by using herbicides, and stubble is retained for erosion control (OECD, 2001). No-tillage is recognized as one of the key conservational agricultural strategies without compromising crop yield (Phillips et al., 1980),

but this statement is controversial. A recent global meta-analysis study synthesizing 678 studies across 50 crops with 6005 paired observations concluded that no-tillage reduces crop yield by 5%, and especially the negative impact of no-tillage was the largest for maize (−7.6%) (Pittelkow et al., 2015).

As a case study, I analyzed maize. Although the largest negative effect was found for maize, it is just a global average across various conditions. I hypothesized that the effect of no-tillage can be positive under some conditions, and the conditions can be identified using interpretable machine learning. If the conditions were discovered, our scientific knowledge would improve: "On average, no-tillage reduces maize yield; however, no-tillage can increase yield if the condition is …." I analyzed the relative yield change in maize (%) that was quantified by comparing no-tillage to conventional tillage in a paired experimental setup.

the most dominant crop type in the dataset with global coverage ($n = 1271$; Fig. 2a). A relative change in crop yield from conventional to no-tillage was random (Fig. 2b; mean = −0.02, standard deviation = 0.25; note that the extreme values of 97.5th percentile or higher were removed), indicating that whether no-tillage increases or decreases crop yield compared to conventional tillage is quite controversial. With machine learning modeling, I explored under which conditions the effect tends to be positive.

2.3. Modeling

A relative change in Maize crop yield from conventional to no-tillage was regressed with 17 variables: Crop yield under conventional tillage as baseline (Yield_CT) [kg/ha]; latitude and longitude of experimental sites accounting for spatial dependence [degree]; Years since no-tillage started accounting for lagged effect (Years_NT); crop rotation with at least three crops involved in conventional tillage and no-tillage for temporal dependency (Crop_rotation_CT and _NT) [yes/no]; soil texture (ST) [seven categories related to sand, silt, clay composition]; soil cover (Soil_cover_CT and _NT) [yes/no/mixed]; weed and pest control (Weed_pest_control_CT and _NT) [yes/no]; Precipitation and potential evapotranspiration over the growing season and their difference for water availability (P, E, PB, respectively) [mm]; average, maximum, and minimum air temperature during the growing season (Tave, Tmax, Tmin, respectively) [degree Celsius].

The modeling process is illustrated in Fig. 3. The sample (n = 1271) was split randomly into a training and test dataset (80:20 split). Four machine learning algorithms were used: linear model with AIC stepwise variable selection, decision tree (conditional inference tree; Hothorn et al., 2006), random forests (Breiman, 2001b), and gradient boosting (Friedman, 2001). The former two algorithms are relatively simple

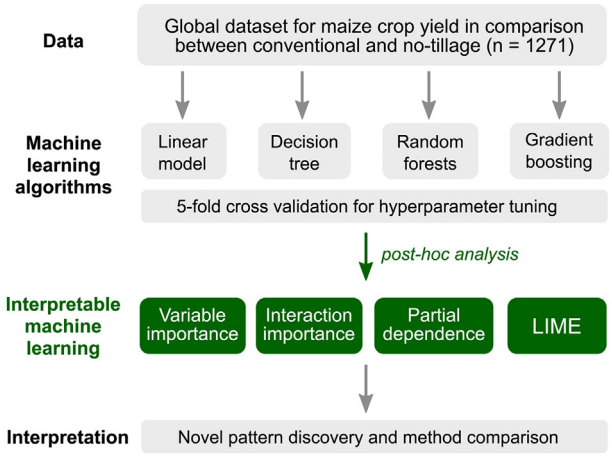


Fig. 3. Study framework for novel pattern discovery using interpretable machine learning methods after implementing machine learning algorithms (i.e., post-hoc analysis). LIME: Local Interpretable Model-Agnostic Explanations.

models with high model-based interpretability. The latter two are complex models combining 100–10,000 models (weak learners), and therefore they require post-hoc interpretable methods for understanding model behavior. These four models were compared to show how the models learn differently. Note that the interpretable machine learning methods I introduce can be used with any other machine learning methods like support vector machines and artificial neural network. A 5-fold cross-validation was employed for finding the best hyperparameter set for decision tree (mincriterion = 0.01), random forests (mtry = 12) and gradient boosting (n.trees = 1000, interaction.depth = 3) in terms of root mean squared error (RMSE). Model performance was evaluated with R-squared (R^2) and RMSE.

2.4. Interpretable machine learning methods

I use a set of post-hoc, model-agnostic methods (3 global and 1 local) so that model behavior can be compared among algorithms in a standard way: Permutation-based variable importance (global), pairwise interaction importance (global), partial dependence plot (global), and LIME local variable importance (Fig. 3).

Permutation-based variable importance measure: This is a measure to rank the relative importance of predictor variables for prediction. The fundamental idea is that if one randomly permutes the values of an important variable in the training data, the model performance would

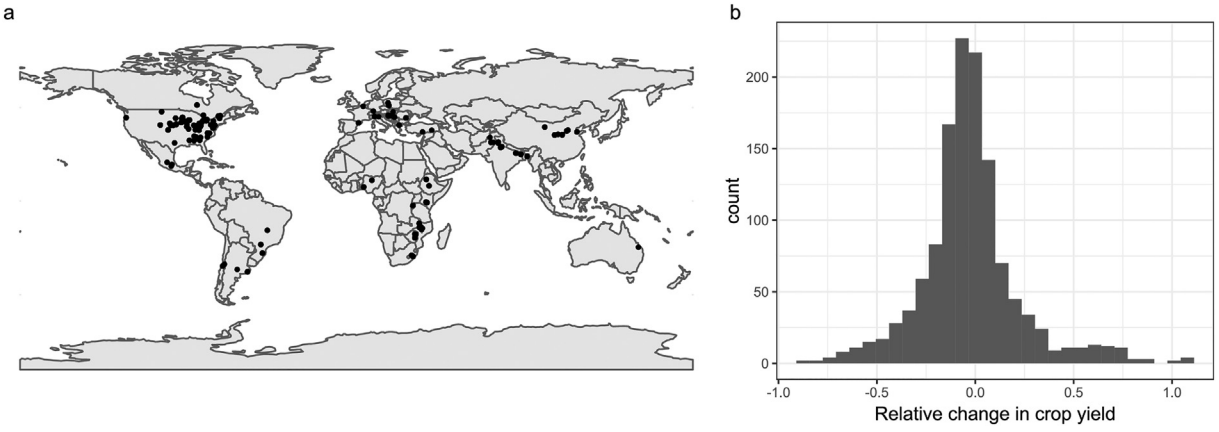


Fig. 2. Collection of experiments comparing Maize crop yield in conventional and no-tillage conditions ($n = 1271$): (a) experimental site distribution and (b) histogram of yield change in no-tillage relative to conventional tillage. Data is available from Su et al. (2021), and extreme values (97.5 percentile) were removed.

degrade because permutation destroys the relationship between the variable and the response variable (Breiman, 2001b). The larger the loss in model performance, the larger its importance. The importance measure is based on the difference between a baseline performance measure (R^2 in this study) and the same performance measure obtained after permuting the values of a particular variable in the training data. To account for random variability due to permutation, I calculated permutation-based importance thirty times and took an average.

Pairwise interaction importance: This measure is used to quantify the strength of two-way interaction effects that affects model prediction. The fundamental idea is that if a certain variable pair (X_i, X_j) has a strong interaction strength, the modeled association between X_i and the response variable would strongly depend on the other variable's value, X_j. I used the method in Greenwell et al. (2018). It evaluates how much the flatness of the modeled association of X_i to the response variable changes by changing the value of X_j, calculating the standard deviation of a flatness score. This procedure is also done by flipping X_i and X_j to take an average. Another popular approach for quantifying interaction strength is Friedman's H-statistic (Friedman and Popescu, 2008). But, I did not use this approach because Greenwell et al. (2018) warned that Friedman's H-statistic may not adequately discover strong interactions (yet, Greenwell et al. did not argue any potential reasons)."

Partial dependence plot: This method helps visualize the modeled association between a subset of the predictors (conventionally, 1–2 variables) and the response while accounting for the average effect of the other predictors in the model (Friedman, 2001). To estimate the association of X_i with the response, the model gives predictions given a fixed value of X_i while changing the values of all the other predictors available in the training dataset. This procedure is done for the entire range of X_i. I refer to the method (Greenwell, 2017), while many other approaches are available. This is because Greenwell (2017) offers the *pdp* package, the most generalized implementation in R with a clear documentation for practical usage.

LIME variable importance: LIME stands for Local Interpretable Model-agnostic Explanations (Ribeiro et al., 2016), a technique to evaluate the variable importance for each prediction. LIME assumes that even though a complex machine learning model shows a nonlinear, non-additive behavior, the behavior can be approximated with a simpler model like a linear model (so-called local surrogate model). I implemented the version by Molnar (2019). In short, when a prediction is made with the machine learning model, LIME generates many data points by slightly perturbing the predicted case. It fits a locally weighted linear regression model with L1-regularization to the points where weights are based on their proximity to the predicted case. Then, the variable importance of the linear model is reported. In this study, I used the Canberra distance with the kernel width of 2 because of a good model fit, while other distance measures with a different width can be used.

2.5. Programming language and reproducibility

All data handling and analysis were done in R version 4.2.1 (R Core Team, 2022) with the following libraries: For data handling and visualization, tydiverse (Wickham et al., 2019), patchwork (Pedersen, 2022), stars (Pebesma et al., 2022), rnaturalearth (South, 2017); for machine learning implementation, caret (Kuhn, 2008); for interpretable machine learning methods, vip (Greenwell et al., 2020), pdp (Greenwell, 2017), iml (Molnar and Schratz, 2022). The script and data are available in the GitHub repository (https://github.com/masahiroryo/2022_IML_Agriculture.git).

3. Results

The model performance revealed random forests as the best algorithm (R^2 = 0.42; RMSE = 0.199), followed by gradient boosting (0.33; 0.200), decision tree (0.18; 0.225), and linear model (0.11; 0.236) (Fig. 4).

In terms of variable importance, the random forests and gradient boosting commonly selected yield in conventional tillage as the best predictor, followed by temperature-related variables (Fig. 5c, d). The decision tree and linear model also selected yield in conventional tillage as one of the top predictors but regarded it as less important than soil texture (Fig. 5a, b). On the contrary, random forests and gradient boosting did not select soil texture within the top ten predictors.

Variable importance was also evaluated for discovering key variable interactions. Interaction strength was investigated for all possible pairs among the variables that were selected within the top 3 in variable importance by at least one algorithm (Fig. 5). In total, six variables were investigated, accounting for the fifteen pairwise combinations: Yield_CT, Tmax, Tave, Tmin, ST, Soil_cover_NT. Overall, different algorithms learned different interactions as important for prediction. The linear model showed no importance for any pairs because no interactions were included in the formula (Fig. 6a). Both random forests and gradient boosting identified the interaction of Yield_CT and Tmax as the strongest one (Fig. 6c, d). The decision tree selected this interaction pair as the top 3 (Fig. 6b). The top 3 pairs identified by each algorithm included one of the top 3 important variables in Fig. 5.

Hereafter, I decided to investigate the effects of Yield_CT and Tmax more because random forests and gradient boosting selected these variables within the top 3 in variable importance (Fig. 5) and the strongest combination (Fig. 6c, d). Partial dependence plots were depicted for diagnosing how the associations between Yield_CT and relative yield change were modeled by each of the four algorithms (Fig. 7a). All models suggest a negative relationship. However, the strength and curve shape differed among the models. The linear model suggested a linear relationship, the decision tree suggested a unimodal curve, and both random forests and gradient boosting suggested a negative but nonlinear relationship where the slope of the curve gets milder along with Yield_CT. The models except the linear one suggested no association with Yield_CT > 15,000 because the data points were scarce (Fig. 7c). In terms of Tmax (Fig. 7b), the linear model, decision tree, and random forests suggested a positive relationship with relative yield change. Both decision tree and random forests identified a sharp stepwise relationship around Tmax = 32 (Fig. 7b). Gradient boosting did not show a clear association.

Two-dimensional partial dependence plots were depicted to visually confirm the interaction effects of Yield_CT and Tmax (Fig. 8). All models except the linear one suggest that relative yield change is conditional to

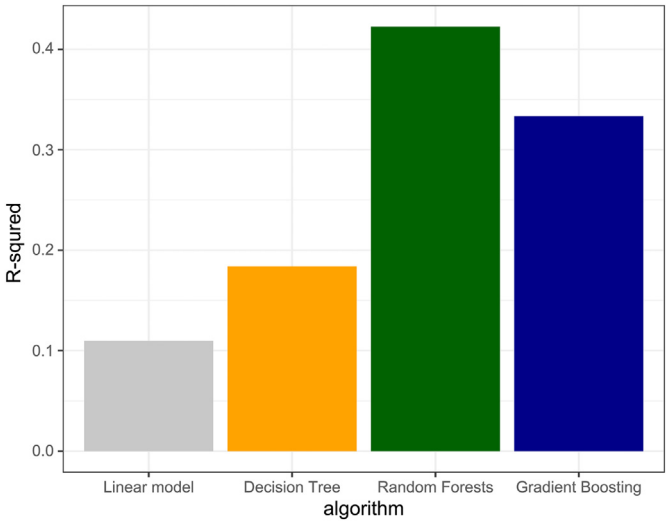


Fig. 4. Model performance.

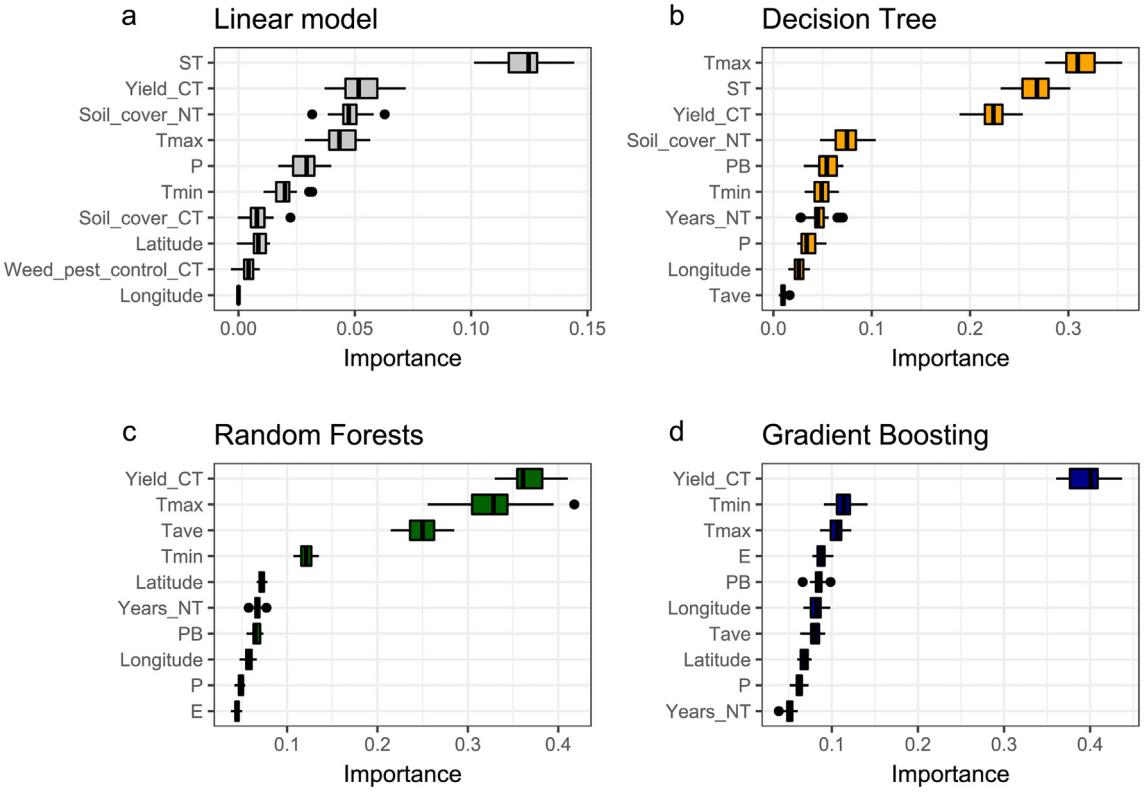


Fig. 5. Permutation-based variable importance.

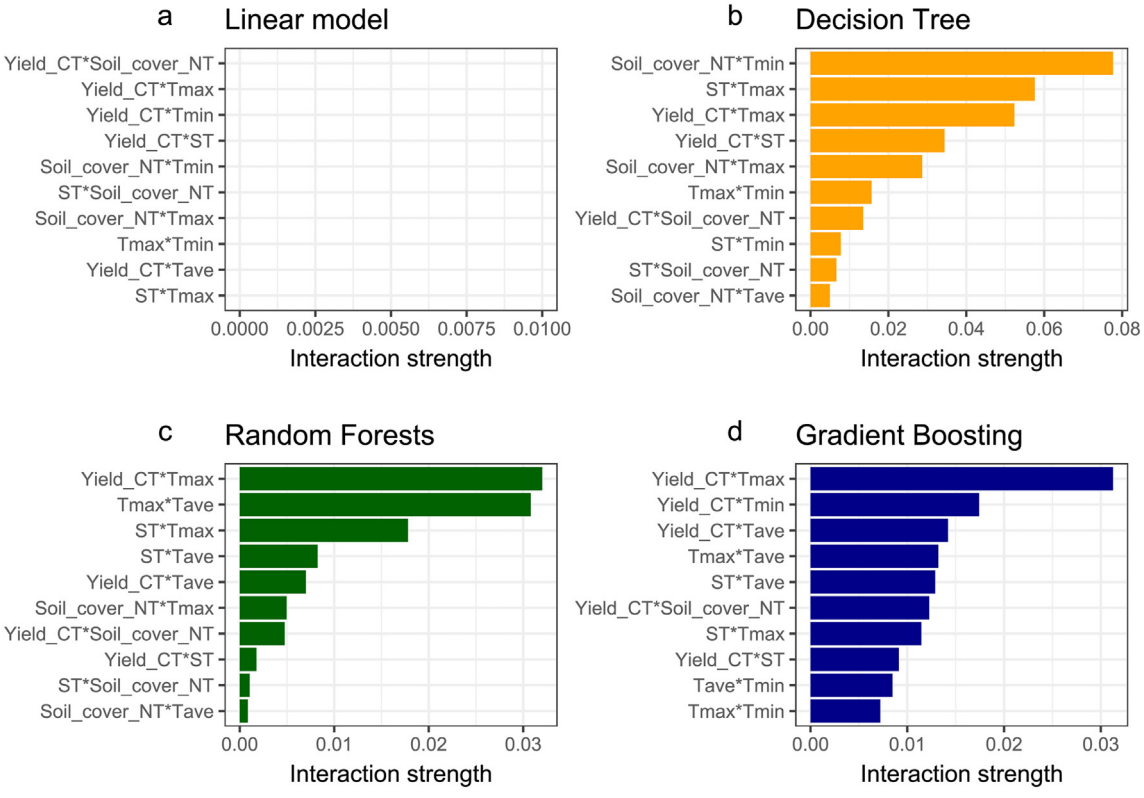


Fig. 6. Pairwise variable interaction importance.

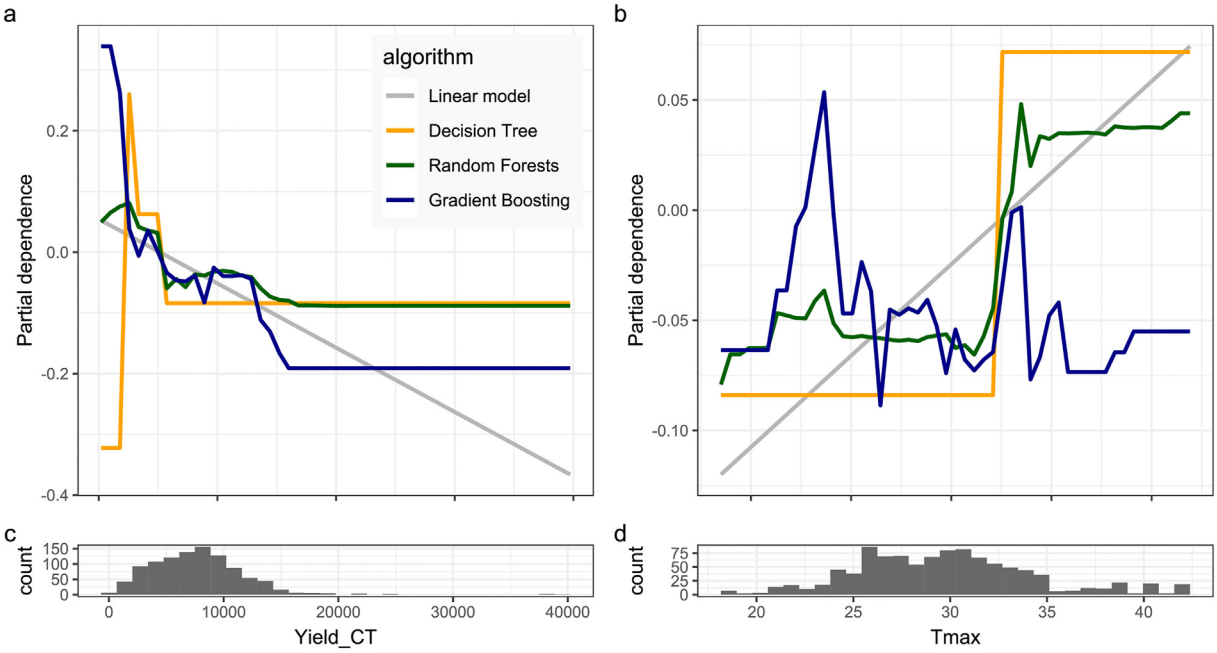


Fig. 7. Partial dependence plots for Yield_CT (a) and Tmax (b) with the data distributions (c, d).

both Tmax and Yield_CT. The patterns identified with all models but linear one show a clear split in the patterns along Tmax around 32° Celsius and Yield_CT around 5000, suggesting the interaction effect of these variables. This interaction effect was further confirmed by depicting partial dependence plots of Yield_CT conditional to a Tmax 32-deg threshold (Fig. 9). It is visible that the association of Yield_CT is stronger if Tmax is higher than 32°. Note that this interaction pattern could be identified only by the previous data analysis procedure. It is not

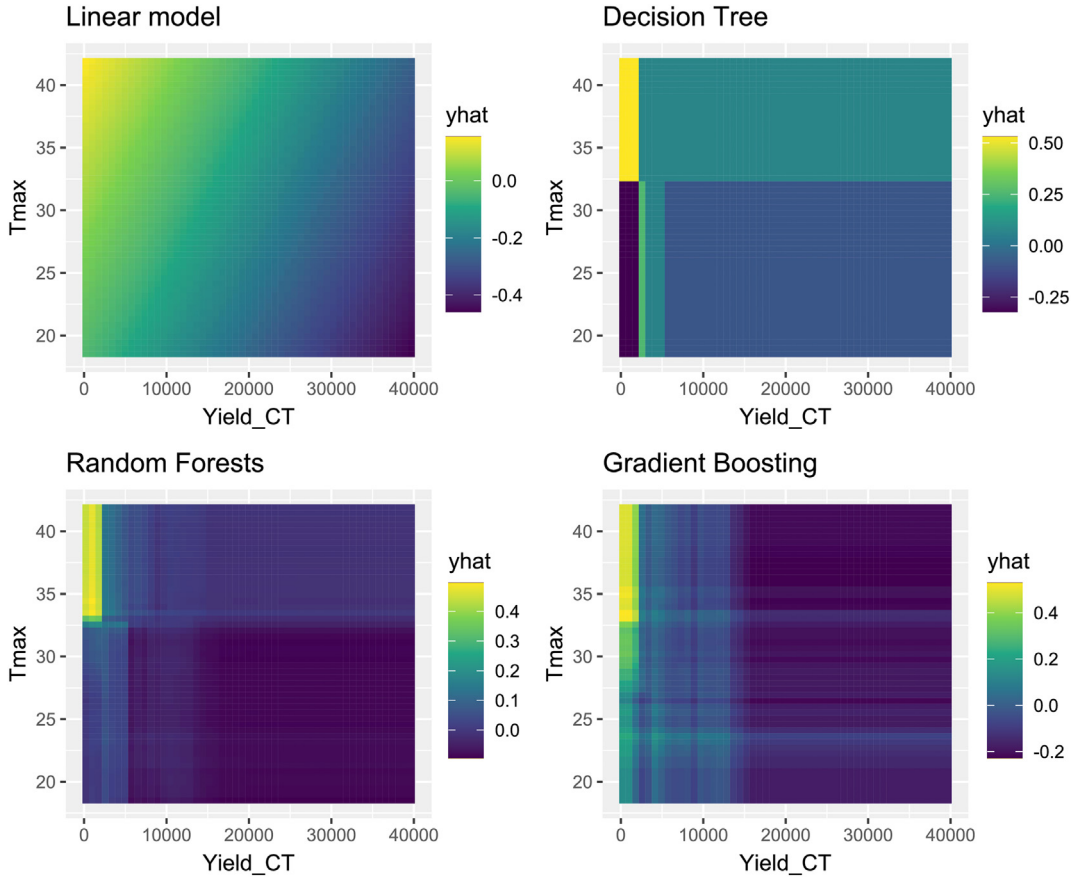


Fig. 8. Partial dependence plot (2D). A brighter yellow region (top-left) indicates that crop yield in no-tillage is higher than conventional tillage, while a darker blue region (bottom-right) indicates the opposite. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

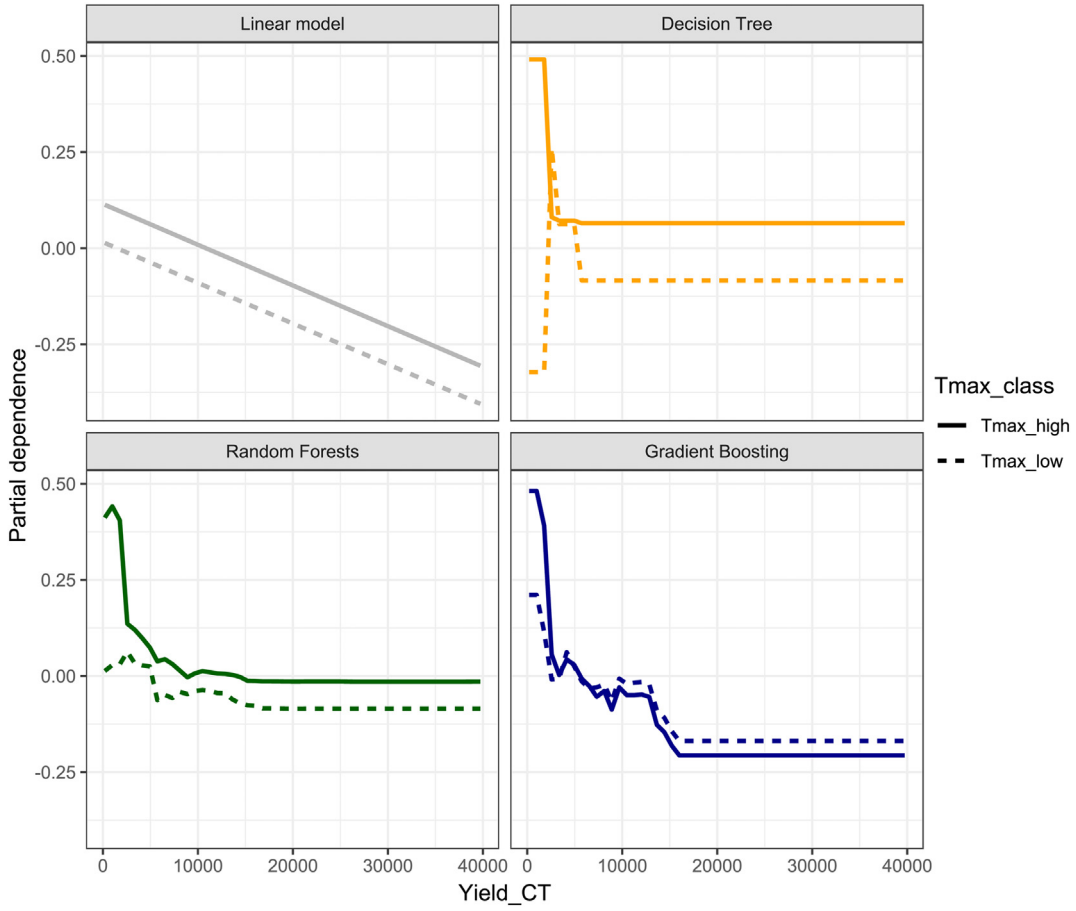


Fig. 9. Partial dependence plot of Yield_CT conditional to Tmax value (higher or lower than 32° Celsius). It suggests that relative yield change becomes higher where yield in conventional tillage is lower than 5000, and the maximum temperature is higher than 32°.

discoverable when only partial dependence plots for a single variable are investigated, as seen in Fig. 7a.

Until here (Figs. 4–9), the focus was to explain global model behavior to understand what the models learned from the data. However, it does not explain local model behavior, which is important for answering what the models consider important when predicting a value given a specific instance. To showcase a local model behavior diagnostic, I used the LIME method for evaluating the variable importance of a randomly selected local site. The site was an experimentation field in Rwanda (Fig. 10e), where the value of relative yield change was −0.239 (Yield_CT = 9200; Yield_NT = 7000). At the site, all models but the linear one suggested that evapotranspiration (E = 520 mm) had a positive effect and soil type of clay (ST = Clay) had a negative effect (Fig. 10a-d). These variables were more important than Yield_CT and Tmax, the most important variables for regulating the global model behavior (Fig. 4), indicating that globally important variables are not necessarily important locally because of context dependence.

4. Discussion

Analyzing the global dataset of maize crop yield as a case study, I demonstrated how a set of interpretable machine learning tools could be used for agricultural data analysis. All methods are post-hoc and model-agnostic, meaning they apply to any machine learning algorithms after training with the data. I used permutation-based variable importance, pairwise variable interaction importance, and partial dependence plot for global model interpretation. I identified that relative yield change can be positive where yield in conventional tillage is smaller than 5000 [kg/ha], and the maximum temperature is higher than 32° Celcius. For local model behavior, I used the LIME method, revealing that locally important variables can differ from the global ones because the conditions are different site by site. While machine learning applications are increasingly popular in agriculture, they often do not use these methods or just a few. These methods can be applied for pattern discovery from any structured (i.e., tabular) dataset and test the reliability of machine learning methods while addressing nonlinearity, variable interactions, and context dependency.

The discovered pattern can be interesting for agronomists working with maize, although explaining the pattern agriculturally is beyond the aim of this study. The most similar work is a global meta-analysis of crop yield under conventional and no-tillage conditions (Pittelkow et al., 2015). Analyzing 6005 paired observations from 678 studies for 50 crops, they concluded that no-tillage reduces yields on average by 5.1%, and the reduction rate was the worst for maize crops (−7.6%; −2.7% in this study). Pittelkow et al. (2015) also explored some reasons based on previous reviews and meta-analyses, concluding that maize yield decreases, especially in cooler climates and areas with high precipitation (Ogle et al., 2012; Rusinamhodzi et al., 2011; Toliver et al., 2012; Van den Putte et al., 2012). My case study analysis suggests yield can decrease in non-hot climates (32° as a threshold), which is in line with them, while precipitation was not a key factor. However, no previous studies found yield in conventional tillage as the strongest factor interacting with temperature: Therefore, this pattern is a new piece of knowledge discovery. Moreover, a meta-analysis tends to focus on the effect (positive or negative) on the mean value of the entire (global) data. Identifying locally specific effects, nonlinear thresholds, and strong interactions from data is a promising avenue for data synthesis and exploration.

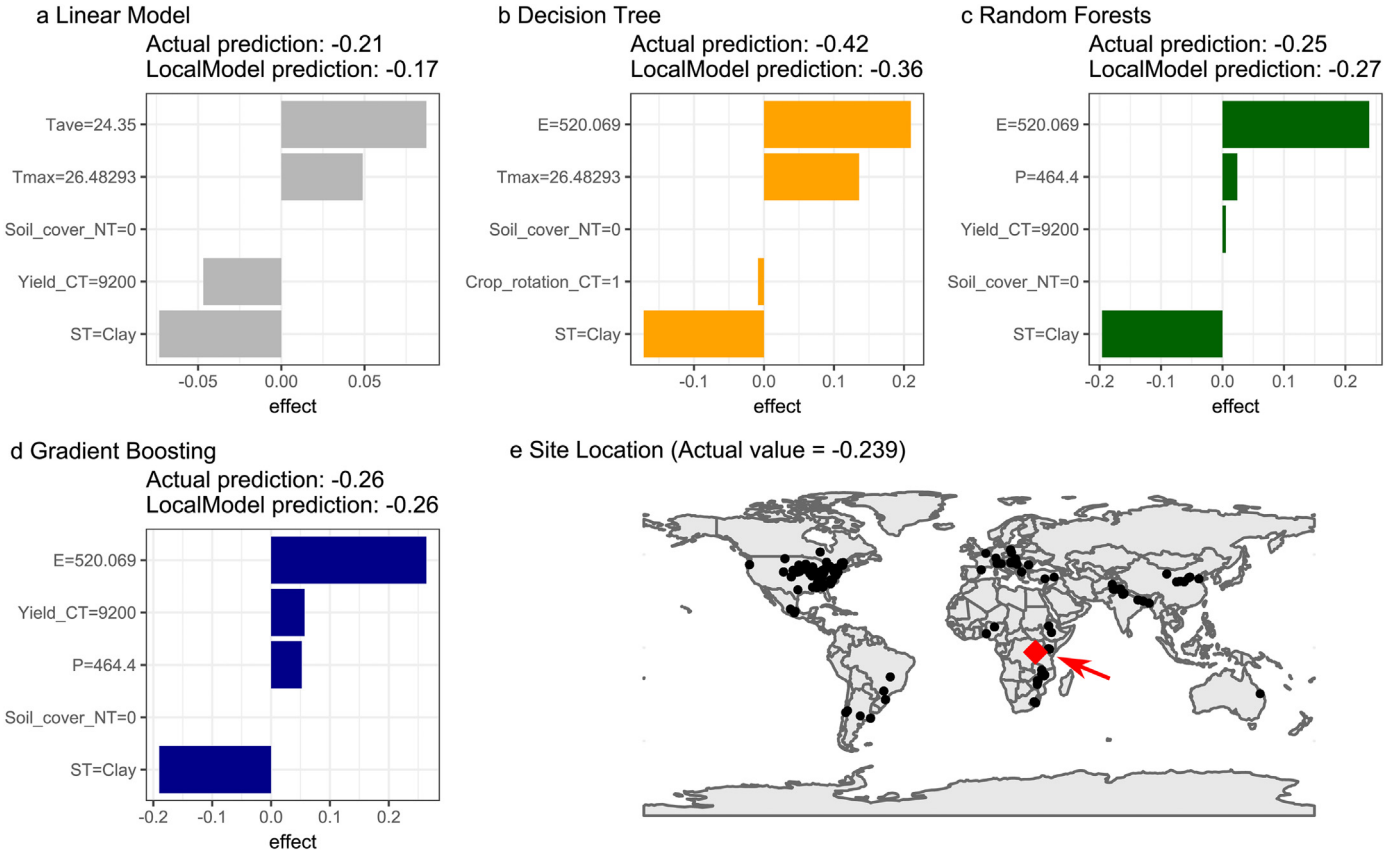


Fig. 10. Local Interpretable Model-Agnostic Explanations (LIME) method for explaining the variable importance at the randomly selected local experimentation site (red point in panel e; Bugusera, Rwanda; latitude = 2°21′S, longitude = 30°15′E). Tave: average temperature; Tmax: maximum temperature; E: Evapotranspiration; P: Precipitation; CT: conventional tillage; NT: no-tillage; ST: Soil type. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As the next step, an emerging, exciting question can be "why" – why do the sites with a lower yield in conventional tillage and a maximum temperature over 32° are more likely to increase Maize crop yield with no-tillage in comparison to conventional tillage? Here, the user of interpretable machine learning needs to communicate with a domain expert to explore the potential reasons behind the pattern. If one comes up with a potential reason (without supporting evidence), it is so-called "hypothesis generation", where the hypothesis can be tested based on an experiment for causality.

Testing causality is necessary for understanding the mechanism regardless of the strength of a discovered pattern because machine learning methods can only explore correlation but not causation (Ryo et al., 2021). Correlation can emerge without causation, and causation can also emerge without correlation. Correlation should be carefully interpreted with the potential existence of any confounding factor. A strong correlation is useful for prediction as a proxy for any underlying mechanisms, but caution is needed because this approach is invalid when the underlying mechanisms change over time (Dormann et al., 2013).

I believe that XAI and interpretable machine learning can bring substantial benefits to agricultural science. However, I also elaborate major caveats. The largest, fundamental question is if we should ever use post-hoc model-agnostic methods for explaining complex models or just use simpler models that can be more directly interpreted (Krishnan, 2020; Molnar et al., 2020; Rudin, 2019). Basically, "explaining the modeled associations" is not the same as "explaining the real causal associations" (Lipton, 2018). In particular, high stakes decision making needs interpretable models instead of explaining black box models (post-hoc) (Rudin, 2019; Rudin et al., 2022). Some post-hoc methods have parameters that affect the results, meaning that the explanation changes quite easily. For instance, the LIME method requires the user to specify the distance measure, kernel width, the number of predictors used, and the proximity method. The result can differ depending on the setting. Also, one needs to pay attention to bias in the data. Globally collected datasets are often biased to information from developed countries or certain regions. Spatially extrapolating the modeled associations to a novel environment can be highly misleading, especially when the condition of the predicted environment does not fit in the probability distribution of the training data (Meyer and Pebesma, 2022).

In conclusion, I hope that this article encourages applications of XAI and interpretable machine learning tools in the agriculture domain. The script is available, so one can learn how the methods were implemented from the code. Opening the black box is a promising next step for AI applications in agriculture.

Credit author statement

Masahiro Ryo: This single author covered all processes from conceptualization to writing.

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.

Acknowledgment

This study was supported by ZALF Integrated Priority Project (IPP2022) "Co-designing smart, resilient, sustainable agricultural

landscapes with cross-scale diversification", Bundesministerium für Bildung und Forschung (BMBF) Land-Innovation-Lausitz project "Landschaftsinnovationen in der Lausitz für eine klimaangepasste Bioökonomie und naturnahen Bioökonomie-Tourismus" (03WIR3017A), BMBF project "Multi-modale Datenintegration, domänenspezifische Methoden und KI zur Stärkung der Datenkompetenz in der Agrarforschung" (16DKWN089), and Brandenburgische Technische Universität Cottbus-Senftenberg GRS cluster project "Integrated analysis of Multifunctional Fruit production landscapes to promote ecosystem services and sustainable land-use under climate change" (GRS2018/19). I thank two anonymous reviewers for constructive comments.

References

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160.
Başağaoğlu, H., Chakraborty, D., Lago, C.D., et al., 2022. A review on interpretable and explainable artificial intelligence in hydroclimatic applications. Water. 14, 1230.
Benos, L., Tagarakis, A.C., Dolias, G., et al., 2021. Machine learning in agriculture: a comprehensive updated review. Sensors 21, 3758.
Boehmke, B., Greenwell, B., 2020. Hands-On Machine Learning with R Available at.
Breiman, L., 2001a. Statistical modeling: the two cultures. Stat. Sci. 16, 199–215.
Breiman, L., 2001b. Random Forests. Mach. Lang. 45, 5–32.
Carvalho, D.V., Pereira, E.M., Cardoso, J.S., 2019. Machine learning interpretability: a survey on methods and metrics. Electronics 8, 832.
De Clercq, D., Wen, Z., Fei, F., et al., 2020. Interpretable machine learning for predicting biomethane production in industrial-scale anaerobic co-digestion. Sci. Total Environ. 712, 134574.
Dormann, C.F., Elith, J., Bacher, S., et al., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27–46.
Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. ArXiv 1702, 08608.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232.
Friedman, J.H., Popescu, B.E., 2008. Predictive learning via rule ensembles. Ann. Appl. Stat. 2 (3), 916–954.
Garrido, M.C., Cadenas, J.M., Bueno-Crespo, A., et al., 2022. Evaporation forecasting through interpretable data analysis techniques. Electronics 11, 536.
Greenwell, B.M., 2017. Pdp: an R package for constructing partial dependence plots. R J. 9, 421–436.
Greenwell, B.M., Boehmke, B.C., McCarthy, A.J., 2018. A simple and effective model-based variable importance measure. ArXiv 1805, 04755.
Greenwell, B.M., Boehmke, B., Gray, B., 2020. vip: Variable Importance Plots. https://cran.r-project.org/web/packages/vip/index.html.
Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Stat. 15 (3), 651–674.
Krishnan, M., 2020. Against interpretability: a critical examination of the interpretability problem in machine learning. Philos. Technol. 33, 487–502.
Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26.
Liakos, K.G., Busato, P., Moshou, D., et al., 2018. Machine learning in agriculture: a review. Sensors 18, 2674.
Lipton, Z.C., 2018. In machine learning, the concept of interpretability is both important and slippery. Queue 16, 28.
Mamalakis, A., Barnes, E.A., Ebert-Uphoff, I., 2022. Investigating the Fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. Artif. Intell. Earth Syst. 1 (4), e220012.
Meske, C., Bunde, E., 2020. Using explainable artificial intelligence to increase trust in computer vision. ArXiv 2002, 01543.
Meyer, H., Pebesma, E., 2022. Machine learning-based global maps of ecological variables and the challenge of assessing them. Nat. Commun. 13, 2208.
Molnar, C., 2019. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. 2nd Ed. https://christophm.github.io/interpretable-ml-book/index.html.
Molnar, C., Schratz, P., 2022. iml: Interpretable Machine Learning. https://cran.r-project.org/web/packages/iml/index.html.
Molnar, C., König, G., Herbinger, J., et al., 2020. Pitfalls to avoid when interpreting machine learning models. ArXiv 2007, 04131.
Murdoch, W.J., Singh, C., Kumbier, K., et al., 2019. Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. 116, 22071–22080.
Newman, S.J., Furbank, R.T., 2021. Explainable machine learning models of major crop traits from satellite-monitored continent-wide field trial data. Nat. Plants 7, 1354.
OECD, 2001. Environmental Indicators for Agriculture– Vol. 3: Methods and Results (glossary: p399-400). OECD Publ. Serv. 409.
Ogle, S.M., Swan, A., Paustian, K., 2012. No-till management impacts on crop productivity, carbon input and soil carbon sequestration. Agric. Ecosyst. Environ. 149, 37–49.
Orynbaikyzy, A., Gessner, U., Mack, B., et al., 2020. Crop type classification using fusion of Sentinel-1 and Sentinel-2 data: assessing the impact of feature selection, optical data availability, and parcel sizes on the accuracies. Remote Sens. 12, 2779.
Pebesma, E., Sumner, M., Racine, E., et al., 2022. Stars: spatiotemporal arrays. Raster Vector Data Cubes. https://cran.r-project.org/web/packages/stars/index.html.
Pedersen, T.L., 2022. Patchwork: the Composer of Plots. https://cran.r-project.org/web/packages/patchwork/index.html.
Phillips, R.E., Thomas, G.W., Blevins, R.L., et al., 1980. No-tillage agriculture. Science 208, 1108–1113.
Pittelkow, C.M., Linquist, B.A., Lundy, M.E., et al., 2015. When does no-till yield more? A global meta-analysis. Field Crop Res. 183, 156–168.
R Core Team, 2022. R: A Language and Environment for Statistical Computing. Austria, Vienna. Retrieved from https://www.R-project.org/.
Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should I trust you?": explaining the predictions of any classifier. ArXiv 1602, 04938.
Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215.
Rudin, C., Chen, C., Chen, Z., et al., 2022. Interpretable machine learning: fundamental principles and 10 grand challenges. Stat. Surv. 16 (1–85), 3.
Rusinamhodzi, L., Corbeels, M., van Wijk, M.T., et al., 2011. A meta-analysis of long-term effects of conservation agriculture on maize grain yield under rain-fed conditions. Agron. Sustain. Dev. 31, 657.
Ryo, M., Rillig, M.C., 2017. Statistically reinforced machine learning for nonlinear patterns and variable interactions. Ecosphere 8, e01976.
Ryo, M., Angelov, B., Mammola, S., et al., 2021. Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. Ecography 44, 199–205.
Sabrina, F., Sohail, S., Farid, F., et al., 2022. An interpretable artificial intelligence based smart agriculture system. Cmc-Comput. Mater. Contin. 72, 3777–3797.
Sihi, D., Dari, B., Kuruvila, A.P., et al., 2022. Explainable machine learning approach quantified the long-term (1981-2015) impact of climate and soil properties on yields of major agricultural crops across CONUS. Front. Sustain. Food Syst. 6, 847892.
South, A., 2017. Rnaturalearth: World Map Data from Natural Earth. https://cran.r-project.org/web/packages/rnaturalearth/index.html.
Su, Y., Gabrielle, B., Makowski, D., 2021. A global dataset for crop production under conventional tillage and no tillage systems. Sci. Data 8, 33.
Toliver, D.K., Larson, J.A., Roberts, R.K., et al., 2012. Effects of no-till on yields as influenced by crop and environmental factors. Agron. J. 104, 530–541.
Van den Putte, A., Govers, G., Diels, J., et al., 2012. Soil functioning and conservation tillage in the Belgian Loam Belt. Soil Tillage Res. 122, 1–11.
Viana, C.M., Santos, M., Freire, D., et al., 2021. Evaluation of the factors explaining the use of agricultural land: a machine learning and model-agnostic approach. Ecol. Indic. 131, 108200.
Wei, K., Chen, B., Zhang, J., et al., 2022. Explainable deep learning study for leaf disease classification. Agron.-Basel 12, 1035.
Wickham, H., Averick, M., Bryan, J., et al., 2019. Welcome to the Tidyverse. J. Open Source Softw. 4, 1686.
Wolanin, A., Mateo-Garcia, G., Camps-Valls, G., et al., 2020. Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. Environ. Res. Lett. 15, 024019.
Zhang, Z., Huang, J., Duan, S., et al., 2022. Use of interpretable machine learning to identify the factors influencing the nonlinear linkage between land use and river water quality in the Chesapeake Bay watershed. Ecol. Indic. 140, 108977.
Zhou, Y., Wu, W., Wang, H., et al., 2022. Identification of soil texture classes under vegetation cover based on Sentinel-2 data with SVM and SHAP techniques. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 3758–3770.