Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

Predicting plant disease epidemics using boosted regression trees

Chun Peng [a, 1], Xingyue Zhang [b, 1], Weiming Wang [a, *]

[a] *School of Mathematics and Statistics, Huaiyin Normal University, Huaian, 223300, PR China*
[b] *École Polytechnique Fédérale de Lausanne, Rte Cantonale, 1015, Lausanne, Switzerland*

ARTICLE INFO

Article history:
Received 1 February 2024
Received in revised form 13 May 2024
Accepted 18 June 2024
Available online 18 June 2024
Handling Editor: Dr Yijun Lou

Keywords:
Plant disease epidemics
Scalar-on-function model
Boosted regression trees

ABSTRACT

Plant epidemics are often associated with weather-related variables. It is difficult to identify weather-related predictors for models predicting plant epidemics. In the article by Shah et al., to predict Fusarium head blight (FHB) epidemics of wheat, they explored a functional approach using scalar-on-function regression to model a binary outcome (FHB epidemic or non-epidemic) with respect to weather time series spanning 140 days relative to anthesis. The scalar-on-function models fit the data better than previously described logistic regression models. In this work, given the same dataset and models, we attempt to reproduce the article by Shah et al. using a different approach, boosted regression trees. After fitting, the classification accuracy and model statistics are surprisingly good.

1. Introduction

One purpose of epidemiology is to predict the outbreak of diseases beforehand. Weather plays a huge role in infectious diseases, whether for human, animal, or for plants. In this report, we attempt to reproduce the article *Predicting Plant Disease Epidemics from Functionally Represented Weather Series* by Shah et al. (Shah et al., 2019). In the study of plant disease epidemics, epidemiologists often investigate how disease outbreaks are correlated with weather patterns (Chakraborty et al., 2000).

We can expect crops to have a low level of disease, which is harmless. A more severe level of diseases can cause epidemics, which leads to reduction in crop yield. Farmers often intervene such situation with the use of crop protection chemicals. These decisions are supported by predictive models which help farmers to forecast disease outbreaks.

To accurately predict plant disease outbreaks, it is crucial to identify weather-based variables. One common approach is to mine a time series of weather variables to recognize time periods and predictors related to disease outbreaks (Carisse et al., 2018). Window-pane analysis is often used to accomplish this. However, window-pane analysis has the flaw of being 'data dredging'. Shah et al. proposed using functional data analysis (FDA) instead. They used scalar-on-function regression, which is one form of FDA, to identify weather variables and time periods associated with epidemics of Fusarium head blight (FHB). FHB

[*] Corresponding author.
E-mail address: wangwm_math@hytc.edu.cn (W. Wang).
Peer review under responsibility of KeAi Communications Co., Ltd.
[1] These authors contributed equally.

https://doi.org/10.1016/j.idm.2024.06.006

is the most economically significant wheat disease in many areas of the world. FDA demonstrates an improvement of prediction accuracy over existing models.

In this work, we use boosted regression trees instead of FDA to reproduce the work by Shah et al., with the objective of improving current FHB predicting models. The prediction accuracy and model efficiency are significantly enhanced.

2. Related work

Early work of predicting FHB epidemics using boosted regression trees In 2014, Shah et al. applied boosted regression trees on the original dataset of 527 FHB observations (Shah et al., 2014). They used BRTs on a training data set of 369 observations and testing data set of 158 observations. The resulting misclassification rate on the testing data is 0.23. Models were simplified, dropping some insignificant models during model fitting. The variable *RESIST* was included in every model. On the simplified 5-, 7-, 10-, 14-, and 15-day pre-anthesis $\mathrm{brt}_{\mathrm{i}}$ models, the cross-validated AUCs were 0.802, 0.832, 0.843, 0.875, and 0.872 respectively. For post-anthesis, the values were 0.867, 0.852, 0.881, 0.851, and 0.879 respectively.

Predicting FHB epidemics using random forests Shah et al. investigated the feasibility of random forests on the prediction of FHB epidemics (Shah et al., 2023a). Predictors were selected as input variables using three random forest variable selection algorithms: Boruta, varSelRF, and VSURF. Compared with the logistic regression models, the random forest models had better performance in general.

Accuracy in the prediction of disease epidemics when ensembling simple but highly correlated models As a case study on FHB epidemics, Shah et al. examined ensembling methods which combine the predictions made by individual component base models to achieve better prediction accuracy (Shah et al., 2023b). Some base models may produce highly correlated predictions. Three ensembling methods were investigated: soft voting, weighted averaging of smaller subsets of the base models, and penalized regression as a stacking algorithm (Shah et al., 2023b). The stacked algorithm has superior performance than the other two.

FHB prediction models from the United States In the US, the first models for FHB prediction were logistic regression models by De Wolf et al. (De et al., 2003). Information used was from 4 states at 50 location-years, representing 3 different wheat production regions in the US. The prediction accuracy of these logistic regression models were from 62% to 85%. These models were effectively employed in an online FHB risk assessment service in 31 states in the US.

FHB prediction models from Argentina A computer program using SAS identified the key meteorological factors correlated with wheat head blight incidence in Pergamino, a region in humid pampeana (Moschini & Fortugno, 1996). Analyzing data from 1978 to 1990 with linear regression, the study found specific humidity and rainfall conditions most strongly linked to the disease. Two predictive models were developed and successfully validated against data from 1991 to 1993, accurately forecasting disease occurrence.

3. Methods

We use the same dataset as in the work by Shah et al. (Shah et al., 2019). There are 999 observations, 273 FHB epidemics ($\mathrm{Y_i} = 1$) and 726 FHB non-epidemics ($\mathrm{Y_i} = 0$). This is a binary classification problem. Since the response variable is binary, logistic regression was naturally applied.

Scalar predictors include *resist* and *wc*. *resist* is the level of cultivar resistance to FHB, a categorical factor (Shah et al., 2019). *wc* describes the combination of wheat type (spring (sw) or winter (ww) wheat) and corn residue presence (corn = 1) or absence (corn = 0). *wc* = 1 if sw. *wc* = 2 if ww and corn = 0. *wc* = 3 if ww and corn = 1.

There are also weather-based predictors measured 5–15 days before or after flowering (Shah et al., 2019). They are derived from temperature (T), relative humidity (RH), and TRH. Infection usually takes place during flowering (anthesis), which makes this period crucial. There are 39 weather-based predictors in total (see Appendix A). For instance, weather variable 3 is the mean RH from 7 days pre-anthesis to anthesis.

In the original work, there are 26 models in total (see Appendix B). Model can be a standard logistic regression (lr) or a scalar-on-function regression model (sof). We apply boosted regression trees on the lr models. Note that models (1, 2, 3), (7, 8, 9) and (11, 12, 13) only differ in window lengths (Shah et al., 2019). Models (7, 8, 9) and (11, 12, 13) differ only by the inclusion of the scalar variable *wc*.

Instead of using standard logistic regression, we turn these models into boosted regression tree models. Boosted regression trees is an ensemble method that combines the advantage of two algorithms: regression trees and boosting (Elith et al., 2008). We can treat the individual terms in the final BRT models as individual trees and they are fitted in a ‘forward, stagewise fashion’ (Elith et al., 2008). Boosted regression trees also has other advantages: No prior data transformation nor elimination of outliers is needed. BRTs are highly effective in modeling complex, non-linear relationships that often exist in ecological and biological data. Unlike scalar-on-function regression, which typically assumes a linear relationship between predictors and the response variable, BRTs can automatically detect and model intricate interactions and non-linearities without requiring prior specification of the form of these relationships. Also, BRTs inherently model interaction effects between variables. This capability is crucial in ecological modeling. BRTs are generally more robust to outliers and extreme values compared to traditional regression methods. Moreover, BRTs are more scalable, accurate, flexible, and efficient.

As for predictive metrics, Boosted Regression Trees (BRTs) generally outperform traditional models across key predictive metrics. By combining multiple weak learners, BRTs achieve lower misclassification rates and enhance both sensitivity and specificity, adapting effectively to complex and non-linear data patterns. This ensemble method also excels in optimizing Area Under the Curve (AUC) values and Cohen's Kappa statistics, providing a robust measure of classification accuracy and agreement that is less susceptible to imbalances and biases in the data. Consequently, BRTs offer a superior alternative in scenarios requiring high accuracy and reliability in predictions.

To implement boosted regression trees in R, we use the *gbm.step* function in the *gbm* package. This function assesses the optimal number of boosting trees using k-fold cross validation. It is an implementation of the cross-validation procedure by Hastie et al. (Hastie et al., 2009). The data is divided into 10 subsets. Then, the function fits a gbm model with increasing complexity. After processing every fold, the average holdout residual deviance and its standard error are used to identify the optimal number of trees (Hastie et al., 2009). We need to adjust parameters such as bag fraction, tree complexity, learning rate, etc. After hyperparameter tuning, we select bag fraction to be 0.5, tree complexity to be 5 or 10, and learning rate to be 0.005 or 0.01.

To compare models, AUC, sensitivity (the proportion of FHB epidemics correctly classified as such), specificity (the proportion of FHB non-epidemics correctly classified as such), and misclassification rate are used (Shah et al., 2019). The Youden Index is required to calculate some of these statistics. More specifically, Area Under the Curve (AUC), which measures the entire area underneath the Receiver Operating Characteristic (ROC) curve, helps assess the model's ability to discriminate between classes across all thresholds. Sensitivity (or true positive rate) quantifies the proportion of actual positives (FHB epidemics) correctly identified, while specificity measures how well the model identifies actual negatives (non-epidemics). The misclassification rate provides the overall proportion of incorrect predictions. The Youden Index, a summary measure of the ROC curve, combines sensitivity and specificity to assess the model's effectiveness, with higher values indicating better performance. Lastly, Cohen's Kappa statistic quantifies the level of agreement between two raters who classify items into categories, adjusting for agreement that occurs by chance, thereby providing a more accurate measure of inter-rater reliability.

Table 1
Logistic regression and boosted regression tree model statistics comparison.

Model ID	Method	Scalars	Period	AUC	YI	Sensitivity	Specificity	Kappa	Misclass
1	lr	resist	pre	0.718	0.27	0.692	0.653	0.29	0.336
	brt			0.978	0.36	0.905	0.939	0.827	0.07
2	lr	resist	pre	0.714	0.31	0.601	0.74	0.312	0.298
	brt			0.940	0.31	0.868	0.865	0.685	0.134
3	lr	resist	pre	0.717	0.24	0.78	0.556	0.259	0.382
	brt			0.826	0.27	0.725	0.725	0.430	0.262
5	lr	resist	pre	0.716	0.32	0.593	0.747	0.314	0.295
	brt			0.974	0.31	0.945	0.888	0.775	0.096
7	lr	resist	post	0.739	0.3	0.67	0.7	0.323	0.308
	brt			0.976	0.35	0.919	0.944	0.845	0.063
8	lr	resist	post	0.732	0.3	0.656	0.716	0.33	0.3
	brt			0.994	0.41	0.956	0.970	0.915	0.034
9	lr	resist	post	0.743	0.3	0.681	0.73	0.366	0.283
	brt			0.992	0.33	0.971	0.944	0.881	0.049
11	lr	resist + wc	post	0.751	0.32	0.656	0.749	0.368	0.276
	brt			0.989	0.32	0.974	0.949	0.893	0.044
12	lr	resist + wc	post	0.748	0.31	0.663	0.738	0.361	0.282
	brt			0.994	0.35	0.967	0.956	0.899	0.041
13	lr	resist + wc	post	0.758	0.3	0.681	0.742	0.38	0.274
	brt			0.991	0.33	0.960	0.949	0.883	0.048
15	lr	resist + wc	pre	0.755	0.27	0.703	0.68	0.327	0.313
	brt			0.992	0.39	0.949	0.959	0.891	0.044
17	lr	resist	pre	0.717	0.3	0.626	0.708	0.297	0.314
	brt			0.987	0.36	0.930	0.959	0.878	0.049
19	lr	resist	post	0.731	0.28	0.663	0.69	0.307	0.317
	brt			0.994	0.36	0.978	0.952	0.900	0.041
20	lr	resist + wc	post	0.749	0.29	0.692	0.702	0.343	0.3
	brt			0.993	0.32	0.989	0.945	0.896	0.043
21	lr	resist	pre	0.733	0.29	0.67	0.696	0.318	0.311
	brt			0.987	0.36	0.941	0.950	0.872	0.052
22	lr	resist + wc	pre	0.756	0.28	0.689	0.683	0.319	0.315
	brt			0.990	0.32	0.967	0.941	0.874	0.052
23	lr	resist	pre	0.7	0.24	0.729	0.567	0.233	0.388
	brt			0.986	0.38	0.945	0.948	0.870	0.053
24	lr	resist + wc	pre	0.712	0.32	0.553	0.773	0.311	0.287
	brt			0.987	0.33	0.967	0.928	0.853	0.061
25	lr	resist	prepost	0.75	0.25	0.762	0.628	0.316	0.335
	brt			0.993	0.34	0.978	0.950	0.898	0.042
26	lr	resist + wc	prepost	0.783	0.28	0.711	0.723	0.381	0.28
	brt			0.986	0.28	0.974	0.910	0.830	0.072

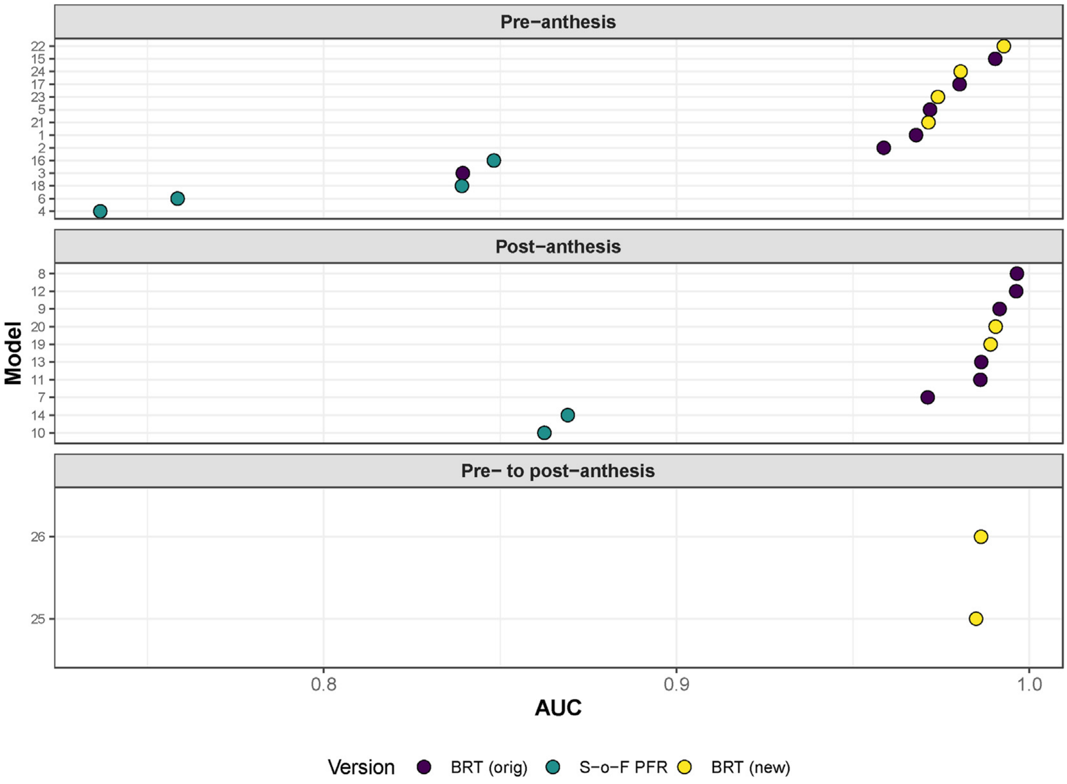

Fig. 1. AUC by model version and period.

Chart 1 provides a visual representation of the systematic process employed in our study.

Below is the R code snippet for *Model 1* as a demonstration:

```
fit1_brt <- gbm.step(data=X2, gbm.x=2:3, gbm.y=1, family='bernoulli', tree.complexity=5, learning.rate=0.01, bag.fraction=0.5)
summary(fit1_brt)
preds <- attr(fit1$terms , "term.labels")[-1]
y_brt <- data.frame(id = 1:nrow(X), actual = Y, fitted.prob = fit1_brt$fitted)
lr.1.brt <- f.stats.brt(id = 1, scalar = "resist", wb = "1", wb.preds = preds, model = "lr.1.brt", version = "orig", period = "pre", y_brt = y_brt)
stats.out.brt(lr.1.brt)
```

4. Results

Among 26 models in the original work, twelve were based on previously reported logistic regression models. Six of these (1, 2, 3, 5, 15, 17) were about pre-anthesis conditions. The rest (7, 8, 9, 11, 12, 13) were about post-anthesis conditions. Later on, eight more logistic regression models (19–26) with newly derived variables were added. Models 21–24 focus on pre-anthesis conditions. Models 19 and 20 focus on post-anthesis condition. Models 25 and 26 cover conditions in windows spanning anthesis (Shah et al., 2019). We can see that models with 2 or 3 weather-based variables generally perform better than models with a single weather-based variables. In the original work, four s-o-f models had better performance that logistic regression models.

Apart from the s-o-f models, we can compare our models' results with the models from the original work by Shah et al. (Shah et al., 2019). The model statistics has significantly improved. Table 1 demonstrates the comparison between our result and the original result. For most models, the misclassification rate has dropped from about 0.3 to below 0.1. Sensitivity (the proportion of FHB epidemics correctly classified as such) and specificity (the proportion of FHB non-epidemics correctly classified as such) have increased from about 0.6 to more than 0.9 (almost reaching 1 in some cases). AUC is almost 1 for many models. Kappa (Cohen's Kappa statistic) has increased to more than 0.8 for many models. We can see that boosted regression trees work really well on the prediction of FHB epidemics.

To illustrate the model statistics more vividly, we plot several figures. Fig. 1 shows the AUC value by period and model version. Purple represents the original logistic regression models which are now BRT models. Yellow means the newly added

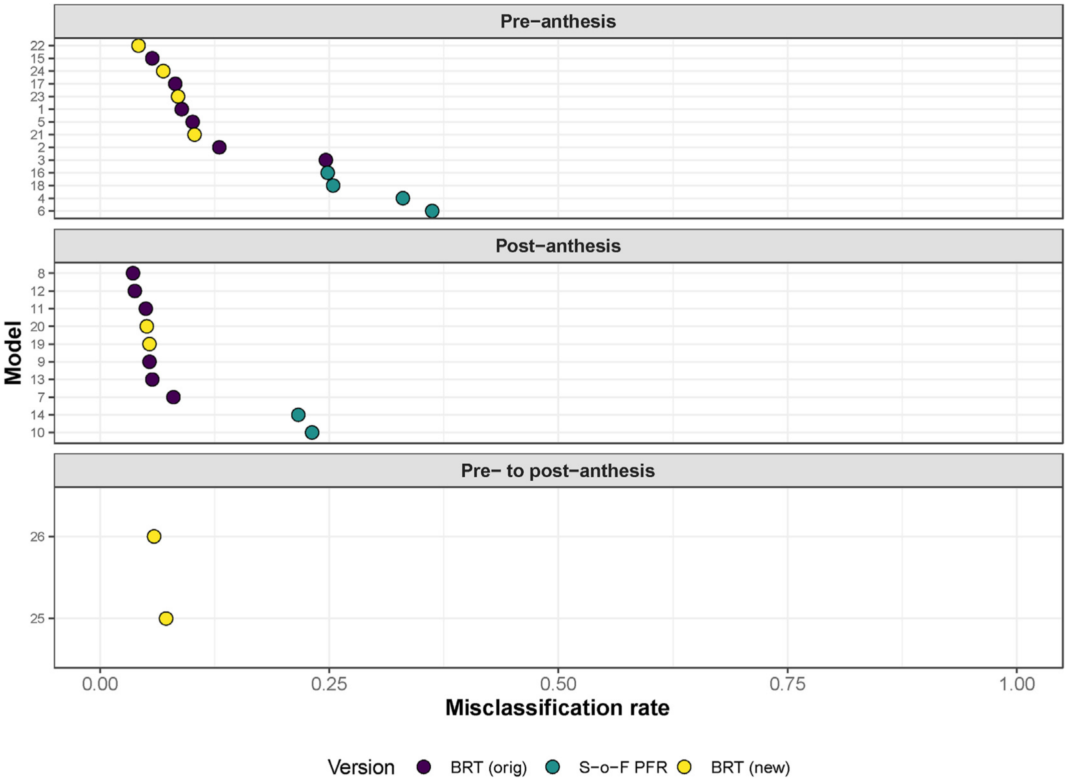

Fig. 2. Misclassification rate by period and model version.

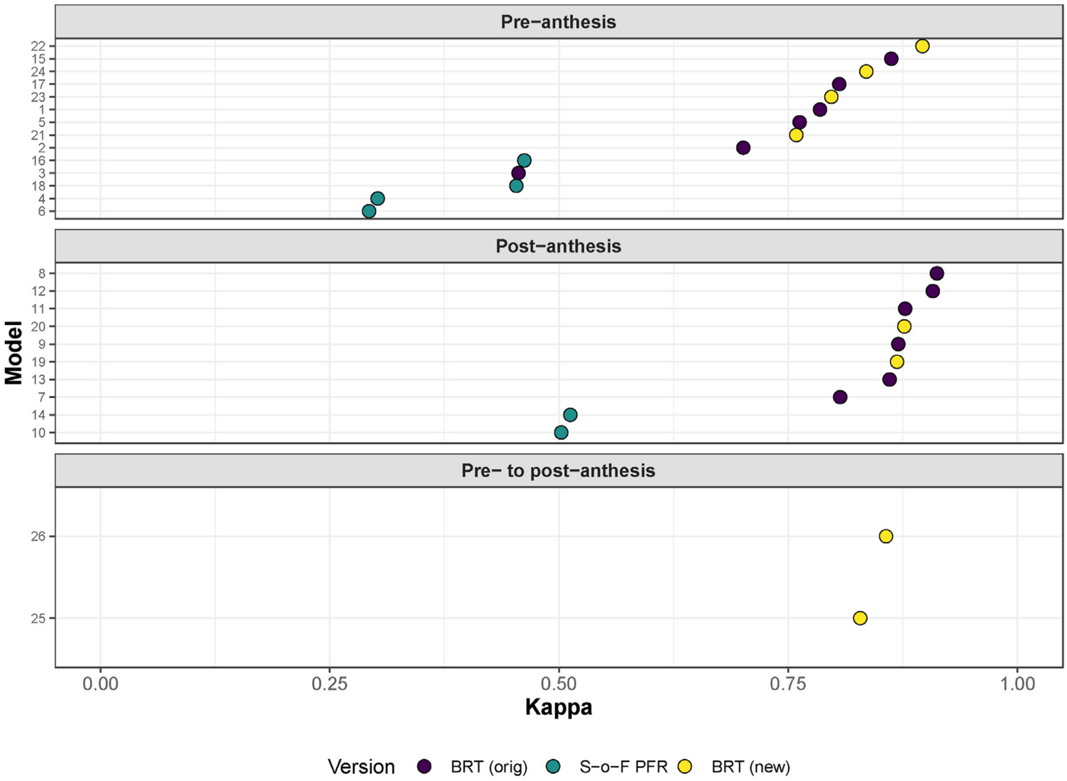

Fig. 3. Kappa by period and model version.

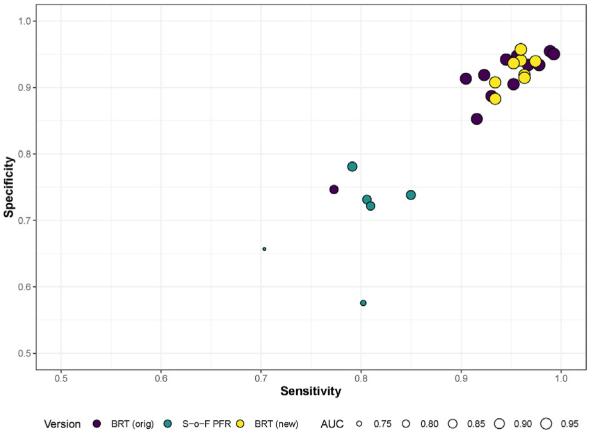

Fig. 4. Sensitivity and specificity by model version and AUC.

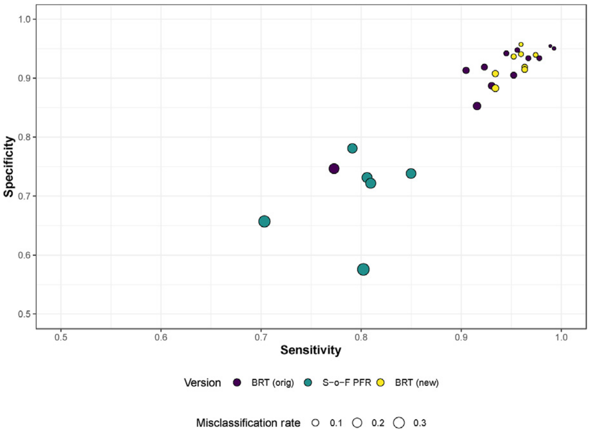

Fig. 5. Sensitivity and specificity by model version and misclassification rate.

logistic regression models which are now BRT models too. Green represents the scalar-on-function models in the original work by Shah et al. We can see that BRT models have higher AUC values than s-o-f models. Fig. 2 shows the misclassification rate by period and model version. BRT models have very low misclassification rate. Fig. 3 demonstrates the Cohen's Kappa statistic by period and model version. Again, BRT models have superior performance.

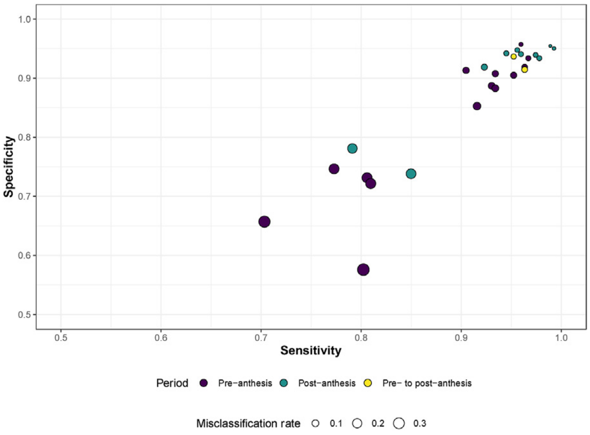

Fig. 6. Sensitivity and specificity by period and misclassification rate.

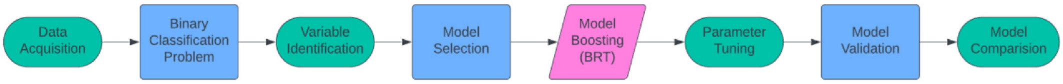

Chart 1. Flow chart of systematic process.

Fig. 4 shows sensitivity and specificity by model version and AUC. Fig. 5 displays sensitivity and specificity by model version and misclassification rate. Fig. 6 demonstrates sensitivity and specificity by period and misclassification rate. BRT models have higher sensitivity and specificity values.

5. Conclusion

In this investigation, we have successfully demonstrated the application of boosted regression trees (BRTs) in predicting Fusarium head blight (FHB) epidemics, building upon the foundational work by Shah et al. (Shah et al., 2019). The original research highlighted the limitations of standard logistic regression models and the advantages of employing scalar-on-function models for this purpose. Our findings suggest a significant leap in predictive accuracy and model efficiency through the use of BRTs, which not only matched but substantially exceeded the performance of the previously established models.

The superior classification performance of BRT models is particularly notable, with misclassification rates significantly reduced to below 0.1 for most models and measures of sensitivity and specificity approaching near perfect scores. These results underscore the robustness of BRTs in handling complex, non-linear relationships between weather-related variables and the incidence of FHB epidemics. The ability of BRTs to effortlessly manage interactions without prior data transformation or the elimination of outliers provides a compelling case for their broader application in epidemiological modeling.

However, while our study marks a considerable advancement in the predictive modeling of plant disease epidemics, it is not without its limitations. The approach's dependency on high-quality, comprehensive datasets and the potential for overfitting in BRT models are challenges that require careful consideration. Furthermore, the interpretability of BRT models can be less straightforward compared to more traditional statistical methods, which may pose challenges for broader adoption among practitioners and policymakers.

To enhance the application of Boosted Regression Trees (BRTs) in predicting plant diseases like Fusarium Head Blight (FHB), it is crucial to improve data quality through partnerships with agricultural and meteorological organizations, incorporate regularization and cross-validation to prevent overfitting, and enhance model interpretability using tools like partial dependence plots. Future research should focus on optimizing data resolution for modeling, developing hybrid models that balance complexity with interpretability, and creating visualization tools to clarify variable interactions and influences.

CRediT authorship contribution statement

Chun Peng: Writing – original draft, Formal analysis, Data curation. **Xingyue Zhang:** Writing – original draft, Formal analysis, Data curation. **Weiming Wang:** Writing – review & editing, Methodology, Funding acquisition, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No. 12071173 and 12171192) and Huaian Key Laboratory for Infectious Diseases Control and Prevention (HAP201704).

Appendix A. Weather-based Predictors in the Original Work

Weather variable ID	Description
Shah et al. (2014) **models**	
3	Mean RH from 7 days pre-anthesis to anthesis
4	Mean RH from 10 days pre-anthesis to anthesis
5	Mean RH from 14 days pre-anthesis to anthesis
6	Mean overnight RH from anthesis to 5 days post-anthesis

(*continued*)

Weather variable ID	Description
7	Mean overnight RH from anthesis to 7 days post-anthesis
8	Mean overnight RH from anthesis to 10 days post-anthesis
13	No. hrs overnight RH $\geq 90\%$ from 10 days pre-anthesis to anthesis
18	Mean T from 7 days pre-anthesis to anthesis
19	Mean T from 15 days pre-anthesis to anthesis
20	Mean T from anthesis to 5 days post-anthesis
21	Mean T from anthesis to 7 days post-anthesis
22	Mean T from anthesis to 10 days post-anthesis
28	No. hrs $T < 9\,^{\circ}C$ from 7 days pre-anthesis to anthesis
29	No. hrs $T < 9\,^{\circ}C$ from 15 days pre-anthesis to anthesis
31	No. hrs $15\,^{\circ}C \leq T \leq 30\,^{\circ}C$ & RH $\geq 80\%$ from 15 days pre-anthesis to anthesis
32	No. overnight hrs $15\,^{\circ}C \leq T \leq 30\,^{\circ}C$ & RH $\geq 80\%$ from anthesis to 5 days post-anthesis
33	No. overnight hrs $15\,^{\circ}C \leq T \leq 30\,^{\circ}C$ & RH $\geq 80\%$ from anthesis to 7 days post-anthesis
34	No. overnight hrs $15\,^{\circ}C \leq T \leq 30\,^{\circ}C$ & RH $\geq 80\%$ from anthesis to 10 days post-anthesis
s-o-f models	
11	Daily mean RH 120 days pre- to 20 days post-anthesis
27	Daily mean T 120 days pre- to 20 days post-anthesis
12	Daily mean overnight RH 120 days pre- to 20 days post-anthesis
14	Cumulative no. overnight hrs in which RH $\geq 90\%$ from 120 days pre- to 20 days post-anthesis
30	Cumulative no. hrs in which $T < 9\,^{\circ}C$ from 120 days pre- to 20 days post-anthesis
38	Cumulative no. overnight hrs in which $15\,^{\circ}C \leq T \leq 30\,^{\circ}C$ & RH $\geq 80\%$ from 120 days pre- to 20 days post-anthesis
39	Cumulative no. hrs in which $15\,^{\circ}C \leq T \leq 30\,^{\circ}C$ & RH $\geq 80\%$ from 120 days pre- to 20 days post-anthesis
New lr models	
9	Mean RH from 20 days pre-anthesis to anthesis
10	Mean RH from 10 days pre-anthesis to 10 days post-anthesis
15	No. hrs in which RH $\geq 90\%$ from 20 days pre-anthesis to anthesis
16	No. hrs in which RH $\geq 90\%$ from anthesis to 10 days post-anthesis
17	No. hrs in which RH $\geq 90\%$ from 50 days pre-anthesis to 10-days post-anthesis
23	Mean T from 25 days pre-anthesis to 15 days pre-anthesis
24	Mean T from 20 days pre-anthesis to 10 days pre-anthesis
25	Mean T from 40 days pre-anthesis to 30 days pre-anthesis
26	Mean T from anthesis to 20 days post-anthesis
35	No. hrs $15\,^{\circ}C \leq T \leq 30\,^{\circ}C$ & RH $\geq 80\%$ from 30 days pre-anthesis to anthesis
36	No. hrs $15\,^{\circ}C \leq T \leq 30\,^{\circ}C$ & RH $\geq 80\%$ from 60 days pre-anthesis to 40 days pre-anthesis
37	No. hrs $15\,^{\circ}C \leq T \leq 30\,^{\circ}C$ & RH $\geq 80\%$ from anthesis to 10 days post-anthesis

Appendix B. The Models in the Original Work

Model ID	Class	Scalars	Weather-based predictors
1	lr	resist	3
2	lr	resist	4
3	lr	resist	5
4	sof	resist	11
5	lr	resist	4 + 13
6	sof	resist	11 + 14
7	lr	resist	6 + 20 + 32
8	lr	resist	7 + 21 + 33
9	lr	resist	8 + 22 + 34
10	sof	resist	12 + 27 + 38
11	lr	resist + wc	6 + 20 + 32
12	lr	resist + wc	7 + 21 + 33
13	lr	resist + wc	8 + 22 + 34
14	sof	resist + wc	12 + 27 + 38
15	lr	resist + wc	3 + 18 + 28
16	sof	resist + wc	11 + 27 + 30
17	lr	resist	19 + 29 + 31
18	sof	resist	27 + 30 + 39
19	lr	resist	16 + 26 + 37
20	lr	resist + wc	16 + 26 + 37
21	lr	resist	9 + 23 + 35
22	lr	resist + wc	9 + 23 + 35
23	lr	resist	15 + 25 + 36
24	lr	resist + wc	15 + 25 + 36

(*continued on next page*)

(*continued*)

Model ID	Class	Scalars	Weather-based predictors
25	lr	resist	10 + 17 + 24
26	lr	resist + wc	10 + 17 + 24

References

Carisse, O., McNealis, V., & Kriss, A. (2018). Association between weather variables, airborne inoculum concentration, and raspberry fruit rot caused by Botrytis cinerea. *Phytopathology, 108*(1), 70–82.

Chakraborty, S., Tiedemann, A. V., & Teng, P. (2000). Climate change: Potential impact on plant diseases. *Environmental Pollution, 108*(3), 317–326.

De, W. E. D., Madden, L. V., & Lipps, P. E. (2003). Risk assessment models for wheat Fusarium head blight epidemics based on within-season weather data. *Phytopathology, 93*(4), 428–435.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology, 77*(4), 802–813.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2, pp. 1–758). New York: springer.

Moschini, R. C., & Fortugno, C. (1996). Predicting wheat head blight incidence using models based on meteorological factors in Pergamino, Argentina. *European Journal of Plant Pathology, 102*, 211–218.

Shah, D. A., De Wolf, E. D., Paul, P. A., & Madden, L. V. (2014). Predicting Fusarium head blight epidemics with boosted regression trees. *Phytopathology, 104*(7), 702–714.

Shah, D. A., De Wolf, E., Paul, P. A., & Madden, L. V. (2023a). Into the trees: random forests for predicting Fusarium head blight epidemics of wheat in the United States. *Phytopathology, 113*(8), 1483–1493.

Shah, D. A., De Wolf, E. D., Paul, P. A., & Madden, L. V. (2023b). Accuracy in the prediction of disease epidemics when ensembling simple but highly correlated models. *PLoS Computational Biology, 17*(3), Article e1008831.

Shah, D. A., Paul, P. A., De Wolf, E. D., & Madden, L. V. (2019). Predicting plant disease epidemics from functionally represented weather series. *Philosophical Transactions of the Royal Society B, 374*(1775), Article 20180273.