ORIGINAL RESEARCH **Open access**

When Simplicity Meets Performance: Evaluating Machine Learning Complexity in Crop Protection Models

Erik Van der Meer[1*], Lieke Jansen[1], Bram de Vries[2], Sophie Claes[3]

Abstract

Plant diseases and pests are major contributors to substantial reductions in crop yields. To mitigate the damage from these issues, farmers apply plant protection products. These substances are known to cause harmful consequences for both the environment and human well-being. Agronomists persistently seek to restrict the deployment of plant protection products exclusively to those scenarios where their application is truly essential. To pinpoint such scenarios, agronomists and policymakers commonly use decision-support systems that simulate and anticipate the progression of plant diseases. These systems draw on either mechanistic models or statistical techniques derived from extensive collections of biotic (e.g., disease incidence and plant phenological stage) and abiotic (e.g., meteorological and soil characteristics) data gathered from agricultural fields. The rapid advancement of robust machine learning (ML) techniques over the past 10 years has established them as a promising avenue for simulating plant disease dynamics. Machine learning models can identify the key factors driving disease and pest outbreaks, provided they remain sufficiently simple to allow human analysis. However, this simplicity can lead to lower predictive performance than more complex models. In this paper, we conduct a detailed examination of how ML models with differing degrees of complexity perform across two specific crop disease forecasting scenarios: downy mildew affecting grapevine and Cercospora leaf spot affecting sugar beet. We evaluate model accuracy alongside complexity through a cross-validation strategy organized by year. Our outcomes show that variations in meteorological conditions across years are a critical factor in forecasting plant diseases. In addition, aligning with findings from the field of interpretable ML, there is a pronounced trade-off between model complexity and accuracy. Consequently, models with moderate complexity emerge as particularly attractive for forecasting crop disease progression, since they provide clear explanations for their forecasts.

Keywords Machine learning, Plant diseases, Farmers, Plant pests, Meets performance, Crop protection

*Correspondence:
Erik Van der Meer
e.vandermeer@wur.nl
[1] Department of AI for Tulip and Bulb Flower Yield Forecasting, Faculty of Agriculture, Wageningen University, Wageningen, Netherlands
[2] Department of Machine Learning for Storage Rot and Moisture Analysis, Faculty of Agriculture, University of Groningen, Groningen, Netherlands
[3] Department of Smart Phenotyping for Ornamentals, Faculty of Agriculture, University of Amsterdam, Amsterdam, Netherlands

Introduction

Protecting crops from plant diseases is vital to safeguarding overall yield levels. In pursuit of this goal, farmers and agronomists turn to plant protection products—commonly known as pesticides—to manage diseases and pests within cultivated areas. Nevertheless, it is widely recognized that using these products entails several significant disadvantages. Beyond their detrimental effects on farmers' health and their contribution to environmental pollution, they impose financial burdens on producers and end-users alike. Their involvement in fostering pesticide-resistant strains compounds this [1]

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

and causes secondary contamination across various phases of the food production chain [2]. As a result, reducing pesticide use in agricultural settings delivers significant benefits. A practical method to reduce dependence on these substances involves tailoring their use to site-specific conditions, including weather and climate patterns, soil properties, and cultivation techniques [3]. Such customization is enabled by predictive models that estimate disease incidence or the likelihood of outbreaks. These instruments enable farmers and agronomists to avoid applying pesticides when they are not required.

Over time, various initiatives have focused on modeling and predicting the probability of outbreaks and the level of plant disease incidence in cultivated fields [4]. Contemporary techniques are generally divided into two primary groups. Mechanistic models, on the one hand, are formulated from accumulated knowledge and earlier observations regarding the biological cycles of pests and diseases. Developing these models requires comprehensive agronomic research and substantial expert involvement, making them the preferred choice for long-term plant disease modeling. A notable shift occurred following the accumulation of vast datasets, encompassing historical records of field diseases (obtained via direct human assessments or imagery) as well as details on abiotic factors such as soil attributes and weather records. This data-rich environment has enhanced the viability of statistical models—especially machine learning approaches—throughout the recent decade, supporting ongoing gains in their precision and refinement.

In the domain of crop protection, ML models are typically developed for a single crop type and a specific disease. This specificity arises because distinct crops exhibit unique growth patterns, just as diseases and pests display varied behaviors. Furthermore, it is understood that such models are typically calibrated for a defined geographic area and exhibit reduced reliability when applied to data from dissimilar locations [5]. Certain methodologies employ deep learning-based image classification [6-9] to identify diseases. Alternative models focus on projecting or estimating disease incidence within a defined seasonal window, for example, before harvesting, using human-collected observations. The resulting projection might manifest as either a quantitative incidence estimate (regression) or a binary outbreak probability assessment (classification) [3]. In these contexts, model outputs guide agronomists in determining whether to apply plant protection products.

In addition to predicting the occurrence of crop diseases, ML models can support agronomists by clarifying the underlying drivers of disease progression. Achieving this requires the model to be adequately straightforward and interpretable for human comprehension. Representative cases of interpretable ML approaches include linear equations and shallow decision trees.

An accessible yet reliable measure of a model's interpretability is its complexity [10]. This is generally assessed by the number of significant parameters influencing the model's outputs, and it correlates strongly with interpretability. This relationship becomes evident when comparing the interpretive challenge of a linear model with 5 variables to that of one with 300 variables.

Even though sophisticated models like neural networks and gradient boosting methods are often less transparent than straightforward techniques such as linear regression or basic trees, there are instances where this added complexity enhances overall predictive effectiveness [10-12]. The balance between complexity and predictive accuracy arises because greater numbers of parameters or relaxed constraints expand models' capacity and adaptability to detect intricate feature relationships. In contrast, basic models operate under presumptions that might overlook the multifaceted nature of actual datasets. Linear models, for example, presuppose a straight-line relationship between the input variables and the outcome variable (the element being predicted or modeled). Such an assumption, among other limitations, prevents the consideration of interactions among input variables when explaining the outcome. Positioned between highly complex and overly simplistic models are pattern-based approaches [10, 13], which deliver a compelling balance by serving as accessible white-box systems that surpass the predictive strength of both linear regression and decision trees.

Prior research applying ML techniques to crop protection has generally overlooked the importance of interpretability along with the associated complexity-interpretability balance [7, 9, 14]. Accordingly, our work advances current knowledge by

© 2024 The Author(s). | Published by Agro Intelligence Research Press

investigating this balance specifically within crop protection contexts. We develop multiple established machine learning models spanning a range of complexity levels for two representative crop protection objectives: (i) forecasting disease incidence and (ii) estimating the date when symptoms first appear. These predictions target downy mildew in grapevine fields and Cercospora leaf spot in sugar beet cultivation, both situated in France. For each case, we incorporate biotic factors (e.g., historical disease levels) and abiotic factors (e.g., weather data) as input variables. Given that our objectives involve standard regression tasks, the selected models encompass (i) black-box ensemble techniques, including random forests and gradient boosting trees; (ii) white-box options, such as linear regression; and (iii) HiPaR [10], which is a pattern-based regression technique featuring intermediate complexity. The experimental outcomes validate a distinct trade-off between complexity and accuracy within our selected scenarios. They also illustrate multiple strategies for deriving agronomic knowledge from both white-box and black-box ML systems. Our findings reveal that, despite variations in predictive precision and structural design, the models share several core insights. Additionally, effects tied to specific years are highly influential, making it particularly challenging to develop a universal model that accurately forecasts disease incidence across years.

Section 2 outlines the datasets used in this research, the modeling approaches applied to them, and the results obtained. This leads into Section 3, which discusses the various agronomic observations derived from the developed ML models. Finally, Section 4 concludes the paper and proposes directions for subsequent studies on disease incidence forecasting in agricultural settings.

Materials and Methods

Here, we present the agronomic datasets used in our investigation, along with the machine learning models developed from them.

Data

This research utilizes four datasets that address two prominent plant diseases recorded in French agricultural settings: Grape downy mildew and Sugar beet Cercosporia.

Sugar beet Cercosporia epidemiologic data

Incidence levels of Sugar beet Cercosporia (SBC) were documented across multiple fields in France through the efforts of several advisory organizations, among them the ITB (Institut Technique de la Betterave). These field observations span 2009 to 2020 and encompass various regions of France.

Within each observed location, a defined portion of the field—subsequently termed the “plot”—underwent monitoring over the course of one particular year. Assessments were conducted weekly, involving visual examinations of leaves sampled from 100 plants to determine disease incidence rates. Incidence was measured as the percentage of sugar beet leaves exhibiting signs of Cercospora leaf spot (*Cercospora beticola*). These weekly evaluations were conducted in each plot from leaf emergence (typically around mid-May) through to harvest, which occurs after mid-September. In total, the compiled dataset includes 1235 distinct plots. It is important to note that no individual plot received monitoring in all years, and that not every plot could be assessed during any one given year.

For each plot, the SBC onset date (corresponding to the year of symptom appearance) is identified as the first day on which the share of infected leaves exceeded 10%. The end-of-season incidence for SBC corresponds to the highest incidence value recorded between the 25th of August and the 15th of September.

Grape downy mildew epidemiologic data

Incidents of Grape downy mildew (GDM) were monitored across various vineyards in France through multiple wine advisory organizations, such as the IFV (Institut Français de la Vigne et du Vin). The collection of this information

© 2024 The Author(s). | Published by Agro Intelligence Research Press

occurred between 2010 and 2017.

In every selected plot, experts tracked an untreated line of vines. This untreated line was flanked on both sides by two more untreated lines to avoid any unintended fungicide application. Weekly visual evaluations were conducted on leaves within the central observed line to quantify disease incidence. The incidence value reflected the fraction of vine leaves exhibiting symptoms of downy mildew triggered by *Plasmopara viticola*. These weekly checks were conducted in each vineyard beginning at budburst (early March) and continued until at least bunch closure (mid to late July), or ceased once incidence neared 100%. Overall, the records include nearly 9407 weekly entries tied to 713 plots.

Per plot, specialists identified the GDM onset date (yearly symptom appearance date) as the opening week in which the fraction of infected vine leaves exceeded 1%. End-of-season incidence for GDM corresponded to the highest incidence level noted for that individual plot.

Meteorological data

The SAFRAN weather database supplied the meteorological indicators; this resource was built and updated by France's national meteorological agency (Météo-France). SAFRAN segments the entire French landscape into an 8×8 km grid and maintains meteorological records for all cells within it [15]. Information collected daily on humidity levels, average temperature, wind speed, rainfall quantity, and solar radiation helped generate multiple meteorological indicators applicable to both diseases under study.

Regarding SBC, every meteorological indicator encompasses a 15-day half-month window ranging from January to June. The features within the dataset obey a consistent naming pattern. The opening segment denotes the time-related detail via the initial three letters of the pertinent month, then appends 'A' to signify the first half of the month or 'B' for the second half. The following segment clarifies the climatic property along with its computation approach. Descriptions of the feature suffixes appear in **Table 1**. For example, the indicator JanA-ndRHm60 indicates the number of days (nd) when relative humidity (RHm60) remained above 60 percent during the first half (A) of January (Jan).

Table 1. Description of the meteorological variables used to model the dynamics of the Sugar beet Cercosporia (SBC). Temperatures are considered inhibiting below 10 °C or above 38 °C.

Name	Rephrased feature
RHmX	Average relative humidity falls below X, where X ∈ {60, 65, 80, 90}.
H87	The humidity index is equal to 87.
H87Y	Humidity index remains at 87 for a minimum of Y hours, where Y ∈ {6, 10}.
TmX	Mean temperature exceeds X, where X ∈ {15, 20}.
TmXTinfYZ	Mean temperature is above X (X = 15) while staying below Y (Y = 10) for at least Z hours, where Z = 3.
TblоX	Count of days during which temperatures were considered restrictive to SBC growth for longer than X hours, where X ∈ {3, 6}.

In the case of GDM, indicators either reflect meteorological states exactly on the recording date or accumulate totals over the prior four weeks. For instance, the input variable ETP reflects evapotranspiration on the actual recording occasion. ETP-4w stands for the total evapotranspiration sum for the four earlier weeks. Exceptions apply to the counts of rainy days and dry days, which start accumulating from the first day of January. Experts recommended these four weeks based on their understanding of how quickly downy mildew spreads.

Four prediction targets

© 2024 The Author(s). | Published by Agro Intelligence Research Press

Merging information from both diseases, along with the linked climatic indicators, produced 4 datasets that match our 4 prediction targets.

• Sugar beet Cercosporia (SBC) end-of-season incidence (% of leaves with disease) with 1235 plots and 367 variables, including one categorical variable and 366 numeric ones. The categorical feature is the risk exposure, an indicator defined by agronomists based on their own knowledge of each plot's sensitivity to SBC. The numerical variables correspond to the ones described in Subsection 2.1.3.

• Sugar beet Cercosporia (SBC) symptoms appearance date (day number of year) with 1235 plots and 367 variables.

• Grape downy mildew (GDM) end-of-season incidence (% of sick leaves) with 359 plots and 22 variables, including two categorical and 20 numeric.

• Grape downy mildew (GDM) symptoms appearance date (week number of year) with the same 359 plots and 22 variables.

Thus, the target variables are numerical. We are thus confronted with a regression problem in all cases.

Regression methods

We assume the goal is to predict the values of a real-valued target variable using observations from other variables, the predictive variables. Examples of target variables are given in Subsection 2.1.4. Conversely, the predictive variables constitute the set of meteorological indicators **(Table 1)**. This scenario constitutes a classical regression problem. We first introduce some notation and then survey the most popular regression methods used in crop protection on tabular data. We extend the discussion with a description of a pattern-aided regression method that addresses the complexity-accuracy trade-off introduced in the previous sections.

Problem formulation and notation

Let us assume that we count on a set of n target observations represented as a column vector y∈R^n. Those target observations are associated with a set of observations on the predictive variables, organized in the matrix X ∈ R^ (n×d). Each row x_i^⊤∈R^d in the matrix stores the observed values of the d predictive variables associated with a target observation yi. From now on, we denote vectors and matrices using boldface names to distinguish them from scalars and functions. Moreover, matrices are denoted with capital letters. If a predictive variable is categorical, e.g., plant variety, we assume its values have been encoded as real numbers, for instance, via one-hot encoding or dimensionality reduction.

The goal of regression analysis is to learn a function f such that y=f(X)+ϵ and ϵ is minimal. The function f is a model of the data designed to predict the target variable for unseen instances x^⊤∈R^d of the predictive variables. The term ϵ is the error of the regression model and accounts for potentially unobserved predictors of y. The model f is learned on a set of training and validation observations.

Classical regression methods

Linear regression

This method assumes that the relation between the target variable y and the predictive variables X is linear, that is,

$$\boldsymbol{y} = \boldsymbol{\beta X'} + \boldsymbol{\epsilon} \text{ with } \boldsymbol{\beta} = \boldsymbol{argmin}_{\hat{\boldsymbol{\beta}}} ||\boldsymbol{y} - \boldsymbol{X'}\hat{\boldsymbol{\beta}}||_2^2 \quad (1)$$

© 2024 The Author(s). | Published by Agro Intelligence Research Press

$$\boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \quad (2)$$

Where $\mathbf{X}' = 1 \oplus \mathbf{X}$, i.e., $\mathbf{X}' \in \mathrm{R}^{\mathrm{n}\times(\mathrm{d}+1)}$ and $\boldsymbol{\beta} \in R^{d+1}$ are the parameters of the model (the operator ⊕ denotes column concatenation), namely the linear coefficients associated with each of the d predictive variables, plus the intercept coefficient β_0. The parameters of the model can be computed by minimizing the loss function $L_l\left(\widehat{\beta}\right) = ||\boldsymbol{y} - \mathbf{X}'\widehat{\beta}||_2^2$ with the method of ordinary least squares (OLS) as illustrated in **Eq. 2**. Linear models are among the most popular regression methods due to their simplicity and interpretability. This is because the magnitude of the coefficients tells us explicitly how much a predictive variable contributes to the model's prediction. On the downside, the linearity assumption may come at the expense of low prediction accuracy, which is why linear models are often used as baseline methods.

Lasso

To reduce the risk of overfitting in linear regression, Lasso [16] proposes L1-regularization of the loss function, which favors models with few nonzero coefficients. This is achieved by minimizing the following objective:

$$\boldsymbol{\beta} = argmin_{\hat{\boldsymbol{\beta}}}\ \mathscr{L}_l(\hat{\boldsymbol{\beta}}) + \theta||\hat{\boldsymbol{\beta}}||_1. \quad (3)$$

By minimizing the L1-norm of β, we can obtain sparse models that not only prevent or mitigate overfitting but are also less complex and therefore easier for humans to inspect. The penalization term θ is a hyperparameter that controls the importance of the sparsity constraint in the optimization process. The Lasso method selects the set of parameters. $\widehat{\beta}$ that achieves the highest cross-validation performance.

Decision/regression trees

A decision tree is a binary tree in which each internal node tests a Boolean condition associated with a single predictive variable. Offspring nodes represent further decision trees linked to each possible evaluation result (true or false). Terminal leaves, also called final nodes, are directly associated with a model prediction for the target variable. In situations where the target variable takes numerical values, the structures are called regression trees [17]. Regression trees are white-box models, since the prediction for any specific instance x⊤∈Rd can be explained by tracing the path from the root to the responsible leaf node. This characteristic supports the interpretability of regression trees, provided their depth remains suitable for human review. Even with this interpretability advantage, decision trees frequently overfit unless their parameters are properly tuned, and ensemble techniques such as random forests or gradient boosting trees often surpass them in predictive performance.

Random forests

Random Forests consist of assemblies of basic decision tree predictors [18]. Overall predictions are obtained by averaging the outputs of all trees in the assembly. Construction of these basic predictors uses bagging and randomized feature selection. During bagging, each tree is trained on samples drawn uniformly with replacement from X and y. Trees also train on different feature subsets, thereby acquiring individual, though partial, perspectives of the full data. These procedures provide random forests with considerable resistance to overfitting, making them a common choice in crop protection work [19]. Their drawback lies in limited interpretability, since the averaging operation renders it hard to connect model outputs directly back to input features, absent later post-hoc analysis tools.

Gradient boosting

© 2024 The Author(s). | Published by Agro Intelligence Research Press

Gradient boosting [20] qualifies as yet another common ensemble approach. It shares with random forests the core idea of building a strong prediction from multiple weak learners. However, it differs by relying on an additive framework in which successive learners hm train specifically on the residuals left by the earlier learner hm−1, effectively addressing the negative gradient associated with the loss function under minimization. Stated another way, each added learner aims at rectifying shortcomings of those preceding it:

$$f_m(\boldsymbol{X}) = f_{m-1}(\boldsymbol{X}) + \gamma_m h_m(\boldsymbol{X}) \quad (4)$$

$$\gamma_i = \mathscr{L}(\boldsymbol{y}, f_m(\boldsymbol{X})) \quad (5)$$

Learners themselves may adopt any form, yet decision trees appear most frequently in this capacity [21]. Gradient boosting systems exhibit robustness to overfitting and operate essentially as black boxes, comparable to random forests.

Hierarchical pattern-aided regression (HiPaR)

Pattern-aided regression

Regression approaches built on patterns involve collections of localized models fitted to particular data segments. These segments are defined by understandable patterns formed by logical conditions applied to predictive variables, such as wind_speed > 50. Localized models typically take interpretable forms, such as linear functions, that uncover region-specific associations between target and predictive variables that are invisible at the global scale. Research has demonstrated [10, 13] that these techniques achieve superior predictive performance compared with linear regression while incurring only a modest increase in complexity. Notable pattern-aided regression examples encompass piecewise regression [21], regression trees [18], model trees [22], Contrast pattern-aided regression (CPXR) [13], and HiPaR [10]. Further explanation of HiPaR follows below.

HiPaR

Hierarchical Pattern-aided Regression [10] produces estimates of the target variable through a compact collection of localized hybrid rules defined over predictive variables. Rules adopt the following format:

$$p = C_1 \wedge \ldots C_m \Rightarrow \boldsymbol{y} = f_p(\boldsymbol{X}_p). \quad (6)$$

Within this format, pattern p joins multiple conditions on predictive variables, for instance, wind-speed>50∧, humidity>30. Conditions of this kind delineate subsets, or zones, denoted Xp⊂X within the overall data. A hybrid rule connects to a localized linear model fp trained exclusively on Xp that enhances outcomes from a global linear model, f, fitted across the entire X. Termed the default model, f supplies predictions in any case where none of the localized hybrid rules match. Following default model construction, HiPaR discovers a limited group of hybrid rules via two successive stages:

1. The enumeration stage has the algorithm traverse the pattern space p using a depth-first hierarchical sequence. Encountering a pattern p leads HiPaR to build a hybrid rule p⇒y=fp(Xp) from observations inside Xp that fulfill the pattern, after which it examines narrower zones within Xp. Because the search space expands exponentially with the number of features, pruning techniques shrink the exploration space by bypassing unpromising, narrower zones; one common approach imposes a minimum support threshold to avoid zones with too few observations.

2. Pruning during enumeration still risks producing an overly large set of hybrid rules. HiPaR therefore applies a selection stage to retain only a small number of strong-performing hybrid rules exhibiting minimal overlap. This stage operates

© 2024 The Author(s). | Published by Agro Intelligence Research Press

under two tunable parameters: support and overlap bias. These govern the preference between highly detailed rules and more general ones, as well as the permitted degree of overlap among retained rules.

In contrast to tree models, HiPaR hybrid rules originate from a hierarchical arrangement allowing overlapping zones, illustrated in **Figure 1**. When an incoming observation satisfies several hybrid rules simultaneously, the ultimate prediction equals a weighted average drawn from those rules' individual outputs. Weight assignment occurs inversely to each rule's error rate calculated on a validation segment. Such a mechanism boosts HiPaR's robustness relative to plain linear functions and regression trees, though it also noticeably increases overall complexity. Still, HiPaR hybrid rules remain white-box models, allowing straightforward examination of the key predictive variables shaping predictions for any observation.

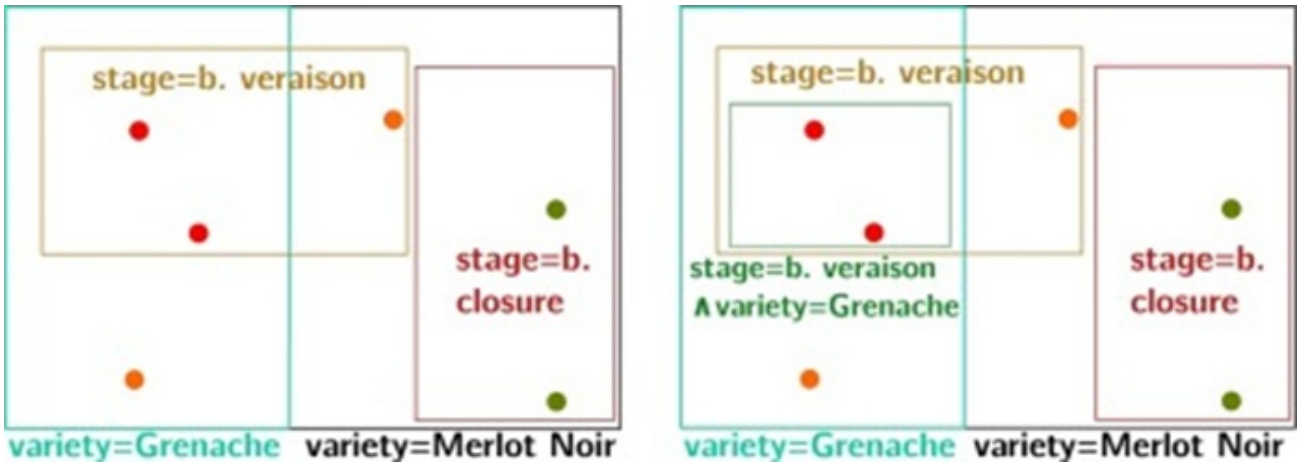


Figure 1. A depiction of the regions explored by HiPaR for two steps of the enumeration phase. Each rectangle defines a region described by a pattern, and HiPaR learns a local regression model over it. Regions can overlap; an example is the regions stage = "b. veraison" and variety = "Grenache". Once a region is explored, e.g., stage = "b. veraison", HiPaR will look at its sub-regions in a depth-first-search manner (figure on the right).

Table 2 summarizes the strengths and weaknesses of the methods discussed in this section.

Table 2. Overview of the machine learning methods used in this study.

Models	Rephrased characteristics	Rephrased advantages	Rephrased disadvantages
Lasso	Linear regression approach with sparse representation	Straightforward and easy to interpret	Serves mainly as a baseline technique
HiPaR	Model relying on pattern extraction	Moderately complex methodology	Requires substantial computation time
Random Forests	Ensemble method based on decision trees	Strong predictive performance with integrated feature importance measures	Functions as a black-box approach
Gradient Boosting	Ensemble learning technique	Provides high predictive accuracy	Considered a black-box model

Training and testing procedures

Optimization and performance evaluation

© 2024 The Author(s). | Published by Agro Intelligence Research Press

A major difficulty in comparing multiple machine learning models is choosing configurations that maintain balanced, informative comparisons. Given the large year-to-year differences typical in agronomy, ordinary cross-validation is ill-suited. We therefore adopt year-wise cross-validation, treating each year as a separate fold. Observations from one selected year comprise the test portion, while those from all remaining years support model training. This setup evaluates how effectively algorithms handle genuinely new situations, such as forecasting for a previously unseen year.

Inside each fold, hyperparameter optimization determines the optimal configuration for each technique. HiPaR's enumeration stage may demand substantial time with extremely low support thresholds. We therefore conduct enumeration once at a support threshold of 30% of the dataset size, skipping exploration of smaller zones, and then optimize only the selection-stage hyperparameters to secure the best rule collection.

We rely on the coefficient of determination (R2) to gauge the quality of predictions. R2 expresses the proportion of variance in the target variable that the predictors account for. In contrast to root mean square error (RMSE), R2 enables straightforward comparison across separate prediction tasks (for instance, disease incidence versus symptom appearance timing). Values of R2 near 1 indicate an excellent fit of the model to the data. Scores around zero match results achievable by always predicting the target average, and negative scores indicate underperformance relative to that elementary mean-based predictor.

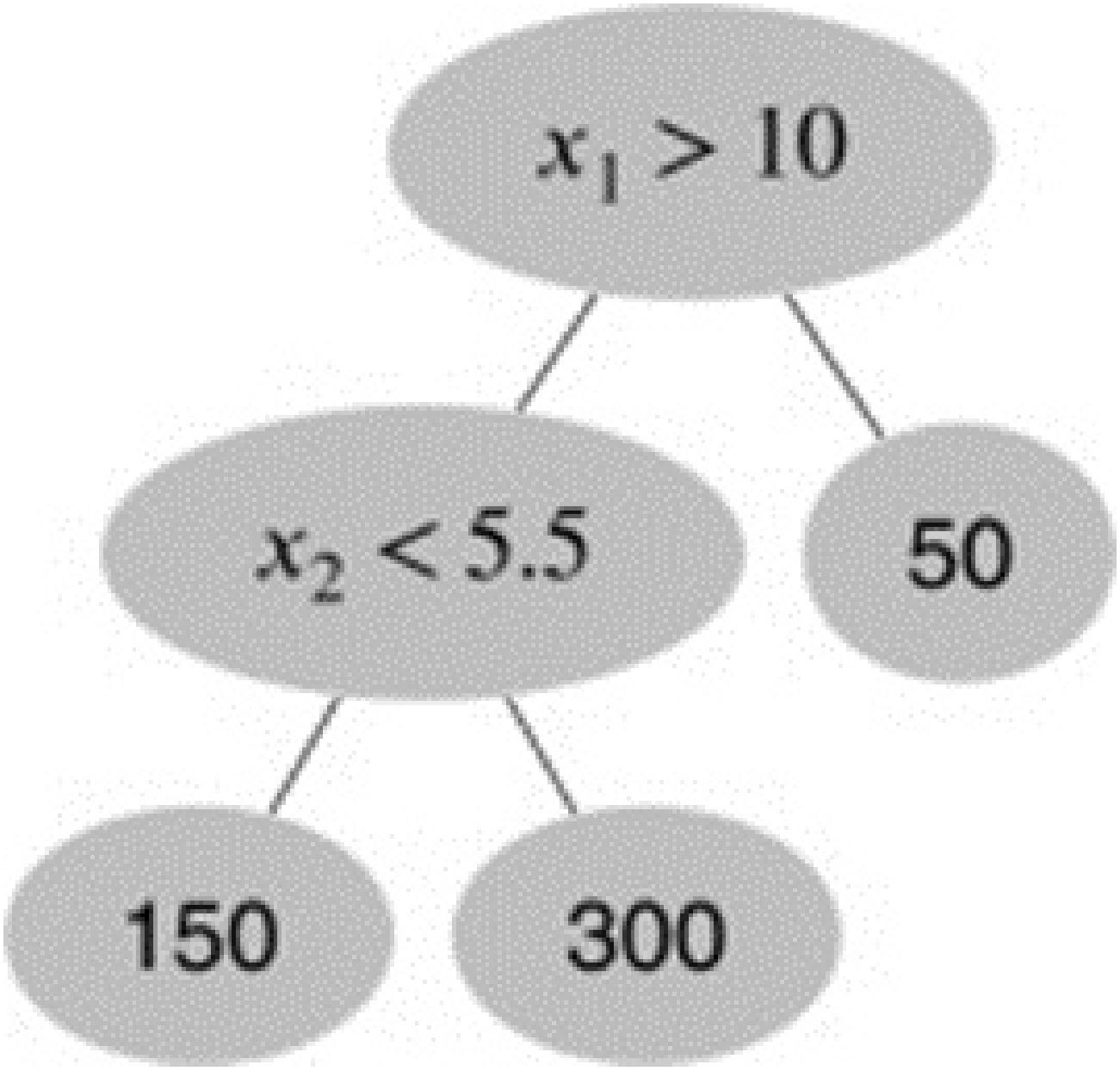
Complexity measure

To assess the complexity of the machine learning models under examination, we use the complexity metric for pattern-based approaches introduced by Galárraga *et al.* [10], which tallies the total number of components in each model. A component consists of either a nonzero coefficient or a condition placed on a predictive variable. Notably, this same metric extends to tree-based techniques, including random forests and gradient boosting trees, since every node inside every tree within the ensemble represents either a condition on one attribute or a linear model – in the case of basic regression trees, that linear model reduces to a single constant value. The overall element count can grow substantially when ensembles contain numerous trees, underscoring the high complexity of such systems.

Following this logic, a Lasso model typically exhibits lower complexity than a HiPaR model containing multiple rules. This difference arises because Lasso counts only the nonzero coefficients in its linear equation, while HiPaR accounts for both the conditions and the coefficients in each localized model.

If we examine the following regression tree T:

© 2024 The Author(s). | Published by Agro Intelligence Research Press



Then its complexity c(T) is 5. Similarly, if we examine the rule R:

$$C_1 \wedge C_2 \wedge C_3 \wedge C_4 \Rightarrow \boldsymbol{y} = 3x_1 + 4x_2 - 4x_3 + 8, \quad (7)$$

Then its complexity c(R) is 8 because the rule includes 4 conditions and 4 linear coefficients.

Results and Discussion

Results

Performance-complexity trade-off

© 2024 The Author(s). | Published by Agro Intelligence Research Press

Figure 2 illustrates the balance between complexity and accuracy across the machine learning techniques examined. The x-axis displays model complexity (using a logarithmic scale). The y-axis shows the median R2 coefficient achieved by each model during cross-validation. Models appearing in the upper-left region of the plot achieve a superior accuracy-complexity balance because they deliver stronger predictions using fewer components. As noted by Galárraga *et al.* [10], highly complex models such as random forests or gradient boosting trees reach top performance levels but only at the cost of substantial complexity. Lasso regression, serving as our reference point, often proves the least accurate option. HiPaR occupies an intermediate position between linear regression and ensemble techniques, offering a notably attractive balance across 3 of the 4 prediction tasks.

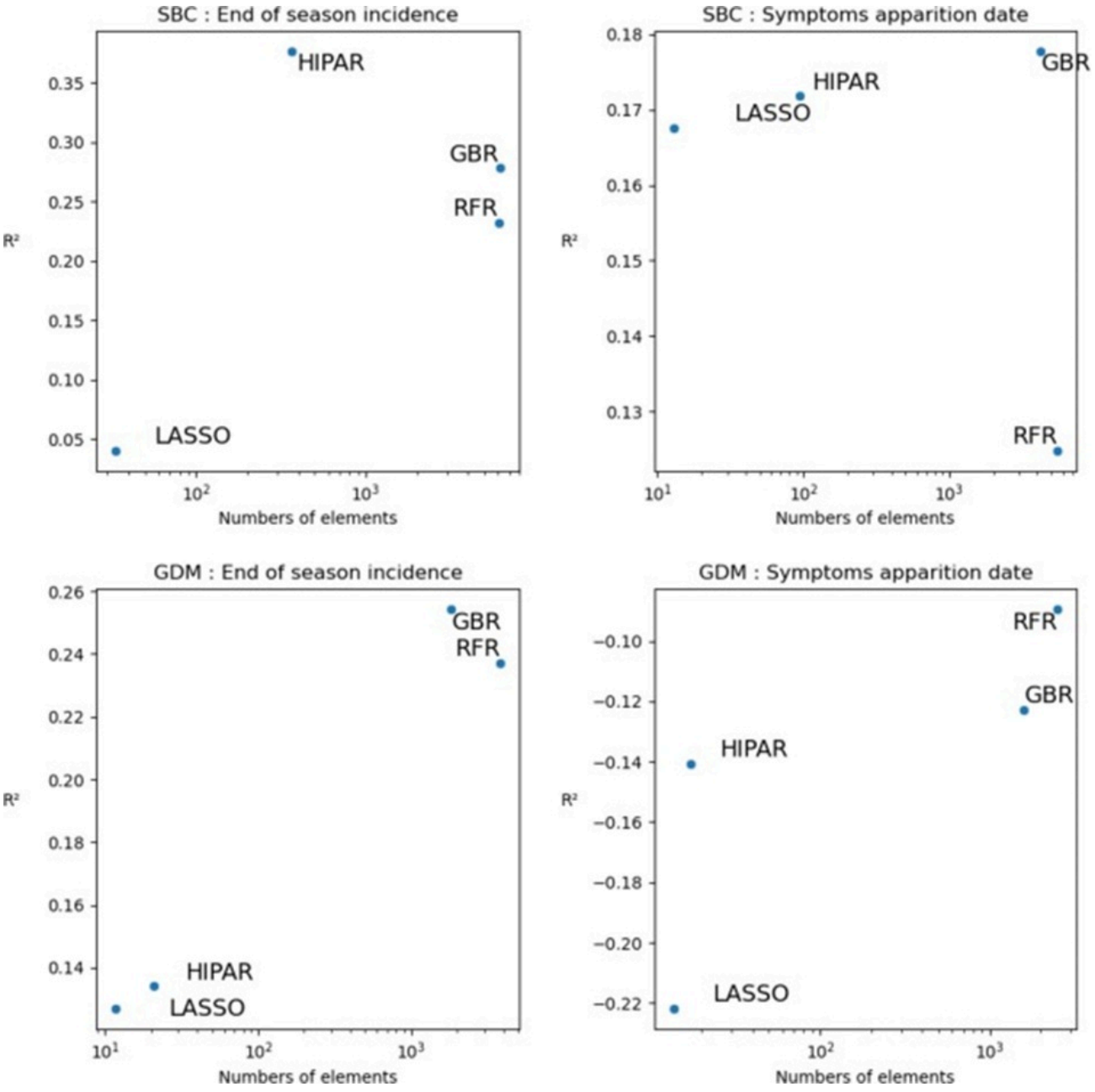


Figure 2. R2 values for different machine learning models, plotted against their complexity. The x-axis corresponds to the number of elements that compose each model (log scale). The y-axis is the median R2 values in cross-validation. GBR stands for gradient boosting regression, and RFR for random forests regression.

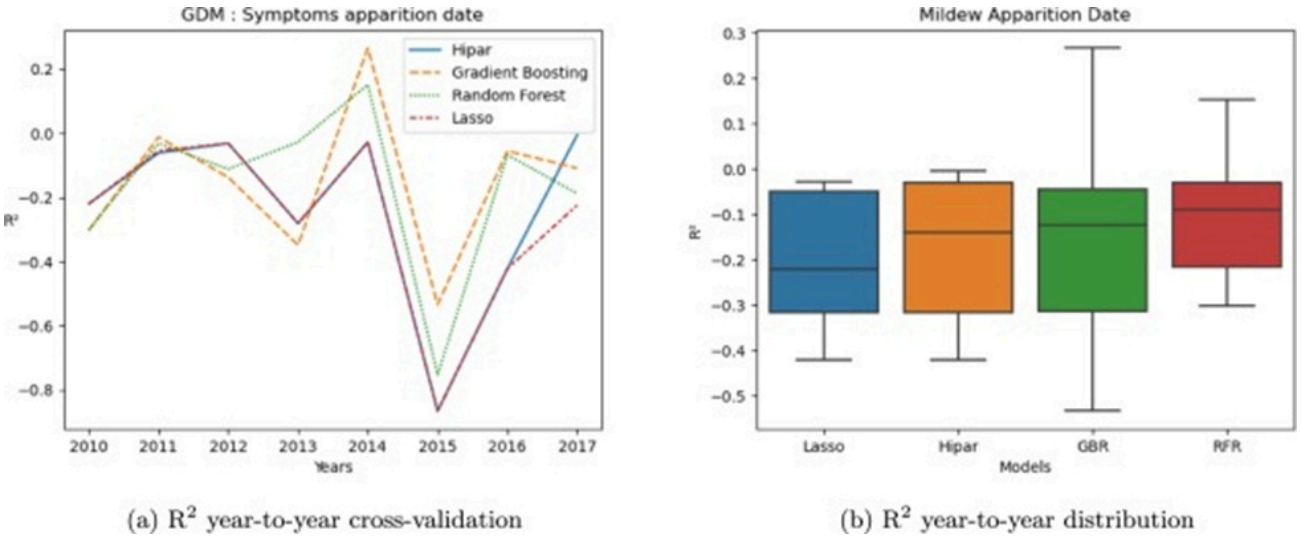
© 2024 The Author(s). | Published by Agro Intelligence Research Press

We note that accuracy levels differ sharply across tasks. Every model struggles to forecast the appearance date of downy mildew within vineyards, as the median R2 across all techniques remains negative (bottom-right panel). We record R2 values ranging from 0.12 to 0.26 for final downy-mildew incidence (bottom-left panel), with gradient boosting achieving the highest R2. HiPaR stays near Lasso here, indicating it identified a few regression rules that offered only slight gains beyond the reference. The performance gap between the two target variables in the downy-mildew data might stem from the comparatively small number of observations available for the symptoms' appearance date (359), compared with 700 for end-of-season incidence.

The influence of the aggregated variables also stays somewhat restricted. This refers to most variables spanning 4 weeks before data collection. Although this reinforces the observed trade-off, the modest R2 values render this dataset less suitable for deeper investigation.

Outcomes for sugar beet Cercospora appear more promising. Median R2 scores for the appearance date range from 0.13 to 0.18, with gradient boosting at the forefront and HiPaR in second place (top-right panel). For end-of-season incidence prediction, performance spans from 0.05 to 0.35. In this scenario, HiPaR surpasses all other techniques and discovers a substantial number of rules that markedly improve performance beyond a single linear model, while avoiding the elevated complexity of ensemble methods.

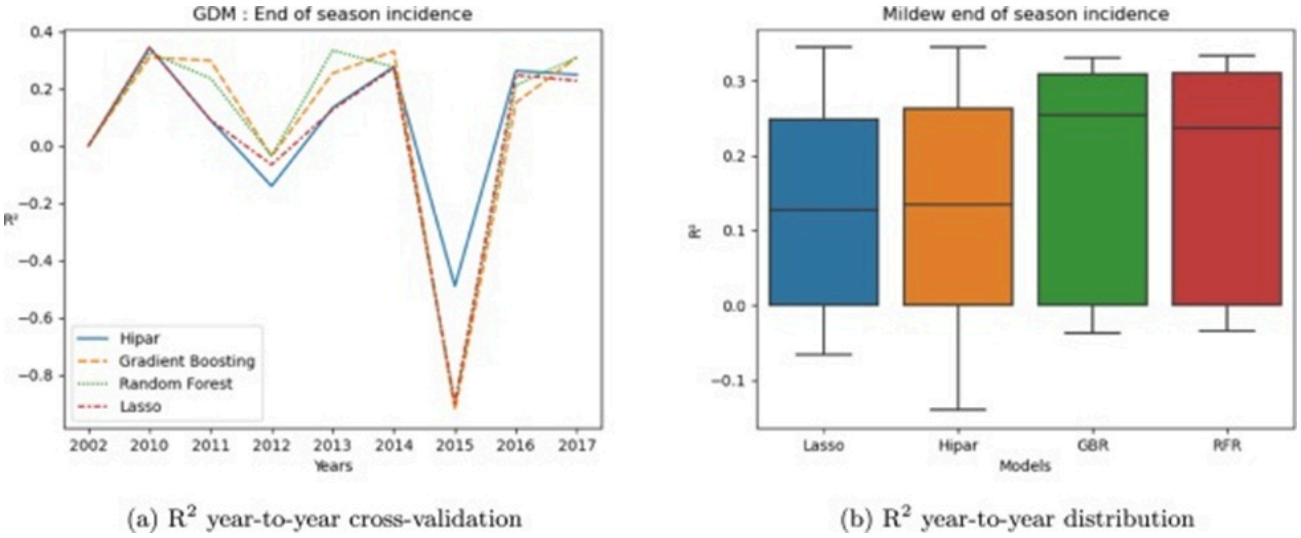
Examination of method performance broken down by year **(Figures 3-6)** reveals that results can fluctuate dramatically from year to year, and that both end-of-season incidence and appearance date are particularly challenging to model in certain years. This pattern holds across all techniques. Overall, Cercospora end-of-season incidence forecasts display a general decline in quality over time. Variability in performance among folds **(Figures 3-6)** remains comparable between methods and shows no clear pattern.



(a) R^2 year-to-year cross-validation (b) R^2 year-to-year distribution

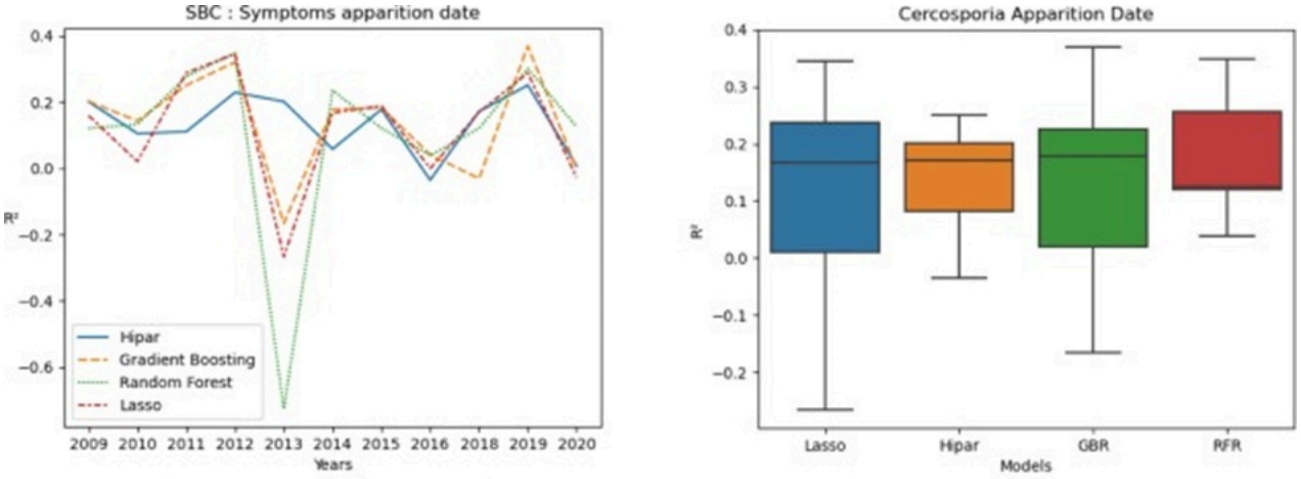
Figure 3. Mildew symptom appearance date.

© 2024 The Author(s). | Published by Agro Intelligence Research Press



(a) R^2 year-to-year cross-validation (b) R^2 year-to-year distribution

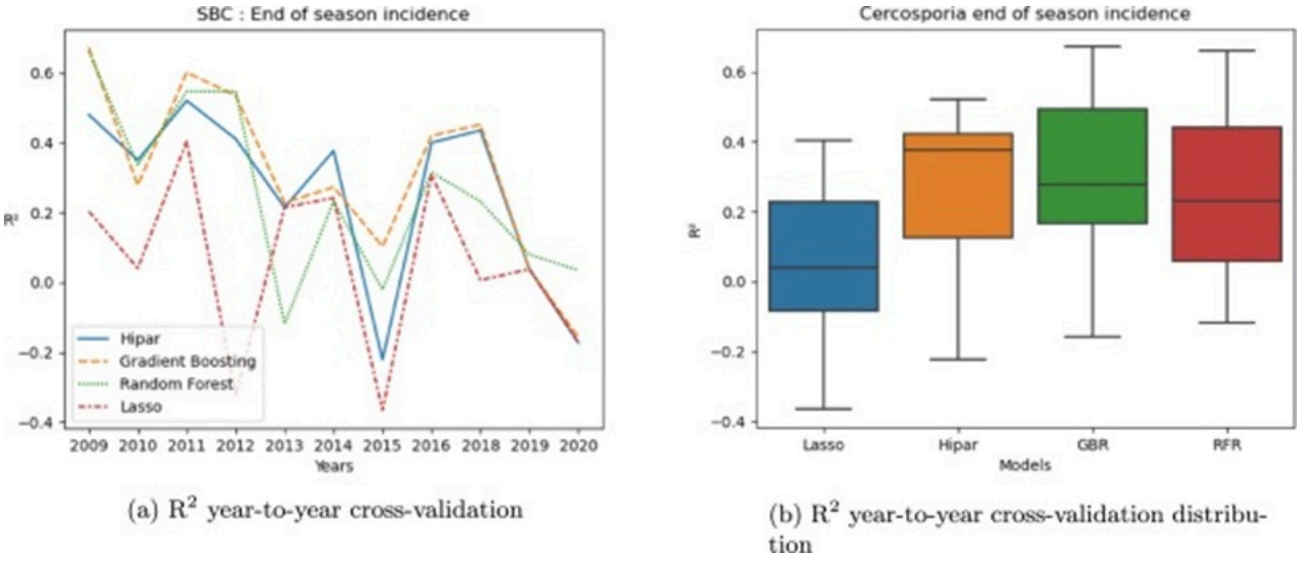
Figure 4. Mildew end-of-season incidence.



(a) R^2 year-to-year cross-validation (b) R^2 year-to-year cross-validation distribution

Figure 5. Cercosporia symptoms date of apparition.

© 2024 The Author(s). | Published by Agro Intelligence Research Press



(a) R^2 year-to-year cross-validation

(b) R^2 year-to-year cross-validation distribution

Figure 6. Cercosporia end-of-season incidence.

Having demonstrated the accuracy-complexity trade-off within our selected scenarios, we now explore the insights embedded in the various methods. To achieve this, we review the models developed for predicting the year 2009 regarding the end-of-season incidence of sugar beet Cercospora, given that these models deliver the strongest explained variance among all years (R2 scores of 0.67 and 0.66 for gradient boosting trees and random forests, 0.3 for Lasso, and 0.47 for HiPaR). For transparent white-box models such as Lasso and HiPaR, we perform a straightforward examination of their components. For the intricate black-box systems, we apply standard model-interpretation tools and check whether the models agree on the connections between predictive and target variables.

Use case: incidence of the sugar beet Cercospora

This section presents an interpretability effort to extract agronomic insights from the experimental machine learning models built to forecast sugar beet Cercospora incidence. These models were trained every year except 2009 and yielded the strongest cross-validation results from our tests. We employ established interpretation tools such as feature importance rankings, partial dependence plots, and straightforward rule examination. The initial tool reveals the most influential variables shaping predictions. PDPs, along with rule examination, help uncover threshold-based effects on predictive variables, i.e., situations in which target-variable behavior shifts across segments according to specific cutoffs on the predictors. Pattern-based regression techniques, such as HiPaR, excel at detecting such effects. In addition, these techniques enable the analysis of more detailed interactions among the predictive variables contained within the rules. Our findings lay the foundation for the discussion presented in Section 3.

Feature importance

One straightforward approach to interpreting a machine learning model's knowledge is to create a feature-importance ranking that indicates the degree to which each input variable influences the model's outputs. This ranking may rely on genuine contributions of variables to model responses, for example, the coefficients from linear regression, or on model-specific scores calculated afterward for black-box models. Following this idea, we compare feature-importance rankings from Lasso, RFR, and GBR as shown in **Figure 7**. Lasso linear coefficients directly represent the contributions of input features to model outputs and therefore carry signs. To convert these into importance measures, we use their absolute values. In contrast, RFR and GBR rely on tree ensembles for which several importance metrics exist. We selected the permutation feature importance approach available in the scikit-learn library. This technique gauges feature importance by randomly shuffling its values across rows within X. The resulting drop in accuracy then quantifies the model's dependence on that feature for generating predictions – larger drops indicate greater informativeness for target-variable forecasting.

© 2024 The Author(s). | Published by Agro Intelligence Research Press



Figure 7. Parallel coordinates chart comparing feature-importance rankings for Lasso, random forests, and gradient boosting trees when predicting end-of-season incidence for the sugar beet Cercospora. For each model, we selected the top 4 most important features. For each model, features below the 4 can be further down the importance order than what is displayed.

As observed, RFR and GBR produce highly comparable rankings – their top-4 variables match exactly, even if ordering differs slightly. The variable threshold-1 emerges as the leading feature for all three models. This variable indicates the day when initial Cercospora symptoms appeared in the field. On the other hand, RFR and GBR performance depend on the risk-zone expert indicator, which holds less weight for linear regression. While importance measures indicate which information draws the model's attention, they do not indicate whether those features raise or lower incidence predictions. We can gather this detail, however, by reviewing the Lasso coefficients.

Table 3 shows the top 5 most important linear coefficients.

Table 3. Top 5 important linear coefficients learned by Lasso.

Variable	Coefficient
Threshold-1	-41
AprA-STm10	27.64
JunA-ndW2	-23.39
JunB-STm10	22.24
AprB-ndRHm60	14.73

We remind the reader of the meaning of these variables:

- **Threshold-1**: The date of the symptoms' appearance
- **AprA-STm10**: The sum of the daily average temperatures of the days above 10 °C during the first half of April.
- **JunB-ndW2**: The number of days in the first half of June such that the average wind speed was higher than 2 m/s.
- **JulB-STm10**: The sum of the daily average temperatures above 10 °C during the second half of June.
- **AprB-ndRHm60**: The number of days in the second half of April such that the relative humidity is higher than 60%.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

Table 3 indicates that delayed symptom onset correlates with reduced final incidence. Predicted incidence rises alongside higher temperatures and humidity during June and April, while stronger winds appear to limit Cercospora spread. These outcomes require cautious interpretation since our reference Lasso model accounts for only 30% of the variance in the target variable. Nevertheless, the same variables appear in stronger models, such as RFR and GBR, confirming their relevance to the target variable.

Threshold effects

Pattern-based regression techniques, as noted previously, are designed specifically to uncover how thresholds on predictive variables influence the target variable. HiPaR expresses these influences directly through its rule conditions. To check whether our models had captured such influences, we conducted a thorough review of the hybrid rules produced by HiPaR in the current scenario. We compared the identified thresholds with those captured by the more advanced models, RFR and GBR. Given that these models rely on extensive collections of threshold-oriented estimators, we investigated the influence of thresholds using partial dependence plots (PDPs). This standard inspection method helps display how model predictions (y-axis) respond to changes in a single predictive variable (x-axis).

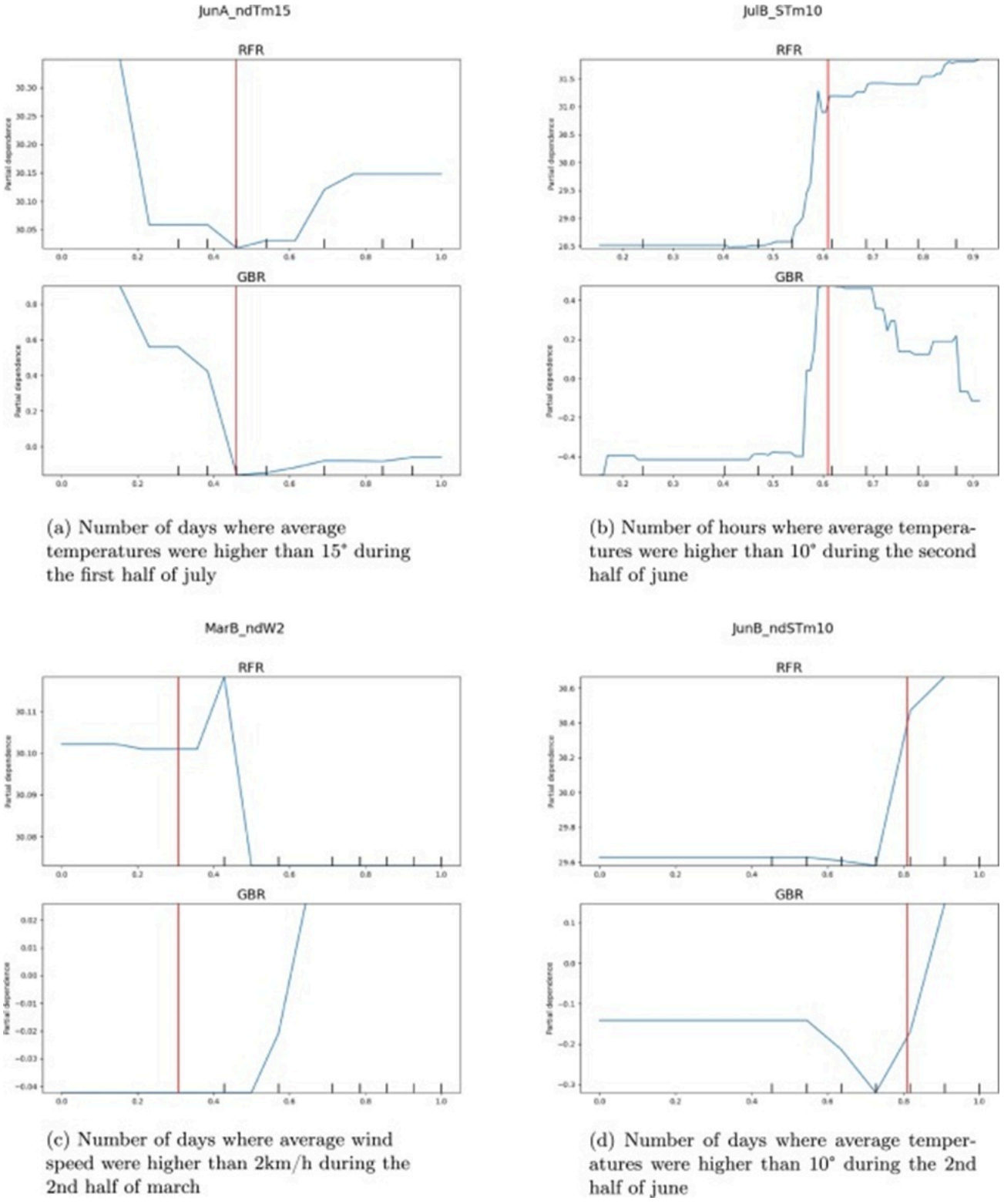
In the scenario examined here, HiPaR generated 3 hybrid rules, whose conditions are presented in **Table 4**. As seen earlier, the thresholds (highlighted in red) within HiPaR's rules correspond fairly well to shifts observed in PDP behavior. Even though they are not the dominant features identified earlier, this correspondence suggests that the thresholds are relevant (as assessed by RFR and GBR). We believe these features may function as stand-ins for other variables or exert subtle indirect effects on outcomes that PDPs cannot easily capture.

Table 4. Conditions of the hybrid rules learned by HiPaR when predicting the end-of-season incidence of the sugar beet Cercospora.

Rule 1	*JunA-ndTm15* < 8, *risk-zone=false*
Rule 2	*JulB-STm10* < 324
Rule 3	*MarB-ndW2* ≥ 4, *JunB-ndTm15* < 13

Put simply, HiPaR identified separate linear response patterns depending on whether a plot belongs to an expert-designated high-risk area (risk-zone) or not, whether hot days in June and July remain under defined limits (JunA-ndTm15, JulB-STm10, JunB-ndTm15), or whether strong winds occurred in the latter part of March (MarB-ndW2). We subsequently created PDPs for the numerical variables in RFR and GBR, which appear in **Figure 8**.

© 2024 The Author(s). | Published by Agro Intelligence Research Press



(a) Number of days where average temperatures were higher than 15° during the first half of july

(b) Number of hours where average temperatures were higher than 10° during the second half of june

(c) Number of days where average wind speed were higher than 2km/h during the 2nd half of march

(d) Number of days where average temperatures were higher than 10° during the 2nd half of june

Figure 8. Partial Dependence Plots for the predicting variables MarB-ndW4, JulB-STm10, JunA-ndTm15, and JunB-STm10 on random forests and gradient boosting trees. The red line represents a threshold learned by HiPaR.

Feature interactions

© 2024 The Author(s). | Published by Agro Intelligence Research Press

Each condition shown in **Table 4** links to a localized linear model (fitted using Lasso). These local models expose region-specific interactions among the conditioned variables and their associated coefficients, aiming to improve upon the baseline linear model (also referred to as the default) trained on the full dataset. From the 368 input features available to the models, Lasso keeps between 25 and 55. This amounts to between 6.7% and 15% of all available features. In addition, the localized models prove consistently simpler than the default model, as detailed in **Table 5**.

Table 5. Number of common nonzero coefficients of the linear models learned by HiPaR for the prediction of the end-of-season sugar beet Cercospora.

Empty cell	Rule 1	Rule 2	Rule 3	Default Model
Rule 1	25	8	3	6
Rule 2		28	6	16
Rule 3			26	12
Default Model				55

We also find limited overlap in coefficients among the various hybrid rules. This suggests each localized model draws upon unique signals when estimating end-of-season incidence. **Figure 9** presents the magnitude and direction of 16 such coefficients across both the default and localized models.

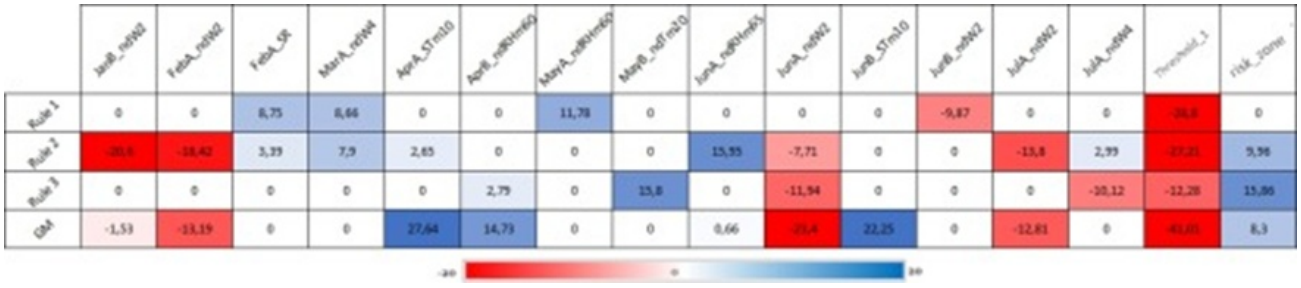


Figure 9. A color encoding for the linear coefficients of the three hybrid rules learned by HiPaR. Cells in white □ denote features with a linear coefficient strictly equal to 0, which means the model doesn't use those features.

The first key finding is that the appearance date (threshold-1) remains highly relevant across all models, consistently showing a negative relationship with predicted incidence. The variables risk-zone and JunA-ndW2 (wind speed during the first half of June) feature in all models except the initial rule, as they form part of that rule's conditional component **(Table 4)**.

This particular rule can be read as follows: In plots featuring reduced disease vulnerability (risk-zone=false) paired with cooler conditions in early June (JunA-ndTm15 < 8), Cercosporia spreads more aggressively when May humidity (MayA-ndRHm60), March wind speed (MarA-ndW4), and February rainfall (FebA-SR) rise. June wind levels (JunB-ndW2) correspond to slower disease advancement.

The second rule implies that cooler July temperatures (JulB-STm10 < 324) increase Cercosporia vulnerability to wind across January, February, June, and July (JanB-ndW2, FebA-ndW2, JunA-ndW2, JulA-ndW2). Meanwhile, a moist June (JunA-ndRHm65) or windy March (MarA-ndW4) acts as a promoting element. Strong July winds (JulA-ndW4), February rain (FebA-SR), and warmer April conditions (AprA-STm10) exert a restraining influence on Cercosporia expansion.

The third rule becomes active under windy March conditions (MarB-ndW4 ≥4) and moderate June heat (JunB-ndTm15 < 13). Here, warmer May temperatures (MayB-ndTm20) and humid April weather (AprB-ndRHm60) are associated with

© 2024 The Author(s). | Published by Agro Intelligence Research Press

greater Cercosporia development. In contrast, June winds (JunA-ndW2) and July winds (JulA-ndW4) show an inverse relationship with growth.

In the end, the default model integrates signals from all localized rules, though it does not always use the same variables. This integration stems from the need to accommodate observations spanning every sub-region during training. As a result, it favors variables (such as JunB-STm10) that provide broad explanatory value across the entire dataset at a global scale, yet contribute little explanatory strength when confined to narrower data segments, for instance, observations from plots outside expert-defined risk areas (risk-zone=false).

Discussions

Our discussion is organized around three central themes: (a) the complexity-accuracy trade-off analyzed in Subsection 2.4.1, (b) how complexity affects interpretability, and (c) the agronomic knowledge generated by the trained models.

Complexity and accuracy

The outcomes presented here align with earlier research on model complexity [10, 13], particularly the general pattern that complex models deliver superior predictive accuracy to simpler ones. Nevertheless, it is important to stress that this pattern assumes that models have been suitably tuned and trained. A complex model trained on scarce data, for example, will almost certainly overfit, especially when the number of parameters matches or exceeds the number of data points. Conversely, when data aligns closely with the assumptions of a simple model (such as linearity), that model can perform exceptionally well regardless of its relative simplicity. Finally, even a well-trained model under valid assumptions will deliver poor results when tested on data that substantially differs from the training set. We encounter exactly this issue with models evaluated on 2013 and 2015 data for both incidence and appearance date predictions in the two crops. Observations from those years proved unusual, with several predictive variables recording values beyond the ranges seen in other years. This led to pronounced underfitting and yielded the lowest R2 scores observed during our entire study.

Interpretability

There is a widespread belief that greater model complexity is associated with reduced interpretability. Our scenario clearly demonstrates this relationship. Linear and pattern-based models permitted straightforward extraction of insights directly from their internal structures. For more complex systems such as random forests and gradient boosting trees, we relied on external tools, including permutation-based accuracy drop tests and partial dependence plots (PDPs). While these tools prove valuable, they come with constraints. Importance metrics do not reveal the direction (positive or negative) of a variable's relationship with predictions. PDPs handle at most two variables simultaneously and rest on independence assumptions that rarely hold in real datasets. Each point on a PDP curve results from averaging model outputs across all possible combinations of the other variables. Because certain combinations may be improbable, PDPs warrant careful interpretation, particularly when predictive variables display correlations. That said, the PDPs produced for RFR and GBR in our tests closely aligned with the threshold effects detected by HiPaR. Importantly, even though RFR, GBR, and HiPaR all employ thresholds on predictive variables, their clear superiority over Lasso strengthens the notion that threshold effects offer a sensible modeling approach for predicting plant diseases using meteorological inputs.

Agronomical insights

From our examination of the sugar beet Cercospora scenario, it becomes clear that organizing meteorological indicators into seasonal categories—specifically winter, spring, and summer—helps capture a meaningful portion of the differences observed in disease incidence rates.

Winter sets the foundational environment

This phase corresponds to the time when the main inoculum of Cercospora persists in the soil as spores. Spring corresponds to the active growth window for both the crops and the Cercospora organism. Summer covers the final

© 2024 The Author(s). | Published by Agro Intelligence Research Press

stretch of the season, when disease symptoms and their consequences become plainly evident.

In broad terms, dry summer weather tends to restrain Cercospora proliferation. This emerges from the notable weight the models gave to wind and temperature elements in June and July. Dry winter conditions likewise appear to curb disease expansion. By contrast, warm, moist spring weather stands out as the chief driver of rising Cercospora incidence. The hybrid rules generated by HiPaR allow us to uncover more refined links between incidence levels and the predictive variables. Rule 2, shown in **Table 4**, suggests that a temperate July directs attention toward the starting conditions (winter), notably wind, sunlight exposure, and temperature, with the latter two factors contributing positively to primary inoculum levels. Furthermore, a breezy spring with moderate June temperatures should focus attention on the growth phase (spring), especially on temperature and humidity, both of which show positive associations with incidence. In every instance, the appearance date functions as the strongest indicator of eventual incidence, underscoring that timely detection remains the most powerful tool against Cercospora.

No substantial insights could be extracted from downy-mildew predictions on vines, as the transparent models account for at most 14% of the observed variance in incidence—the outcomes for appearance date fall even lower. We attribute this disparity to the dataset, which relies exclusively on meteorological records spanning only the four weeks before the season's end. Put differently, this dataset lacks the extensive meteorological signals available in the sugar beet Cercospora dataset. This result underscores the critical role of detailed, comprehensive meteorological data in simulating plant disease dynamics. We also regard the exploration of how varying levels of detail in meteorological indicators influence such modeling efforts as a valuable direction for continued research.

Conclusion

In this paper, we have illustrated the benefits of examining the trade-off between model complexity and forecasting plant disease incidence. It is widely recognized that in many domains, advanced models such as neural networks and gradient boosting often yield superior results compared to basic approaches like linear regression. Yet this improvement entails a loss in interpretability, which (a) holds great importance when seeking to derive insights from the prediction model, and (b) encourages openness, potentially enhancing user confidence. While post-hoc explanation techniques can support insight extraction from precise black-box models, they do not represent the only path forward, as evidenced by this study: models with moderate complexity via pattern-aided regression can attain competitive predictive capabilities while preserving clarity and interpretability. In addition, our evaluations using post-hoc explainability tools, such as partial dependence plots, demonstrate that pattern-aided regression identifies threshold effects, similarly to those leveraged by higher-accuracy black-box ensemble methods. With these models, we have further established that moderate-complexity techniques excel at retrieving more relevant details than simpler alternatives. At the same time, moderate-complexity models prove simpler to interpret than highly elaborate ones. This underscores the practical value of pattern-aided regression, making it a compelling option for crop forecasting. Since interpretability directly correlates with user acceptance, evaluating model complexity is a significant consideration that should not be overlooked. This matter has been explored previously through learning complexity [23] and data complexity [24], though seldom with respect to the complexity of the final model. In summary, our investigation reveals that year-to-year meteorological fluctuations make disease incidence forecasting particularly difficult, and that reliable predictions for arbitrary years will require further study and larger volumes of high-quality historical data.

In the future, we envision studying whether increasing the temporal and spatial granularity of meteorological attributes can improve the quality of our predictions. An interesting research avenue is to apply representation learning techniques to learn novel, useful meteorological indicators that more accurately predict disease incidence. Given the inter-annual variations of weather patterns, future approaches should be able to categorize prediction models based on the meteorological profile of the data used to train them. We believe that unsupervised learning techniques could be adapted in that regard. Such approaches may even be necessary in light of a climate that will continue to change in the coming years.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

Acknowledgements

We thank the French applied agricultural research organization for sugar beet (ITB - Institut Technique de la Betterave) (ITB) and the French Wine and Vine Institute (IFV – Institut Français de la Vigne et du Vin) for providing the meteorological and agronomical data used in this study.

We also thank all the experts from each institute who helped us interpret our results with their insights: Fabienne Maupas, Ghislain Malatesta, Gouwie Céline (ITB), and Marc Raynal, Christian Debord, Xavier Burgun, and Marc Vergnes (IFV). We thank Lucile Vallet (Acta) for her work and for preparing the sugar beet dataset.

Conflict of interest

None

Financial support

This work was funded by the DigitAg institute (ANR-16-CONV-0004) and the RegEpi project (ECOPHYTO R&D program, Ecophyto-2019-REGEPI grant, French Biodiversity Agency – OFB). This project data was also part of the network data science and modeling methods for agriculture and the agri-food sector (www.modelia.org, funded by CASDAR grants of the French Ministry of Agriculture, RMT-SDMAA-19WRT034).

Ethics statement

None

Received: 18 Jun 2023 Revised: 23 Aug 2023 Accepted: 08 Sep 2023
Published online: 25 November 2024

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Heap I. Global perspective of herbicide-resistant weeds. Pest Manag Sci. 2014;70(9):1306-15. https://doi.org/10.1002/ps.3696.

2. Parsons P, Freeman E, Weidling R, Williams GL, Gill P, Byron N. Using existing knowledge for the risk evaluation of crop protection products in order to guide exposure driven data generation strategies and minimise unnecessary animal testing. Crit Rev Toxicol. 2021;51(7):600-21. https://doi.org/10.1080/10408444.2021.1987384.

3. Chen M. Analyse du risque de mildiou de la vigne dans le Bordelais à partir de données régionales et d'informations locales collectées en cours de saison. PhD thesis. Université Paris-Saclay (ComUE); 2019.

4. Edwards-Jones G. Knowledge-based systems for crop protection: Theory and practice. Crop Prot. 1993;12(8):565-78. https://doi.org/10.1016/0261-2194(93)90119-4.

5. Velasquez-Camacho L, Otero M, Basile B, Pijuan J, Corrado G. Current trends and perspectives on predictive models for mildew diseases in vineyards. Microorganisms. 2023;11(1):1-19. https://doi.org/10.3390/microorganisms11010073.

6. van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: Systematic literature review. Comput Electron Agric. 2020;177:105709. https://doi.org/10.1016/j.compag.2020.105709.

7. Ip Li Minn Ang RHL, Seng KP, Broster JC, Pratley JE. Big data and machine learning for crop protection. Comput Electron Agric. 2018;151:376-83. https://doi.org/10.1016/j.compag.2018.06.008.

8. Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine learning in agriculture: A review. Sensors. 2018;18(8):1-29. https://doi.org/10.3390/s18082674.

9. van Evert FK, Fountas S, Jakovetic D, Crnojevic V, Travlos I, Kempenaar C. Big data for weed control and crop protection. Weed Res. 2017;57(4):218-33. https://doi.org/10.1111/wre.12255.

10. Galárraga L, Pelgrin O, Termier A. HiPaR: Hierarchical pattern-aided regression. In: Advances in Knowledge Discovery and Data Mining. Cham: Springer International Publishing; 2021. p. 320-32.

11. Mori T, Uchihira N. Balancing the trade-off between accuracy and interpretability in software defect prediction. Empir Softw Eng. 2019;24(2):779-825. https://doi.org/10.1007/s10664-018-9638-1.

12. Johansson U, Sönströd C, Norinder U, Boström H. Trade-off between accuracy and interpretability for predictive in silico modeling. Future Med Chem. 2011;3(6):647-63. https://doi.org/10.4155/fmc.11.23.

13. Dong G, Taslimitehrani V. Pattern-aided regression modeling and prediction model analysis. IEEE Trans Knowl Data Eng. 2015;27(9):2452-65. https://doi.org/10.1109/TKDE.2015.2411609.

14. Fenu G, Malloci FM. Review forecasting plant and crop disease: An explorative study on current algorithms. Big Data Cogn Comput. 2021;5(1):1-24. https://doi.org/10.3390/bdcc5010002.

15. Quintana-Seguí P, Le Moigne P, Durand Y, Martin E, Habets F, Baillon M, et al. Analysis of near-surface atmospheric variables: Validation of the SAFRAN analysis over France. J Appl Meteorol Climatol. 2008;47(1):92-107. https://doi.org/10.1175/2007JAMC1636.1.

16. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Series B. 1994;58:267-88.

17. Kramer S. Structural regression trees. In: AAAI/IAAI. vol. 1. Citeseer; 1996. p. 812-9.

18. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. https://doi.org/10.1023/A:1010933404324.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

19. Elavarasan D, Vincent DR, Sharma V, Zomaya AY, Srinivasan K. Forecasting yield by integrating agrarian factors and machine learning models: A survey. Comput Electron Agric. 2018;155:257-82. https://doi.org/10.1016/j.compag.2018.10.024.

20. Mason L, Baxter J, Bartlett P, Frean M. Boosting algorithms as gradient descent. Adv Neural Inf Process Syst. 1999;12.

21. McGee VE, Carleton WT. Piecewise regression. J Am Stat Assoc. 1970;65(331):1109-24. https://doi.org/10.1080/01621459.1970.10481152.

22. Wang Y, Witten IH. Inducing model trees for continuous classes. In: ECML Poster Papers. 1997.

23. Kearns MJ. The computational complexity of machine learning. Cambridge (MA): MIT Press; 1990.

24. Dwivedi R, Singh C, Yu B, Wainwright MJ. Revisiting complexity and the bias-variance tradeoff. arXiv preprint arXiv:2006.10189. 2020.

© 2024 The Author(s). | Published by Agro Intelligence Research Press