PCE-BP: Polynomial Chaos Expansion-Based Bagging Prediction Model for the Data Modeling of Combine Harvesters

Liangyi Zhong[1], Mengnan Deng[2], Maolin Shi[3]*, Ting Lou[4], Shaoyang Zhu[5], Jingwen Zhan[6], Zishang Li[7], Yi Ding[8]
School of Agricultural Engineering, Jiangsu University, Zhenjiang, China, 212013[1, 2, 3, 4, 5, 6, 7, 8]
School of Mechanical Engineering, Jiangsu University, Zhenjiang, China, 212013[3]

***Abstract*—With the rapid developments of measurement and monitoring techniques, massive amounts of in-situ data have been recorded and collected from the measurement system of combine harvesters in their working process and/or field experiments. However, the relationship between the operation parameters and the performance index such as clearing loss usually changes greatly in different sample subspaces, which makes it difficult for conventional prediction models to model the in-situ data, since most of them assume that the relationship is the same or similar throughout the whole sample space. Therefore, a polynomial chaos expansion-based bagging prediction model (PCE-BP) is proposed in this article. A polynomial chaos expansion-based decision tree is constructed to divide the sample space such that the relationship between the operation parameters and the performance index in the same part is more similar than the others, and bagging is used to ensemble the polynomial chaos expansion-based decision trees to reduce the perturbation and provide robust predictions. The experiments on the mathematical functions show that the proposed prediction model outperforms polynomial chaos expansion, polynomial chaos expansion-based decision tree, and the conventional bagging prediction model. The proposed prediction model is validated through two monitoring datasets from a combine harvester. The experimental results show that the PCE-BP model provides better cleaning loss and impurity rate prediction results than the other prediction models in most experiments, showing the advantages of sample space partitioning and bagging in the data modeling of combine harvesters.**

***Keywords*—Combine harvester; data modeling; polynomial chaos expansion; decision tree; bagging**

I. INTRODUCTION

A combined harvester is a critical type of agricultural machinery that has been widely used to harvest grain crops. In the working units of a combine harvester, the grain crops are divided into grains and other materials by cutting, feeding, threshing, and cleaning. Since the interactions between the working units (header, conveying trough, cleaning fan, and sieve) and crops are very complex, it is difficult to construct an accurate theoretical or simulation models to accurately describe the working process of combine harvesters [1], [2], [3], [4]. In recent decades, numerical simulation methods such as computational fluid dynamics have been used to predict and analyze the working process of combine harvesters, which provides useful advice and references for the design, analysis, and optimization of working units [5], [6]. However, the computational cost of numerical simulations is too high to be accepted. For instance, the computational fluid dynamics simulation of a cleaning fan takes approximately four hours. If 100 simulations are conducted to obtain the mapping function between the design variables and the flow rate, it would take out 400 hours, around 17 days. In the past few years, the intelligent techniques of combine harvester since more operation parameters can be monitored and measured in the operation process and/or field experiments of combine harvesters. The interaction mechanisms and information among the working units and those between the crops and working units are involved in the measured data. In addition, the computational cost of the data-driven model is usually much lower than that of the corresponding numerical simulation. Therefore, the data-driven design, analysis, optimization, and control of combine harvesters are being the topics of interest in recent years [7], [8], [9], [10].

In the data mining tasks of combine harvesters, the first and most crucial step is constructing a prediction model for the response of interest, such as cleaning loss and grain impurity. Compared with hyperparameter prediction models (such as artificial neural networks and support vector regression), polynomial regression-based prediction models offer the advantages of lower computational cost and higher interpretability ,which have been widely used in the data modeling of combine harvesters.For example, Zareei and Abdollahpour [7] applied polynomial regression to identify the primary factors influencing header loss and determined the optimal factor combination through experimental design. Mirzazadeh et al. [11] constructed a semi-threshed cluster prediction model using polynomial regression and then optimized the feeding rate, fan speed, and sieve open rate to reduce the impurity rate. However, the interplay between operational factors and the associated performance indicators often changes greatly in different sample subspaces, as the interactions between the working units and crops are very complex, as discussed above. Most conventional polynomial regression-based prediction models assume that the regression relationship is the same or similar throughout the whole sample space, so it is challenging to assess the complex relationship between the operation parameters and the performance index of combine harvesters. On the other hand, the conventional polynomial regression method lacks nonlinear learning ability [12]. To this end, we proposed a prediction method in this work, aiming to solve the first problem by sample space partition based on the interplay between operational factors and the associated

*Corresponding Author.

performance indicators and to solve the second problem by introducing a Gaussian stochastic process to polynomial regression (polynomial chaos expansion, PCE).

Many sample space partition methods have been proposed in the area of machine learning, such as decision trees [13], the k-means algorithm [14], and the fuzzy c-means algorithm [15]. Since the k-means algorithm and fuzzy c-means algorithm cannot provide the partition rules directly, the sample space partition strategy proposed here is developed under the framework of decision tree. In a decision tree, the sample space is recursively partitioned so that the samples in the same subspace at each leaf node have similar/identical classification labels (classification decision tree) or responses (regression decision tree). Decision trees for classification purposes play a prominent role in various applications, including fault detection and identifying patterns. Chaitanya and Yadav [16] proposed a fault identification and location approach for multi-terminal lines based on a decision tree. The proposed method has been applied and validated based on series-compensated transmission lines and double-ended transmission lines. Liu et al. [17] designed a void detection method to assist in the health monitoring of sandwich-structured immersed tube tunnels, where the void classifier was constructed using a decision tree based on the characteristics of impact elastic waves. Muralidharan and Sugumaran [18] used continuous wavelet transform to represent the vibration signals of monoblock centrifugal pumps and applied a decision tree to predict different types of faults. In a regression decision tree, the sample space is recursively partitioned so that the continuous responses in the same subspace are similar. The average response of the samples at the same node is considered as the predicted value for new points. Liang et al. [19] used a regression decision tree to predict the uniaxial compressive strength based on the material parameters and indicated that the regression decision tree outperformed multiple regression in most experimental cases. Waruru et al. [20] analyzed the near infrared diffuse reflectance spectroscopy data of air-dried soil and then used a regression decision tree to estimate the soil aggregation level based on the spectral data. Nieto et al. [21] collected the filter pressure drop data of a micro irrigation system and constructed a pressure drop prediction model using a regression decision tree. In addition, the importance of the input variables is ranked based on the nodes and splitting values of the regression decision tree. In conventional decision trees, the choice of the splitting input variable and the determination of the partitioning threshold for each node hinge on either the classification labels or the mean response exhibited by the samples. Put simply, within each leaf node's subspace, samples share identical classification labels or comparable responses, yet there's no consistent correlation between input variables and the output. The prediction performance of the decision tree frequently undergoes significant variations due to the perturbation in the splitting feature optimization process as well. To solve the first problem, a new decision tree based on polynomial chaos expansion is proposed here, in which the sample space at each node is divided according to the regression relationship of samples. Then, bootstrap aggregation [22], [23], [24], also called bagging, is used to improve the robustness of the prediction results to solve the second problem.

Here's how the remainder of this document is structured. In Section II, the related works of sample space partitioning and polynomial chaos expansion are reviewed, and the motivation and framework of the proposed method are discussed as well. The details of the proposed polynomial chaos expansion-based bagging prediction model are presented in Section III. In Sections IV, several mathematical functions and two in-situ datasets of a combine harvester are used to validate the proposed prediction model. Section VI summarizes the conclusions and viewpoints.

II. Polynomial Chaos Expansion-based Bagging Prediction Model (PCE-BP)

A. Polynomial Chaos Expansion

In this study, Polynomial Chaos Expansion (PCE) is employed to assess the correlation between operational parameters and performance indices. In a PCE model, the response of interest y is estimated as follows [25].

$$y = \sum_{\alpha \in N^n} \beta_\alpha \Psi_\alpha(\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{x}$ is the vector of input variables, $\alpha = (\alpha_1, \dots, \alpha_n)$ is an n-dimensional index, β_α is the coefficients, and Ψ_α is the tensor product of normalized univariate orthogonal polynomials as follows.

$$\Psi_\alpha(x) = \prod_{i=1}^{n} \Psi_{\alpha_i}^{i}(\boldsymbol{x}_i) \tag{2}$$

Usually, only the p-degree is considered in Eq. (1) to reduce the computation cost, and the response of interest in Eq. (1) is revised as follows.

$$y \cong \sum_{\alpha \in A^{p,n}} \beta_\alpha \Psi_\alpha(x)$$
$$A^{p,n} = \{\alpha \in N^n : \alpha = \sum_{i=1}^{n} \alpha_i \le p\} \tag{3}$$

Eq. (1) can be rewritten as

$$y = \Psi\beta \tag{4}$$

where, y is the vector composed of the responses for the n samples, Ψ is the matrix of Hermite normalized univariate orthogonal polynomials, and β is the vector of the polynomial chaos coefficients. Upon examining the aforementioned equation, it becomes apparent that acquiring knowledge of the coefficients β allows for the derivation of the PCE model. Notably, when the quantity of samples is at least as many as the model's degree, the coefficients β can be estimated utilizing the least squares approach, as outlined below.

$$\beta = (\Psi^T\Psi)^{-1}\Psi^T y \tag{5}$$

B. Proposed Prediction Model

As discussed in Introduction, we proposed a new prediction model based on polynomial chaos expansion, named the polynomial chaos expansion-based bagging prediction model (PCE-BP), in which the sample space is partitioned to enhance prediction precision, while bagging techniques are employed to bolster the stability of the prediction outcomes. In the proposed prediction model, m PCE-based decision trees are generated. The main difference between the proposed PCE-based decision tree and the conventional decision tree is that the node is split according to on the regression relationship, but not the

classification labels or the mean response. At all nodes, the samples (D_{train}) are categorized into the left subset $D_{train-L}$ and the right subset $D_{train-R}$. Based on D_{train}, $D_{train-L}$, and $D_{train-R}$, three PCE models are constructed, named PCE_t, PCE_L, and PCE_R. The training error before and after partition is used to calculate the splitting criterion S.

$$S = R^2_{after} - R^2_{before} + \theta$$

$$R^2_{before} = 1 - \left(\frac{\sum_{i=1}^{n_t} (y_{i,t} - \widehat{y_{i,t}})^2}{\sum_{i=1}^{n_t} (y_{i,t} - \underline{y})^2} \right) \quad (6)$$

$$R^2_{before} = 1 - \left(\frac{\sum_{i=1}^{n_L} (y_{i,L} - \widehat{y_{i,L}})^2 + \sum_{i=1}^{n_R} (y_{i,R} - \widehat{y_{i,R}})^2}{\sum_{i=1}^{n_t} (y_{i,t} - \underline{y})^2} \right)$$

where $y_{i,t}$, $y_{i,L}$, and $y_{i,R}$ are the real responses of D_{train}, $D_{train-L}$, and $D_{train-R}$, respectively; $\underline{y}$ is the mean response of D_{train}; $\widehat{y_{i,L}}$, $\widehat{y_{i,R}}$, and $\widehat{y_{i,t}}$ are the PCE predicted responses; n_t, n_L, and n_R are the sample sizes of D_{train}, $D_{train-L}$, and $D_{train-R}$, respectively; and θ is the adjustment coefficient. When $S > 0$, the current node is a leaf node. The construction process of the polynomial chaos expansion-based decision tree is summarized in Fig. 1. Utilizing the classification rules derived from the PCE-based decision tree, samples for prediction undergo classification until they arrive at leaf nodes, at which point the PCE models positioned as those leaf nodes provide the predictive responses.

A popular heuristic algorithm, the gray wolf optimizer [26], is modified to optimize the splitting feature at each node. In the modified optimization algorithm, the split input variable is first transformed into a latent variable. Given d input variables, the latent variable represents each alternative splitting input variable through the following Table I.

After that, the new quantitative variable (q) and the division point (p_d) are combined into a vector $z = [q, p_d]$ and optimized as follows. In the optimization process, the solution with the best splitting criterion S is set as the alpha (z_α), the beta (z_β) and the delta wolf (z_δ) are worse than z_α, and the other wolves are omega wolves (z_ω). During each iteration, the solution will undergo an update as detailed below.

$$z(t+1) = z(t) - a \cdot d \quad (7)$$

where $z(t+1)$ is the updated solution, $z(t)$ is the current solution, a is a coefficient vector, and d is the motion of the wolf relative to the prey (z_{prey}), which is defined as follows:

$$d = \left| c \cdot z_{prey}(t) - z(t) \right| \quad (8)$$

where:

$$a = 2a \cdot r_1 - \tau \quad (9)$$

$$c = 2 \cdot r_2 \quad (10)$$

$$\tau = 2 - t\left(\frac{2}{T}\right) \quad (11)$$

TABLE I. THE INPUT VARIABLES AND LATENT VARIABLES

Input variable	1	2	…	$d-1$	d
Quantitative variable	$[0, \frac{1}{d})$	$[\frac{1}{d}, \frac{2}{d})$	…	$[\frac{d-2}{d}, \frac{d-1}{d})$	$[\frac{d-1}{d}, 1)$

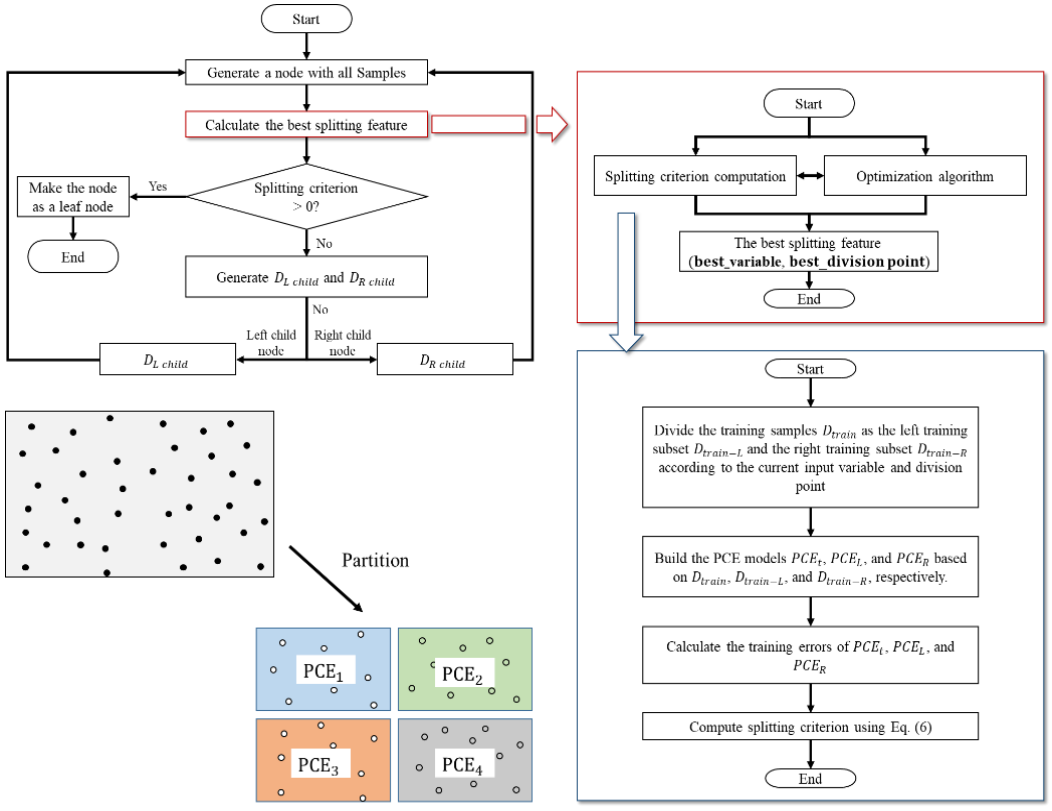


Fig. 1. Polynomial chaos expansion-based decision tree.

Where, T is the maximum number of iterations, and t is the current iteration. The positions of the other wolves (ω) are adjusted based on the three best solutions (z_α, z_β, and z_δ) as follows:

$$\begin{cases} d_\alpha = |c_1 \cdot z_\alpha(t) - z| \\ d_\beta = |c_2 \cdot z_\beta(t) - z| \\ d_\delta = |c_3 \cdot z_\delta(t) - z| \end{cases} \quad (12)$$

$$\begin{cases} z_1 = z_\alpha - a_1 \cdot (d_\alpha) \\ z_2 = z_\beta - a_2 \cdot (d_\beta) \\ z_3 = z_\delta - a_3 \cdot (d_\delta) \end{cases} \quad (13)$$

$$z(t+1) = \frac{z_1(t+1)+z_2(t+1)+z_3(t+1)}{3} \quad (14)$$

The above process of Eq. (7)-Eq. (14) repeats until the termination criterion is fulfilled, and the z_α of the last iteration is considered the best splitting feature.

From the above equations, it can be found that the random generation of initial solutions have effect on the construction process of the PCE-based decision tree. In other words, the generated decision tree might be slightly different even if the settings are same. Thus, Bagging is introduced to solve this problem, in which where m PCE-based decision trees are generated simultaneously and the final prediction result is estimated by the generated PCE-based decision trees.

$$\widehat{y^*} = \frac{\sum_{i=1}^{m} \widehat{y_i^*}}{m} \quad (15)$$

Where $\widehat{y^*}$ represents the final estimated response, and $\widehat{y_i^*}$ refers to the response produced by the iii-th decision tree based on PCE.

III. Experiments on Mathematical Functions

The effects of the parameter settings including the number of trees (m) and the adjustment coefficient (θ) on the proposed PCE-BP model are studied through two mathematical functions. The proposed model (PCE-BP) is compared with PCE, PCE-based decision tree (PCET), and random forest.The effectiveness of the aforementioned methods is assessed using R-square (R^2) , calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \underline{y})^2} \quad (16)$$

where y_i is the i-th real response, $\widehat{y_i}$ is the corresponding predicted value, n is the number of testing points, and $\underline{y}$ is the mean of the real responses. The closer R^2 is to 1, the better the performance of the prediction model.

A. Experiments on a Single Two-Dimensional Function

A two-dimensional function is utilized to validate the proposed PCE-BP model, maintaining consistency across the entire space. The function is defined as follows:

$$y = x_1^2 - 5cos(2\pi x_2)\ x \in [-1, 1] \quad (17)$$

The impact of the number of trees is examined first. For each number of trees (5, 10, 15, …., 45, and 50), 20 experiments are conducted, where the parameter θ is set as 0.05. 100 samples are generated using Latin hypercube sampling, and another 2,000 samples are used to validate the prediction models. The obtained mean and variance of R^2 are shown in Fig. 2.

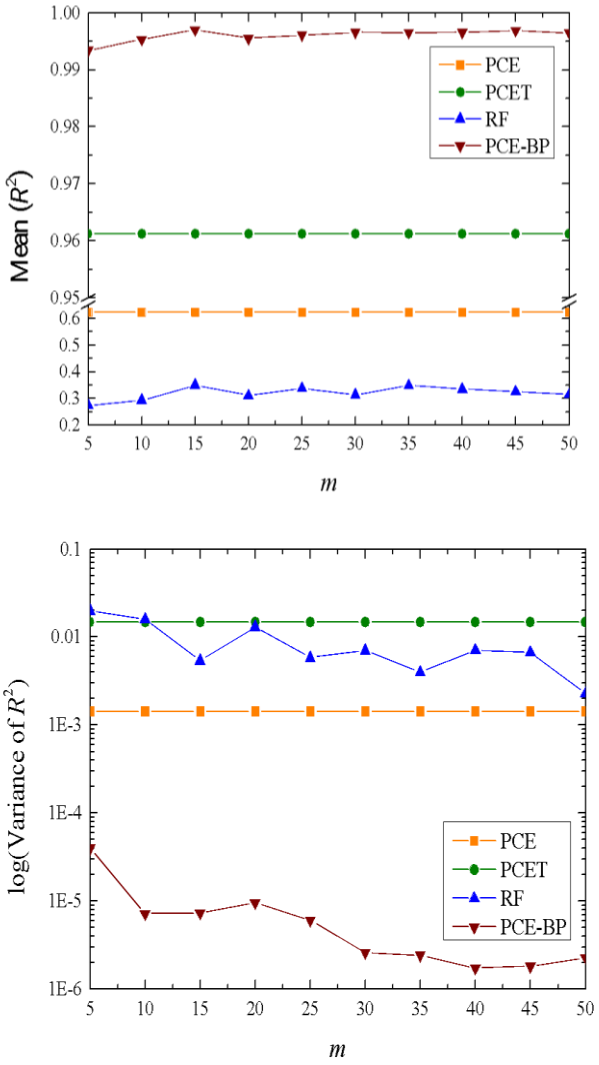


Fig. 2. The prediction results with different *m*.

From Fig. 2, it is found that the PCE-BP method outperforms the other prediction models in terms of the mean R^2 for all the values of m. The mean R^2 of PCET is higher than that of PCE, which is mainly because the PCET model partitions the sample space into several subspaces so that the regression relationship between the input variables and the response of interest is similar. Compared with the PCET model, the PCE-BP model produces better results with R^2 higher than 0.99. In each PCE-based decision tree, the sample space splitting at each node is influenced by the training samples and the randomly generated initial potential solutions for the splitting feature. The PCE-BP model uses bagging strategy to solve this issue, in which the predicted response is averaged by several PCE-based decision trees. The performance of the PCE-based decision tree varies greatly (the highest variance of R^2 for most experiments is shown in Fig. 2.), so its mean R^2 is smaller than that of the PCE-BP model. The PCE-BP model classifies the samples based on the regression relationship, but the RF model is based on the

mean responses. Thus, the PCE-BP model is also better than RF. The PCE-BP model is much better than the other three models in term of the variance of R^2 as well. The highest value is smaller than 0.0001, showing its robust prediction performance. With the parameter m increasing, the mean R^2 of the PCE-BP model first increases and then changes slightly when m exceeds 10. The variance of R^2 first decreases as parameter m increases and tends to remain stable. A larger m means that more PCE-based decision trees are generated in the PCE-BP model so that the perturbation brought by the splitting feature optimization is eliminated. As a result, the prediction accuracy is increased, as shown in Fig. 3. From the results and analysis above, it can be determined that the adverse effect of the splitting feature optimization is effectively reduced when m exceeds 25. In other words, the proposed PCE-BP model provides competitive prediction results for the single two-dimensional mathematical function tested here when the number of trees exceeds 25.

From Eq. (6), it can be found that the parameter θ is directly correlated with the splitting criterion, which would have an important effect on the prediction results of the PCE-BP model. The number of trees is set as 30, and the parameter θ is set as 0.04, 0.042, ..., 0.058, and 0.06. The mean and variance of R^2 over 20 experiments for each θ are presented in Fig. 3. The PCE-BP model is still better than that of the PCE, PCET, and RF models, showing the advantages of sample space partitioning and bagging. With the parameter θ increasing from 0.04 to 0.058, the mean R^2 increases and then tends to be stable. When the parameter θ is higher than 0.058, the mean R^2 decreases slightly. From Section III, it is known that the larger θ is, the higher the regression error before and after partitioning at each node.Samples within the same subspace show a more similar regression relationship between the input variables and the response of interest than those in different subspaces. The prediction accuracy of the PCE-BP model increases as parameter θ increases. However, when the parameter θ is too high, it is sample space is hard to be divided by the PCE-BP model, so the performance of the PCE-BP model decreases. The variance of R^2 decreases with increasing θ, indicating that the performance of the PCE-BP model is more robust. As a conclusion, the PCE-BP model produces competitive prediction performance when the parameter θ is approximately 0.05.

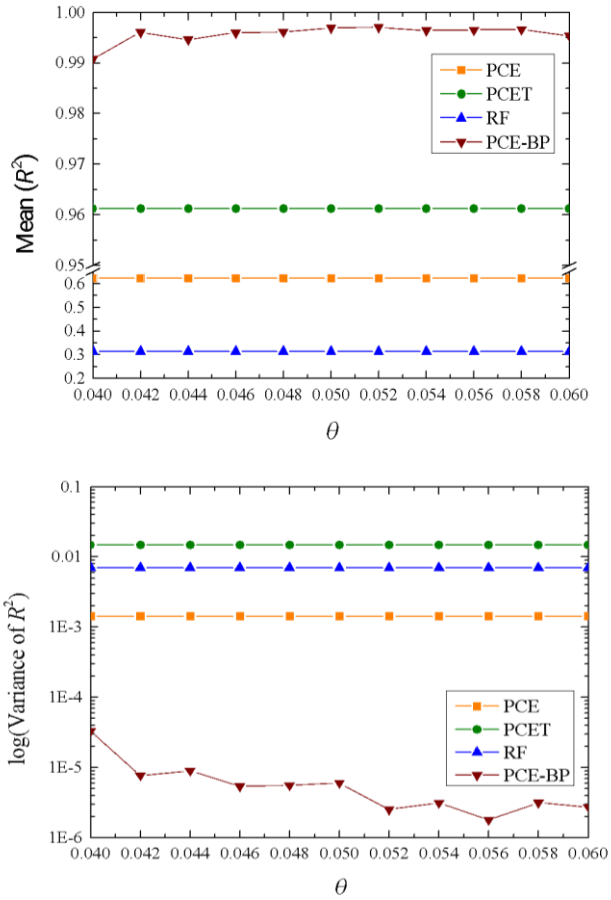


Fig. 3. The prediction results with different θ.

B. Piecewise Four-Dimensional Function

A piecewise four-dimensional function is used to validate the PCE-BP model, in which the mathematical function changes in different subspace, as defined as follows:

$$\begin{cases} y = sin\ sin\ (2\pi x_1) + x_2^2 + x_3 + x_4, x_1 \in [0, 0.5], x_2, x_3, x_4 \in [0, 1] \\ y = cos\ cos\ (2\pi x_1) + x_2 + x_3 + x_4, x_1 \in [0.5, 1], x_2, x_3, x_4 \in [0, 1] \end{cases} \quad (18)$$

The parameter θ is 0.05, and m is set as 5, 10, …, 45, and 50. Two hundred points are generated for the prediction models, then then the model accuracy is validated through another 4,000 samples. The obtained mean and variance of R^2 for 20 experiments are shown in Fig. 4. The mean R^2 of PCET PCE-BP is higher than that of PCE in all experiments. With the help of sample space partition according the relationship between the input variables and the response of interest, the PCE models at the leaf nodes of the PCET and PCE-BP can accurately evaluate the relationship, and the overall prediction accuracy is improved as well. The PCE-BP model outperforms the RF model as well, indicating that the prediction model of each subspace of the PCE-BP model can accurately evaluate the relationship. In addition, the prediction accuracy of the PCET and PCE-BP models both surpass the RF model. The variance of R^2 of the PCE-BP and RF models is smaller than those of the PCE and PCET models, which is mainly because of the introduction of bagging. With the parameter m increases, the average performance of the PCE-BP model increases and then tends to be stable. When m increases from 35 to 50, the variance of R^2 changes slightly. A larger m means that more PCET trees are generated in the PCE-BP model, so the perturbation brought by the splitting feature optimization can be more effectively eliminated. Thus, as m increases, the prediction accuracy increases, and the performance variance decreases, as shown in Fig. 4. When the parameter m is too high, the perturbation brought by the splitting feature optimization cannot be further reduced, so the mean and variance of the prediction results tend to be stable. Overall, the proposed PCE-BP model outperforms

the other three models and produces competitive prediction performance when the number of trees surpasses 25.

In the following experiments, the parameter m is set as 25; θ is set as 0.04, 0.042, ..., 0.058, and 0.060. The experiments are conducted 20 times for each value of sita, and the mean and variance of R^2 are shown in Fig. 5. From this figure, it is found that the proposed model outperforms the other models. With the parameter θ increasing from 0.040 to 0.046, the mean of R^2 of the proposed model increases, but the variance of R^2 decreases. As θ continuously increases to 0.06, both the mean and variance of R^2 change slightly. From the introduction of the PCE-BP model, it can be found that the initial space at the node is more likely to be partitioned when parameter θ is relatively smaller. In other words, the PCE-based decision tree is deeper, which results in the tendency of decision tree overfitting [13]. Thus, the prediction performance of the proposed model tends is being better when θ is larger. On the other hand, the variation brought by the splitting feature optimization is also reduced since the space at the node is more difficult to partition. Therefore, the variance of R^2 decreases as θ increases. The PCE-BP model provides competitive results for the piecewise four-dimensional mathematical function when θ is approximately 0.05.

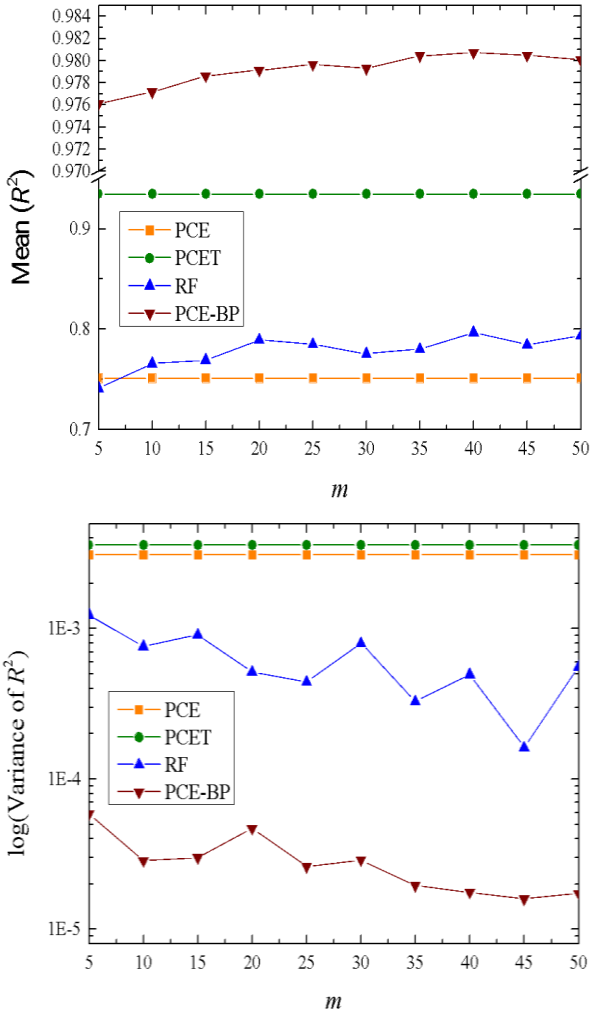


Fig. 4. The prediction results with different m.

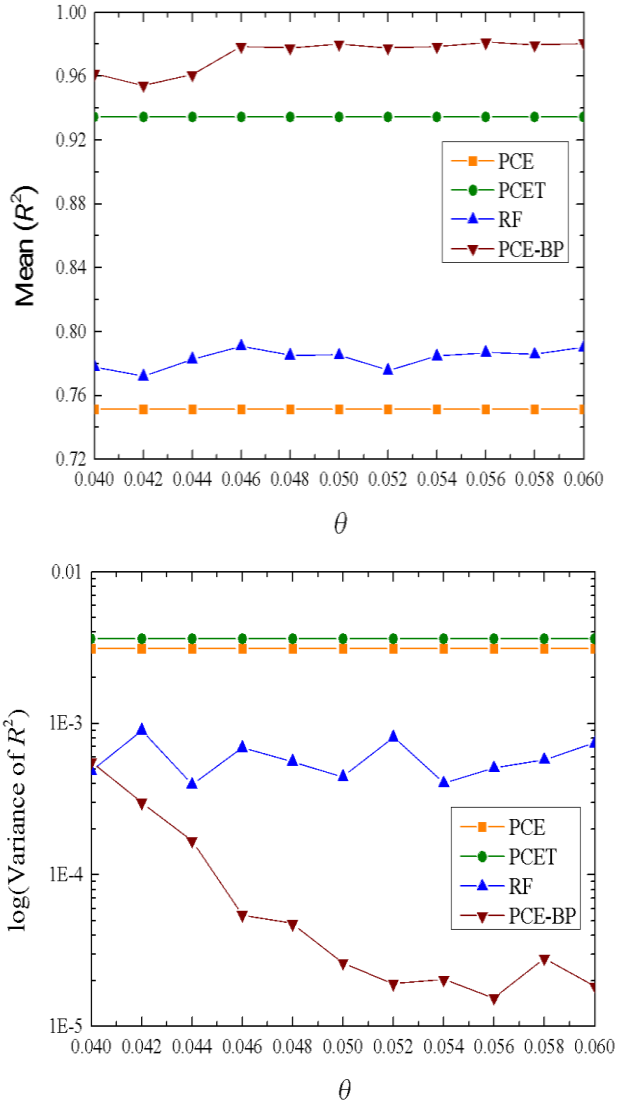


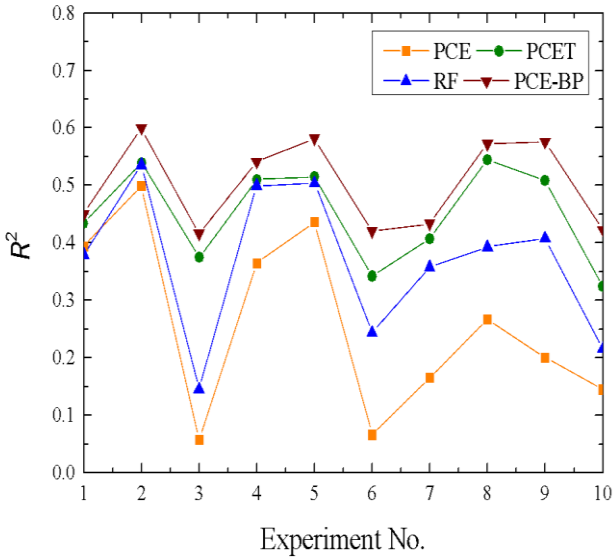
Fig. 5. The prediction results with different θ.

IV. Validation Based on the Datasets of a Combine Harvester

A. Case 1 (Cleaning Loss)

The PCE-BP model is applied to a while-feed combine harvester manufactured in Jiangsu, China (World Ruilong 4LZ-6.0A). The dataset used here comes from the field experiment (Fig. 6), containing 750 samples, including the header height, the open rate of the cleaning fan, the rotation speed of the cleaning fan, the open rate of the sieve, the angle of the guide plate, the rotation speed of the threshing drum, the gap of the threshing drum, and the cleaning loss. In each experiment, ten folds cross-validation experiments are conducted (in each experiment, a segment of the data serves as the basis for verifying the accuracy of the cleaning loss prediction models, whereas the remaining nine segments are dedicated to building the prediction model). The parameters m and θ of the PCE-BP model are set as 30 and 0.05, respectively. Fig. 7 shows the results of ten experiments.

Fig. 6. Field experiment of harvesting rapes.



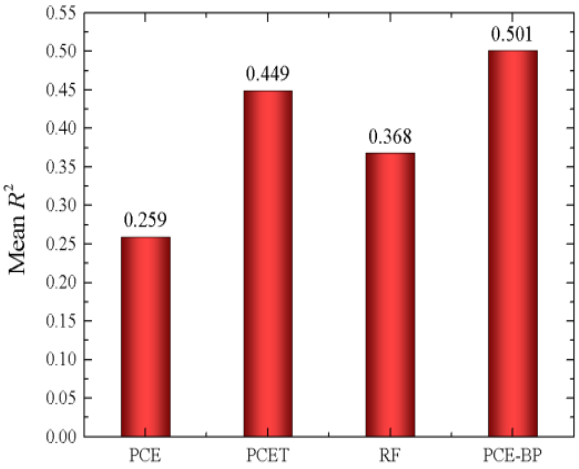
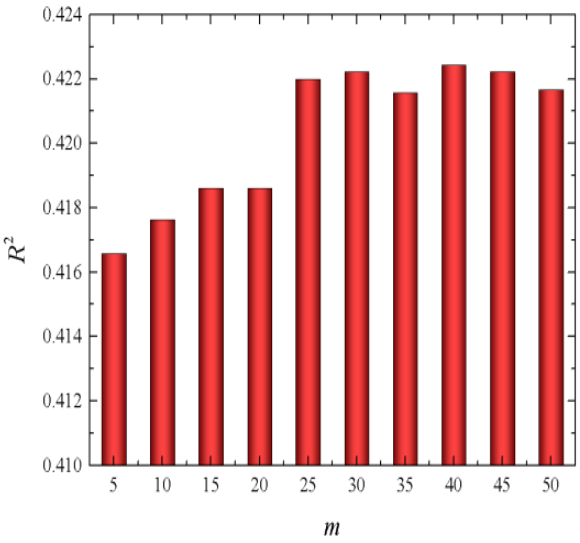


Fig. 7. Experimental results of Case 1.



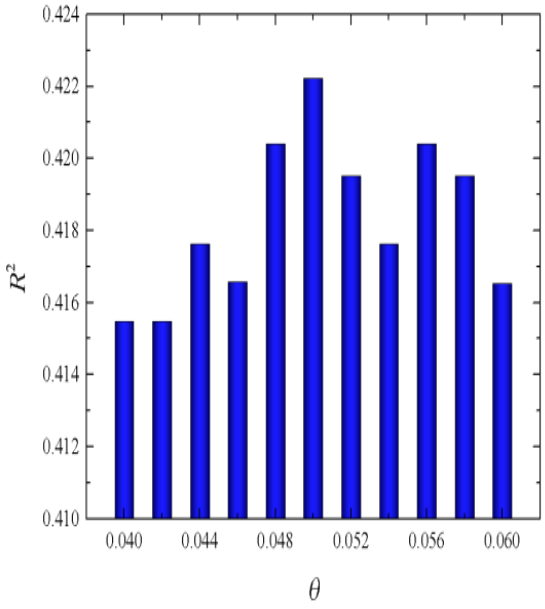


Fig. 8. Effect of parameters m and θ on the PCE-BP model.

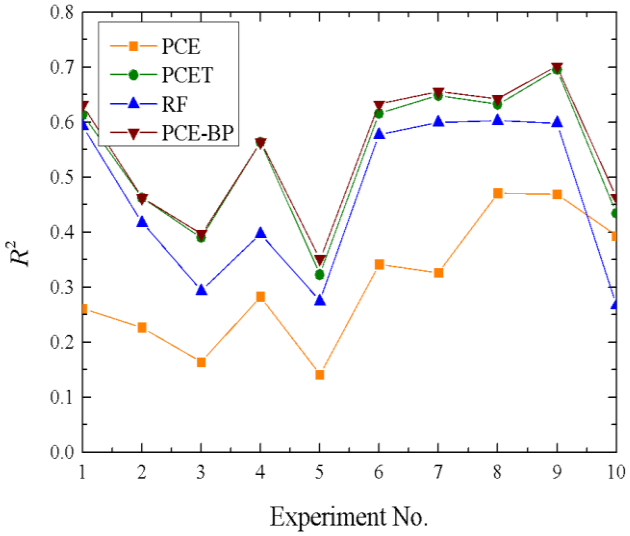
It is found that the PCE-BP model is better than the other three models. The R^2 of the PCET model is higher than that of the conventional PCE model, where the PCET's average R^2 is 0.449 and PCE is 0.259. In the PCET model, the sample space is divided into different parts such that the relationship between the operating parameters and the cleaning loss in the same part is more similar than those in the other parts, thus improving the cleaning loss prediction accuracy. Similarly, RF is also better than the PCE model, which can be attributed to the sample space partitioning. With the help of the bagging strategy, the perturbation brought by the feature splitting in PCET is eliminated. Additionally, the RF model divides the sample space according to the mean responses. Thus, the PCE-BP model outperforms the PCET and RF models in all experiments.

The training and testing datasets in the tenth experiment are used to study the number of trees m on the proposed model. The parameter θ is set as 0.05, and the parameter m is set as 5, 10,

…, 45, and 50. Fig. 8 shows the experimental results. It is found that the prediction performance of PCE-BP increases as m increases from 5 to 25. A larger m means that more PCE-based decision trees are generated in the PCE-BP model, which means that the perturbation brought by the splitting feature optimization can be effectively reduced, thus increasing the prediction accuracy. As m continually increases to 50, the R^2 of the PCE-BP model changes slightly, which is mainly because the perturbation cannot be reduced further. The effect of the parameter θ is studied as well, and the results are shown in Fig. 8, where the number of tree is 30. It is observed that R^2 tends to increase and then decrease with increasing parameter θ. From the introduction of the PCE-BP model, it can be found that a deeper PCE-based decision tree would be constructed when the parameter θ is relatively smaller. The generated tree is easy to overfit, so R^2 tends to be lower. When the parameter θ is relatively larger, the space at the node is more difficult to partition, so R^2 is lower. Overall, the PCE-BP model can provide competitive performance for different values of m and θ.

B. Case 2 (Impurity Rate)

Another dataset is used here, which includes 337 samples with recorded operational parameters and impurity rates. Ten folds cross-validation experiments are conducted as well in this subsection. The parameters m and θ are set to 30 and 0.05, respectively. Fig. 9 shows the experiments. It is found that the PCE-BP model outperforms the PCE and RF models, which can be attributed to sample space partitioning based on the relationship between the input variables and the response of interest. The R^2 of the PCE-BP is higher than PCET in most experiments. In experiments 2, 3, and 9, the performance of the PCE-BP model is very close to that of the PCET model. From Fig. 10, it can be found that the mean R^2 of PCE-BP is 0.550, which exceeds those of the other three models, highlighting the benefits of sample space partitioning and bagging.



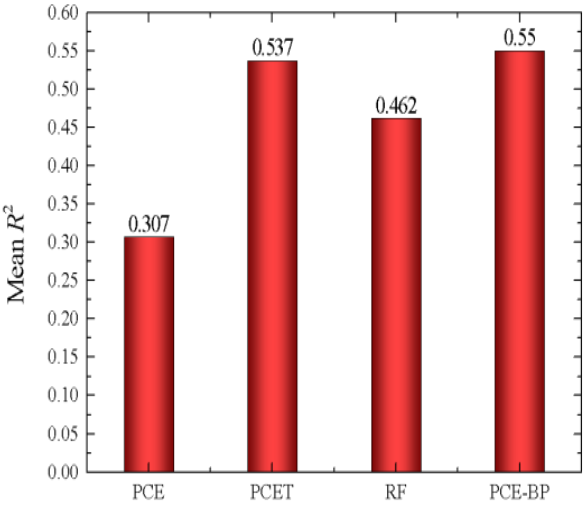
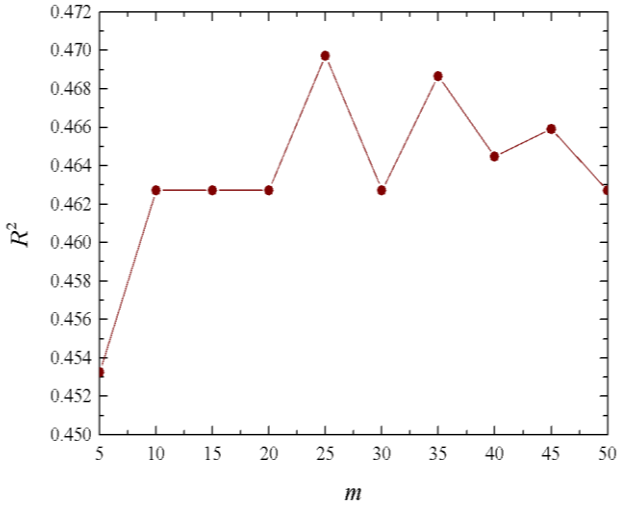


Fig. 9. Experimental results of Case 2.



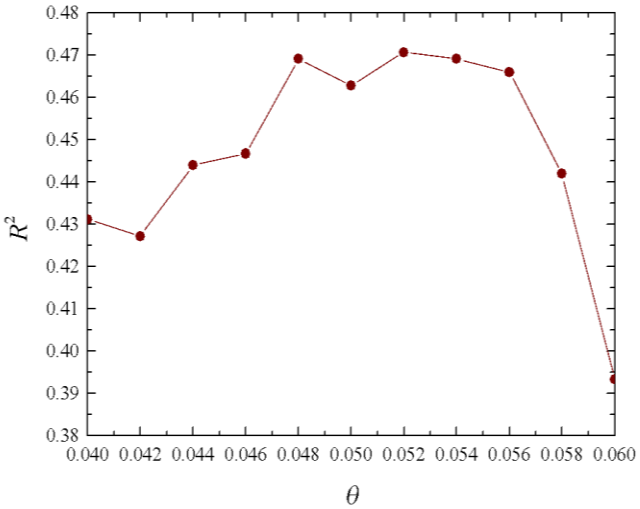


Fig. 10. Effects of parameters m and θ on the PCE-BP model.

The training and testing datasets from the tenth experiment are used to study the effects of the parameters m and θ on the PCE-BP model. The obtained results are shown in Fig. 10 (θ is set as 0.05). It is found that the R^2 with $m = 25\sim50$ is higher than $m = 5\sim20$. The parameter m is set as 30, the parameter θ is set as 0.04, 0.042, …, 0.058, and 0.06, and the results are shown in Fig. 10 as wll. R^2 first increases and then decreases as θ increases. When the factor θ is relatively small, a deeper PCE-based decision tree is constructed, which means that the prediction model easily overfits. On the other hand, when the parameter θ is relatively larger, the space at the node is more difficult to partition, so R^2 is lower as well. Overall, the PCE-BP model provides better results than the other three models for most values of parameters m and θ, as shown in Fig. 9 and Fig. 10.

V. Conclusion

In this paper, a polynomial chaos expansion-based bagging prediction model (PCE-BP) is proposed for modeling the field data of a combine harvester. In the proposed model, a polynomial chaos expansion-based decision tree is designed to partition the sample space, and bagging is used to ensemble the polynomial chaos expansion-based decision trees. The efficiency of the proposed prediction model is first validated through single and piecewise mathematical functions. The results show that the proposed prediction model outperforms polynomial chaos expansion, polynomial chaos expansion-based decision tree, and the conventional bagging prediction model functions. The proposed model demonstrates excellent prediction performance with 25 trees and the adjustment coefficient of 0.05. The proposed prediction model is further validated through two in-situ datasets of a combine harvester. The PCE-BP model provides more accurate cleaning loss and impurity rate prediction results than the conventional prediction models in most experiments. The experimental results show the advantages of sample space partitioning and bagging in the data modeling of combine harvesters.

From the results of the experiments, we found that the construction time of the proposed model is higher than that based on the conventional polynomial chaos expansion. In future work, the cost reduction of the proposed model will be our research topic. In addition, the other ensemble strategy such as boosting would be introduced into the proposed model as well.

Acknowledgment

This work is supported by the Project funded by China Postdoctoral Science Foundation (Grant No. 2022M711388); the Natural Science Foundation of Jiangsu Province (Grant No. BK20210777); Jiangsu Province and Education Ministry Co-Sponsored Synergistic Innovation Center of Modern Agricultural Equipment (Grant No. XTCX2014); and the Funding of Jiangsu University (Grant No. 20JDG068).

References

[1] I. Badretdinov, S. Mudarisov, R. Lukmanov, V. Permyakov, R. Ibragimov and R. Nasyrov. “Mathematical modeling and research of the work of the grain combine harvester cleaning system,” Computers and Electronics in Agriculture, 2019, vol. 165, p. 104966.

[2] Z. Liang, Y. Li, J. D. Baerdemaeker, L. Xu and W Saeys. “Development and testing of a multi-duct cleaning device for tangential-longitudinal flow rice combine harvesters,” Biosystems Engineering, 2019, vol. 182, p. 95-106.

[3] J. Pang, Y. Li, J. Ji, & L. Xu. “Vibration excitation identification and control of the cutter of a combine harvester using triaxial accelerometers and partial coherence sorting,” Biosystems Engineering, 2019, vol. 185, p. 25-34.

[4] Z. Qiu, G. Shi, B. Zhao, X. Jin, and L. Zhou. “Combine harvester remote monitoring system based on multi-source information fusion,” Computers and Electronics in Agriculture, 2022, vol. 194, p. 106771.

[5] C. Fan, T. Cui, , D. Zhang and Qu, Z. “Design of Feeding Head Spiral Angle Longitudinal Axis Corn Threshing Separation Device Based on EDEM,” 2019 Boston, Massachusetts July 7- July 10, 2019: n. pag.

[6] Tang, H., Xu, C., Zhao, J., and Wang, J. “Formation and steady state characteristics of flow field effect in the header of a stripping prior to cutting combine harvester with CFD” Computers and Electronics in Agriculture, 2023, vol. 211, p. 107959.

[7] S. Zareei and S. Abdollahpour. “Modeling the optimal factors affecting combine harvester header losses,” Agricultural Engineering International: CIGR Journal, 2016, vol. 18(2), p. 60-65.

[8] Z. Guan, Y. Li, S. Mu, M. Zhang, T. Jiang, H. Li, G. Wang and C. Wu. “Tracing algorithm and control strategy for crawler rice combine harvester auxiliary navigation system,” Biosystems Engineering, 2021, vol. 211, p. 50-62.

[9] L. Nádai, F. Imre, S. Ardabili, T. M. Gundoshmian, P. Gergo and A. Mosavi. “Performance analysis of combine harvester using hybrid model of artificial neural networks particle swarm optimization,” In 2020 RIVF International Conference on Computing and Communication Technologies (RIVF) (pp. 1-6). IEEE.

[10] Chen, M., Jin, C., Ni, Y., Yang, T., and Zhang, G “Online field performance evaluation system of a grain combine harvester,” Computers and Electronics in Agriculture, 2022, vol. 198, p. 107047.

[11] A. Mirzazadeh, S. Abdollahpour and M. Hakimzadeh. “Optimized Mathematical Model of a Grain Cleaning System Functioning in a Combine Harvester using Response Surface Methodology,” Acta Technologica Agriculturae, 2022, vol. 25(1), p. 20-26.

[12] Torre, E., Marelli, S., Embrechts, P., and Sudret, B. “Data-driven polynomial chaos expansion for machine learning regression,” Journal of Computational Physics, 2019, vol. 388, p. 601-623.

[13] Costa, V. G., and Pedreira, C. E. “Recent advances in decision trees: An updated survey,” Artificial Intelligence Review, 2023,vol. 56(5), 4765-4800.

[14] Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., and Heming, J. “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,” *Information Sciences*, 2023, vol. *622*, p. 178-210.

[15] M. Shi, Z. Liang, J. Zhang, L. Xu and Song, X. “A robust prediction method based on Kriging method and fuzzy c-means algorithm with application to a combine harvester,” Structural and Multidisciplinary Optimization, 2022, vol. 65(9), p. 1-18.

[16] B. K. Chaitanya, & A. Yadav. “Decision tree aided travelling wave based fault section identification and location scheme for multi-terminal transmission lines,” Measurement, 2019, vol. 135, p.312-322.

[17] R. Liu, S. Li, G. Zhang and W. Jin. “Depth detection of void defect in sandwich-structured immersed tunnel using elastic wave and decision tree,” Construction and Building Materials, 2021, vol,.305, p. 124756.

[18] V. Muralidharan and V. Sugumaran. “Feature extraction using wavelets and classification through decision tree algorithm for fault diagnosis of mono-block centrifugal pump,” Measurement, 2013, vol.46(1), p. 353-359.

[19] M. Liang, E. T. Mohamad, R. S. Faradonbeh, D. J. Armaghani and S. Ghoraba. “Rock strength assessment based on regression tree technique,” Engineering with Computers, 2016, vol. 32(2), p. 343-354.

[20] B. K. Waruru, K. D. Shepherd, G. M. Ndegwa and A. M. Sila. “Estimation of wet aggregation indices using soil properties and diffuse reflectance near infrared spectroscopy: An application of classification and regression tree analysis,” Biosystems Engineering, 2016, vol. 152, p. 148-164.

[21] P. J. G. Nieto, E. García–Gonzalo, G. Arbat, M. Duran–Ros, F. R. Cartagena and J. Puig-Bargues. “Pressure drop modelling in sand filters

in micro-irrigation using gradient boosted regression trees," Biosystems engineering, 2018, vol. 171, p. 41-51.

[22] R. E. Banfield, L. O. Hall, K. W. Bowyer and W. P. Kegelmeyer. "A comparison of decision tree ensemble creation techniques," IEEE transactions on pattern analysis and machine intelligence, 2006, vol. 29(1), p. 173-180.

[23] P. Yariyan, S. Janizadeh, T. Van Phong, H. D. Nguyen, R. Costache, H. Van Le and J. P. Tiefenbacher. "Improvement of best first decision trees using bagging and dagging ensembles for flood probability mapping," Water Resources Management, 2020, vol. 34(9), p. 3037-3053.

[24] S. Moral-García, C. J. Mantas, J. G .Castellano, M. D. Benítez and J. Abellan. "Bagging of credal decision trees for imprecise classification," Expert Systems with Applications, 2020, vol. 141, p. 112944.

[25] J. Zhang, X. Yue, J. Qiu, L. Zhuo and J. Zhu. "Sparse polynomial chaos expansion based on Bregman-iterative greedy coordinate descent for global sensitivity analysis," Mechanical Systems and Signal Processing, 2021, vol. 157, p. 107727.

[26] S. Mirjalili, S.M. Mirjalili and A. Lewis. "Grey wolf optimizer," Advances in engineering software, 2014, vol. 69, p. 46-61.