# A machine learning model for studying the seasonality of aphids in wheat-based cropping systems of the terai zone of Darjeeling, West Bengal, India

**Biwash Gurung[1], Suprakash Pal[2], Md. Wasim Reza[3], Bishal Gurung[4,*] and Achal Lama[5]**

[1]School of Agricultural Sciences, GD Goenka University, Gurugram 122 103, India
[2]Directorate of Research (RRS-TZ), Uttar Banga Krishi Viswavidyalaya, Pundibari, Cooch Behar 736 165, India
[3]Regional Research Sub-station (Terai Zone) Kharibari, Uttar Banga Krishi Viswavidyalaya, Pundibari, Cooch Behar 736 165, India
[4]Department of Statistics, North-Eastern Hill University, Shillong 793 022, India
[5]ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India

**The primary goal of this study is to determine the effect of weather variables on aphid populations and development of a weather-based forewarning model using a powerful machine learning technique called random forest. The developed model could be employed to formulate proper management strategies to help the farming community control aphid infestation.**

**Keywords:** Aphid infestation, forewarning model, machine learning, random forest, weather parameters, wheat-based cropping system.

WHEAT *Triticum aestivum* (Linnaeus) is a crop belonging to the family Gramineae. It is a commercially important crop grown worldwide in a range of climatic conditions. Wheat is the second most important food crop, which contributes about 35% of the total food grain production. It is thus a major contributor to the agrarian economy of India[1]. Wheat is also a staple food, predominantly in the northern and northwestern parts of India. It is the most important cereal crop in temperate areas of the world and a staple food for more than 35% of the world's population, covering at least 43 countries, and occupies 23% of the global cultivated area[2]. According to FAOSTAT[3], wheat production has increased from 235 million tonnes (mt) in 1961 to an estimated 733 mt in 2015. One of the major constraints limiting the yield of agricultural products is the attack by insect pests. Wheat production regularly suffers from threats due to diseases and pests, with the annual capital loss due to insect pests amounting to around Rs 413.68 billion. Insect pests are known to attack wheat crops worldwide[4]. In India, several insect pests infesting wheat crops have been reported from planting until the harvest stage[5]. Among the various biotic stresses reported on this crop, aphids are one of the most important and destructive pests in West Bengal and many parts of India. Gurung *et al.*[6] studied the effect of weather parameters on the population dynamics of coccinellids in different crop ecosystems and established the relationship between populations of predatory coccinellids and abiotic factors. Gurung *et al.*[7] have also developed forewarning models based on weather variables for tomato leaf curl infestation employing beta regression methodology.

Previously, the problem of insect-pests was not severe in wheat, but with fluctuating climate, monocropping, and promotion of new crop production technologies like conservation agriculture technologies, minor and random pests are now becoming consistent and major pests of this croprequire regular monitoring. Keeping this in mind, a statistical model employing machine learning (ML) technique has been developed in this study for modelling the seasonality of aphids in wheat (*Triticum aestivum* L.)-based cropping systems of the terai zone of Darjeeling, West Bengal.

The ML technique was used to select important weather variables that are related to aphid populations. The random forest (RF) methodology has been employed to model the variability in pest infestation data. The developed aphid forewarning model could be used to formulate proper management strategies to help the farming community.

## Materials and methods

### *Meteorological data and study location*

Field experiments were conducted at experimental plots in the research station farm at the Regional Research Substation (Terai Zone), Kharibari, Uttar Banga Krishi Viswavidyalaya, Darjeeling, West Bengal during the *rabi* season of 2018–19 and 2019–20. Geographically, the farm is located at an elevation of 113 m amsl, at 26.55°N lat. and 88.19°E long. The area comes under the sub-Himalayan terai agroclimatic zone and the administrative jurisdiction of Darjeeling. The terai zone comprises the entire portion of Jalpaiguri and Cooch Behar districts, the Islampur sub-division of

*For correspondence. (e-mail: bishalgurung@nehu.ac.in)

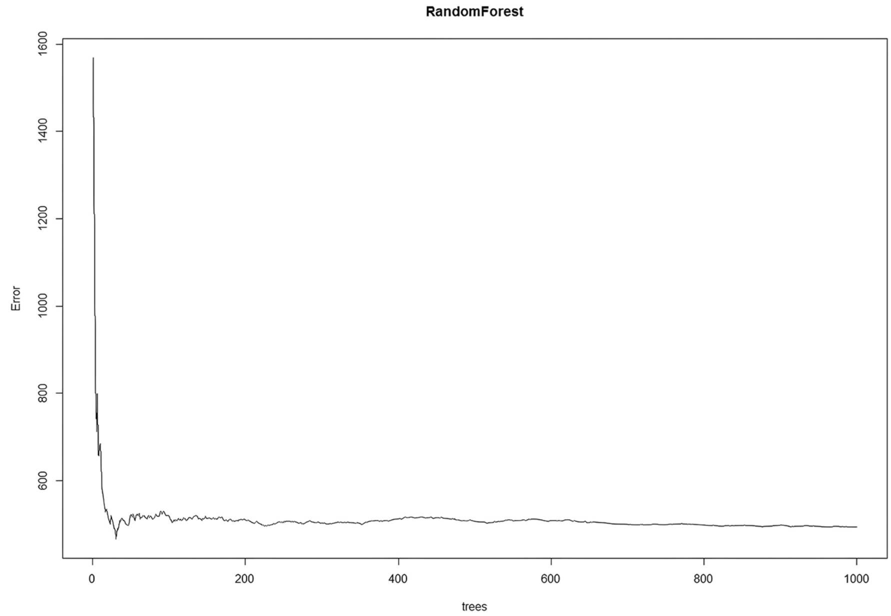

**Figure 1.** Plot of the random forest technique.

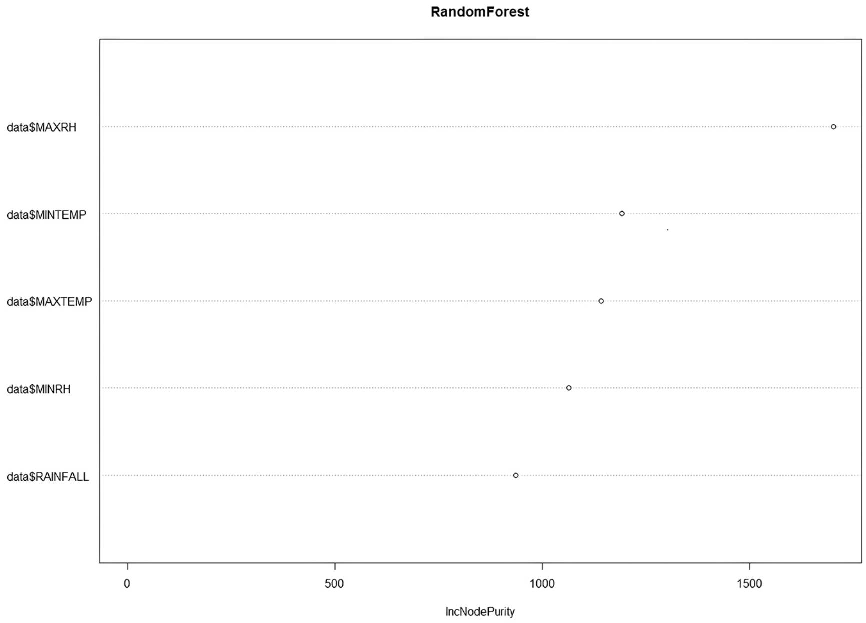

**Figure 2.** Plot of feature selection using increase in purity importance criterion.

North Dinajpur district and the Siliguri sub-division of Darjeeling district in West Bengal. The experiment was conducted to detect, categorize and document the pests and natural enemies associated with wheat, as this helps adopt preventive measures and timely management strategies.

## *Weather parameters*

The available meteorological data on weather variables, viz. rainfall (mm), maximum relative humidity ($\mathrm{RH}_{\max}$), minimum relative humidity ($\mathrm{RH}_{\min}$), maximum temperature ($T_{\max}$), minimum temperature ($T_{\min}$) and their difference ($T_{\max} - T_{\min}$) were collected from All India Coordinated Research Project (AICRP) on Agro-Meteorology, Uttar Banga Krishi Viswavidyalaya, Darjeeling.

## *Model description: random forest regression*

There are several ML techniques reported in the literature. One important technique is the RF, which is an ensemble

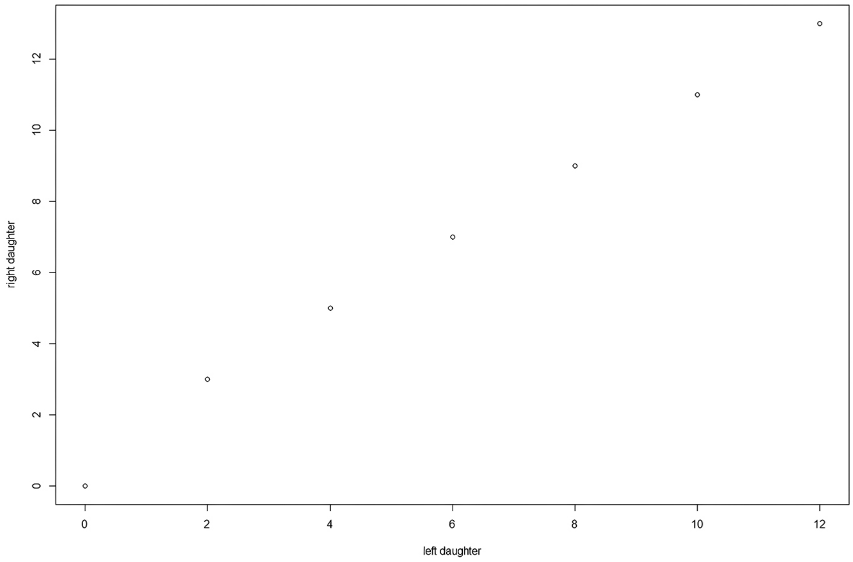

**Figure 3.** Tree diagram.

**Table 1.** Fitted model employing feature selection through random forest

| Variable | Parameter estimate | Standard error | $t$-value | Pr > $|t|$ | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | 1220.02 | 425.65 | 2.86 | 0.011 | 312.75 | 2127.28 |
| MAXRH | –10.48 | 3.96 | –2.64 | 0.018 | –18.93 | –2.03 |
| MAXTEMP | –12.39 | 3.64 | –3.40 | 0.003 | –20.14 | –4.63 |
| MINTEMP | 6.65 | 2.53 | 2.62 | 0.018 | 1.25 | 12.04 |

**Table 2.** Fitted model employing feature selection through stepwise regression

| Variable | Parameter estimate | Standard error | $t$-value | Pr > $|t|$ | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | –22.71 | 17.33 | –1.31 | 0.20 | –58.28 | 12.84 |
| MINTEMP | 3.78 | 1.511 | 2.50 | 0.01 | 0.68 | 6.88 |

of decision trees. RF is made up of various trees that are assembled in an unambiguous ‘random’ manner. Each tree is built on a distinct sample of rows, and each node is fragmented into various sets of features. Prediction is obtained from each tree, and the predictions thus obtained are averaged to give a single result. In RF variable selection, which is also called feature selection in ML jargon, is carried out by estimating the importance of each variable or feature.

The usual method of estimating variable importance is the average decrease in impurity, also known as Gini importance. Using RF algorithm, it is possible to determine how much each variable or feature decreases the impurity. The less a feature decreases the impurity, the less important it is, and vice versa. The final variable importance in the random forest is determined by averaging the decrease in impurity from each variable or feature across all the trees.

*Accuracy-based importance:* For each tree, an out-of-bag (OBB) sample of data is not utilized during the development process. The OBB sample is used to decide the importance of a particular feature by checking the prediction accuracy of the sample. The shuffling of values of the variable in the OBB sample is done at random, while all other variables remain unchanged. Then, the decrease in the accuracy of the shuffled data is noted. These values of this measure help us in estimating the reduction in accuracy when a particular variable is eliminated and, conversely, how the accuracy increases by including a variable. Accuracy is calculated by:

$$\text{Accuracy} = \frac{\text{Number of correctly identified members of a class}}{\text{Total no. of times the model predicted that class}}.$$

*Gini-based importance:* The Gini impurity (GI) criterion is used in selecting the variable to be split at each node when assembling a tree. Every time a variable is selected to split a node, the sum of the Gini reduction over all trees of the forest is determined for that particular variable.

**Table 3.** Comparison of goodness-of-fit performance

| | Aphid population | |
|---|---|---|
| Criterion | Stepwise regression | Random forest regression |
| MAE | 15.30 | 14.83 |
| MSE | 336.32 | 333.17 |

**Table 4.** Forecast performance for hold-out data

| | Aphid population | |
|---|---|---|
| Criterion | Stepwise regression | Random forest regression |
| MAPE | 25.24 | 25.10 |
| MSPE | 778.15 | 757.08 |
| RMAPE | 147.23 | 139.26 |

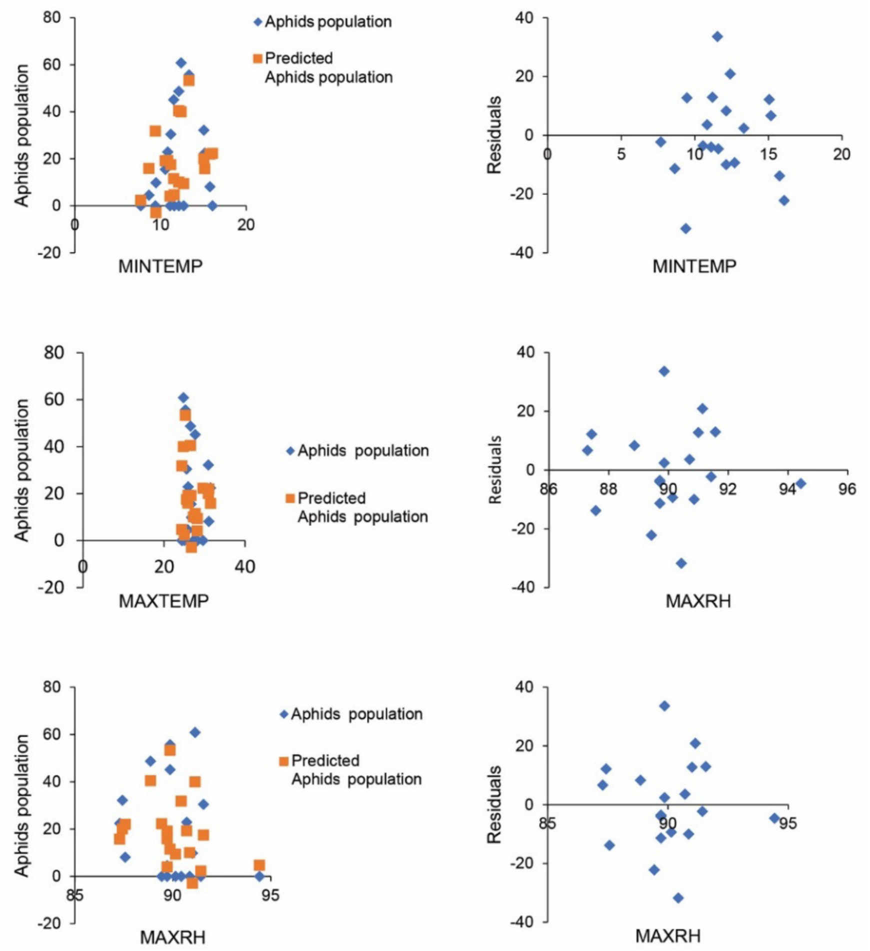

**Figure 4.** (Left) Line-fit plot and (right) residual plot of independent weather variables.

The functional form of Gini impurity is given by:

$$\mathrm{GI} = 1 - \sum_{i=1}^{n} (p)^2.$$

RF is usually employed for regression and discrimination. In discrimination, the goal is to predict the group label of each sample in the dataset. In regression, the goal is to predict the dependent variable (e.g. yield or infestation of pests/diseases) based on the independent variables of the data. RF is widely used because it is easy to train, can be applied for high-dimensional data, and is extremely accurate. RF can also handle missing values and can be employed for imbalanced datasets. After training the data using RF

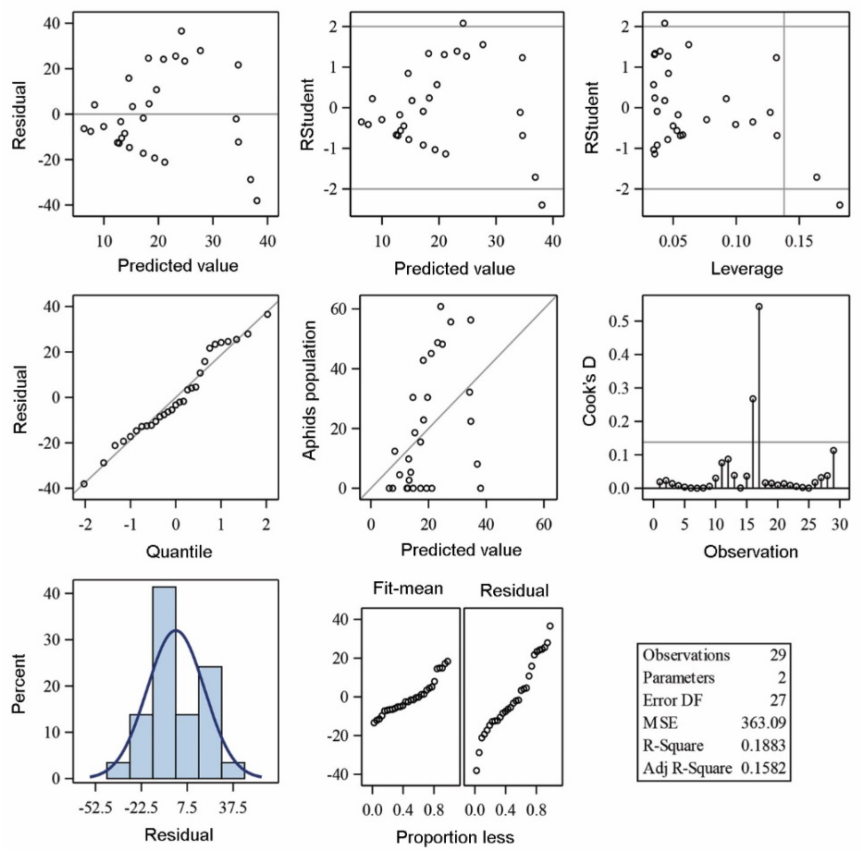

**Figure 5.** Fit diagnostics for the aphid population.

algorithm, it can be used to make predictions. For this, RF uses the predictions of each decision tree and combines them by averaging.

The SAS code and R scripts used for the analysis are given in Annexure 1 for ease of implementation.

The feature selection algorithms assist in selecting only those relevant features in the predictive algorithm. Instead of a complete set of features, feature subsets give better results for the same algorithm with less computational time. Feature selection algorithms improve the performance of software defect prediction (SDP) models. There is no single best feature selection method because the performance of different methods varies according to the datasets and models used for prediction. Low-importance variables, on the other hand, may be eliminated from a model, making it easier and faster to fit and predict. The correlation-based method gave the best results for alfalfa yield prediction. RF has been reported best for sugarcane (*Saccharum* spp.) yield modelling with data obtained from a sugarcane mill[8–13].

## Results and discussion

The feature selection of the available meteorological data on weather variables was carried out using the powerful RF technique, and important variables were selected.

Figure 1 shows the number of trees utilized in the RF regression along with its error values. This number should not be set too small to ensure that every input row gets predicted at least a few times.

From the analysed data, we found that maximum RH, minimum temperature and maximum temperature significantly affected the aphid population. We have considered both the criteria, i.e. percentage increase in mean squared error (MSE) and increase in node purity, for feature selection (Figures 2 and 3). Thereafter, the variable selected were used to develop a forecasting model for the seasonality of aphids. Table 1 shows the fitted model after feature selection using the RF technique.

From the fitted model, we conclude that feature selection provides statistically significant variables, as all three variables selected using RF have a significant effect on aphid infestation.

Further, we compared the usual regression methodology using variables selected through stepwise regression (Table 2). The comparison was made with respect to modelling as well as forecasting the performance of the two competing models, viz. RF and stepwise regression. Using stepwise regression, we developed a regression model from the available regressor variables by selecting and removing these variables based on their *P*-values stepwise untill there was no variable left to be selected or removed.

**Annexure 1.** SAS code for stepwise regression.

```
data regression;
input RAINFALL    MAXRH     MINRH     MAXTEMP  MINTEMP
        Aphidspopulation;
cards;
;
.
.
.

ods rtf;


proc reg;

model Aphidspopulation = RAINFALL    MAXRH     MINRH     MAXTEMP
        MINTEMP/selection=STEPWISE;

run;

ods rtf close;
```

**R Code for random forest variable selection**

```
install.packages("randomForest")

library("randomForest")

data=read.csv(file.choose())

RandomForest=randomForest(data$Aphidspopulation~data$RAINFALL+data$MAXRH+data$
        MINRH+data$MAXTEMP+data$MINTEMP, data, ntree=1000, keep.forest=TRUE,
        importance=TRUE)

plot(RandomForest)

importance(RandomForest)

varImpPlot(RandomForest,              sort=TRUE,                n.var=min(30,
        nrow(RandomForest$importance)),type=2,       class=NULL,       scale=TRUE,
        main=deparse(substitute(RandomForest)))

getTree(RandomForest,k=10,labelVar=FALSE)
```

For comparison, we determined the modelling and forecasting performance of the fitted model. For forecasting purposes, we took the last five points of the datasets.

Further, the goodness-of-fit performance of the fitted models was compared using MSE and mean absolute error (MAE) criteria (Table 3). The forecasting performance was also determined using mean square prediction error (MSPE) and mean absolute prediction error (MAPE) (Table 4). Tables 3 and 4 indicate the superiority of RF regression over stepwise regression for modelling as well as forecasting the dataset under consideration.

Figure 4 demonstrate the predicted values of the aphid population with respect to the each independent weather variables. The fit diagnostic of the RF based feature selection model for predicting aphid population has been depicted by Figure 5.

## Conclusion

The present investigation was conducted for feature selection in aphid populations using the RF ML technique. A comparison was also made to check the improvement of RF over the usual variable selection method from modelling as well as forecasting points of view. Thus, the model developed can be employed by various plant protection agencies to formulate management strategies against aphids in the terai zone of Darjeeling. Future work may explore the possibility of employing XGboost, support vector machine, Gaussian process regression and balanced repeated replication.

1. Nagarajan, S., Wheat production in India: a success story and future strategies. *Indian Farm.*, 2000, **9**, 915.
2. Khakwani, A. A., Dennett, M. D., Muni, M. and Abid, M., Growth and yield response of wheat varieties to water stress at booting and anthesis stages of development. *Pak. J. Biotechnol.*, 2012, **44**, 879–886.
3. FAOSTAT, Food and Agriculture Organization of the United Nations Statistics Division, FAO, Rome, 2014.
4. Hatchett, A. H., Stacks, K. J. and Webster, J. A., Insect and mite pests of wheat. In *Wheat and Wheat Important* (ed. Heyne, E. G.), Madison, Wisconsin, USA, 1987, p. 625.
5. Pal, B. P., *Wheat*, Indian Council of Agriculture Research, New Delhi, 1996, pp. 244–246.
6. Gurung, B., Ponnusamy, N. and Pal, S., Effect of weather parameters on population dynamics of coccinellids on different crop ecosystems. *J. Agrometeorol.*, 2018, **20**(3), 254–255.
7. Gurung, B., Dutta, S., Singh, K. N., Lama, A., Vennila, S. and Gurung, B., Development of weather-based forewarning model for tomato leaf curl infestation. *J. Agrometeorol.*, 2022, **24**(4), 424–426.
8. Balogun, A. O., Basri, S., Abdulkadir, S. J. and Hashim, A. S., Performance analysis of feature selection methods in software defect prediction: a search method approach. *Appl. Sci.*, 2019, **9**(13), 2764.
9. Bocca, F. F. and Rodrigues, L. H. A., The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. *Comp. Electron. Agric.*, 2016, **128**, 67–76.
10. Breiman, L., Random forests. *Mach. Learn.*, 2001, **45**, 5–32; doi: 10.1023/A:1010933404324.
11. Gopal, P. M. and Bhargavi, R., Optimum feature subset for optimizing crop yield prediction using filter and wrapper approaches. *Appl. Eng. Agric.*, 2019, **35**, 9–14.
12. Oreski, D., Oreskib, S. and Klicek, B., Effects of dataset characteristics on the performance of feature selection techniques*. Appl. Soft Comp.*, 2017, **52**, 109–119.
13. Whitmire, C. D., Vance, J. M., Rasheed, H. K., Missaoui, A., Rasheed, K. M. and Maier, F. W., Using machine learning and feature selection for alfalfa yield prediction. *AI*, 2021, **2**, 71–88.

Received 24 April 2023; revised accepted 31 August 2023

doi: 10.18520/cs/v125/i11/1244-1249