International Journal of Future Generation Communication and Networking
Vol. 13, No. 4 (2020), pp.71-78
http://dx.doi.org/10.33832/ijfgcn.2020.13.4.07

PLANT DISEASE DETECTION FOR HIGH DIMENSIONAL IMBALANCED DATASET USING AN ENHANCED DECISION TREE APPROACH

Anshul Bhatia[1], Anuradha Chug[2] and Amit Prakash Singh[3]

USIC&T, Guru Gobind Singh Indraprastha University
Sector – 16C, Dwarka, New Delhi-110078, India
anshul.usict.127164@ipu.ac.in[1], anuradha@ipu.ac.in[2], amit@ipu.ac.in[3]

***Abstract*— The purpose of the research is to find a robust and efficient model for plant disease detection. Therefore, the current study proposes an enhanced-DTC (Decision Tree Classifier) approach for high dimensional imbalanced dataset in plant disease diagnosis. In this approach, instead of just using traditional decision tree algorithm, its capabilities are enhanced with Random Over (RO) sampling method for class balancing and three well-known feature selection techniques, i.e., Consistency (Cons), Correlation-based Feature Selection (CFS), and Random Forest Importance (RFI) filter for dimensionality reduction. The proposed methodology aims to enhance the performance of the five most commonly used decision tree algorithms, namely, C4.5, Classification and Regression Tree (CART), Bagging CART (Bag-CART), Partial Decision Tree (PART-DT), and Boosted C5.0 (B-C5.0). Results specify that the enhanced-DTC approach performs superior to the existing decision tree algorithms for the multiclass Soybean Large (SBL) dataset. It has been observed that the enhanced-DTC approach with both RFI and C4.5 method performed the best with an Accuracy (ACC) of 98.10% and Area Under Curve (AUC) of 97.79%. A real-time application of the proposed model can be used by the agricultural experts to take preventive measures in the most sensitive areas that are prone to a particular disease. Hence, timely intervention would help in reducing the loss in productivity of plants which will further benefit the global economy, agricultural production, and the food industry.**

***Keywords*— Plant Disease, High Dimensional, Imbalanced Dataset, Enhanced Decision Tree Approach, Feature Selection, Random Over Sampling**

1. INTRODUCTION

Plant diseases, pathogens, and meteorological factors severely affect the quality and productivity of crops. Many countries witness yield losses in agricultural production due to pests and diseases. 70% of yield losses occur because of plant diseases, which further affect the global economy and agriculture production. The degree of financial losses due to plant diseases is very high as compared to worldwide yield losses of 600 million US dollars [1]. Soybean diseases are center of attention because they can decrease the worldwide production of soybean crop by 11% annually [2]. Various pathogens as well as viruses are responsible for the development of diseases in soybean crop. Some of these diseases cause extensive financial losses whereas some of them are not significant in present scenario. Some of the most severe soybean diseases are charcoal rot and stem canker caused by macrophomina phaseolina and diaporthe phaseolorum var. caulivora

Received: August 14, 2020
Reviewed: November 25, 2020
Accepted: December 4, 2020

ISSN: 2207-9645 IJFGCN

pathogens, respectively [3,4]. Precise diagnosis of these diseases is necessary for developing disease forecasting model for soybean. Classification techniques that could predict and distinguish different types of soybean diseases based on meteorological conditions and symptoms would help in controlling the diseases. The authors have already developed, tested, and validated various classification algorithms for the accurate diagnosis of different plant diseases such as potato early blight, potato late blight, tomato powdery mildew, tomato late blight, *etc.*, [5–11]. In the current study, authors have used publicly available dataset namely Soybean Large (SBL) for soybean disease diagnosis.

SBL dataset is a high dimensional multiclass dataset with non-uniform (imbalanced) class distribution. High dimensional dataset always suffer from problem of irrelevant features which may reduce the performance of a classifier. Similarly, imbalanced distribution of classes can also affect the performance of a classifier in learning phase. Therefore, in this research, an enhanced-Decision Tree Classifier (DTC) approach has been proposed for soybean disease detection. The proposed approach enhances the performance of different decision tree algorithms (C4.5, Partial Decision Tree (PART-DT), Classification and Regression Tree (CART), Bagging CART (Bag-CART), and Boosted C5.0 (B-C5.0)) by using a combination of feature selection algorithms (Consistency (Cons), Correlation-based Feature Selection (CFS), and Random Forest Importance (RFI) filter) for dimensionality reduction and Random Over (RO) sampling method for class balancing. Although, decision tree algorithm has already been successfully used for precise detection of plant diseases by Revathi et al. [12] and Corrales *et al.* [13] in their studies. Since, proposed approach enhances the capabilities of decision tree algorithms for better prediction. Performance of models developed from the proposed approach has been evaluated and compared using two most popular performance metrics, *i.e.*, Accuracy (ACC) and Area Under Curve (AUC).

The organization of the remaining paper is done as follows: Section 2 provides us with a literature review advancing towards materials and methods in Section 3. A detailed description of enhanced-DTC approach has been given in Section 4. Section 5 shows analysis of the results obtained through this experimentation followed by conclusion and direction of future work in Section 6.

2. RELATED WORK

In literature, several classification algorithms based on machine learning have been used by different researchers for plant disease prediction, and some of them are explained here. Chakraborty *et al.*, [14] have proposed a diseased forecasting model to analyze the severity of anthracnose disease of the tropical pasture legume “Stylosanthes scabra” using Artificial Neural Network (ANN) algorithm. After three years, a neural network based model was developed by Klem *et al.*, [15] for prediction of deoxynivalenol content in wheat grain. In one of the studies, Rumpf *et al.*, [16] examined the Support Vector Machine (SVM) classifier for discrimination between diseased and non-diseased sugar beet leaves. In the paper, they have also used multiple classification algorithms like Decision Tree, ANN and SVM to differentiate between the diseases found in sugar beet leaves named powdery mildew, leaf rust, and Cercospora leaf spot. Further, Bauer et al. [17] have used *k*-Nearest Neighbour (*k*NN) classifier to categorize healthy and infected sugar beet leaves having diseases named as "Cercosporabeticola" and "Uromycesbetae." Later, Fuentes *et al.*, [18] have used different deep learning methods to detect tomato diseases. Further, Verma *et al.*, [5–7] have used various machine learning techniques and statistical techniques on image as well as sensor data for tomato plant disease investigation. In current study, we have also tried to give our contribution towards plant disease prediction by proposing a new approach named as “enhanced-DTC” approach for soybean disease forecasting.

3. MATERIALS AND METHODS

The current section discusses the techniques and datasets used in the present study. The subsequent subsection 3.1 highlights the SBL dataset. Afterward, subsection 3.2 elaborates the feature selection techniques followed by RO sampling method in subsection 3.3. Further, decision tree algorithms and performance evaluation measures are explained in subsections 3.4 and 3.5, respectively. R-Studio Version 1.1.463 has been used to implement the proposed approach.

3.1 DATASET

The proposed approach is applied on SBL dataset taken from the UCI Repository of Machine Learning [19]. This dataset contains information regarding many soybean diseases based on various meteorological factors and plants' global and local attributes of plants. Overall, SBL dataset includes 683 instances with some missing values, which were handled using random forest imputation method. This method uses random forest algorithm to predict missing values after training the classifier on observed values [20]. Finally, SBL dataset has 35 features and one target disease class, which contains 19 soybean disease classes.

3.2 FEATURE SELECTION TECHNIQUES

Feature selection techniques are used to remove subset of redundant and irrelevant features from the datasets in order to minimize unpredictability and inconsistencies from the trained model [21]. The three feature selection techniques used in this study are explained here. CFS filter algorithm works on correlation based heuristic estimation function for the ranking of features, RFI filter uses RF algorithm to find weights of attributes, whereas, Cons filter uses best fit search algorithm to identify the proper feature subset [22]. Table I shows the selected features from SBL dataset for each of the three feature selection techniques. It also includes total number of features present in SBL dataset.

Table I. Feature Selection Table for Enhanced-DTC Approach

Total Number of Features (35)	
Crop-hist, Precipitation, temperature, Leaf spot-size, Hail, Plant stand, Leaf-shredding, Time of incidence, Seed Germination, Seed Treatment, Plant Growth, Leaves, Leaf mildew growth, Stem, Presence of Lodging, Area-damaged, Canker lesion color, Stem cankers, Leaf spots-margin, External decay, Mycelium on stem, Fungal fruiting body on stem, Sclerotia – internal or external, Internal discoloration, Leaf malformation, Fruits Pods, Fruits Spots, Seed, Seed discoloration, Seed shriveling, Mold growth, Seed size, Roots, Severity, Leaf spots-halo	
Feature Selection Algorithm	**Selected Features**
RFI Filter (30)	Time of incidence, canker lesion color, internal discoloration, leaf mildew growth, precipitation, fruits spots, leaf spot-size, temperature, fungal fruiting body on stem, fruits pods, leaf malformation, stem cankers, mold growth, external decay, leaf spots-margin, seed, leaf spots-halo, stem, plant growth, area-damaged, roots, seed discoloration, leaves, severity, leaf-shredding, sclerotia, hail, seed size, plant stand, seed shriveling
CFS Filter (17)	Leaf spot-size, time of incidence, external decay, precipitation, leaf mildew growth, temperature , area-damaged, leaf spots-halo, fungal fruiting body on stem, leaf spots-margin, internal discoloration, fruits pods, fruits spots, seed, canker lesion color, roots, leaf malformation
Cons Filter (14)	Hail, time of incidence, precipitation, temperature, seed treatment, leaf spot-size, leaf-shredding, germination, leaf mildew growth, canker lesion color, area-damaged, fruits spots, seed size, crop-hist

It is evident from Table I that there are 8 such features, *i.e.*, precipitation, temperature, leaf spot size, time of incidence, leaf mildew growth, area-damaged, canker lesion color, and fruits spot, which are selected by all the three feature selection algorithms. These features of SBL dataset can be assumed as the most relevant features during the development of disease prediction models. However, 4 features out of 35, namely, seed germination, presence of loading, mycelium on stem, and sclerotia – internal or external can be claimed as the least relevant features for disease prediction, because these are not selected by any of the three feature selection algorithms.

3.3 RANDOM OVER (RO) SAMPLING

RO sampling [23] is used to handle the imbalance dataset. This technique randomly copies existing data samples of minor classes to increase the training data observations for balancing it with major classes. In the current study, RO sampling is used to generate random subset of the SBL dataset and balance its class distribution, as shown in Table II.

3.4 DECISION TREE ALGORITHMS

A decision tree is a tree type structure which divides the whole dataset into mutually exclusive spaces and each space has a class label which describes all the data points associated with the dataset. Five decision tree algorithms used in this study are: CART [24], C4.5 [25], PART-DT [26], Bag-CART [27], and B-C5.0 [28].

3.5 PERFORMANCE EVALUATION METRICS

Two commonly used performance metrics namely ACC and AUC [29] have been used in the current study for evaluating the performance of proposed models. ACC is defined as the number of accurate predictions out of the total number of predictions, whereas; AUC is also a performance metric which sum up the well-known Receiver Operating Characteristic (ROC) curve in a specific value. ROC is a plot between the Sensitivity/Recall/True Positive Rate (TPosR) and False Positive Rate (FPosR).

Table II. Class Distribution Before and After Applying RO Sampling on SBL Dataset

Class labels	Before RO Sampling	After RO Sampling	Class labels	Before RO Sampling	After RO Sampling
1. herbicide-injury	8	92	11. diaporthe-stem-canker	20	92
2. cyst-nematode	14	92	12. charcoal-rot	20	92
3. diaporthe-pod-&-stem-blight	15	92	13. purple-seed-stain	20	92
4. 2-4-d-injury	16	92	14. anthracnose	44	92
5. phyllosticta-leaf-spot	20	92	15. brown-stem-rot	44	92
6. downy-mildew	20	92	16. phytophthora-rot	88	92
7. powdery-mildew	20	92	17. alternarialeaf-spot	91	92
8. rhizoctonia-root-rot	20	92	18. frog-eye-leaf-spot	91	92
9. bacterial-pustule	20	92	19. brown-spot	92	92
10. bacterial-blight	20	92			

4. ENHANCED-DTC APPROACH

Enhanced-DTC approach uses a decision tree algorithm, a feature selection technique and RO sampling. The main objective of this approach is to enhance the performance of decision tree algorithm for soybean disease diagnosis. Research methodology for enhanced-DTC approach is shown in Fig. 1. The proposed methodology is applied on the SBL dataset. First of all, a feature selection technique (RFI, CFS, and Cons) is selected and applied on the SBL dataset to get the appropriate features for the soybean disease detection. After feature selection, RO sampling is successfully applied for balancing the imbalanced SBL dataset. The balanced data obtained after RO sampling is divided into

70% train-set and 30% test-set. Subsequently, a decision tree algorithm out of CART, C4.5, PART-DT, Bag-CART, and B-C5.0 is picked and applied on the train-set to get a prediction model for soybean disease detection. Afterwards, the performance of resultant model is evaluated on test-set by ACC and AUC metrics. Lastly, the performance of the enhanced-DTC approach is observed with regard to every feature selection technique and decision tree algorithm.

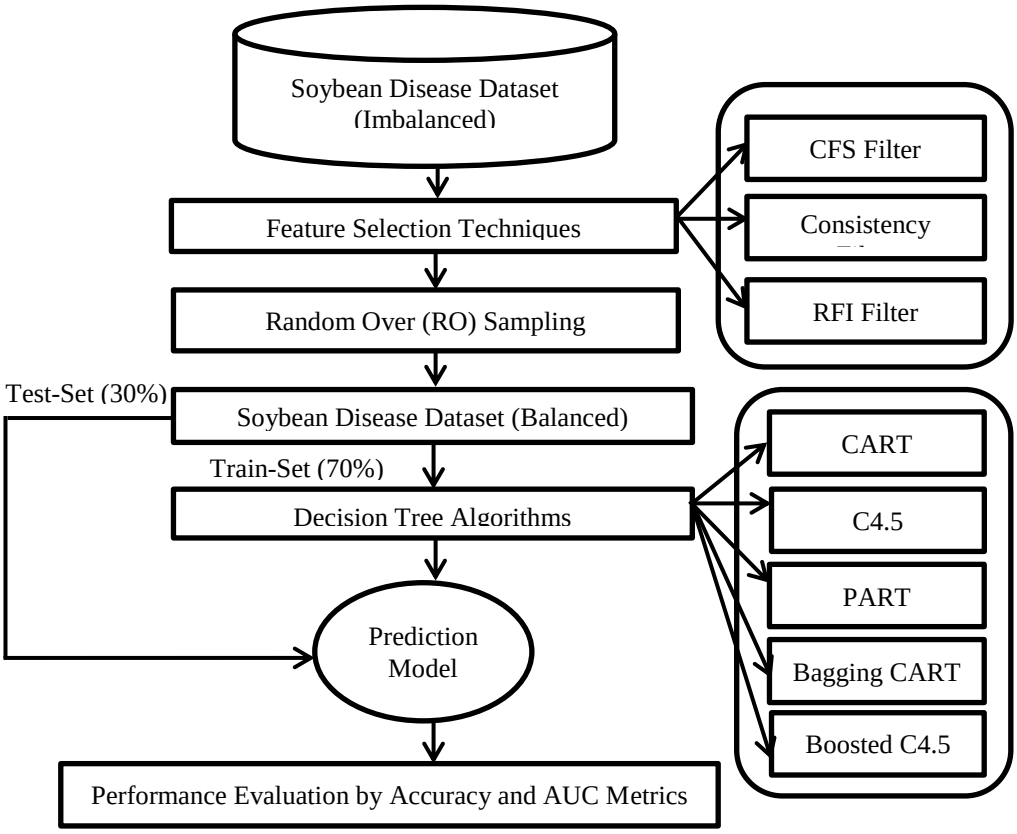

Fig. 1 Research Design of Enhanced-DTC Approach

5. RESULTS AND DISCUSSIONS

This section highlights the results of the proposed approach applied on SBL dataset. At first, a feature selection algorithm is selected from Cons, CFS, and RFI filters. It is evident from Table III that SBL dataset was imbalanced before applying RO sampling. It contained minority classes such as herbicide-injury, cyst-nematode, and diaporthe-pod-&-stem-blight. It also included some of the majority classes such as brown-spot, frog-eye-leaf-spot, and alternarialeaf-spot. Further, RO was applied for balancing the dataset. Subsequently, balanced SBL dataset was divided into 70-30 train-test ratio. Later, a decision tree algorithm was selected out of CART, C4.5, PART-DT, Bag-CART, and B-C5.0 algorithms and applied on 70% train-set to get the prediction model. Performance of the obtained model from enhanced-DTC approach was tested using 30% test-set in terms of ACC and AUC metrics for all the three feature selection algorithms. The proposed approach is tested with respect to each of the decision tree algorithms. Resultantly, the enhanced-DTC approach performs better than the existing decision tree algorithm for SBL dataset.

Table III. Performance Comparison of Decision Tree Algorithm and Enhanced-DTC Approach for Soybean Disease Diagnosis in Terms of ACC and AUC Metrics

Metrics	**Decision Tree Algorithm**			
	Existing Approach	Enhanced DTC Approach using RO sampling and Feature Subset Selection		
	CART	CFS-CART	Cons-CART	RFI-CART
ACC	77.56%	93.91%	*93.90%*	94.29%
AUC	83.59%	94.74%	*94.72%*	94.76%
	C4.5	CFS-C4.5	Cons-C4.5	RFI-C4.5
ACC	88.78%	97.52%	97.71%	**98.10%**
AUC	91.08%	97.17%	97.48%	**97.79%**
	PART-DT	CFS- PART-DT	Cons- PART-DT	RFI- PART-DT
ACC	85.73%	96.76%	97.71%	97.33%
AUC	89.43%	96.67%	97.58%	97.39%
	Bag-CART	CFS- Bag-CART	Cons- Bag-CART	RFI- Bag-CART
ACC	91.71%	97.33%	97.33%	97.33%
AUC	93.61%	96.82%	96.88%	96.88%
	B-C5.0	CFS- B-C5.0	Cons- B-C5.0	RFI- B-C5.0
ACC	90.24%	97.33%	97.14%	97.52%
AUC	92.79%	97.19%	96.97%	97.24%

Table III shows the performance comparison of existing decision tree algorithms and enhanced-DTC approach in terms of ACC and AUC metrics. In this table, decision tree algorithm column has been divided into two parts, where first part shows the existing decision tree approach and the second part indicates the enhance decision tree approach with the corresponding feature selection algorithm. For example, in CFS-CART approach, CFS shows the feature selection algorithm and CART is type of decision tree approach. It can be seen from Table III that the enhanced-DTC algorithm with RFI and C4.5 technique performed the best with 98.10% ACC and 97.79% AUC (marked in bold), whereas, the enhanced-DTC with Cons filter along with CART algorithm performed the worst with 93.90% ACC and 94.74% AUC (marked in italic). It is also evident from Table III that in case of CART algorithm, the enhanced-DTC approach with RFI technique among all the feature selection techniques performed the best with 94.29% ACC and 94.76% AUC. Similarly, in case of C4.5 algorithm, the RFI technique again overruled other feature technique algorithms with 98.10% ACC and 97.79% AUC. However, in case of PART-DT algorithm, Cons filter gave the best results for the proposed approach in terms of ACC and AUC, having the values 97.71% and 97.58%, respectively. Further, In terms of Bag-CART algorithm, the proposed approach achieved better results with both Cons and RFI filter with ACC 97.33% and AUC 96.88%. Again, in case of B-C5.0 algorithm, the enhanced-DTC approach performed superior with RFI filter in terms of ACC and AUC, having the values 97.52% and 97.24%, respectively.

6. CONCLUSION AND FUTURE DIRECTIONS

The current study discusses an enhanced-DTC approach for improving the performance of decision tree algorithm for soybean disease diagnosis. The proposed algorithm is successfully applied on high dimensional multi-class SBL dataset with imbalanced class distribution. The enhanced-DTC approach with Cons, CFS, and RFI filter indicates the improvement in performance of soybean diagnosis as compared to normal decision tree algorithm. C4.5 decision tree algorithm with RFI filter performed the best with an accuracy of 98.10 %, whereas, the CART algorithm with cons filter performed the worst with 93.90% accuracy. Therefore, it is concluded that the enhanced-DTC approach is a good substitute for soybean diseases diagnosis and multi-class classification problem. In future, this work can be extended with more feature selection

algorithms. In addition to this, a hybrid solution of classification algorithms along with soft computing techniques like genetic or fuzzy algorithms can also be proposed and tested on SBL dataset. More resampling techniques can be used to do better comparative analysis.

ACKNOWLEDGMENT

We are indebted to the Department of Science and Technology (DST) for their financial support in implementation of this research work under the project titled "Application of Internet of Things (IoT) in Agriculture Sector", DST/Reference.No.T-319/2018-19.

REFERENCES

[1] Chaudhary, A., Kolhe, S. and Kamal, R., "An improved random forest classifier for multi-class classification", Inf. Process. Agric., vol. 3, (2016), pp. 215-222.
[2] Ryley, M., "Effects of some diseases on the quality of culinary soybean seed", Proc. 12th Australian Soybean Conf., Toowoomba. Northern Australian Soybean Industry Association, Toowomba, (2003).
[3] Keeling, B. L., "A seedling test for resistance to soybean stem canker caused by Diaporthe phaseolorum var", Caulivora, Phytopathology, vol. 72, (1982), pp. 807-809.
[4] Su, G., Suh, S.-O., Schneider, R. W. and Russin, J. S., "Host specialization in the charcoal rot fungus", Macrophomina phaseolina, Phytopathology, vol. 91, (2001), pp. 120-126.
[5] Verma, S., Chug, A. and Singh, A. P., "Prediction Models for Identification and Diagnosis of Tomato Plant Diseases", 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), (2018), pp. 1557-1563.
[6] Verma, S., Chug, A., Singh, A. P., Sharma, S. and Rajvanshi, P., "Deep Learning-Based Mobile Application for Plant Disease Diagnosis: A Proof of Concept With a Case Study on Tomato Plant", in Applications of Image Processing and Soft Computing Systems in Agriculture, IGI Global, (2019), pp. 242-271.
[7] Verma, S., Bhatia, A., Chug, A. and Singh, A. P., "Recent Advancements in Multimedia Big Data Computing for IoT Applications in Precision Agriculture: Opportunities, Issues, and Challenges", in Multimedia Big Data Computing for IoT Applications, Springer, (2020), pp. 391-416.
[8] Verma, S., Chug, A. and Singh, A. P., "Exploring capsule networks for disease classification in plants", J. Stat. Manag. Syst., vol. 23, (2020), pp. 307-315.
[9] Verma, S., Chug, A. and Singh, A. P., "Application of convolutional neural networks for evaluation of disease severity in tomato plant", J. Discret. Math. Sci. Cryptogr., vol. 23, (2020), pp. 273-282.
[10] Bhatia, A., Chug, A. and Singh, A. P., "Application of extreme learning machine in plant disease prediction for highly imbalanced dataset", J. Stat. Manag. Syst., (2020), pp. 1-10.
[11] Bhatia, A., Chug, A. and Singh, A. P., "Hybrid SVM-LR Classifier for Powdery Mildew Disease Prediction in Tomato Plant", in 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN), (2020), pp. 218-223.
[12] Revathi, P., Revathi, R. and Hemalatha, M., "Comparative Study of Knowledge in Crop Diseases Using Machine Learning Techniques", Int. J. Comput. Sci. Inf. Technol., vol. 2, (2011), pp. 2180-2182.
[13] Corrales, D. C., Corrales, J. C. and Figueroa-Casas, A., "Towards detecting crop diseases and pest by supervised learning, Ing. y Univ., vol. 19, (2015), pp. 207-228.
[14] Chakraborty, S., Ghosh, R., Ghosh, M., Fernandes, C. D., Charchar, M. J. and Kelemu, S., "Weather-based prediction of anthracnose severity using artificial neural network models", Plant Pathol., 53 (2004), pp. 375-386.
[15] Klem, K., Vanova, M., Hajslova, J., Lancová, K. and Sehnalová, M., "A neural network model for prediction of deoxynivalenol content in wheat grain based on weather data and preceding crop", Plant Soil Environ, vol. 53, (2007), pp. 421.
[16] Rumpf, T., Mahlein, A.-K., Steiner, U., Oerke, E.-C., Dehne, H.-W. and Plümer, L., "Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance", Comput. Electron. Agric., vol. 74, (2010), pp. 91-99.
[17] Bauer, S. D., Korč, F. and Förstner, W., "The potential of automatic methods of classification to identify leaf diseases from multispectral images", Precis. Agric., vol. 12, (2011), pp. 361-377.
[18] Fuentes, A., Yoon, S., Kim, S. and Park, D., "A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition", Sensors, vol. 17, (2017), pp. 2022.
[19] {UCI} Machine Learning Repository, (2017).
[20] Stekhoven, D. J., "missForest: Nonparametric missing value imputation using random forest, Astrophys", Source Code Libr. (2015).
[21] Das, S., "Filters, wrappers and a boosting-based hybrid for feature selection", Icml, vol. 1, (2001), pp. 74-81.

[22] Romanski, P., Kotthoff, L. and Kotthoff, M. L., "Package 'FSelector', Repos", CRAN, (2018).
[23] Branco, P., Ribeiro, R. P. and Torgo, L., "UBL: an R package for utility-based learning", arXiv Prepr. arXiv1604.08079, (2016).
[24] Steinberg, D. and Colla, P., "CART: classification and regression trees, top ten algorithms data Min", vol. 9, (2009), pp. 179.
[25] Quinlan, J. R., "C4. 5: Programs for Machine Learning", Elsevier, (2014).
[26] Frank, E. and Witten, I. H., "Generating accurate rule sets without global optimization", (1998).
[27] Cutler, A. and Zhao, G., "Pert-perfect random tree ensembles", Comput. Sci. Stat., vol. 33, (2001), pp. 490-497.
[28] Pashaei, E., Ozen, M. and Aydin, N., "Improving medical diagnosis reliability using Boosted C5. 0 decision tree empowered by Particle Swarm Optimization", in 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 7230-7233.
[29] Huang, J. and Ling, C. X., "Using AUC and accuracy in evaluating learning algorithms", IEEE Trans. Knowl. Data Eng., vol. 17, (2005), pp. 299-310.

