JOURNAL OF AEROSPACE INFORMATION SYSTEMS
Vol. 18, No. 12, December 2021

Technical Notes

Comparison of Machine Learning Models for Data-Driven Aircraft Icing Severity Evaluation

Sibo Li* and Roberto Paoli†
University of Illinois at Chicago, Chicago, Illinois 60607

https://doi.org/10.2514/1.I011047

I. Introduction

ICE formations on aircraft surface might lead to performance degradation because they modify the wing profile, which results in reduced lift and stall characteristics, and increased drag [1]. The main aerodynamic and environmental factors that affect the physical formation process of ice are flight speed, angle of attack (AOA), exposure time, liquid water content (LWC), droplet median volumetric diameter (MVD), and freestream temperature [2]. Numerical simulation approach has been widely applied to investigate the ice accretion process [3–6], and it generally involves the following procedures: solving the airflow field, tracking the water droplet trajectories, solving the icing thermodynamic model, and modifying the mesh. This process often requires significant computing resources, which limits the application of the numerical icing models in real-time ice accretion prediction [7–9]. To address this challenge, the authors previously studied adapting the machine learning model extreme gradient boosting model (XGBoost) [10] for aircraft icing severity evaluation based on six flight conditions (flight speed, angle of attack, exposure time, LWC, MVD, and freestream temperature) to represent a real flight situation [11]. The three icing features maximum ice thickness, icing area, and icing severity level [2] are predicted with reasonable accuracy. However, in the previous study, the XGBoost model was only compared to the classical methods multiple linear regression (MLR) [12] and ordinal logistic regression (OLR) [13]. Due to the interactions of multiple aerodynamic and environmental factors, aircraft icing is considered as a complex phenomenon, and the mapping relationship between the input flight conditions and the output aircraft icing severity features is strongly nonlinear [11]. Therefore, the effectiveness of different machine learning models on the icing application is worth investigating. In this study, the conventional machine learning models, including MLR, OLR, decision tree [14], naive Bayes [15], K-nearest neighbors (KNN) [16], and support vector machine (SVM) [17], and ensemble models, including random forest (RF) [18], adaptive boosting (AdaBoost) [19], and XGBoost, are tested and compared in predicting the maximum ice thickness, icing area, and icing severity level. A performance error analysis method containing various components is established to determine the accuracy of the studied methods. The workflow of this paper is illustrated in Fig. 1.

Received 4 August 2021; revision received 16 September 2021; accepted for publication 21 October 2021; published online 22 November 2021.

*Ph.D., Department of Mechanical and Industrial Engineering.
†Assistant Professor, Department of Mechanical and Industrial Engineering; also Argonne National Laboratory, Lemont, Illinois 60439.

The rest of this paper is organized as follows: Section II introduces the data collection that was obtained in a previous paper [11] and the machine learning models tested in this study. Section III elaborates the performance comparison of the tested machine learning models and discusses the corresponding modeling results. Finally, conclusions are presented in Sec. IV.

II. Data-Driven Methods

A. Data Collection

The dataset collected in the authors' previous paper [11] is adopted in the current study as well, which contains 1890 samples. For the sake of brevity, the reader is referred to [11] for a detailed description. The statistical parameters of the six flight conditions (flight speed, angle of attack, exposure time, LWC, MVD, and freestream temperature) are given in Table 1. The models are trained to predict three icing severity features based on the NACA0012 airfoil, including the size of the area covered by ice, the maximum ice thickness, and the icing severity level (Table 2). The statistical parameters of the icing area and maximum ice thickness in the prepared dataset are summarized in Table 3. The number of samples corresponding to light, moderate, heavy, and severe icing severity levels are 822, 497, 403, and 168, respectively.

B. Machine Learning Models

1. *Conventional Techniques*

The conventional techniques, including MLR, OLR, KNN, SVM, decision tree, and naive Bayes, are implemented using scikit-learn [20]. The OLR model is an extension of a logistic regression, which is used when the dependent variable has three or more levels with a natural ordering to the levels [13]. KNN is implemented through the instance-bases learning with parameter k; it uses a majority voting mechanism [16]. For regression problems, the prediction is given based on the mean or the median of the k-most similar instances. For classification problems, the output is the class with the highest frequency from the k-most similar instances. The models of SVM consist of two main groups: a) the SVM classifier models [17] and b) a support vector regression (SVR) model [21]. SVM seeks a line that best separates two classes. Its implementation is based on *libsvm*, and the default radial basis function (RBF) is selected as the kernel function. For SVR, the linear, polynomial, and RBF kernels can be applied. In this study, the nonlinear SVR with an RBF kernel is used. Decision tree constructs a binary tree from the training data [14]; a tree can be seen as a piecewise constant approximation. Split points are chosen greedily to minimize a cost function. Gini index is chosen as the cost function. For each given input value, Naive Bayes calculates the probability of each attribute, conditional on the class value [22]. It is implemented using the Gaussian naive Bayes algorithm, and a Gaussian distribution is assumed to estimate the probabilities for input variables using the Gaussian probability density function.

2. *Ensemble Methods*

Ensemble methods are techniques that create multiple models and then combine them to improve the generalizability and robustness over a single model. The current work studies adapting RF, AdaBoost, and XGBoost to the aircraft icing severity evaluation. RF is operated by constructing a set of decision trees at training time, each individual tree gives a class prediction, and the class that has most votes becomes the model's prediction [18]. AdaBoost [19] and XGBoost [10] are both boosting ensemble models, where the base models are built sequentially and one tries to reduce the bias of the

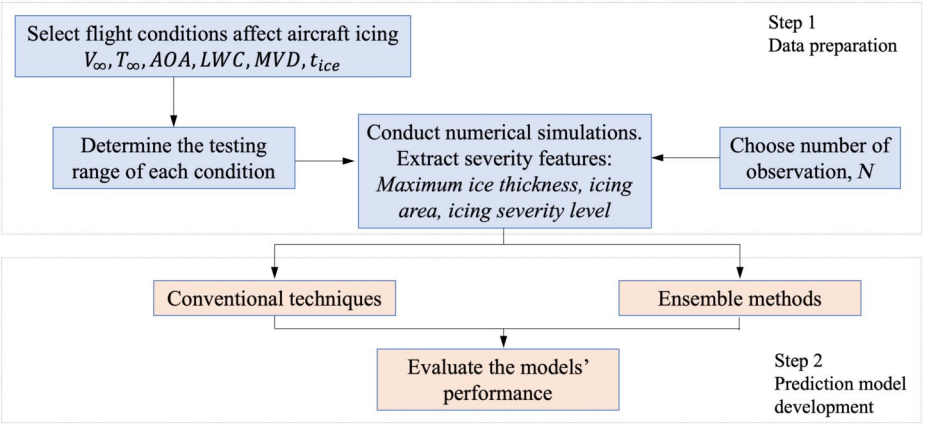

Fig. 1 Workflow of the proposed approaches.

Table 1 Statistics of flight conditions [11]

Feature	V_∞, knots	T_∞, K	AOA, °	LWC, g/m^3	MVD, μm	t_{ice}, min
Maximum	250.0	265	9	1.50	50	30.0
Minimum	100.0	253	0	0.50	5	1.0
Step size	37.5	3	3	0.25	5	9.5

Table 2 Icing severity level based on icing thickness [2]

Icing severity level	Light	Moderate	Heavy	Severe
Maximum thickness, mm	0.1–5.0	5.1–15	15.1–30	>30

Table 3 Statistics of icing severity features

Feature	Icing area, m^2	Maximum ice thickness, m
Maximum	4.718	0.2130
Minimum	0.732	0.0001
Mean	2.563	0.0212

combined model. The motivation is to combine several weak models to produce a powerful ensemble [20].

C. Performance Evaluation Measures

To evaluate the performance of the developed models, multiple statistical measures were employed. For predicting the icing area and maximum ice thickness, root mean squared error (RMSE) [23], coefficient of determination R^2 [23], mean absolute error (MAE) [19], and Taylor diagram [24] are applied. For predicting the icing severity levels, the models are quantitatively evaluated by using model evaluation indicators, such as accuracy and confusion matrix.

III. Results and Discussion

Evaluations of the conventional techniques and ensemble methods on the icing area, maximum ice thickness, and icing severity level are given in this section. All the parameter settings for the machine learning models are set to obtain the models' best performance by using a scikit-learn class called "GridSearchCV" [20]. Specifically, the first step is to create a grid that contains all the possible combinations of tuning parameters. Multiple values of the tuning parameters are chosen within reasonable ranges. Then, cross-validation is applied to identify the optimal hyperparameter sets for different models. The reader is referred to [11] for detailed operations. Instead of using the simple train/test split method, 10-fold stratified cross-validation is applied to all the classifiers and 10-fold cross-validation is applied to all the regressors to avoid overfitting and achieve a less-biased estimate of the model performance [20,22].

A. Icing Area Prediction

1. *Evaluation of Conventional Techniques*

In icing area prediction, conventional methods, including MLR, KNN, SVR, and decision tree, are implemented with scikit-learn using Python. From the 10-fold cross-validation, the average RMSE, R^2, and MAE on testing data are shown in Table 4. In the KNN method, k is usually a small and odd integer; in the current study, k is set to be 5. The RBF is selected as the kernel function for SVR, kernal parameter gamma is 1.0, and convergence epsilon is 0.01. The parameters in decision tree adopt the regressor's default values. It can be seen that SVR works the best among the conventional methods, and MLR has the worst performance. Although the MLR has the advantage of providing clearly interpretable coefficients, it failed to handle nonlinearities in the icing area prediction. Also, it is worth mentioning that, as a simple model, KNN has relatively good performance. Because KNN uses Euclidean distance [16] to compare examples, in order to assign equal importance to all the features when calculating the distance, the features must have the same range of values. Therefore, feature scaling is a crucial component in the training process of KNN. Specifically, normalization is applied in this study.

2. *Evaluation of Ensemble Methods*

For icing area prediction, the RMSE, R^2, and MAE of the ensemble methods are reported in Table 5. The optimal hyperparameter settings for the ensemble models are given by the GridSearchCV. For RF, the number of trees is 100, maximum depth is 10, and subset ratio is set to be 0.2. The maximum number of estimators for AdaBoost is set to be 300. For XGBoost, the number of trees is 700, interaction

Table 4 Accuracy of conventional methods in predicting icing area

Model	RMSE	R^2	MAE
MLR	0.559	0.613	0.208
KNN	0.408	0.798	0.201
SVR	0.316	0.871	0.170
Decision tree	0.401	0.813	0.195

Table 5 Accuracy of ensemble methods in predicting icing area

Model	RMSE	R^2	MAE
RF	0.181	0.946	0.067
AdaBoost	0.246	0.908	0.079
XGBoost	0.083	0.991	0.022

depth is 5, shrinkage factor is 0.1, and minimum child weight is 11. It can be seen that the coefficients of determination of all three models are above 0.9. XGBoost with $R^2 = 0.991$, RMSE $= 0.181$, and MAE $= 0.022$ shows better performance than other models.

3. Overall Comparison of Models for Icing Area

For a more comprehensive presentation, all the predictive models are examined using the graphical demonstration of the Taylor diagram [24] as shown in Fig. 2. Taylor diagram is a practical tool for summarizing how closely a pattern matches observations and understanding the performance of studied models. It provides the correlation coefficient and normalized standard deviations of each model, and the distance from the observation point is a measure of the centered RMSE. The observation point on the axis represents the perfect prediction, which has the correlation coefficient as 1. Therefore, the position of each model symbol appearing in the diagram quantifies how closely that model's predicted icing area matches observations, and the performance of the applied data-driven models can be visualized. It can be seen that the XGBoost outperforms all other models, RF and AdaBoost have the similar performance, and KNN and MLR have low accuracy in predicting the icing area in comparison to other models. Overall, the ensemble methods outperform the conventional techniques in predicting the icing area.

B. Maximum Ice Thickness Prediction

1. Evaluation of Conventional Techniques

Similar to the icing area predictions, conventional methods, including KNN, SVR, and decision tree, are applied to the maximum ice thickness evaluation. The parameter settings remain the same for all the models as in the icing area prediction cases. The averaged RMSE, R^2, and MAE from the 10-fold cross-validation are summarized in Table 6. It can be observed that the three predictive models exhibit different levels of accuracy, and because of the low RMSE and also the higher R^2, decision tree works the best among the conventional models.

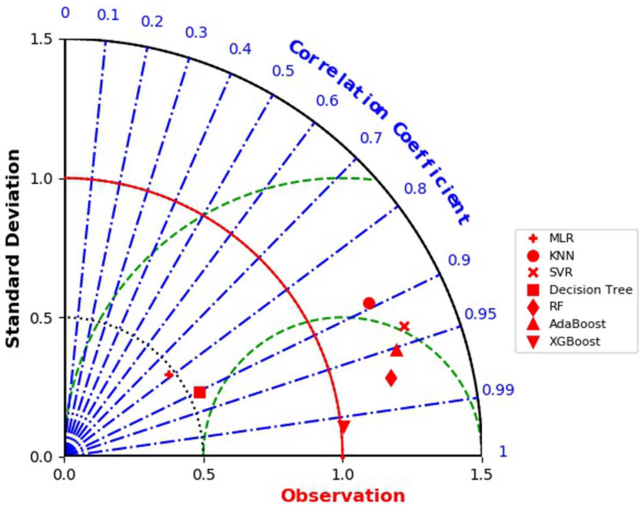

Fig. 2 Taylor diagram graphical presentation for the seven predictive models (MLR, KNN, SVR, decision tree, RF, AdaBoost, and XGBoost) in estimating the icing area.

Table 6 Accuracy of conventional methods in predicting maximum ice thickness

Model	RMSE	R^2	MAE
KNN	0.034	0.701	0.022
SVR	0.025	0.803	0.010
Decision tree	0.016	0.925	0.008

Table 7 Accuracy of ensemble methods in predicting maximum ice thickness

Model	RMSE	R^2	MAE
RF	0.027	0.801	0.0110
AdaBoost	0.017	0.911	0.0063
XGBoost	0.004	0.995	0.0017

2. Evaluation of Ensemble Methods

The three ensemble methods (RF, AdaBoost, and XGBoost) are also used to predict the maximum ice thickness. Again, from the GridSearchCV, for RF, the number of trees is 80, maximum depth is 10, and subset ratio is set to be 0.2. The maximum number of estimators for AdaBoost is set to be 300. For XGBoost, the number of trees is 200, interaction depth is 11, shrinkage factor is 0.1, subsample ratio is 1, and minimum child weight is 1. The statistical performance of the prediction of the investigated ensemble models is provided in Table 7. XGBoost with the lowest RMSE $= 0.004$, the lowest MAE $= 0.0017$, and the highest $R^2 = 0.995$ shows the best results among the ensemble models. RF has low accuracy in predicting maximum ice thickness relative to other models, whereas AdaBoost presents acceptable estimates.

3. Overall Comparison of Models for Maximum Ice Thickness

For a more comprehended presentation, the predictive models (conventional models and ensemble models) are examined using Taylor diagram, as shown in Fig. 3. It can be seen that the most accurate model in predicting the maximum ice thickness is XGBoost because it has the highest correlation coefficient and lowest RMSE. Decision tree and AdaBoost also have a higher level of accuracy in comparison to other models. RF and SVR have similar performance. KNN has the lowest correlation coefficient and the highest RMSE.

C. Icing Severity Level Prediction

1. Evaluation of Conventional Techniques

In icing severity level prediction, conventional methods, including OLR, KNN, SVM, and decision tree, are implemented. The classifiers' default parameter settings are adopted. The evaluation metric in the icing severity level prediction is accuracy, which is defined as the fraction of the amount of correct classifications. From the 10-fold stratified cross-validation, the average accuracy on testing data is

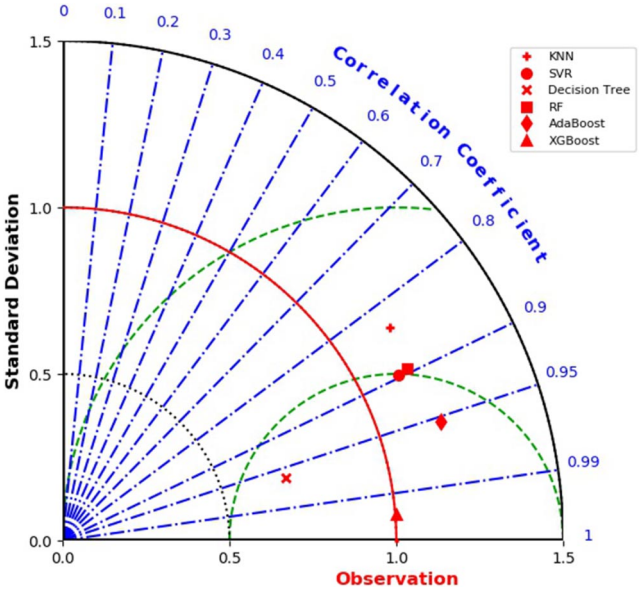

Fig. 3 Taylor diagram graphical presentation for the six predictive models (KNN, SVR, decision tree, RF, AdaBoost, and XGBoost) in estimating the maximum ice height.

shown in Table 8. It can be seen that KNN and SVM work the best among the conventional methods with the testing accuracy of 83%.

2. *Evaluation of Ensemble Methods*

Ensemble methods, including RF, AdaBoost, and XGBoost, are also used to predict the icing severity level. Based on GridSearchCV results, for RF, the number of trees is 25. The maximum number of estimators for AdaBoost is set to be 50. For XGBoost, the number of trees is 80, interaction depth is 10, shrinkage factor is 0.1, subsample ratio is 1, and minimum child weight is 0.1. The averaged accuracy of all the ensemble models on the testing dataset is summarized in Table 9. XGBoost works the best with the testing accuracy of 94%.

3. *Overall Comparison of Models for Icing Severity Level*

The performance of all models on the testing dataset is summarized in Fig. 4. The error bars represent each model's mean and variation of accuracy on all the 10-folds of data. XGBoost exceeds all the other models. On average, the ensemble models perform better than the conventional models. To further compare the performance between the conventional models and ensemble models, the confusion matrices generated by the two representative models (SVM and XGBoost) are summarized in Table 10. In the matrix, each row represents the actual category, and each column represents the predicted category. It can be observed that the diagonal values are much higher than the nondiagonal values for the XGBoost model. However, SVM generates a larger number of wrong prediction and some extreme error cases.

Table 8 Accuracy of conventional methods in predicting icing severity level

Model	Accuracy
OLR	0.74
KNN	0.83
SVM	0.83
Decision tree	0.81
Naive Bayes	0.67

Table 9 Accuracy of conventional methods in predicting icing severity level

Model	Accuracy
RF	0.87
AdaBoost	0.85
XGBoost	0.94

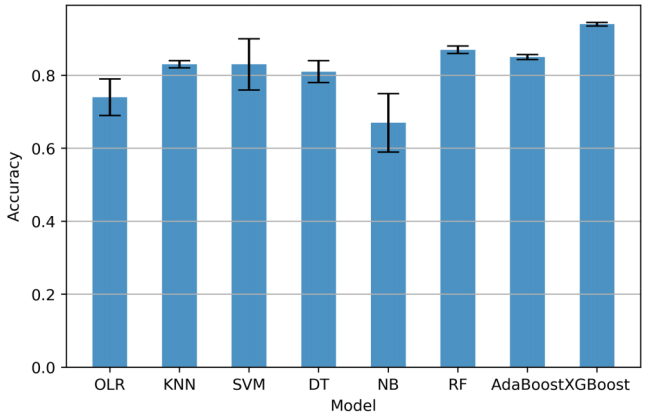

Fig. 4 Overall accuracy comparison between the eight predictive models (OLR, KNN, SVM, decision tree, naive Bayes, RF, AdaBoost, and XGBoost) in predicting the icing severity level.

IV. Conclusions

This paper investigates the application of multiple machine learning models in predicting aircraft icing severity under different flight conditions. Conventional models (MLR, OLR, KNN, SVM, decision tree, and naive Bayes) and ensemble models (RF, AdaBoost, and XGBoost) are investigated and compared in predicting three icing severity features: icing area, maximum ice thickness, and icing severity level. The parameter settings for the tested machine learning models are set to obtain the models' best performance. Various statistical performance measures are applied to explore the predictive capability of the applied models. It is found that RF exhibits superior predictive capability in the icing area and icing severity level evaluation cases, and decision tree generates a high level of accuracy in the maximum ice thickness prediction case. However, XGBoost achieves the best overall predictive performance among all the applied models in all three cases. It is concluded that the XGBoost model is able to extract a valuable nonlinear mapping relationship between flight conditions and icing severity features. It offers therefore a reliable and accurate approach for predicting the aircraft icing severity and should be explored in other engineering properties prediction in aircraft icing. However, it should be mentioned that the range of the predictions is limited to the range of the dataset. It cannot make accurate predictions if data inputted are out of the range of current given dataset. In the future, the plan is to extent the applicability of XGBoost model on aircraft icing by developing hybrid machine learning and computational fluid dynamics system with the aid of graphical processing unit parallelization to achieve accurate and fast evaluations of aircraft icing severity and aircraft performance degradation.

Acknowledgments

The funding from the Argonne National Laboratory for multiscale modeling of complex flows under grant number #ANL 0J-60008-0019A and from National Science Foundation under grant number #1854815 are gratefully acknowledged.

References

[1] Mclean, J., "Determining the Effects of Weather in Aircraft Accident Investigations," *Proceedings of the 24th AIAA Aerospace Sciences Meeting*, AIAA Paper 1986-0323, 1986. https://doi.org/10.2514/6.1986-323

[2] Cao, Y., Tan, W., and Wu, Z., "Aircraft Icing: An Ongoing Threat to Aviation Safety," *Aerospace Science and Technology*, Vol. 75, 2018, pp. 353–385. https://doi.org/10.1016/j.ast.2017.12.028

[3] Li, S., and Paoli, R., "Modeling of Ice Accretion over Aircraft Wings Using a Compressible OpenFOAM Solver," *International Journal of Aerospace Engineering*, Vol. 2019, 2019, p. 11. https://doi.org/10.1155/2019/4864927

Table 10 Confusion matrix results of SVM and XGBoost in predicting icing severity level

	SVM				XGBoost			
Category	Light	Moderate	Heavy	Severe	Light	Moderate	Heavy	Severe
Light	212	31	5	0	241	7	0	0
Moderate	23	109	18	0	1	142	7	0
Heavy	2	13	101	6	0	4	112	6
Severe	0	2	6	42	0	0	3	47

[4] Cao, Y., Ma, C., Zhang, Q., and Sheridan, J., "Numerical Simulation of Ice Accretions on an Aircraft Wing," *Aerospace Science and Technology*, Vol. 23, No. 1, 2012, pp. 296–304. https://doi.org/10.1016/j.ast.2011.08.004

[5] Wright, W. B., "User Manual for the NASA Glenn Ice Accretion Code LEWICE, Ver. 2.2.2," NASA CR-2002-211793, 2002, https://ntrs.nasa.gov/citations/20020080990.

[6] Li, S., and Paoli, R., "Numerical Study of Ice Accretion over Aircraft Wings Using Delayed Detached Eddy Simulation," *Bulletin of the American Physical Society*, No. Q23, 2019, Paper 00009, https://meetings.aps.org/Meeting/DFD19/Session/Q23.9.

[7] Ogretim, E., Huebsch, W., and Shinn, A., "Aircraft Ice Accretion Prediction Based on Neural Networks," *Journal of Aircraft*, Vol. 43, No. 1, 2006, pp. 233–240. https://doi.org/10.2514/1.16241

[8] Li, S., Paoli, R., and D'Mello, M., "Scalability of OpenFOAM Density-Based Solver with Runge-Kutta Temporal Discretization Scheme," *Scientific Programming*, Vol. 2020, 2020, Paper 9083620. https://doi.org/10.1155/2020/9083620

[9] Li, S., and Qiao, H., "Development of a Fast Fluid Dynamics Model Based on PISO Algorithm for Simulating Indoor Airflow," *ASME 2021 Heat Transfer Summer Conference*, ASME, New York, June 2021, Paper V001T01A010, https://asmedigitalcollection.asme.org/HT/proceedings-abstract/HT2021/84874/V001T01A010/1115079. https://doi.org/10.1115/HT2021-63909

[10] Chen, T., and Guestrin, C., "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Assoc. for Computing Machinery, New York, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785

[11] Li, S., Qin, J., and Paoli, R., "Data-Driven Machine Learning Model for Aircraft Icing Severity Evaluation," *Journal of Aerospace Information Systems*, Vol. 18, No. 11, 2021, pp. 876–881. https://doi.org/10.2514/1.I010978

[12] Darlington, R. B., and Hayes, A. F., *Regression Analysis and Linear Models: Concepts, Applications, and Implementation*, Guilford Publ., New York, 2016, https://books.google.com/books?id=l-8 KtAEACAAJ.

[13] Harrell, F., "Ordinal Logistic Regression," *Regression Modeling Strategies*, Springer Series in Statistics, Springer, Switzerland, 2015, pp. 311–325. https://doi.org/10.1007/978-3-319-19425-7

[14] Kamiński, B., Jakubczyk, M., and Szufel, P., "A Framework for Sensitivity Analysis of Decision Trees," *Central European Journal of Operations Research*, Vol. 26, 2018, pp. 135–159. https://doi.org/10.1007/s10100-017-0479-6

[15] Webb, G., Boughton, J., and Wang, Z., "Not So Naive Bayes: Aggregating One-Dependence Estimators," *Machine Learning*, Vol. 58, 2005, pp. 5–24. https://doi.org/10.1007/s10994-005-4258-6

[16] Altman, N. S., "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *American Statistician*, Vol. 46, 1992, pp. 175–185. https://doi.org/10.1080/00031305.1992.10475879

[17] Cortes, C., and Vapnik, V., "Support-Vector Networks," *Machine Learning*, Vol. 20, 1995, pp. 273–297. https://doi.org/10.1007/BF00994018

[18] Breiman, L., "Random Forests," *Machine Learning*, Vol. 45, 2001, pp. 5–32. https://doi.org/10.1023/A:1010933404324

[19] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning*, Springer, New York, 2009. https://doi.org/10.1007/978-0-387-84858-7

[20] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research*, Vol. 12, 2011, pp. 2825–2830, http://jmlr.org/papers/v12/pedregosa11a.html.

[21] Vapnik, V., *The Nature of Statistical Learning Theory*, Springer, New York, 2000, https://www.springer.com/gp/book/9780387987804.

[22] Brownlee, J., "Naive Bayes Classifier from Scratch in Python," 2019, https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/.

[23] Li, S., Qin, J., He, M., and Paoli, R., "Fast Evaluation of Aircraft Icing Severity Using Machine Learning Based on XGBoost," *Aerospace*, Vol. 7, No. 4, 2020, pp. 36–54. https://doi.org/10.3390/aerospace7040036

[24] Taylor, K. E., "Summarizing Multiple Aspects of Model Performance in a Single Diagram," *Journal of Geophysical Research: Atmosphere*, Vol. 106, 2001, pp. 7183–7192. https://doi.org/10.1029/2000 JD900719

P. Wei
Associate Editor